Mestrado em Econometria Aplicada e Previsão

Digit Analysis Using Benford's Law: A Bayesian Approach

Pedro Miguel Teles da Fonseca

Orientação: Prof. Dr. Rui Miguel Batista Paulo

Instituto Superior de Economia e Gestão

Universidade de Lisboa

Portugal

15 de Outubro de 2016

Digit Analysis Using Benford's Law: A Bayesian Approach

Pedro Miguel Teles da Fonseca*

Advisor: Prof. Dr. Rui Miguel Batista Paulo†

Abstract

According to Benford's law, many of the collections of numbers which are generated without human intervention exhibit a logarithmically decaying pattern in leading digit frequencies. Through digit analysis, this empirical regularity can help identifying erroneous or fraudulent data. Due to the power that classical significance tests with fixed dimension attain in large samples, they produce small p-values and, if the sample is big enough, are able to identify any deviation from Benford's law, no matter how tiny, as statistically significant. This may result in the rejection of Benford's law in samples where the deviations from it are without practical importance, and consequently samples which are legit are likely to be classified as erroneous or fraudulent. This dissertation proposes a Bayesian model selection approach to digit analysis. An empirical application with macroeconomic statistics from Eurozone countries demonstrates the applicability of the suggested methodology and explores the conflict between the p-value and Bayesian measures of evidence (Bayes factors and posterior probabilities) in the support they provide to the presence of Benford's law in a given sample. It is concluded that classical significance tests often reject the presence of Benford's law in samples which are deemed to be in conformance to it by Bayesian measures, and that even lower bounds on such measures over wide classes of prior distributions often provide more evidence in favour of Benford's law than the p-value and classical significance tests seem to suggest.

Keywords: Bayes Factor, Bayesian Model Selection, Benford's Law, Conditional Measures of Evidence, Digit Analysis, Fraud Detection, Goodness-of-fit, Hypothesis Testing, Lower Bounds, P-Value, P-Value Calibration, Macroeconomic Statistics, Point Null Hypothesis, Posterior Probability.

* Contact: pedro.teles.fonseca@outlook.com

† ISEG Lisbon School of Economics & Management-Department of Mathematics and CEMAPRE. Contact: rui@iseg.ulisboa.pt

Digit Analysis Using Benford's Law: A Bayesian Approach

Pedro Miguel Teles da Fonseca*

Orientação: Prof. Dr. Rui Miguel Batista Paulo †

Resumo

A lei de Benford, regularidade empírica segundo a qual muitos dos conjuntos de números gerados sem intervenção humana exibem um padrão de decaimento logarítmico nas frequências de ocorrência de primeiros dígitos, pode ser utilizada para, através da análise da frequência de dígitos, identificar conjuntos de números potencialmente erróneos ou fraudulentos. Devido ao elevado nível de potência alcançado pelos testes de hipóteses clássicos de dimensão fixa em amostras grandes, espera-se que, se a amostra for suficientemente grande, estes consigam identificar qualquer desvio em relação à lei de Benford, por mais pequeno que seja, como sendo estatisticamente significativo. Isto pode levar à rejeição da presença da lei de Benford em amostras onde o desvio em relação à mesma não tem significância prática e à identificação de amostras legítimas como sendo fraudulentas. Esta dissertação sugere uma abordagem baseada na seleção bayesiana de modelos. A metodologia proposta é aplicada num estudo empírico que utiliza estatísticas macroeconómicas de países da Zona Euro e explora o conflito entre o valor-p e as medidas bayesianas de evidência (fator de Bayes e probabilidades *a posteriori*) a nível do suporte por elas fornecido à presença da lei de Benford numa amostra. Conclui-se que os testes clássicos rejeitam frequentemente a presença da lei de Benford em amostras onde as medidas bayesianas são favoráveis à sua presença, e que mesmo limites inferiores destas medidas sobre largas famílias de distribuições *a priori* frequentemente fornecem bastante mais suporte à presença da lei de Benford do que o valor-p e os testes clássicos.

Palavras-Chave: Análise da Frequência dos Dígitos, Bondade do ajustamento, Calibração do Valor-p, Detecção de Fraude, Seleção Bayesiana de Modelos, Estatísticas Macroeconômicas, Factor de Bayes, Hipótese Nula Precisa, Lei de Benford, Limites Inferiores, Medidas Condicionais de Evidência, Probabilidade Posterior, Testes de Hipóteses, Valor-p.

* Contacto: pedro.teles.fonseca@outlook.com

† ISEG Lisbon School of Economics & Management-Departmento de Matemática e CEMAPRE. Contacto: rui@iseg.ulisboa.pt

Dedicated to my beloved parents

Acknowledgements

I would like to thank my advisor for having accepted to guide throughout this project and for being willing to share his knowledge, for the availability, for having suggested this interesting topic and for all other suggestions and corrections. I would also like to thank my parents for giving me the opportunity to continue my studies this far.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Pedro Miguel Teles da Fonseca

January 31, 2017

"It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge."

Pierre Simon Laplace (1820)

Contents

List of Figures

List of Tables

Acronyms

BF Bayes Factor.

BL Benford's Law.

BL1 Benford's Law for First Digits.

BL2 Benford's Law for Second Digits.

BLDA Benford's Law based Digit Analysis.

BMS Bayesian Model Selection.

CHT Classical Hypothesis Testing.

CME Conditional Measures of Evidence.

DA Digit Analysis.

DGP Data Generating Process.

LTP Law of Total Probability.

PDF Probability Density Function.

PMF Probability Mass Function.

PP Posterior Probability.

SGPC Stability and Growth Pact Criteria.

US United States.

Chapter 1

Introduction

> "*The difficulty lies, not in the new ideas, but in escaping from the old ones, which ramify, for those brought up as most of us have been, into every corner of our minds.*"
> John Maynard Keynes (1937)

1.1 Motivation and Goals

Contrary to what one might intuitively think, the observed frequencies of leading digits in numbers from many naturally occurring collections of numbers are not uniform. Instead, smaller numbers are more likely to occur as first digits than larger numbers. In many such datasets, about 30% of the entries start with a 1, 18% start with a 2, and so on up to the less likely leading digit (9), occurring only about 5% of the time. Those are the frequencies postulated by Benford's Law (BL).

Digit Analysis (DA) consists in using empirical regularities regarding the occurrence of digits in numbers, such as BL, to screen numerical datasets for anomalies like erroneous or fraudulent data. It relies on goodness-of-fit tests, where a point null hypothesis represents conformance to the expected law. The classical paradigm of hypothesis testing [Classical Hypothesis Testing (CHT)] is the predominant approach. Conformance to BL is used as

proxy to normal behaviour, but because models are supposed to be approximations of the reality and one can not realistically expect the data to perfectly fit the postulated models (even when they are true) in all samples, Benford's Law based Digit Analysis (BLDA) is a problem where economic and practical significance of the deviation from the expected law is more important than its statistical significance. Therefore, CHT with fixed dimension, which according to Pericchi and Torres (2011) over-reject the null hypothesis in large samples due to the high power they attain, making statistical significance prevail over economic significance may be inappropriate for DA. Wasserstein and Lazar (2016) note that if a sample is big enough then CHT can identify any deviation from the hypothesised law, no matter how tiny, as statistically significant. BLDA is then likely to produce many false positives, as it identifies very small deviations from BL, without practical importance, as statistically significant.

One goal of this dissertation is to propose an alternative BLDA methodology, based on Bayesian Model Selection (BMS). Two model selection environments are presented, one where conformance to BL frequencies is assessed jointly and is meant to be an alternative to the classical joint goodness-of-fit tests used in DA, such as the chi-square test, and another one where agreement to each BL postulated frequency is assessed individually, and is meant to be an alternative to the classical z-test. Conditional Measures of Evidence (CME) [Bayes Factors (BFs) and posterior probabilities (PPs)] will be used to quantify the evidence if favour of the null hypothesis (conformance to BL), instead of the classical and widely used p-value, which besides being more easily misinterpreted is difficult to perceive in a probability scale and quantify as the strength of the evidence provided by the data against the null hypothesis.

The other goal is to explore the conflict between CHT and BMS in precise null hypothesis testing, and its impact to BLDA. Delampady and Berger (1987, 1990) show that CME often support precise hypotheses with tiny p-values, and that even lower bounds on CME over wide classes of prior distributions often provide more support to the null hypothesis than the p-value, suggesting that CHT frequently underestimate the evidence provided by the data in favour of the null hypothesis.

1.2 Structure

This dissertation begins with a brief literature review of the relevant topics, where preliminary concepts are introduced (chapter 2): first, BL is defined and its empirical evidence is reviewed (section 2.1), then DA is introduced, with particular focus on BLDA (section 2.5), section 2.6 addresses the conflict between the two main paradigms of hypothesis testing (the Classical and the Bayesian) and section 2.7 details the motivation for the choice of the Bayesian approach in the particular problem being addressed. Chapter 3 details the theoretical foundations of the methodology that will be applied: the BFs are derived in sections 3.2 and 3.3, prior distribution specification is discussed in section 3.4, PPs calculation is discussed in section 3.5 and the lower bounds on BFs and PPs are addressed in section 3.6. Chapter 4 presents an empirical application of the methodology suggested in chapter 3 using real life data: the data is described in section 4.2, the study design in section 4.3, the study results are presented in 4.4 and discussed in section 4.5. The conclusions and the limitations of the approach are presented chapter 5.

1.3 Notation and Terminology

A number's first digit, which may also be referred to as that number's most significant digit, leading digit or mantissa, is the first element of the number's floating point representation and will be denoted D_1. Likewise, D_k represents the k^{th} most significant digit in a number: the k^{th} entry of the number's floating point representation. The base 10 logarithm of x will be denoted $\log(x)$, and its natural logarithm as $\ln(x)$. The CME should be interpreted as measures of evidence conditioned by the data, not the ones conditioned by the truth of the null hypothesis: this includes BFs and PPs and excludes p-values. All BFs in this work are BFs in favour of the null hypothesis. Bold Greek letters represent vectors, capital Greek letters represent parameter spaces and lower case Greek or Latin letters represent parameters. In situations of no ambiguity, the null hypothesis may be referred to as just "the null", the alternative hypothesis as "the alternative" and the prior distribution as "the prior". The Bayesian model combining the prior distribution $h(\theta)$ with the likelihood $f(x|\theta)$ will be textually denoted as $f(x|\theta) \wedge h(\theta)$, for example: Multinomial $\wedge$ Dirichlet Model or Binomial $\wedge$ Beta Model.

Chapter 2

Literature Review

"*The law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable.*"
Simon Newcomb (1881)

2.1 Benford's Law: History and Definition

BL is due to Simon Newcomb and Frank Benford. Newcomb (1881) noticed that books logarithmic tables were more worn out on the first pages and progressively cleaner throughout, suggesting that the larger a number's starting digit the less looked up to that number was. Based on this observation, he made the conjecture on the epigraph of this chapter, which implies the logarithmic relation in equation 2.1 (BL) and the probabilities in table 2.1. Benford (1938), through a dataset of 20229 observations from 20 different variables, showed that Newcomb's conjecture did fit many real life collections of numbers.*

$$P(D_1 = d_1) = \log(d_1 + 1) - \log(d_1) = \log\left(1 + \frac{1}{d_1}\right), d_1 \in \{1, \ldots, 9\} \tag{2.1}$$

Benford's (1938) study consisted in collections of numbers from such diverse sources as river surface areas, population sizes, physical constants, numbers in newspapers front pages, all the

* Diaconis and Freedman (1979) provided convincing evidence that Benford manipulated round-off errors to obtain a better fit. Nevertheless, according to Hill (1995-b), even the unmanipulted data remarkably agrees to the logarithmic relation in equation 2.1.

d_1	1	2	3	4	5	6	7	8	9
$P(D_1 = d_1)$	0.3031	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0512	0.0459

Table 2.1: Benford's Law for First Digits (BL1) probabilities

numbers inside a Reader's Digest issue, the heat of chemical compounds, molecules weight, drainage rates, death rates, atomic weights, baseball statistics, addresses from the first 342 people listed on the American Men of Science, sequences of powers, factorials and square roots, among others.

Through equation 2.1 and table 2.1 it is easy to see that, according to BL, the distribution of leading digits in numbers is far from uniform. Instead, it shows a logarithmic decay. Because DA often uses frequencies of digits other than the first, it is necessary to generalize BL for digits beyond the first (equations 2.4 and 2.5), as well as for combinations of digits (equations 2.2 and 2.3)*:

$$P(D_1 = d_1, D_2 = d_2) = \log\left(1 + \frac{1}{10\,d_1 + d_2}\right) \tag{2.2}$$

$$P(D_1 = d_1, \ldots, D_n = d_n) = \log\left(1 + \frac{1}{\sum_{i=1}^{n} 10^{n-i} d_i}\right) \tag{2.3}$$

$$P(D_2 = d_2) = \sum_{d_1=1}^{9} \log\left(1 + \frac{1}{10\,d_1 + d_2}\right) \tag{2.4}$$

$$P(D_n = d_n) = \sum_{d_1=1}^{9} \sum_{d_2=0}^{10} \cdots \sum_{d_{n-1}=0}^{10} \log\left(1 + \frac{1}{\sum_{i=1}^{n} 10^{n-i} d_i}\right) \tag{2.5}$$

where $d_i \in \{0, 1, \ldots, 9\}$ for digits beyond the first. Equation 2.2 is the joint Probability Mass Function (PMF) of D_1 and D_2, the first two digits, and equation 2.3 is the joint PMF of $D_1, \ldots, D_n$, the first n digits. The marginal PMF of D_2 (equation 2.4) is obtained by using the lLaw of Total Probability (LTP) on equation 2.2 to sum across all possible values of d_1, and the marginal PMF of the n^{th} most significant digit (equation 2.5) is obtained by using the LTP on equation 2.3. Benford's Law for Second Digits (BL2) probabilities, resulting from equation 2.4, can be consulted in table 2.2. There is still a decreasing pattern, although less

* See Benford (1938) or Jamain (2001) for details on the derivation of equations 2.2 and 2.3.

d_2	0	1	2	3	4	5	6	7	8	9
$P(D_2 = d_2)$	0.1197	0.1139	0.1088	0.1043	0.1003	0.0967	0.0934	0.0904	0.0876	0.0850

Table 2.2: Benford's Law probabilities for the second digit.

evident. For the third digit the probability is nearly uniform and for the forth and following the deviation from uniformity is inappreciable [see Berger and Hill (2015, p.2)]. Diaconis (1977) demonstrated that as n gets larger the distribution of D_n converges in exponential time to uniformity.

2.2 Benford's Law: Empirical Evidence

Subsequent to Benford's (1938) work, an abundance of additional empirical evidence has been found in many different domains, such as physics, biology, demographics, and computer science. Some examples of conformance to BL that can be found in the scientific literature are: lists of physical constants [Knuth (1981), Burke and Kincanon (1991)]*, decimal parts of failure (hazard) rates (Becker, 1982), radioactive half-lives [both measured and calculated (Buck, Merchant, and Perez, 1993)], long series of floating point numbers from scientific calculations [Knuth (1981), Hamming (1970)], sequences of factorials (Sarkar, 1973), powers of integers†, Fibonacci numbers (Washington, 1981) and Lucas numbers (Giles, 2007), repeated calculations with real numbers (Knuth, 1981), powers of random numbers (or their reciprocals) as the exponent gets larger, products of random numbers as the number of terms in the product gets higher‡ (Adhikari and Sarkar, 1968), prime numbers in large finite intervals (Luque and Lacasa, 2009)§, the distribution of cells per colony in certain cyanobacterium (Costas et al., 2008), basic genome data (Friar, Goldman, and Pérez–Mercader, 2012), daily pollen counts in European cities (Docampo et al., 2009), population sizes [Nigrini and Wood

* Jamain (2001) warned to the fact that these results regarding physical constants may not be very convincing as the samples are usually not large enough to allow strong statistical conclusions.

† See Raimi (1976) for the powers of two. A generalization for powers of higher order is a consequence of the equidistribution theorem.

‡ Schatte (1988) extended this idea for sufficiently long computations in floating-point arithmetic.

§ Although for this case it is a generalized version of the Benford's law [see Pietronero, Tosatti, Tosatti, and Vespignani (2001)]. Furthermore, the larger the interval, the less logarithmic and more uniform the distribution is. Diaconis (1977) showed that in the infinite limit the distribution of leading digits in prime numbers is uniform.

(1995), Jamain (2001), Hill (1995-b)], numbers (Varian, 1972) and regression coefficients (Tödter, 2009) in scientific publications, vote counts in electoral processes [Torres (2006), Pericchi and Torres (2011)] and the set of all numbers on the World Wide Web (Berger and Hill, 2015, p. 4-6). More important for the purpose of this dissertation are the findings in the fields of economics, finance and accounting: gross domestic product growth rates (Nye and Moul, 2007), many macroeconomic time series such as banking statistics, national financial statistics and balance of payment statistics (Gonzalez-Garcia and Pastor, 2009), business invoices and financial forecasts (Varian, 1972), most of the accounting data [Nigrini (1992, 1995 1999, 1997, 2012)]*, reported income tax data (Nigrini, 1996), interest received in United States (US) tax returns (Berger and Hill, 2015), 1-day returns on the Dow-Jones Industrial Average Index and on the Standard and Poor's Index for stock prices (Ley, 1996), the main Chinese stock market indices (Shengmin and Wenchao, 2010), and the Madrid, Vienna and Zurich stock market prices (Pietronero et al., 2001).

2.3 Benford's Law: Invariance†

Other distinctive properties of BL conforming datasets is that their leading digit distributions are scale invariant (Pinkham, 1961), base-invariant (Hill, 1995-a), inversion invariant‡ (Benford, 1938) and sum invariant (Allaart, 1997). This means that if one begins with a BL conforming dataset and either multiply all entries by a constant, divide one by each entry, or convert all entries to another base§, the observed frequencies of leading digits will remain approximately constant. For other invariances of BL see Jamain (2001). It is demonstrated that BL is the only possible leading digit distribution with such properties, that is, if the frequencies of leading digits in a numerical dataset are either scale invariant (Pinkham, 1961)¶, base

* Nigrini found that lists of items such as accounts receivable or payable, transactions, inventory accounts, fixed asset acquisitions, daily sales, refunds and disbursements all follow BL.

† For a comprehensive and rigorous review of BL invariance properties see Berger and Hill (2011).

‡ Some tabulations of data are given in reciprocal form, such as candles per watt and watts per candle, as Benford (1938) exemplified. If one form of the tabulation follows BL then its reciprocal also does.

§ When converting numbers to another base, the set of possible first digits will differ. Therefore, the leading digits frequencies can not remain the same. However, changing the base will preserve the logarithmic decay in frequencies if the original dataset follows BL (Smith, 2002). The generalization of BL1 for base b is $f_{D_1}(d_1) = \log_b(1 + \frac{1}{d_1})$, for d_1 from 1 up to $b - 1$ (Hill, 1995-b).

¶ Knuth (1981) and Hill (1995-a) accused Pinkham (1961) of making unwarranted assumptions about the distribution of numbers when deriving BL trought scale-invariance. Also, note that scale invariance is

invariant (Hill, 1995-a) or sum invariant (Allaart, 1997) then BL must hold.

2.4 Where can the Benford's Law be Found?

There has been a lot of attempts at explaining the emergence of BL [see Raimi (1976), Hill (1995-b), Scott and Fasli (2001), Jamain (2001), Smith (2002), Fewster (2009), Nigrini (2012) or Berger and Hill (2015) for reviews]. It is accepted that the first rigorous explanation was due to Hill (1995-b), who demonstrated that if random samples from different randomly-selected (in any unbiased way) probability distributions are combined, then the leading digit frequencies in the pooled sample converge to BL. This result helps explaining why BL arises so often. While numbers describing some phenomena are under the control of a single distribution (for example: the height of adult men behave according to a normal distribution), many others are dictated by a random mix of all kinds of distributions. A good example is the dataset resulting from pooling together all the numerical values in a firm's financial statement: the numbers will respect to many different variables, each behaving according to it's own Data Generating Process (DGP). The same principle applies for the set of all numbers in a Census form, tax report, scientific article or magazine. This is congruent with Benford's (1938) findings: he used numbers from 20 different domains, and the pooled sample fitted BL very well, even if some of the datasets did not when considered separately.

Some rules of thumb can help assessing whether a dataset should be expected to conform to BL. The dataset's mean should be larger than its median and the data's histogram should be positively skewed (Wallace, 2002). The graphical representation of the data should resemble a geometric sequence (without artificial truncation) and the logarithm of the difference between the largest and smallest values should be close to an integer value (Nigrini, 2012). The larger the ratio of the mean divided by the median, the most likely the fit is (Durtschi, Hillison, and Pacini, 2004). The numbers should represent quantities, amounts or sizes, should be free of imposed limits* (bounded sequences with restricted significant digits

actually not so rare. Quoting Jones (2002): "In the search for order and laws in complex systems there has been the realisation that much of life is scale invariant.".

* The "free of imposed limits" assumption may be relaxed as long as the data spans two orders of magnitude (Nye and Moul, 2007).

like hours of the day, months or years, human age or weight and the set of all integers are not good fits), should not be assigned sequentially (like phone numbers, checks and lottery numbers), and should not be influenced by human thought (ATM withdrawals, donations, prices or values set at psychological thresholds such as rounded quantities). Datasets where each entry consists in the arithmetical combination of multiple numbers, are also very likely to behave according to BL.*. Scott and Fasli (2001) consider that the best candidate datasets to reproduce BL are the ones with only positive values, uni-modal (and non zero mode) positively skewed distribution in which the median is no more than half of the mean. This implies a lognormal distribution with scale parameter larger than 1.2 for the data.†

Considering the discussion above, most economical and financial datasets are obvious candidates to fit BL: They consist in the aggregation of observations from several different variables (such as the set of all numbers in a firm's financial statement or in a government's budget report) and their entries consist in quantities or amounts that can be interpreted (and are generated) as the mathematical combination of several other variates‡.

2.5 Benford's Law and Digit Analysis

DA consists in using empirical regularities regarding the occurrence of significant digits in numbers to detect erroneous or fraudulent data. The idea is to model a baseline frequencies

---

* Recall the already mentioned findings of Adhikari and Sarkar (1968) about products of random numbers, Hamming (1970) and Knuth (1981) about long series of floating point numbers from scientific calculations, Sarkar (1973) about factorials, Raimi (1976) about powers of integers, Knuth (1981) about repeated calculations with real numbers and Schatte (1988) about long computations in floating point arithmetic. Also, Raimi (1969) and Boyle (1994) argue that multiplying random numbers produces conformance to BL and Boyle (1994) showed that BL is the limiting distribution of leading digits when random variates are repeatedly multiplied, divided or raised to integer powers. Scott and Fasli (2001) showed that when each number in a dataset is a product of many terms, the first digit distribution converges to BL as the number of terms in the product increases. This holds for products of random variates, successive multiplication by a new realization of the same random variable and for successive multiplication by a constant. For products with fewer terms it is possible that full convergence is not reached but a monotonic decay in the frequencies is still likely to be present.

† The "only positive values" condition is not as restrictive as it may look. Note that taking the absolute value of all entries in a dataset leaves the leading digit distribution unchanged.

‡ Trivial example: Revenue=price × quantity. Every variable that can be modelled by an equation, can be interpreted as an arithmetical combination of variates. The amount of terms in such an arithmetical combination is larger than it may look by inspection of the equation, because the equation has an error term that is itself interpreted as an arithmetical combination of all non included variates that affect the value of the dependent variable.

distribution representing normal behaviour and then attempt to detect if some particular dataset significantly departs from it (Bolton and Hand, 2002). According to Durtschi, Hillison, and Pacini (2004) various forms of DA have long been used by auditors when performing analytical procedures, such as checking transaction records for duplicate payments. BL, when applied to detect fraudulent or erroneous data, is just a more complex form of DA. BLDA is only applicable to usually BL conforming datasets, and as seen in section 2.4, this includes most of the economic and financial data.

Varian (1972) was the first to suggest the application of BL to DA. The idea is that in datasets of naturally generated numbers (i.e. without intervention) where digit frequencies conform BL, replacing numbers with fabricated ones typically results in deviation from BL. As discussed in section 2.4, numbers influenced by human thought usually do not conform BL, and hence manipulating numbers from a BL conforming DGP leaves a detectable trace in the data. This may happen for many reasons, like the fact that numbers influenced by human thought are usually tied to psychological thresholds. Durtschi, Hillison, and Pacini (2004) note that someone creating false numbers usually (and subconsciously) favours certain numbers, and may also be biased against certain numbers in an attempt to conceal their actions. Nigrini and Mittermaier (1997) note that when entering fraudulent data, people tend to use the same (or similar) amounts often, moving the observed digit frequencies away from BL. Also, fraudsters are usually unaware of the properties of the DGP behind the data they are manipulating, and consequently tend to distribute the made-up entries leading digits more uniformly than a BL law conforming DGP would. Cho and Gaines (2007) find it very unlikely that someone manipulating numbers would seek to preserve conformance to BL, because even though it is widely applicable it is not widely known. Moreover, experimental research has shown that people do a poor job in replicating random data even when they are told what the DGP is (Camerer, 2003, pp. 134-138). Bolton and Hand (2002) consider the premise behind fraud detection using tools such as BL to be the fact that fabricating data which conforms to BL law is difficult. Diekmann and Jann (2010) consider that in order to ascertain the validity of a BLDA it is necessary to demonstrate that the true data is in accordance to BL and the manipulated data is not, but Rauch et al. (2011) consider the fact that the probability of BL emergence is higher for non-manipulated than for manipulated data to be sufficient.

A wide literature exists on the application of BLDA. Carslaw (1988) analysed New Zealand firm's earnings and found that the numbers contained more zeros in the second digit than expected according to BL, suggesting that firms were manipulating (rounding up) their earnings. Thomas (1989) studied US firms earnings and found that firms reporting losses exhibit the reverse pattern (rounded down numbers), and also found evidence of manipulation (through rounding of numbers) in earnings per share data. Other studies where BL is used in the detection of earnings manipulation are: Niskanen and Keloharju (2000), Kinnunen and Koskela (2003), Caneghem (2002, 2004), Skousen, Guan, and Wetzel (2004), Nigrini (2005) and Guan, He, and Yang (2006). Nigrini (1992, 1996) was the first to extensively apply BL to accounting data with the goal of detecting fraud. He also used it to help identifying tax evaders, and so did Watrin, Struffert, and Ullmann (2008) and Möller (2009). Diekmann (2007) applied BL to scientific fraud detection*, and Asllani and Naco (2014) used it to screen hospital spendings for numerical anomalies. Nye and Moul (2007), Gonzalez-Garcia and Pastor (2009), Judge and Schechter (2009), Tödter (2009), Rauch, Göttsche, Brähler, and Engel (2011) used BL to assess the quality of economic data and macroeconomic statistics. Marchi and Hamilton (2006) found evidence of manipulation in self-reported regulatory data in the Toxic Release Inventory: while reported emissions of some chemicals did not fit BL, the measured values of the same chemicals did. They concluded that manipulation in the data may be the reason why large drops in air emissions reported by firms are not always matched by similar reductions in measured concentrations by pollution monitors. Giles (2007) studied a dataset of winning bids from eBay auctions and because the numbers fitted BL he found no evidence of collusion among bidders nor shill among sellers. Prudêncio (2015) analysed the financial statements of three Portuguese commercial banks from 2007 to 2013 and found significant deviations from BL. One of those banks, Banco Espírito Santo, went Bankrupt in 2014 and its former president, board of directors and other high level employees are facing several charges including manipulation of accounting numbers. Haynes (2012) analysed financial statements from bankrupt municipal governments and found overall nonconformity to BL. He concluded that such screening, had it been done earlier, could have identified that

* Recall the findings of Varian (1972) and Tödter (2009) mentioned in section 2.2: regression coefficients and other numbers from scientific publications are in conformance to BL. Hence, BLDA can help detecting fraud in scientific publications.

something was amiss. DA is also being used to screen and validate numbers from electoral processes [Mebane (2006-a, 2006-b, 2007), Torres (2006), Pericchi and Torres (2004, 2011), and Torres, Fernandez, Gamero, and Sola (2007)].

2.6 Paradigms of Hypothesis Testing

In the classical approach to hypothesis testing [Fisher (1925), Neyman and Pearson (1933)], a significant finding is declared when the value of a test statistic exceeds a specified threshold, with values of the test statistic above that threshold defining the rejection region. The significance level (also known as dimension) of the test is defined as the maximum probability that the test statistic falls into the rejection region when the null is true. Fisher (1925) proposed the p-value (the probability, conditioned on the null hypothesis being true, of obtaining a test statistic which is at least as unlikely as the one actually observed) as a measure of discrepancy between observed data and null hypothesis. For its simplicity and apparent objectivity, the p-value became the standard of measure evidence against an hypothesis (Schervish, 1996).

In the Bayesian approach, initially developed by Jeffreys (1935, 1967), statistical models represent the DGP of the data under each of two competing hypotheses, the BF compares the predictive density of the observed data under one model with that of the alternative model, and the Bayes (1763) theorem is used to compute the PP of the hypotheses.

As Kass and Raftery (1995) note, Bayes theorem updates prior probabilities into PPs through consideration of the data. The update represents the evidence provided by the data, and it is the same regardless of the prior probabilities. In the odds scale, the update corresponds to the BF, and represents the relative predictive density of the data under one of the hypotheses compared with that of the alternative. The BF is a measure of change in support, as it measures the change in prior odds in favour of one hypothesis after the data is observed (Lavine and Schervish, 1999). Bernardo and Smith (1994) intuitively describe the BF as a measure of whether (and by how much) the data have increased or decreased the odds in favour of one of the hypothesis. Unlike the p-value, BFs depend only on the predictive density of observed data, not on long run unobserved results (Goodman, 1999-b). On the other hand, BFs require the specification of an alternative hypothesis. Because in BLDA the

hypotheses being compared form a partition of the parameter space, with the alternative corresponding to the bilateral composite hypothesis of divergence from a point null which represents conformance to BL, the specification of the alternative is automatic, and therefore one of the main critiques to this approach [the subjectivity involved in the specification of the alternative when there is no objective choice (Johnson, 2013)] does not apply. However, the specification of a prior distribution for the alternative is unavoidable. This drawback can be mitigated by finding lower bounds on the CME over wide classes of prior distributions.

These two paradigms of hypothesis testing often produce seemingly incompatible results. Edwards, Lindman, and Savage (1963), Berger and Delampady (1987, 1990), Berger and Sellke (1987), and Lin and Yin (2015) show that evidence against the null hypothesis provided by CME can differ by an order of magnitude from the p-value, when testing a precise null hypotheses, raising concerns about the routine use of moderately small p-values and significance levels. A p-value of 0.05, conventionally labelled as a significant result and considered strong evidence against the null in the classical approach, can result in a PP of at least 0.3 in favour of the null hypothesis*. To solve this, Johnson (2013) considers that, in CHT, evidence thresholds should be decreased to 0.005 for the declaration of a significant finding and to 0.001 for a highly significant finding†. This contradicts Fisher (1925), for whom a p-value below 0.05 was a safe indicator of a significant result, but agrees with Taleb (2016), who warns that due the skewness and volatility of a p-value's meta-distribution (across repetitions of the same experience), to get what people mean by 5% confidence level, a p-value almost one order of magnitude smaller than conventional is needed.

The apparent discrepancy between the two paradigms is due to the fact that they rely on the calculation of different probabilities. p-values and significance levels are conditioned on the null hypothesis being true, and so they can not be a direct measure of the probability of that hypothesis (Goodman, 1999-b). PPs represent actual the probability of the hypotheses being true conditioned on the observed data, which is more straightforward to interpret,

* To highlight the conflict between the p-value and the CME, Berger and Delampady (1987, 1990) and Berger and Sellke (1987) showed that, in precise hypothesis testing, even the lower bounds on BFs and PPs that are found over wide classes of prior distributions are often much larger than the corresponding p-values. Therefore, one can not dismiss this conflict by arguing that the discrepancies are due to the specific prior distribution that was chosen.

† In terms of BFs, Johnson's (2013) revised standards for statistical evidence correspond to values from 25 to 50 for the declaration of a significant result, and 100 to 200 for a highly significant one.

as it is more natural to think in terms of the probability of an hypothesis given the data than in terms of the probability of the data given that the hypothesis is true. Still, a lot of practitioners incorrectly interpret a p-value of 0.05 as the null hypothesis having a 5% probability of being true, or as a 5% error rate on the rejection of the null, misinterpretations which Goodman (1999-a) popularized as "the p-value fallacy"*.

2.7 Why the Bayesian Approach?

Conformance to BL is a goodness-of-fit problem and hence BLDA relies on goodness-of-fit statistical tests. Large samples are always preferable, to give the DGPs a chance to reveal its true properties. Unfortunately, according to Pericchi and Torres (2011), the usefulness and interpretation of the p-value on classical test statistics is drastically affected by sample size, and Goodman (1999-a), Wasserstein and Lazar (2016) warn that the p-value and statistical significance do not take into account the size of the observed deviation from the null: any deviation, no matter how small, can produce a small p-value if the sample size or measurement precision is high enough. Consequently, very small deviations from the null, with no practical importance, are likely to be considered statistically significant†. Conversely, large deviations can produce large p-values if the sample is small or if the measurements are imprecise, similar deviations can have different p-values for different sample sizes and similar p-values may correspond to different deviations in different samples.

Despite the p-value being able to indicate how incompatible the data is with a specified hypothesis (Sellke, Bayarri, and Berger, 2001), knowing the data to be rare under one hypothesis is of little use unless one determines how rare it is under the alternative hypotheses. Hence, although in classical hypothesis testing the smaller the p-value, the more significant the deviation from the null is, it is difficult to perceive the p-value in a probability scale and quantify it as the strength of the evidence in the data against the null, as the p-value only provides one side of the information [Lin and Yin (2015), Wasserstein and Lazar (2016)]‡.

---

* Other references on the susceptibility of the p-value to be misinterpreted are: Gibbons and Pratt (1975), Schervish (1996), Matthews (1998), O'Hagan and Luce (2003) and Hubbard and Bayarri (2003).

† As Wasserstein and Lazar (2016) note, statistical significance is not equivalent to scientific, human, or economic significance.

‡ A famous situation in which considering only how unlikely the evidence was under one of the hypothesis

The practice of summarizing results into either statistically significant or non-significant and drawing a sharp distinction between them, standard in classical hypothesis testing, can also be misleading: besides encouraging the dismissal of potentially important evidence in favour of null, any particular threshold separating significance from non-significance is arbitrary*, and even large changes in observed significance levels can correspond to small, non-statistically nor practically significant changes in the underlying test statistics (Gelman and Stern, 2006). Wasserstein and Lazar (2016) consider this dichotomy to distort the scientific process and warn that scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

In BLDA, economic and practical significance of the deviation from BL is more important than it's statistical significance, as one cannot realistically expect the observed data to perfectly conform BL in all samples, even when BL hypothesis is true. Therefore, CHT with fixed dimension, in which statistical significance is known to overweight economic significance and which according to Pericchi and Torres (2011) are known to over-reject the null hypothesis (which in this case represents conformance to BL) in large samples, may not be adequate, as they are likely to produce many false positive results. According to Ley (1996), the over-rejecting nature of such tests is due to the huge power they attain in large samples, with the acceptance region shrinking with sample size, for a given significance level. Leamer (1983) considers this to be a weakness of the classical method, as models are to be considered mere approximations to reality instead of perfect DGPs.

Aware of all this, Ley (1996) addressed BLDA in a Bayesian way. Despite his posterior distribution based parameter estimates being very close to BL theoretical ones, the classical likelihood-ratio and chi-square tests would reject BL hypothesis in all datasets†. Torres (2006) and Pericchi and Torres (2011) used the Jeffreys model selection approach, based on BFs, PPs and respective lower bounds, and concluded that datasets with apparently very good fit to BL could have nearly zero p-values.

---

resulted in a miscarriage of justice is the Sally Clark trial. See Green (2002) or Bram (2014, p.55)

* Only a very small change in some test statistic is required to move from a observed significance level of 0.51 (non-significant) to 0.49 (significant). According to Matthews (1998), even Fisher when asked why his figure of 0.05 was a safe threshold at which to declare a result as significant, admitted he did not know at all, and that he simply chose 0.05 because it was convenient.

† The same hypothesis was not rejected in the smaller samples resulting from considering only the last years of data.

Chapter 3

Bayesian Digit Analysis: Theoretical Framework

> "*If we have no information relevant to the actual value of the parameter, the (prior) probability must be chosen so as to express the fact that we have none.*"
> Harold Jeffreys (1967)

3.1 Introduction

In a collection of N nonzero numbers there are N first digits, assumed to have been generated according to some 9-variate multinomial density $f(\boldsymbol{x}|\boldsymbol{\theta})$, and M second digits ($M \leq N$) assumed to have been generated according to some 10-variate multinomial density $f(\boldsymbol{y}|\boldsymbol{\xi})$. Because this dissertation focus in BL1 and BL2 based DA, what we want is to check whether or not, in a given sample, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ (the parameter vectors from the multinomial DGPs responsible for the occurrence of first and second digits) are as postulated by BL (tables 2.1 and 2.2, respectively).

Section 3.2 introduces the Multinomial $\wedge$ Dirichlet model, a Bayesian alternative to the classical joint goodness-of-fit statistical tests commonly used in BLDA.* Because joint

* The most common are: the chi-square test [see Murteira and Antunes (2012, p.460)], the

goodness-of-fit tests evaluate conformance to BL frequencies as a whole but do not identify which frequencies are in agreement to BL in a given sample and which are not, Nigrini (2000) incorporated the z-test in DA, which applies individually to each frequency. Section 3.3 introduces the Binomial ∧ Beta model, a univariate version of the Multinomial ∧ Dirichlet model, as an alternative to the z-test in DA.

3.2 Bayes Factor: Multinomial ∧ Dirichlet Model

Consider the random vectors $\boldsymbol{X} = (X_1, \ldots, X_k)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, respectively defined in the subspaces $\chi = \{(x_1, \ldots, x_k) : x_i \in \mathbb{N}_0, \sum_{i=1}^{k} x_i \leq N\}$ and $\Theta = \{(\theta_1, \ldots, \theta_k) : \theta_i \in (0,1), \sum_{i=1}^{k} \theta_i < 1\}$, where N is the fixed sample size and Θ is the k-dimensional simplex, $\mathcal{S}_k$. Let $\boldsymbol{X}$ follow a k-category multinomial distribution with unknown parameter vector $\boldsymbol{\theta} \in \Theta$: $X|\boldsymbol{\theta} \sim \mathcal{M}_k(N, \boldsymbol{\theta})$:

$$f(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{N!}{\prod_{i=1}^{k+1} x_i!} \prod_{i=1}^{k+1} \theta_i^{x_i} \tag{3.1}$$

with $\boldsymbol{x} \in \chi$, $x_{k+1} = N - \sum_{i=1}^{k} x_i$ and $\theta_{k+1} = 1 - \sum_{i=1}^{k} \theta_i$.

Let the observed counts of significant digits be represented by $\boldsymbol{x} = (x_1, \ldots, x_k)^*$, a realization of $X|\boldsymbol{\theta}$ which is assumed to have arisen under one of two possible and mutually exclusive states of the world (hypothesis/models) relative to $\boldsymbol{\theta}$: H_0 with prior probability $P(H_0) = \pi_0$ or H_1 with prior probability $P(H_1) = 1 - \pi_0$. Because the null hypothesis usually represents the established theory, in the specific problem being addressed H_0 represents conformity to BL. Let $\boldsymbol{\theta_0} = (\theta_{01}, \ldots, \theta_{0k}) \in \Theta$ be the parameter vector of the multinomial PMF (3.1) under H_0. Considering the subspace $\Theta_1 = \Theta \setminus \{\boldsymbol{\theta_0}\}$, the hypotheses being compared are:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0} \quad \text{vs} \quad H_1 : \boldsymbol{\theta} \in \Theta_1 \tag{3.2}$$

Kolmogorov-Smirnov [see Massey Jr (1951) or Conover and Conover (1999, p.428)] and the Kuiper test (Kuiper, 1960), both used with Stephens (1970) correction factor and with Morrow's (2014) BL specific critical values, the Conover (1972) test, and BL specific m-statistic (Leemis, Schmeiser, and Evans, 2000) and d-statistic (Cho and Gaines, 2007) to whom Morrow (2014) also computed critical values. For a review of the properties of these tests in BLDA see Morrow (2014).

* In BL1 analysis $k = 8$, x_1 represents the count of ones as the first digit, x_2 represents the count of twos and so on. In BL2 analysis $k = 9$, x_1 represents the count of ones as the second digit, x_2 the count of twos and so on.

From Berger and Pericchi (2001), we know that the BF in favour of H_0 is obtained through $B_{01}(\boldsymbol{x}) = \frac{m_0(\boldsymbol{x})}{m_1(\boldsymbol{x})}$, where $m_i(\boldsymbol{x})$ is the marginal density of the data under H_i:

$$m_i(\boldsymbol{x}) = \int_{\Theta_i} f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|H_i)\, d\boldsymbol{\theta} \quad (i = 0, 1) \tag{3.3}$$

where $\pi(\boldsymbol{\theta}|H_i)$ is the prior distribution of $\boldsymbol{\theta}$ under H_i, Θ_i is the parameter space of $\boldsymbol{\theta}$ under H_i and $f(\boldsymbol{x}|\boldsymbol{\theta})$ is the likelihood of $\boldsymbol{\theta}$ for a given $\boldsymbol{x}$. Note that $\Theta_0 = \boldsymbol{\theta_0}$ and as already defined $\Theta_1 = \Theta \setminus \{\boldsymbol{\theta_0}\}$. Because H_0 is a point null hypothesis,

$$\pi(\boldsymbol{\theta}|H_0) = 1_{\boldsymbol{\theta_0}}(\boldsymbol{\theta}) \tag{3.4}$$

where $1_{\boldsymbol{\theta_0}}(\boldsymbol{\theta})$ is the indicator function of $\boldsymbol{\theta}$ in $\{\boldsymbol{\theta_0}\}$. Because the Dirichlet family of distributions is the conjugate prior of the Multinomial distribution (Turkman and Paulino, 2015), and also because there is a particular distribution in the Dirichlet family that corresponds to the Berger, Bernardo, and Sun (2015) overall objective prior, a k-variate Dirichelet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k+1}) \in {\mathbb{R}_+}^{k+1}$ will be assumed for the prior distribution of $\boldsymbol{\theta}$ under H_1, i.e. $\boldsymbol{\theta}|H_1 \sim \text{Dir}_k(\boldsymbol{\alpha})$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}, H_1) = \frac{\Gamma(\sum_{i=1}^{k+1} \alpha_i)}{\prod_{i=1}^{k+1} \Gamma(\alpha_i)} \prod_{i=1}^{k+1} {\theta_i}^{\alpha_i - 1} d\boldsymbol{\theta} \tag{3.5}$$

where $\Gamma(\cdot)$ is the Gamma function. This prior is centered on $E(\boldsymbol{\theta}|\boldsymbol{\alpha}, H_1) = \frac{\boldsymbol{\alpha}}{\sum_{i=1}^{k+1} \alpha_i}$ and is symmetric if $\alpha_i = \alpha\ \forall i$. Using 3.1, 3.4 and 3.5 on 3.3:*

$$m_0(\boldsymbol{x}) = \frac{N!}{\prod_{i=1}^{k+1} x_i!} \prod_{i=1}^{k+1} {\theta_{0i}}^{x_i} \tag{3.6}$$

$$m_1(\boldsymbol{x}) = \frac{N!}{\prod_{i=1}^{k+1} x_i!} \frac{B(\boldsymbol{\alpha} + \boldsymbol{x})}{B(\boldsymbol{\alpha})} \tag{3.7}$$

where $B(\boldsymbol{\alpha}) = \int_{\Theta_1} \prod_{i=1}^{k+1} {\theta_i}^{\alpha_i - 1} d\boldsymbol{\theta}$ and $B(\boldsymbol{\alpha} + \boldsymbol{x}) = \int_{\Theta_1} \prod_{i=1}^{k+1} {\theta_i}^{\alpha_i + x_i - 1} d\theta$ are multivariate Beta functions. Finally, dividing 3.6 by 3.7†

$$B_{01}(\boldsymbol{x}) = \frac{\prod_{i=1}^{k+1} ({\theta_{0i}}^{x_i}) \prod_{i=1}^{k+1} [\Gamma(\alpha_i)] \Gamma[\sum_{i=1}^{k+1} (\alpha_i + x_i)]}{\Gamma(\sum_{i=1}^{k+1} \alpha_i) \prod_{i=1}^{k+1} \Gamma(\alpha_i + x_i)} \tag{3.8}$$

* Derivation of 3.7 in appendix A.1.
† Derivation of 3.8 in appendix A.1.

with $\theta_{0k+1} = 1 - \sum_{i=1}^{k} \theta_{0i}$. Besides the digit counts, $x_1, \ldots, x_{k+1}$, and the null hypothesis multinomial probabilities $\theta_{01} \ldots, \theta_{0k+1}$, this BF depends on the Dirichlet hyperparameters, $\alpha_1, \ldots, \alpha_{k+1}$, whose specification will be discussed in section 3.4.

3.3 Bayes Factor: Binomial $\wedge$ Beta Model

From the $X|\boldsymbol{\theta} \sim \mathcal{M}_k(N, \boldsymbol{\theta})$ assumption, it follows that each element in $\boldsymbol{X}$ is Binomially distributed: $X_i \sim \text{Bin}(N, \theta_i) \, \forall i \in \{1, \ldots, k\}$. When assessing conformance to one of BL frequencies individually, a sample of N numbers corresponds the realization of N Bernoulli trials, where a success is the occurrence of the first digit whose frequency is being assessed.

Consider $\boldsymbol{Y} = (Y_i, \ldots, Y_N)$, a vector of N independent Bernoulli random variables, $Y_i \sim \text{Bin}(1, \theta)$, for some unknown $\theta \in \Theta = (0, 1)$. Then $X = \sum_{i=1}^{N} Y_i \sim \text{Bin}(N, \theta)$. Let $\boldsymbol{y} = (y_i, \ldots, y_N)$ be a realization of $\boldsymbol{Y}$ and $x = \sum_{i=1}^{n} y_i$ the corresponding realization of X. The likelihood of θ for a given x in a sample of fixed size N is:

$$f(x|\theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x} \tag{3.9}$$

The data, x, is assumed to have arisen under one of two possible and mutually exclusive hypothesis relative to θ: H_0 with prior probability $P(H_0) = \pi_0$ or H_1 prior probability $P(H_1) = 1 - \pi_0$. Again, H_0 represents conformity to BL. Let $\theta_0 \in (0, 1)$ be the parameter of the binomial PMF (3.9) under H_0, and define $\Theta_1 = (0, 1) \setminus \{\theta_0\}$. The hypothesis being compared are:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1 \tag{3.10}$$

Because the Beta family is the conjugate prior of the binomial distribution (Turkman and Paulino, 2015), and considering the fact that H_0 is a point null hypothesis, the assumed prior distribution for θ is:

$$\pi(\theta|H_i) = \begin{cases} 1_{\theta_0}(\theta) & \text{if } i = 0 \\ \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} & \text{if } i = 1 \end{cases} \tag{3.11}$$

where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(b+b)} = \int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta$ is the beta function, $1_{\theta_0}(\theta)$ is the indicator function of θ_0 in Θ, $\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$ is the Probability Density Function (PDF) of a Beta(a, b)

distribution and $E(\theta|a,b,H_1) = \frac{a}{a+b}$. The marginal densitiy of the data under each hypotheses is obtained through a univariate version of 3.3:

$$m_i(x) = \int_0^1 f(x|\theta)\pi(\theta|H_i)\,d\theta \;\; (i = 0, 1) \tag{3.12}$$

Combining 3.11 with 3.12 and considering the hypothesis in 3.10:

$$m_0(x) = \binom{N}{x} {\theta_0}^x (1-\theta_0)^{N-x} \tag{3.13}$$

$$m_1(x) = \binom{N}{x} \frac{B(x+a, n-x+b)}{B(a,b)} \tag{3.14}$$

and finally, dividing 3.13 by 3.14, the BF in favor of H_0 is:*

$$B_{01}(x) = \frac{\theta_0(1-\theta_0)^{N-x}\,\Gamma(a)\,\Gamma(b)\,\Gamma(n+a+b)}{\Gamma(a+b)\,\Gamma(n+a-x)\,\Gamma(x+a)} \tag{3.15}$$

The specification of the hyperparameters a and b will be discussed in the next section.

3.4 Prior Distribution Specification

The criticism that Bayesian methods require subjective prior specification has been effectively answered by the development of objective Bayesian methods based on non informative priors (Kass and Wasserman, 1996). However, Berger and Delampady (1987) note that such methods are not always available, and that precise null hypothesis testing is an example of a situation where objective procedures do not exist, because even thought one can avoid prior probabilities specification by focusing on BFs, there is no prior distribution specification that can claim to be objective. Nevertheless, there are properties to impose on a prior distribution for it to be considered adequate. Berger and Delampady (1987, 1990) argue that an adequate prior for precise null hypothesis testing should be uni-modal, symmetric about the null parameter value or at least centered on it, and non increasing about that same point†, to acknowledge the central role of the null parameter value (representing the established theory), and avoid treating parameter values other than that as special.

---

* Derivations of 3.13, 3.14 and 3.15 in appendix A.2.

† i.e. symmetric about (or centered in) $\boldsymbol{\theta_0}$, non increasing around $\boldsymbol{\theta_0}$ for the Multinomial $\wedge$ Dirichlet Model and symmetric about (or centered in) θ_0, non increasing around θ_0 for the Binomial $\wedge$ Beta Model.

In the Multinomial $\wedge$ Dirichlet Model, if $\boldsymbol{\alpha} = \mathbf{1}$ is set [like in Pericchi and Torres (2011) and Torres (2006)], 3.5 reduces the PDF of a uniform distribution on $\mathcal{S}_k$, which is in conformance to the Bayes-Laplace principle of indifference: assuming equiprobability of events in the absence of prior knowledge (Syversveen, 1998). For a uniform prior to be obtained in the Binomial $\wedge$ Beta Model set $a = b = 1$. The fact that uniform priors are not invariant to re-parametrizations (Paulino, Turkman, and Murteira, 2003) is not problematic in DA as we are only concerned about parameters, not functions of them. Uniform priors are symmetric, and non increasing in all their domain, hence being non increasing around the null parameter values. For BLDA they are not centered on the null parameter values*.

Instead, if $\boldsymbol{\alpha} = \mathbf{1}\left(\frac{1}{k+1}\right)$ is set in the Multinomial $\wedge$ Dirichlet Model, 3.5 becomes the PDF of a $\text{Dir}_k\left[\mathbf{1}(\frac{1}{k+1})\right]$, which is the overall objective prior (Berger, Bernardo, and Sun, 2015) for the multinomial distribution. Equivalently, if $a = b = \frac{1}{2}$ is set on the Binomial $\wedge$ Beta Model, the reference prior for the binomial distribution is obtained (Yang and Berger, 1996), and because the Binomial $\wedge$ Beta is a single parameter model, the reference prior is equivalent to the Jeffreys prior, which unlike the uniform prior is invariant to reparametrizations (Yang, 1995). The overall objective prior and the reference prior obtained above are symmetric†, but centered in points other than BL null‡. Moreover, because they are symmetric about their mean, they also fail being non increasing around the null parameter value (for $a = b = \frac{1}{2}$ the Beta distribution is U-shaped and symmetric about $\frac{1}{2}$). Therefore, even though the reference prior has the appealing conceptual interpretation of materializing a truly non-informative prior [Bernardo (1979), Berger and Bernardo (1992-a, 1992-b), Berger, Bernardo and Sun (2009)] and although both the reference and the overall objective prior have desirable properties for posterior distribution-based inference (Berger, Bernardo, and Sun, 2015), they are designed for estimation and may not be adequate for this particular BMS problem, as they spread prior density around parameter values other than the null parameter values.

To center a Dirichlet prior on $\boldsymbol{\theta_0}$ is only possible assuming $\boldsymbol{\alpha} = c\,\boldsymbol{\theta_0}$ for some $c > 0$ (Delampady and Berger, 1990). However, if $\boldsymbol{\theta_0}$ is the vector with BL1 or BL2 probabilities

---

* The uniform prior is centered on $\mathbf{1}(\frac{1}{k+1})$ in the Multinomial $\wedge$ Dirichlet Model and on $\frac{1}{2}$ in the Binomial $\wedge$ Beta Model.

† The Dir_k distribution is symmetric when $\alpha_i = \alpha\, \forall i \in \{1, \ldots, k+1\}$, and the Beta (a,b) when $a = b$.

‡ Because $\pi(\boldsymbol{\theta}) \sim \text{Dir}_k(\boldsymbol{\alpha}) \Rightarrow E(\boldsymbol{\theta}) = \frac{\boldsymbol{\alpha}}{\sum_{i=1}^{k+1} \alpha_i}$, the overall objective prior is centered on $\mathbf{1}(\frac{1}{k+1})$, and because $\theta \sim \text{Beta}(a, b) \Rightarrow E(\theta) = \frac{a}{a+b}$ the reference prior is centered on $\frac{1}{2}$.

and $\boldsymbol{\alpha} = c\,\boldsymbol{\theta_0}$ is assumed, the resulting Dirichlet prior is not symmetric, as $\alpha_i \neq \alpha_j$ for $i \neq j$. Equivalently, to center the Beta prior on θ_0, the necessary assumption are $a = s\,\theta_0$ and $b = s\,(1-\theta_0)$ for some $s > 0$, which again results in a asymmetrical prior for the specific case of BL1 and BL2 analysis*. Hence, to use the BFs in 3.8 or 3.15, we have to give up either on the prior being symmetric or in the prior being centered on the null parameter value. Giving up on symmetry seems to have a lower cost: although in the absence of prior knowledge it is desirable to have a symmetric prior, as there is no particular reason to skew the prior density towards any particular region in the parameter space, in BLDA the null corresponds to a established theory, so it might be acceptable to skew the prior distribution towards a region in parameter space suggested by that theory. On the other hand, there is no justification for the prior density to be centered around a point other than the null parameter value.

Because the Dirichlet parameters affect the dispersion of the prior density and consequently define how informative the prior distribution is, so does the choice of c. Because a $\text{Dir}_k(\boldsymbol{\alpha})$ distribution for $\boldsymbol{\theta}$ implies a $\text{Beta}\left(\alpha_i, \sum_{j=1}^{k+1}\alpha_j - \alpha_i\right)$ marginal distribution for each parameter in $\boldsymbol{\theta}$, if $\boldsymbol{\theta} \sim \text{Dir}_k(c\,\boldsymbol{\theta_0})$ then $\theta_i \sim \text{Beta}\left(c\,\theta_{0i}, c - c\,\theta_{0i}\right)$ and $\text{Var}(\theta_i) = \frac{\theta_{0i}(1-\theta_{0i})}{c+1}$. Note that $\text{Var}(\theta_i)$ is a decreasing function of c, and hence smaller values of c are preferable, for the prior to be as least informative as possible†. Equivalently, the values of a and b affect the shape of the beta prior, and so does the choice of s. For the beta prior to have an adequate shape (unimodal, centered on θ_0 and non-increasing around θ_0), besides $a = s\,\theta_0$ and $b = s\,(1-\theta_0)$ it is necessary to have $s\,\theta_0 > 1$ and $s\,(1-\theta_0) > 1$. The smallest value of $s \in \mathbb{N}$ verifying those conditions for BL1 analysis is $s = 22$‡. For BL2 analysis $s = 12$ is enough. Again, smaller values of s are preferable.

An additional property that might be useful to impose in the Dirichlet prior is that after marginalizing it, the beta marginal distribution for each $\theta_i\,(i = 1, \ldots, k+1)$ have the already discussed properties making them adequate prior distributions for $\theta_i|H_1$ in the Binomial $\wedge$ Beta model. This constrains even more the choice of the Dirichlet parameters. What is necessary

---

* The Beta(a, b) is only symmetric if $a = b$. When $a = s\,\theta_0$ and $b = s\,(1-\theta_0)$, it is only possible to have $a = b$ if $\theta_0 = \frac{1}{2}$. No parameter is assumed to have such value under the BL1 or BL2 null hypotheses.

† The Dirichlet parameters can be interpreted as prior multinomial pseudo counts that subsequently will be added to the observed counts, smoothing the weight of the likelihood. The larger c is, the larger all the parameters in α are and consequently the more informative the prior distribution is.

‡ To confirm, just note that multiplying any entry from table 2.1 by 22 yields a number greater than one.

is a $\text{Dir}_k(\boldsymbol{\alpha})$ distribution with $\boldsymbol{\alpha} = c\,\boldsymbol{\theta_0}$, and because $\text{Dir}_k(c\,\boldsymbol{\theta_0}) \Rightarrow \theta_i \sim \text{Beta}\,(c\,\theta_{0i}, c - c\,\theta_{0i})$ it is also necessary that c is such that $c\,\theta_{0i} > 1$ and $c\,(1 - \theta_{0i}) > 1$, $\forall i \in \{1, 2, \ldots, k+1\}$. A Dirichlet distribution fulfilling these requirements is a conjugate prior (for the multinomial likelihood in the Multinomial $\wedge$ Dirichlet model), centered on $\boldsymbol{\theta_0}$, non-increasing around $\boldsymbol{\theta_0}$ and implies a unimodal, centered on θ_{0i}, non-increasing around θ_{0i} beta marginal distribution for each θ_i, which are conjugate prior to the binomial likelihood in the Binomial $\wedge$ Beta model. The smallest such value of c should be considered, which from the discussion above is $c = 22$ for BL1 analysis and $c = 12$ for BL2 analysis.

3.5 Posterior Probabilities

One can avoid the specification of prior probabilities for the hypotheses by focusing solely on BFs. However, to compute PPs for the hypotheses, prior probabilities have to be assumed. The BFs can be used to compute the PP of the null hypothesis being true [see Berger and Sellke (1987)]:

$$P(H_0|\boldsymbol{x}) = \left[1 + \frac{1 - \pi_0}{\pi_0}\, B_{01}(\boldsymbol{x})^{-1}\right]^{-1} \tag{3.16}$$

where $\pi_0 = P(H_0)$ and $1 - \pi_0 = P(H_1)$ are the prior probabilities of the null and of the alternative, respectively. Berger and Sellke (1987) consider that the objective choice of π_0 is $\frac{1}{2}$, even tough some might argue that π_0 should be larger, as H_0 usually represents the established theory. Torres (2006) and Pericchi and Torres (2011) used $\pi_0 = \frac{1}{2}$ when testing for BL in their works mentioned in section 2.7. To set such a value for π_0 is in conformance to the principle of indifference, and results in the BF being equal to the posterior odds of H_0 relative to H_1. The relation in 3.16 applies directly to the BF in 3.8, and if $\boldsymbol{x}$ is replaced by x it also applies to the one in 3.3.

3.6 Lower Bounds

Because the BFs in 3.8 and 3.15 require the specification of a prior distribution, the PPs obtained by using 3.16 on them are affected by prior specification. Although Berger and Delampady (1987) consider that in precise null hypothesis testing there is no choice of prior

distribution that can claim to be objective, it is possible to impose objective restrictions on the family of prior under consideration, and find objective lower bounds on BFs and PPs:

$$\underline{B}_{\Pi}(\boldsymbol{x}) = \inf_{\pi \in \Pi} B^{\pi}(\boldsymbol{x}) \tag{3.17}$$

$$\underline{P}_{\Pi}(H_0|\boldsymbol{x}) = \left(1 + \frac{1-\pi_0}{\pi_0}\, \underline{B}_{\Pi}(\boldsymbol{x})^{-1}\right)^{-1} \tag{3.18}$$

where $B^{\pi}(\boldsymbol{x})$ is the BF (3.8) that is obtained when a distribution π from a family of candidate distributions Π is considered as the prior, $\underline{B}_{\Pi}(\boldsymbol{x})$ is the lower bound on $B^{\pi}(\boldsymbol{x})$, obtained when the prior $\pi \in \Pi$ that maximizes $m_1(\boldsymbol{x})$ is considerer, and $\underline{P}_{\Pi}(\boldsymbol{x})$ is the lower bound on $P(H_0|\boldsymbol{x})$, resulting from using 3.16 on $\underline{B}_{\Pi}(\boldsymbol{x})$. For the BF in 3.15, just replace $\boldsymbol{x}$ by x. For 3.17 and 3.18 to be interpreted as objective lower bounds, the family Π should be large enough as to contain all reasonable prior distributions, and thus minimizing specification subjectivity, but should also have restrictions to exclude nonsensical distributions that would bias the lower bounds against H_0. Berger and Sellke (1987) showed that the family $\Pi = \{\text{all distributions}\}$ unduly biases conclusions against H_0*, and so does $\Pi = \{\text{all symmetric distributions}\}$. They propose as objective restrictions on Π that π is unimodal or (equivalently in the presence of symmetry) non increasing around the null parameter value, so that no parameter values other than that are favoured. Sellke, Bayarri, and Berger (2001) argue that the family $\Pi_{US} = \{\text{Unimodal } \pi, \text{symmetric about the null parameter value}\}$ contains virtually all objective priors, and that no density in this class is absurd. They also consider the family $\Pi_{CU} = \{\text{Conjugate prior } \pi, \text{which under } H_1 \text{ are centered on the null parameter value}\}$ to produce satisfactory results, even though it is more restricted than Π_{US} and may exclude some reasonable distributions. Delampady and Berger (1990) considered both Π_{US} and Π_{CU} to be objective classes because they acknowledge the central role of the null parameter value and spread prior density around it in ways not biased towards particular alternatives. For testing multinomial model parameters, they showed that these classes produce very similar lower bounds and derived formulas to compute $\underline{B}_{\Pi}$. Berger and Delampady (1987) derived the formulas to compute lower bounds in the Binomial case. If Π is representative enough,

* See Edwards, Lindman, and Savage (1963): even with this choice of Π, $\underline{P}_{\Pi}(H_0|\boldsymbol{x})$ is still often larger than the p-value, indicating that even extreme bias towards H_1 in a Bayesian analysis often results in less evidence against H_0 than would appear to have been obtained with the p-value.

the lower bounds on the PP of the null can claim to be objective, and a large lower bound indicates that the data does not constitute strong evidence against the null, even if the p-value is small. As Berger and Sellke (1987) state, a small p-value does not necessarily indicate the presence of strong evidence against the null. Goodman (1999-b) considers lower bounds on CME to be even more objective than p-values, because they are unaffected by hypothetical long run frequentist results that make the p-values uncertain.

Sellke, Bayarri, and Berger (2001) developed a p-value calibration which allows p-values to be interpreted in either a Bayesian or a frequentist way. Those calibrations can also be used to compute lower bounds on BFs and PPs. Instead of depending directly on the sample like the lower bounds in Berger and Delampady (1987, 1990), those lower bounds require only a p-value (p_{obs}) that is valid as input:

$$\underline{B}(p_{obs}) = \begin{cases} -e\, p_{obs}\, \ln(p_{obs}) & \text{if } p_{obs} < \frac{1}{e} \\ 1 & \text{if } p_{obs} \geq \frac{1}{e} \end{cases} \tag{3.19}$$

$$\underline{P}(p_{obs}) = \begin{cases} [1 + \underline{B}(p_{obs})^{-1}]^{-1} & \text{if } p_{obs} < \frac{1}{e} \\ \frac{1}{2} & \text{if } p_{obs} \geq \frac{1}{e} \end{cases} \tag{3.20}$$

where $\underline{B}(p_{obs})$ is interpreted as the lower bound on the BF in favour of H_0 and $\underline{P}(p_{obs})$ can be interpreted as either a lower bound on the type I error conditional probability in the rejection of H_0 or as the lower bound on the PP of H_0 arising from the use of 3.16 on the BF in 3.19 together with the assumption that $\pi_0 = \frac{1}{2}$. The lower bounds on BFs and PPs obtained with this p-value calibration are very similar the ones obtained in Berger and Sellke (1987) with Π_{US} (the family argued to contain all objective priors), giving strong support to the appropriateness of the calibration, which also has the advantage of converting p_{obs} into a more intuitive scale. With this calibration one need not fear misinterpretation of a Frequentist error rate probability as the probability of the hypothesis being true, as they coincide. Pericchi and Torres (2011) describe 3.20 as a useful way to calibrate p-values under a robust Bayesian perspective but warn that because this calibration does not depend on sample size, for large samples it may be very conservative. A full p-value correction requires a BF and the corresponding PP of H_0.

Chapter 4

Bayesian Digit Analysis: Empirical Application

> "*The ancients knew very well that the only way to understand events was to cause them.*"
> Nassim Taleb (2010)

4.1 Overview

This empirical application is meant to show that the concerns raised in section 2.7 regarding the classical hypothesis testing framework do arise in DA, as well as to demonstrate the applicability of the alternative methodologies suggested in chapter 3. The focus is on macroeconomic data, which like accounting data is susceptible of being manipulated. Rauch, Göttsche, Brähler, and Engel (2011) warn that macroeconomic statistics can be used by governments to portray a more favourable picture of their countries economic situation, either to archive preferential conditions in capital markets, or just for popularity purposes. The pressure for European Union's governments to comply with the Stability and Growth Pact Criteria (SGPC) is an additional incentive for them to manipulate macroeconomic statistics, so that sanctions are avoided. There is the necessity to develop effective methods to screen such data for manipulation, and BLDA is one of such possible methods.

4.2 Data

The data was extracted from the Eurostat Database* in July 2016 through the directory: Database by themes → Economy and finance → Government statistics → Government finance statistics → Government deficit and debt → Government deficit/surplus, debt and associated data. For each country being analysed, all numbers from the 38 tables in this category were aggregated in a dataset. Each of those datasets provides two samples: one of first digits and one of second digits. This category was selected because, as Rauch, Göttsche, Brähler, and Engel (2011) note, it is related to public deficit and public debt, which are variables that are used in the calculation of the SGPC relevant criteria such as deficit ratio and debt ratio. The period under consideration is from 1999 to 2015, with 1999 being the starting point because it is the year in which the Euro was introduced as book money. The unit of measurement is Million Euro for all entries. Only countries that joined the Eurozone prior to 2006 were selected, so that at least 10 years of data were available. Samples sizes and some descriptive statistics of the data can be found in table C.2. Matthews (1999) warns that a sample of numbers should be big enough to give the predicted proportions a chance to assert themselves. The considered datasets should be big enough. Divergence in sample sizes between countries is due to missing data, and also because entries equal to zero were removed from the samples. For the same country, the samples used for BL1 and BL2 analysis may differ in size, because numbers with only one digit are considered in BL1 analysis but not in BL2 analysis. Considering the discussion in section 2.4, all sampless that will be subject to analysis have the properties that typically result in the emergence of BL: each sample consists in the aggregation of data from 38 different tables respecting to different economic and financial variables which are measured in amounts and are free of imposed limits, every sample's average is larger than its median, all dataset histograms are positively skewed, and the numbers are supposed to have been generated without human intervention.

* http://ec.europa.eu/eurostat/data/database

4.3 Study Design

Because these datasets are expected to conform to BL, large deviations from it should raise concern about the process that generated them, namely it may suggest that they were not generated by a natural process (i.e. without human intervention). Hence, BLDA can help detecting which datasets are most likely to have been manipulated.

Conformance to BL will be evaluated on each sample from table C.2, using the Multinomial $\wedge$ Dirichlet model. Three BFs in favour of the null hypothesis (3.2) will be computed for each sample (using 3.8): one with $\alpha_i = 1\,\forall i$, resulting in a uniform prior [as in Ley (1996), Torres (2006) and Pericchi and Torres (2011)], one with $\alpha_i = \theta_{0i}$ resulting in a $\text{Dir}_k(\boldsymbol{\theta_0})$ prior, the least informative Dirichlet centered on $\boldsymbol{\theta_0}$, and because of the discussion in 3.4 about the properties of the marginal distributions of the joint prior, one with either $\alpha_i = 22\,\theta_{0i}$ (BL1 samples) or $\alpha_i = 12\,\theta_{0i}$ (BL2 samples) resulting in a $\text{Dir}_k(22\,\boldsymbol{\theta_0})$ prior for BL1 analysis and a $\text{Dir}_k(12\,\boldsymbol{\theta_0})$ prior for BL1 analysis. Equation 3.16 will be applied to each BF in order to compute the corresponding PP. For each sample, the p-value [see Murteira, Ribeiro, et al. (2010, p.416)] from the chi-square test on the null hypothesis in 3.2 will be provided. Those p-values will be calibrated into lower bounds on the PPs using 3.20.

Samples with $P(H_0|\boldsymbol{x}) = 0$ will be analysed with the Binomial $\wedge$ Beta model and with the z-test to identify which frequencies diverge from BL and to which extent. For each such sample, two BFs (using 3.15) in favour of the null (3.10) will be computed for each leading digit frequency: one with $a = b = 1$ resulting in a uniform prior, and one with either $a = 22\,\theta_0$ and $b = 22 - 22\,\theta_0$ (for BL1 samples) or $a = 12\,\theta_0$ and $b = 12 - 12\,\theta_0$ (BL2 samples), resulting in the Beta marginal distributions implied by the $\text{Dir}_k(22\,\boldsymbol{\theta_0})$ and $\text{Dir}_k(12\,\boldsymbol{\theta_0})$ priors used in the Multinomial $\wedge$ Dirichlet model. PPs will be obtained using 3.16, and the lower bounds on them using 3.20 on the p-value from Nigrini's (2000) z-test, based on the statitics:

$$z_i = \frac{|\theta_i - \theta_{0i}| - \frac{1}{2N}}{\sqrt{\frac{\theta_{0i}(1-\theta_{0i})}{N}}} \overset{a}{\sim} \text{Normal}(0,1) \tag{4.1}$$

where $\frac{1}{2N}$ is the continuity correction factor.

BFs will be presented in logarithms, rounded to two decimal figures, and will be interpreted their according to the scale in table C.1. PPs and respective lower bounds, p-values,

variances and standard deviations will be rounded to five decimal figures. Hyperparameter standard deviations will be the preferential measure of how informative a prior distribution is. Unless otherwise stated, CHT is conducted at the dimension of 0.05. BMS prefers the null hypothesis if $P(H_0|\,\text{data}) > 0.5$ [or $B_{01} > 1$ or equivalently $\ln(B_{01}) > 0$] and prefers the alternative otherwise. Because in BLDA the null hypothesis corresponds to a established theory and is used as proxy for the status quo (absence of fraudulent or erroneous data), stronger evidence against the null than $P(H_0|\,\text{data}) < 0.5$ should be required for the null to be rejected. The rejection rule that will be used is $P(H_0|\,\text{data}) < 0.05$, which is the rule people incurring in the p-value fallacy think they are applying.

This study has two goals. The first is to compare the results of the Classical and the Bayesian approaches to BLDA. Bayesian CME will be compared to their Classical counterparts (p-values) to look for divergences in the conclusions drawn. Because of the discussion in section 2.7, CHT is expected to reject BL in samples where both Bayesian methods and graphical inspection suggest otherwise. The second is to try BLDA using prior distributions centered on the null parameter value and non-increasing around it, as Berger and Delampady (1987, 1990) consider adequate for precise hypothesis testing, and in particular try the unified approach discussed in section 3.4 (using as prior in the Binomial $\wedge$ Beta Model the marginal distributions implied by the joint prior of the Multinomial $\wedge$ Dirichlet Model). All hyperparameter variances and standard deviations will be computed.

4.4 Study Results

We developed a VBA macro to perform the computations on the data (see appendix D). A graphical comparison between each dataset's observed first and second digit counts and BL postulated counts can be found in Section B.1. The p-values from the chi-square tests and the results from the Multinomial $\wedge$ Dirichlet model can be found in tables C.3 ($\alpha_i = 1$), C.4 ($\alpha_i = \theta_{0i}$) and C.5 (for $\alpha_i = 22\,\theta_{0i}$). Hyperparameter variances and standard deviations from the Multinomial $\wedge$ Dirichlet can be found in tables C.14 (for $\alpha_i = 1$), C.15 ($\alpha_i = \theta_{0i}$) and C.16 (for $\alpha_i = 22\,\theta_{0i}$). The datasets that will be analysed with the Binomial $\wedge$ Beta model and with the z-test are Austria BL1, Ireland BL1, Luxembourg BL1 and Portugal BL1. The z-test

p-values and the Binomial $\wedge$ Beta model results with the Beta $(1, 1)$ prior (i.e. uniform) can be found in tables C.6 (Austria), C.7 (Ireland), C.8 (Luxembourg) and C.9 (Portugal) and with the Beta $(22\,\theta_0, 22 - 22\theta_0)$ prior in tables C.10 (Austria), C.11 (Ireland), C.12 (Luxembourg) and C.13 (Portugal). The Binomial $\wedge$ Beta model hyperparameter variances and standard deviations can be found in tables C.17 [Beta $(1, 1)$ prior] and C.18 [Beta $(22\,\theta_0, 22 - 22\theta_0)$].

4.5 Discussion of the Results

The pictures in appendix B.1 support the idea that the first digit frequencies in the analysed samples are not uniform. All samples show a decreasing pattern in first digit counts, although not always monotonically decreasing like BL postulates. Considering Hill's (1995-b) theorem mentioned in section 2.4, it is not a surprise that the pooled samples exhibit the best fit to BL1 and BL2. This may also be due to the fact the pooled samples are much larger than the countries samples, giving the true frequencies a better chance to assert themselves.

Let's now discuss prior distribution specification. In tables C.10, C.11, C.12 and C.13 we can see that $P(H_0|data) = 1$ in all lines of the four tables, signalling that the Beta $(22\,\theta_0, 22 - 22\theta_0)$ prior is too informative. Prior density is so concentrated around θ_0 that the data is not able to shift posterior density away from θ_0 in any of the samples that were analysed with the Binomial $\wedge$ Beta model. In table C.18 we can see that the hyperparameter standard deviations look too small with this prior and when compared to the hyperparameter standard deviations of the uniform prior (table C.17) they are, roughly speaking, 3 to 4 times smaller. Emphasis will therefore be given to the Binomial $\wedge$ Beta model results obtained with the uniform prior and to the Multinomial $\wedge$ Dirichlet model results obtained with the uniform and $\text{Dir}_k(\boldsymbol{\theta_0})$ priors. Comparing tables C.14, C.15 and C.16 we can see that in the Multinomial $\wedge$ Dirichlet model the $\text{Dir}_k(\boldsymbol{\theta_0})$ is the least informative prior and the $\text{Dir}_k(22\,\boldsymbol{\theta_0})$ the most informative. The uniform prior is in the middle, but closer to the $\text{Dir}_k(22\,\boldsymbol{\theta_0})$.

Lets now analyse the Multinomial $\wedge$ Dirichlet model results and the p-values from the chi-square test. First, consider the uniform prior (see table C.3). Comparing each PP with the corresponding p-value, there is agreement in the rejection of the null at the traditional dimensions $(0.01, 0.05, 0.1)$ in the datasets of Austria BL1, Belgium BL1, Ireland

BL1, Luxembourg BL1 and Portugal BL1, and agreement in not rejecting the null in the datasets of Belgium BL2, Germany BL2, Ireland BL2, Portugal BL2 and Spain BL2.

In Austria BL2 and Finland BL2 the p-values indicate a significant deviation from BL at the dimension of 0.05 and not significant at 0.1, despite the high PP of the null (≈ 1) and the BFs claiming decisive evidence in favour of the null. Both datasets have a lower bound on the PP of the null greater than 0.3. In Greece BL2 and France BL2 we have $PPs \approx 1$ and BFs indicating decisive evidence in favour of the null but the p-values only reject the null at the dimension of 0.05 and not at 0.01. Again, the lower bounds on the PP indicate weak evidence against the null. In Netherlands BL2 and Pooled Sample BL2 the p-values indicate barely statistically significant deviations from BL at the dimension of 0.05 (not significant at 0.01) but the BFs indicates decisive evidence in favour of the null, $P(H_0|\,\text{data}) \approx 1$ for both datasets and the lower bounds on the PP of the null in both datasets is grater than 0.3.

The last paragraph illustrates the conflict between the p-value and the CME: a small p-value apparently signals strong evidence against the null but when BMS is used the same dataset may generate a large PP of the null and even the lower bound on the PP can be much larger than the p-value. Note that lower bounds are biased towards the alternative hypothesis and hence the true probability of the null is almost certainly larger than the lower bound. Also, notice by how much one underestimates the probability of the null when incurring in the p-value fallacy [interpreting the p-value as the $P(H_0|\,\text{data})$]. These results also illustrate the impact of the statistical significance dichotomy to DA: samples with good fit to BL may have the null rejected just because of an arbitrarily chosen evidence threshold.

In the Finland BL1, France BL2, Germany BL1, Greece BL1, Italy BL1, Italy BL2, Luxembourg BL2, Netherlands BL1, Spain BL1 and Pooled BL1 samples, the conclusions drawn with CHT at the traditional dimensions differ from those of BMS. All these samples have zero or nearly zero p-values and $P(H_0|\,\text{data}) > 0.9$ (except Spain BL1 which has $P(H_0|\,\text{data}) = 0.8079$). In most of these datasets the lower bounds on the PPs are small and seem to agree with the p-value, reinforcing the idea that the evidence in favour of the null is weak. Yet, one must be careful as these lower bounds are based on a p-value calibration which does not depend on sample size and hence can be very conservative in large samples. Nevertheless, the Netherlands BL1 sample is a good example of the inadequacy of the p-value

for precise hypothesis testing: if CHT is used, the p-value of 0.00916 rejects the null at the traditional significance levels but the three prior distributions used in this study suggest a PP of the null very close to one, which is decisive evidence in favour of the null. The lower bound indicates that there is at least a nearly 0.1 PP that the null is true. A similar situation occurs with Greece BL2. As for the Pooled Sample BL2, despite having a PP of the null of at least 0.30444 and $P(H_0|\,\text{data}) \approx 1$ in the three prior specifications, it has a p-value of 0.05579, which is close to the 0.05 rejection threshold and rejects the null at the dimension of 0.1.

Now consider the Pooled Sample BL1 dataset. It is the largest sample and the shows the best graphical fit to BL. Nevertheless, a p-value equal to zero indicates that the classical method finds strong evidence against the null, and the deviation from BL is considered significant at all significance levels, despite the high PP of the null obtained with the three prior specifications. The fact that $\underline{P}(H_0|\,\text{data})$ is also equal to zero is either due to the conservative nature of the p-value calibration in large samples or to the fact that the observed p-value which is input in the calibration is very close or equal to zero.*

With the $\text{Dir}_k(\boldsymbol{\theta_0})$ prior there are even more samples where the conclusions drawn from BMS differ from those of CHT: Austria BL1, Belgium BL1, Finland BL1, France BL1 and BL2, Germany BL1, Greece BL1, Ireland BL1, Italy BL1 and BL2, Luxembourg BL2, Netherlands BL1 and Spain BL1. All those samples have either zero or nearly zero p-values, $P(H_0|\,\text{data}) \approx 1$ and BFs claiming decisive evidence in favour of the null†. CHT and BMS agree in the rejection of the null in Luxembourg BL1 and Portugal BL1 and agree in not rejecting it in Belgium BL2, Germany BL2, Ireland BL2 and Portugal BL2. For Austria BL2, Finland BL2, Greece BL2, Netherlands BL2 and Pooled Sample BL2 the conclusions may or may not coincide depending on the dimension of the chi square test. Again, note that small p-values can be obtained in samples with good conformance to BL and that samples in which BMS finds strong evidence in favour of the null (high PPs or even high lower bounds on PPs) may have the null rejected because of an arbitrarily chosen evidence threshold.

---

* Some zeros in the result tables are actually decimal numbers rounded down to zero according to the rules defined in section 4.3. In such situations, the p-value that is input in the calibration is the decimal number, as the macro uses more decimal places than the five we are using. If the observed p-value is actually zero (or a decimal number sufficiently small for the macro to round it to zero), the macro is programmed to return a lower bound equal to zero, as it is the limit of the calibration as the p-value goes to zero. Otherwise the macro would return an error message as it couldn't compute the logarithm of zero.

† With the exception of Austria BL1 where the evidence in favour of the null is "just" very strong.

Lets now discuss the Binomial $\wedge$ Beta model and z-test results. For Austria BL1 (see table C.6) the classical method finds the most significant deviations from BL in the counts of fives and sevens (both p-values equal to zero). CME agree, as those are the counts with lowest PP of having been generated by BL, both with nearly zero probability. CHT also rejects BL in the count of ones and fours. CME disagrees and assigns a PP of conformance to BL of 0.54162 to the count of ones and 0.19382 to the count of fours. If lower bounds are considered the PPs are now 0.1682 and 0.0244, respectively. This reduces the gap between the conclusions drawn with Classical and the Bayesian method but still results in considerable evidence in favour of the null in the count of ones, which is vehemently rejected with CHT.

In the Ireland BL1 dataset, CHT rejects the null in the counts of threes, fours, fives, sixes and nines. There is a conflict between the p-value and the CME in the count of fives: a p-value of 0.00994 results in the rejection of the null at the traditional dimensions and apparently is strong evidence against the null but the PP of the null is 0.45297 and its lower bound is 0.11018. In the count of nines the situation is even worse: a p-value of 0.03934 is sufficient to reject the null at the 0.05 significance level in CHT although the obtained PP is 0.80347 with a lower bound of 0.25707. The same goes for the count of threes: BL is rejected in CHT as the p-value of 0.04486 is on the 0.05 rejection region, but the uniform prior finds a PP of 0.77214 for BL, and it's lower bound is 0.27458. In the counts of sixes the conflict is less evident [p-value $= 0.00018$, $P(H_0|\,\text{data}) = 0.05526$, $\underline{P}(H_0|\,\text{data}) = 0.004139$].

In the Luxembourg BL1 dataset (see table C.8), CHT rejects the null in the counts of ones, twos and fives. The lowest p-values (equal to zero for the counts of ones and fives) are backed by very tiny PPs of the null, even when lower bounds are considered. In count of twos, the nearly zero p-value may conflict with the 0.04634 PP of the null, although if the lower bound is considered the PP is only 0.01442. Note that the cont of eights, despite having a p-value not far from the rejection threshold and which would result in the rejection of the null at the dimension of 0.1, has nearly 0.9 PP of the null, and a lower bound of 0.35 on that PP.

The Portugal BL1 sample seems to be less problematic (see table C.9). There is conformance in the rejection of BL in the counts of ones, twos and fours. The only conflict to highlight is in the count of sevens: a p-value that would call for a rejection of the null at the dimension of 0.1 contrasts with 0.97175 PP of the null, with a 0.5 lower bound.

Chapter 5

Conclusion

> "*Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise*"
> John Tukey (1962)

5.1 Conclusion

A Bayesian approach to Benford's law based digit analysis was suggested. The conflict between classical hypothesis testing and Bayesian model selection, and its consequences to Benford's law based digit analysis were explored through an empirical application using macroeconomic statistics from Eurozone countries. Combining the ideas exposed in chapters 2 and 3 with the results from the empirical application in chapter 4, the main conclusions to be drawn are:

- Graphical inspection suggests that, as suspected prior to the experience, the analysed samples of macroeconomic statistics show a decreasing pattern in leading digit frequencies. Digit analysis can therefore be an useful tool in the detection of erroneous or fraudulent collections of such numbers.
- The conflict between classical hypothesis testing and Bayesian model selection does arise in Benford's law based digit analysis and can be of severe consequences if one

is not aware of it: large samples in which both graphical inspection and conditional measures of evidence suggest agreement with Benford's law are likely to have the null (i.e. Benford's law) rejected by classical hypothesis testing with fixed dimension.

- This conflict is not due to prior distribution specification, as even the lower bounds on conditional measures of evidence over wide classes of prior distributions often provide much more support to Benford's law than the p-values when compared to the typical classical hypothesis testing evidence thresholds, suggesting that to consider only the p-value may cause an underestimation of the evidence in favour of Benford's law.
- Models are just approximations of reality and one can not realistically expect the data to perfectly fit the postulated models in all samples even when they are true. Similarly, one cannot realistically expect collections of numbers to perfectly fit Benford's law even when the underlying data generating process is indeed Benford's law conforming. The usefulness and interpretation of p-values is then drastically affected by sample size, and classical hypothesis testing with fixed dimension is of limited usefulness in Benford's law based digit analysis: due to the high power that classical tests attain in large samples they are likely to produce small p-values and reject Benford's law in samples with very tiny (and without practical importance) deviations from it, producing many false positives results (i.e. classifying legit samples as fraudulent or erroneous).
- Although Bayesian methods are often subject to criticism for relying on subjective procedures, the classical method is also not without an element of subjectivity, as the threshold separating statistical significance from non-significance is itself arbitrary. It was shown that when classical hypothesis testing is used in Benford's law based digit analysis the conclusions drawn are sensible to that arbitrarily chosen evidence threshold. It can therefore be misleading to draw sharp conclusions based solely on the statistical significance of a deviation from Benford's law.
- Because in precise null hypothesis testing the conditional measures of evidence are often much larger than the p-value, and it was shown that Benford's law based digit analysis is no exception, one incurring in the p-value fallacy will typically largely overestimate the practical importance of an observed deviation from Benford's law.

- Instead of reporting Benford's law based digit analysis results in a binary significant/not significant way, as typically done in classical hypothesis testing , it is recommended to report a measure of evidence quantifying the extent of the deviation from Benford's law. Because the p-value is an incomplete measure of the evidence provided by the data, quantifying only how unlikely it is to observe data as or more extreme than the one actually observed under one hypothesis without taking into consideration how unlikely it is under the alternatives, it is advisable to report conditional measures of evidence. Posterior probabilities measure the probability of the hypotheses provided by the data and are more safe and straightforward to interpret.

5.2 Limitations

- Benford's law is just a proxy for normal behaviour, and even a deviation from it with both economic and statistical significance does not ensure that the data is erroneous or fraudulent. No accusations should be made based solely on Benford's law based digit analysis and further investigation is always required after the identification of a suspect dataset. Linville (2011) warns that there are case-specific factors that may legitimately skew the frequencies of digits, like some pricing schemes or discount campaigns. For example, a firm that sets all prices ending up in 99 has transaction records not in conformance to Benford's law. Before applying Benford's law based digit analysis it is important to think if the specific situation being analysed has any particularity that may legitimately skew the observed frequencies away from Benford's law.
- In chapter 3 the possible values of c and s were restricted to the set of natural numbers when they need only be restricted for the set of positive real numbers. This choice was made purely on convenience, as it would be more difficult to find the smallest real positive number satisfying the restrictions that were imposed.
- Classical hypothesis testing was criticized for underestimating the evidence provided by the data in favour of the null. That critique may also apply to Bayesian model selection. The lower bounds on the conditional measures of evidence are also biased against the null, as from the family of priors under consideration they choose the prior

which maximizes the predictive density of the alternative hypothesis. The fact that the p-value calibration that was used does not depend on sample size also contributes to the underestimation of the true posterior probability of the null, as the calibration may be conservative in large samples. Sellke, Bayarri, and Berger (2001) note that the fact that such lower bounds are often still much larger than p-values indicates the severe nature of the bias against a precise null that can arise due to the p-value fallacy.

- The true Bayes factors and posterior probabilities of the null are almost certainly larger than the lower bounds on them that were computed. Therefore, it should be safe not to reject the null when the lower bounds are high. On the other hand, if the lower bound are small the conclusion is ambiguous, as it does not necessarily implies that the true conditional measures of evidence are themselves small.
- To allow Bayes factors and posterior probabilities to be computed analytically, only conjugate priors were used. Unfortunately, the Dirichlet and Beta families were not versatile enough to allow for the simultaneous imposition of all the desired restrictions on the hyperparameters without making the priors too informative.

5.3 Further Research

- The p-value calibrations that were used are just approximations to the exact lower bounds on the conditional measures of evidence. It would be interesting to code a routine to compute the exact lower bounds using the formulas derived by Berger and Sellke (1987) and Delampady and Berger (1990).

Bibliography

Adhikari, A and B Sarkar (1968). Distribution of most significant digit in certain functions whose arguments are random variables. *Sankhy: The Indian Journal of Statistics, Series B*, 47–58 (Cited on pages 6, 9).

Allaart, PC (1997). An invariant-sum characterization of Benford's law. *Journal of Applied Probability*, 288–291 (Cited on pages 7, 8).

Asllani, A and M Naco (2014). Using Benford's Law for Fraud Detection in Accounting Practices. *Journal of Social Science Studies* **2**(1), 129 (Cited on page 11).

Bayes, T and R Price (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (1683-1775)*, 370–418 (Cited on page 12).

Becker, PW (1982). Patterns in listings of failure-rate & MTTF values and listings of other data. *Reliability, IEEE Transactions on* **31**(2), 132–134 (Cited on page 6).

Benford, F (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572 (Cited on pages 4–8).

Berger, A and TP Hill (2015). *An Introduction to Benford's Law*. Princeton University Press (Cited on pages 6–8).

Berger, A, TP Hill, et al. (2011). A basic theory of Benford's Law. *Probability Surveys* **8**, 1–126 (Cited on page 7).

Berger, JO and JM Bernardo (1992-b). Ordered group reference priors with application to the multinomial problem. *Biometrika* **79**(1), 25–37 (Cited on page 21).

Berger, JO, JM Bernardo, and D Sun (2009). The formal definition of reference priors. *The Annals of Statistics*, 905–938 (Cited on page 21).

Berger, JO and M Delampady (1987). Testing precise hypotheses. *Statistical Science*, 317–335 (Cited on pages 2, 13, 20, 23–25, 29).

Berger, JO and T Sellke (1987). Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American statistical Association* **82**(397), 112–122 (Cited on pages 13, 23–25, 37).

Berger, JO, JM Bernardo, et al. (1992-a). On the development of reference priors. *Bayesian statistics* **4**(4), 35–60 (Cited on page 21).

Berger, JO, JM Bernardo, D Sun, et al. (2015). Overall objective priors. *Bayesian Analysis* **10**(1), 189–221 (Cited on pages 18, 21).

Berger, J and LR Pericchi (2001). Objective Bayesian Methods for Model Selection: Introduction and Comparison. *IMS Lecture Notes - Monograph Series* **38** (Cited on page 18).

Bernardo, J and A Smith (1994). Bayesian Theory Wiley. *New York* (Cited on page 12).

Bernardo, JM (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 113–147 (Cited on page 21).

Bolton, RJ and DJ Hand (2002). Statistical Fraud Detection: A Review. *Statistical Science* **17**(3), 235–255 (Cited on page 10).

Boyle, J (1994). An application of Fourier series to the most significant digit problem. *The American Mathematical Monthly* **101**(9), 879–886 (Cited on page 9).

Bram, U (2014). *Thinking statistically*. Kuri Books (Cited on page 15).

Buck, B, A Merchant, and S Perez (1993). An illustration of Benford's first digit law using alpha decay half lives. *European Journal of Physics* **14**(2), 59 (Cited on page 6).

Burke, J and E Kincanon (1991). Benford's law and physical constants: the distribution of initial digits. *American Journal of Physics* **59**(10), 952 (Cited on page 6).

Camerer, C (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press, pp. 134–138 (Cited on page 10).

Carslaw, CA (1988). Anomalies in income numbers: Evidence of goal oriented behavior. *Accounting Review*, 321–327 (Cited on page 11).

Cho, WKT and BJ Gaines (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The american statistician* **61**(3), 218–223 (Cited on pages 10, 17).

Conover, WJ (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association* **67**(339), 591–596 (Cited on page 17).

Conover, WJ and W Conover (1999). Practical nonparametric statistics (Cited on page 17).

Costas, E, V Lopez-Rodas, FJ Toro, and A Flores-Moya (2008). The number of cells in colonies of the cyanobacterium Microcystis aeruginosa satisfies Benford's law. *Aquatic Botany* **89**(3), 341–343 (Cited on page 6).

Delampady, M and JO Berger (1990). Lower bounds on Bayes factors for multinomial distributions, with application to chi-squared tests of fit. *The Annals of Statistics*, 1295–1316 (Cited on pages 2, 13, 20, 21, 24, 25, 29, 37).

Diaconis, P (1977). The distribution of leading digits and uniform distribution mod 1. *The Annals of Probability*, 72–81 (Cited on page 6).

Diaconis, P and D Freedman (1979). On rounding percentages. *Journal of the American Statistical Association* **74**(366a), 359–364 (Cited on page 4).

Diekmann, A (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientif ic Data. *Journal of Applied Statistics* **34**(3), 321–329 (Cited on page 11).

Diekmann, A and B Jann (2010). Benford's Law and Fraud Detection: Facts and Legends. *German Economic Review* **11**(3), 397–401 (Cited on page 10).

Docampo, S, M del Mar Trigo, MJ Aira, B Cabezudo, and A Flores-Moya (2009). Benford's law applied to aerobiological data and its potential as a quality control tool. *Aerobiologia* **25**(4), 275–283 (Cited on page 6).

Durtschi, C, W Hillison, and C Pacini (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting* **5**(1), 17–34 (Cited on pages 8, 10).

Edwards, W, H Lindman, and LJ Savage (1963). Bayesian statistical inference for psychological research. *Psychological review* **70**(3), 193 (Cited on pages 13, 24).

Fewster, RM (2009). A simple explanation of Benford's Law. *The American Statistician* **63**(1), 27–32 (Cited on page 8).

Fisher, RA (1925). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd (Cited on pages 12, 13).

Friar, JL, T Goldman, and J Pérez–Mercader (2012). Genome sizes and the Benford distribution. *PloS one* **7**(5), e36624 (Cited on page 6).

Gelman, A and H Stern (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician* **60**(4), 328–331 (Cited on page 15).

Gibbons, JD and JW Pratt (1975). P-values: interpretation and methodology. *The American Statistician* **29**(1), 20–25 (Cited on page 14).

Giles, DE (2007). Benford's law and naturally occurring prices in certain ebay auctions. *Applied Economics Letters* **14**(3), 157–161 (Cited on pages 6, 11).

Gonzalez-Garcia, J and MGC Pastor (2009). *Benford's law and macroeconomic data quality*. 9-10. International Monetary Fund (Cited on pages 7, 11).

Goodman, SN (1999-a). Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of internal medicine* **130**(12), 995–1004 (Cited on page 14).

Goodman, SN (1999-b). Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of internal medicine* **130**(12), 1005–1013 (Cited on pages 12, 13, 25).

Green, P (2002). Letter from the President to the Lord Chancellor regarding the use of statistical evidence in court cases, 23 January 2002. In: *The Royal Statistical Society* (Cited on page 15).

Guan, L, D He, and D Yang (2006). Auditing, integral approach to quarterly reporting, and cosmetic earnings management. *Managerial auditing journal* **21**(6), 569–581 (Cited on page 11).

Hamming, RW (1970). On the distribution of numbers. *Bell System Technical Journal* **49**(8), 1609–1625 (Cited on pages 6, 9).

Haynes, A (2012). "Detecting Fraud in Bankrupt Municipalities Using Benford's Law". Scripps Senior Theses. Scripps College (Cited on page 11).

Hill, TP (1995-a). Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society* **123**(3), 887–895 (Cited on pages 7, 8).

Hill, TP (1995-b). A statistical derivation of the significant-digit law. *Statistical Science*, 354–363 (Cited on pages 4, 7, 8, 30).

Hubbard, R and M Bayarri (2003). P values are not error probabilities. *Institute of Statistics and Decision Sciences, Working Paper* (03-26), 27708–0251 (Cited on page 14).

Jamain, A (2001). "Benford's Law". Ecole Nationale Supérieure d'Informatique et Mathematiques Appliquées de Grenoble-Imperial College of London (Cited on pages 5–8).

Jaynes, ET (2003). *Probability theory: The logic of science*. Cambridge university press (Cited on page 68).

Jeffreys, H (1935). Some tests of significance, treated by the theory of probability. In: *Proceedings of the Cambridge Philosophical Society*. Vol. 31. 02. Cambridge Univ Press, pp.203–222 (Cited on page 12).

Jeffreys, SH (1967). *Theory of Probability: 3d Ed*. Clarendon Press (Cited on pages 12, 16, 57).

Johnson, VE (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences* **110**(48), 19313–19317 (Cited on page 13).

Jones, BK (2002). Logarithmic distributions in reliability analysis. *Microelectronics Reliability* **42**(4), 779–786 (Cited on page 8).

Judge, G and L Schechter (2009). Detecting problems in survey data using Benford's Law. *Journal of Human Resources* **44**(1), 1–24 (Cited on page 11).

Kass, RE and AE Raftery (1995). Bayes factors. *Journal of the american statistical association* **90**(430), 773–795 (Cited on pages 12, 57).

Kass, RE and L Wasserman (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**(435), 1343–1370 (Cited on page 20).

Keynes, JM (1937). The general theory of employment. *The quarterly journal of economics*, 209–223 (Cited on pages 1, 70).

Kinnunen, J and M Koskela (2003). Who is miss world in cosmetic earnings management? A cross-national comparison of small upward rounding of net income numbers among eighteen countries. *Journal of International Accounting Research* **2**(1), 39–68 (Cited on page 11).

Knuth, DE (1981). The Art of Computer Programming 2: Seminumerical Programming. *AddisonWesley, Reading, MA*, 239–249 (Cited on pages 6, 7, 9).

Kuiper, NH (1960). Tests concerning random points on a circle. In: *Indagationes Mathematicae (Proceedings)*. Vol. 63. Elsevier, pp.38–47 (Cited on page 17).

Laplace, PSMd (1820). *Théorie analytique des probabilités*. V. Courcier (Cited on page vi).

Lavine, M and MJ Schervish (1999). Bayes Factors: What They Are and What They Are Not. *The American Statistician* **53**(2), 119–122 (Cited on page 12).

Leamer, EE (1983). Model choice and specification analysis. *Handbook of econometrics* **1**, 285–330 (Cited on page 15).

Leemis, LM, BW Schmeiser, and DL Evans (2000). Survival distributions satisfying Benford's law. *The American Statistician* **54**(4), 236–241 (Cited on page 17).

Ley, E (1996). On the peculiar distribution of the US stock indexes' digits. *The American Statistician* **50**(4), 311–313 (Cited on pages 7, 15, 28).

Lin, R and G Yin (2015). Bayes factor and posterior probability: Complementary statistical evidence to p-value. *Contemporary clinical trials* **44**, 33–35 (Cited on pages 13, 14).

Linville, M (2011). The Problem Of False Negative Results In The Use Of Digit Analysis. *Journal of Applied Business Research (JABR)* **24**(1) (Cited on page 36).

Luque, B and L Lacasa (2009). The first-digit frequencies of prime numbers and Riemann zeta zeros. In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. The Royal Society, pp.rspa–2009 (Cited on page 6).

Marchi, S and JT Hamilton (2006). Assessing the accuracy of self-reported data: an evaluation of the toxics release inventory. *Journal of Risk and uncertainty* **32**(1), 57–76 (Cited on page 11).

Massey Jr, FJ (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* **46**(253), 68–78 (Cited on page 17).

Matthews, R (1998). Bayesian Critique of Statistics in Health: The great health hoax. *The Sunday Telegraph* **13** (Cited on pages 14, 15).

Matthews, R (1999). The power of one. *New Scientist* **163**(2194), 26–30 (Cited on page 27).

Mebane Jr, WR (2006-a). Election forensics: the second-digit Benford's law test and recent American presidential elections. In: *Election Fraud Conference*. Citeseer (Cited on page 12).

Mebane Jr, WR et al. (2006-b). Detecting attempted election theft: vote counts, voting machines and Benford's law. In: *Annual Meeting of the Midwest Political Science Association, Chicago, IL, April*. Citeseer, pp.20–23 (Cited on page 12).

Mebane Jr, WR (2007). Statistics for digits. In: *Summer Meeting of the Political Methodology Society, Pennsylvania State University, July*, pp.18–21 (Cited on page 12).

Möller, M (2009). Measuring the Quality of Auditing Services with the Help of Benford's Law-An Empirical Analysis and Discussion of this Methodical Approach. *Available at SSRN 1529307* (Cited on page 11).

Morrow, J (2014). Benford's Law, families of distributions and a test basis (Cited on page 17).

Murteira, BJF and M Antunes (2012). *Probabilidades e Estatistica-Volume II*. Escolar Editora (Cited on page 16).

Murteira, B, CS Ribeiro, JA Silva, and C Pimenta (2010). *Introdução à Estatatística*. Escolar Editora (Cited on page 28).

Newcomb, S (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics* **4**(1), 39–40 (Cited on page 4).

Neyman, J and ES Pearson (1933). "On the problem of the most efficient tests of statistical hypotheses". In: *Philosophical Transactions of the Royal Society, Series A*. Springer, pp.289–337 (Cited on page 12).

Nigrini, M (1996). A taxpayer compliance application of Benford's law. *The Journal of the American Taxation Association* **18**(1), 72 (Cited on pages 7, 11).

Nigrini, M (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*. Vol. 586. John Wiley & Sons (Cited on pages 7, 8).

Nigrini, MJ (1999). Adding value with digital analysis. *Internal Auditor* **56**(1) (Cited on page 7).

Nigrini, MJ (2000). Digital Analysis Using Benford's Law: Tests & Statistics for Auditors. *Global Audit Publications* (Cited on pages 17, 28).

Nigrini, MJ (2005). An assessment of the change in the incidence of earnings management around the Enron-Andersen episode. *Review of Accounting and Finance* **4**(1), 92–110 (Cited on page 11).

Nigrini, MJ and LJ Mittermaier (1997). The use of Benford's law as an aid in analytical procedures. *Auditing* **16**(2), 52 (Cited on pages 7, 10).

Nigrini, MJ and W Wood (1995). Assessing the integrity of tabulated demographic data. *Preprint, Univ. Cincinnati and St. Mary's Univ* (Cited on pages 6, 7).

Nigrini, MJ (1992). *The detection of income tax evasion through an analysis of digital distributions*. UMI (Cited on pages 7, 11).

Niskanen, J and M Keloharju (2000). Earnings cosmetics in a tax-driven accounting environment: evidence from Finnish public firms. *European Accounting Review* **9**(3), 443–452 (Cited on page 11).

Nye, J and C Moul (2007). The political economy of numbers: on the application of Benford's law to international macroeconomic statistics. *The BE Journal of Macroeconomics* **7**(1) (Cited on pages 7, 8, 11).

O'Hagan, A and B Luce (2003). A primer on Bayesian statistics in health economics and outcomes research. *London: Medtap International Inc* (Cited on page 14).

Paulino, CDM, MAA Turkman, and B Murteira (2003). *Estatística bayesiana* (Cited on page 21).

Pericchi, LR and D Torres (2004). La Ley de Newcomb-Benford y sus aplicaciones al referéndum revocatorio en Venezuela. *Reporte Técnico no-definitivo 2a. versión: Octubre* **1**, 2004 (Cited on page 12).

Pericchi, L and D Torres (2011). Quick Anomaly Detection by the Newcomb—Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela. *Statistical Science*, 502–516 (Cited on pages 2, 7, 12, 14, 15, 21, 23, 25, 28).

Pietronero, L, E Tosatti, V Tosatti, and A Vespignani (2001). Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A: Statistical Mechanics and its Applications* **293**(1), 297–304 (Cited on pages 6, 7).

Pinkham, RS (1961). On the distribution of first significant digits. *The Annals of Mathematical Statistics* **32**(4), 1223–1230 (Cited on page 7).

Prudêncio, ARG (2015). "Aplicação da Lei de Benford para o controlo das demonstrações financeiras de entidades bancárias". MSc thesis. Instituto Superior de Economia e Gestão (Cited on page 11).

Raimi, RA (1969). The peculiar distribution of first digits. *Scientific American* **221**, 109–120 (Cited on page 9).

Raimi, RA (1976). The first digit problem. *The American Mathematical Monthly* **83**(7), 521–538 (Cited on pages 6, 8, 9).

Rauch, B, M Göttsche, G Brähler, and S Engel (2011). Fact and Fiction in EU-Governmental Economic Data. *German Economic Review* **12**(3), 243–255 (Cited on pages 10, 11, 26, 27).

Sarkar, B (1973). An observation on the significant digits of binomial coefficients and factorials. *Sankhy: The Indian Journal of Statistics, Series B*, 363–364 (Cited on pages 6, 9).

Schatte, P (1988). On mantissa distributions in computing and Benford's law. *Journal of Information Processing and Cybernetics* **24**(9), 443–455 (Cited on pages 6, 9).

Schervish, MJ (1996). P values: what they are and what they are not. *The American Statistician* **50**(3), 203–206 (Cited on pages 12, 14).

Scott, P and M Fasli (2001). Benford's law: An empirical investigation and a novel explanation. *Unpublished manuscript* (Cited on pages 8, 9).

Sellke, T, M Bayarri, and JO Berger (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician* **55**(1), 62–71 (Cited on pages 14, 24, 25, 37).

Shengmin, Z and W Wenchao (2010). Does Chinese Stock Indices Agree with Benford's Law? In: *Management and Service Science (MASS), 2010 International Conference on*. IEEE, pp.1–3 (Cited on page 7).

Skousen, CJ, L Guan, and TS Wetzel (2004). Anomalies and unusual patterns in reported earnings: Japanese managers round earnings. *Journal of international financial management & accounting* **15**(3), 212–234 (Cited on page 11).

Smith, S (2002). The Scientist and Engineer's Guide to Digital Signal Processing, Chapter 34: Explaining Benford's Law. *Republished in softcover by Newnes* (Cited on pages 7, 8).

Stephens, MA (1970). Use of the Kolmogorov-Smirnov, Cramér-Von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 115–122 (Cited on page 17).

Stone, JV (2013). *Bayes' rule: a tutorial introduction to Bayesian analysis*. Sebtel Press (Cited on page 50).

Syversveen, AR (1998). Noninformative bayesian priors. interpretation and problems with construction and applications. *Preprint Statistics* **3** (Cited on page 21).

Taleb, NN (2010). *The bed of Procrustes: Philosophical and practical aphorisms*. Vol. 4. Random House (Cited on page 26).

Taleb, NN (2016). The Meta-Distribution of Standard P-Values. *Fat Tail Research Program Working Papers* (Cited on page 13).

Thomas, JK (1989). Unusual patterns in reported earnings. *Accounting Review*, 773–787 (Cited on page 11).

Tödter, KH (2009). Benford's Law as an Indicator of Fraud in Economics. *German Economic Review* **10**(3), 339–351 (Cited on pages 7, 11).

Torres, J, S Fernandez, A Gamero, and A Sola (2007). How do numbers begin?(The first digit law). *European journal of physics* **28**(3), L17 (Cited on page 12).

Torres, N (2006). "Newcomb–Benford's Law Applications to Electoral Processes, Bioinformatics, and the Stock Index." PhD thesis. MS thesis (Cited on pages 7, 12, 15, 21, 23, 28).

Tukey, JW (1962). The future of data analysis. *The Annals of Mathematical Statistics* **33**(1), 1–67 (Cited on page 34).

Turkman, MAA and CD Paulino (2015). *Estatística Bayesiana Computacional*. Sociedade Portuguesa de Estatística (Cited on pages 18, 19).

Van Caneghem, T (2002). Earnings management induced by cognitive reference points. *The British Accounting Review* **34**(2), 167–178 (Cited on page 11).

Van Caneghem, T (2004). The impact of audit quality on earnings rounding-up behaviour: some UK evidence. *European Accounting Review* **13**(4), 771–786 (Cited on page 11).

Varian, HR (1972). Benfords law. *American Statistician* **26**(3), 65 (Cited on pages 7, 10, 11).

Von Neumann, J (1947). The mathematician. *The works of the mind* **1**(1), 180–196 (Cited on page 52).

Von Neumann, J (1955). Method in the physical sciences. *Collected Works* **6**, 491–498 (Cited on page 57).

Wallace, WA (2002). Assessing the quality of data used for benchmarking and decision-making. *The Journal of Government Financial Management* **51**(3), 16 (Cited on page 8).

Washington, L (1981). Benford Law for Fibonacci and Lucas-Numbers. *Fibonacci Quarterly* **19**(2), 175–177 (Cited on page 6).

Wasserstein, RL and NA Lazar (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician* (Cited on pages 2, 14, 15).

Watrin, C, R Struffert, and R Ullmann (2008). Benford's Law: an instrument for selecting tax audit targets? *Review of managerial science* **2**(3), 219–237 (Cited on page 11).

Yang, R and JO Berger (1996). *A catalog of noninformative priors*. Institute of Statistics and Decision Sciences, Duke University (Cited on page 21).

Yang, Y (1995). Invariance of the reference prior under reparametrization. *Test* **4**(1), 83–94 (Cited on page 21).

Appendices

Appendix A

Derivations

"Bayes' rule is a rigorous method for interpreting evidence in the context of previous experience or knowledge"

James Stone (2013)

A.1 Multinomial ∧ Dirichlet Model

Note that 3.3 together with 3.1, 3.4 and 3.5 implies :

$$
\begin{aligned}
m_0(\boldsymbol{x}) &= f(\boldsymbol{x}|\boldsymbol{\theta_0}) = \frac{N!}{\prod_{i=1}^{k+1} x_i!} \prod_{i=1}^{k+1} {\theta_i}^{x_i} \\
m_1(\boldsymbol{x}) &= \int_{\Theta_1} f(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} = \int_{\Theta_1} \frac{N!}{\prod_{i=1}^{k+1} x_i!} \prod_{i=1}^{k+1} {\theta_i}^{x_i} \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{k+1} {\theta_i}^{\alpha_i - 1} d\boldsymbol{\theta} = \\
&= \frac{N!}{\prod_{i=1}^{k+1} x_i! B(\boldsymbol{\alpha})} \int_{\Theta_1} \prod_{i=1}^{k+1} {\theta_i}^{\alpha_i + x_i - 1} d\boldsymbol{\theta} = \\
&= \frac{N!}{\prod_{i=1}^{k+1} x_i!} \frac{B(\boldsymbol{\alpha} + \boldsymbol{x})}{B(\boldsymbol{\alpha})}
\end{aligned}
$$

where $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{k+1} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k+1} \alpha_i)}$ and $B(\boldsymbol{\alpha} + \boldsymbol{x}) = \int_{\Theta_1} \prod_{i=1}^{k+1} {\theta_i}^{\alpha_i + x_i - 1} = \frac{\prod_{i=1}^{k+1} \Gamma(\alpha_i + x_i)}{\Gamma(\sum_{i=1}^{k+1} \alpha_i + x_i)}$. Then, by the definition of BF in favour of H_0:

$$B_{01}(\boldsymbol{x}) = \frac{m_0(\boldsymbol{x})}{m_1(\boldsymbol{x})} = \frac{\frac{N!}{\prod_{i=1}^{k+1} x_i!}\prod_{i=1}^{k+1} {\theta_{0i}}^{x_i}}{\frac{N!}{\prod_{i=1}^{k+1} x_i!}\frac{B(\boldsymbol{\alpha}+\boldsymbol{x})}{B(\boldsymbol{\alpha})}} = \frac{\prod_{i=1}^{k+1} {\theta_{0i}}^{x_i}}{\frac{B(\boldsymbol{\alpha}+\boldsymbol{x})}{B(\boldsymbol{\alpha})}} = \frac{\prod_{i=1}^{k+1} {\theta_{0i}}^{x_i} B(\boldsymbol{\alpha})}{B(\boldsymbol{\alpha}+\boldsymbol{x})} =$$

$$= \left(\frac{\prod_{i=1}^{k+1} ({\theta_{0i}}^{x_i})\prod_{i=1}^{k+1}[\Gamma(\alpha_i)]}{\Gamma[\sum_{i=1}^{k+1}(\alpha_i)]}\right) \Big/ \left(\frac{\prod_{i=1}^{k+1}\Gamma(\alpha_i+x_i)}{\Gamma(\sum_{i=1}^{k+1}\alpha_i+x_i)}\right) = \frac{\prod_{i=1}^{k+1}({\theta_{0i}}^{x_i})\prod_{i=1}^{k+1}[\Gamma(\alpha_i)]\Gamma[\sum_{i=1}^{k+1}(\alpha_i+x_i)]}{\Gamma(\sum_{i=1}^{k+1}\alpha_i)\prod_{i=1}^{k+1}\Gamma(\alpha_i+x_i)}$$

A.2 Binomial $\wedge$ Beta Model

Note that 3.12 together with 3.9 and 3.11 implies:

$$m_0(x) = f(x|\theta_0) = \binom{N}{x}\theta_0^x(1-\theta_0)^{N-x}$$

$$m_1(x) = \int_0^1 f(x|\theta)\pi(\theta|a,b)d\theta = \int_0^1 \binom{N}{x}\theta^x(1-\theta)^{N-x}\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}d\theta =$$

$$= \binom{N}{x}\int_0^1 \frac{\theta^{(x+a)-1}(1-\theta)^{(N+b-x)-1}}{B(a,b)}d\theta = \binom{N}{x}\frac{B(x+a,N-x+b)}{B(a,b)}$$

because $\int_0^1 \theta^{(x+a)-1}(1-\theta)^{(n+b-x)-1} = B(x+a,N-x+b)$. Next, to compute the Bayes Factor:

$$B_{01}(x) = \frac{m_0(x)}{m_1(x)} = \frac{\binom{N}{x}\theta_0^x(1-\theta_0)^{N-x}}{\binom{N}{x}\frac{B(x+a,N-x+b)}{B(a,b)}} = \frac{\theta_0^x(1-\theta_0)^{N-x}B(a,b)}{B(x+a,N-x+b)} =$$

$$= \frac{\theta_0^x(1-\theta_0)^{N-x}\Gamma(a)\Gamma(b)\Gamma(N+a+b)}{\Gamma(a+b)\Gamma(N+a-x)\Gamma(x+a)}$$

because $B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $B(x+a,N-x+b) = \frac{\Gamma(x+a)\Gamma(N-x+b)}{\Gamma(N+a+b)}$.

Appendix B

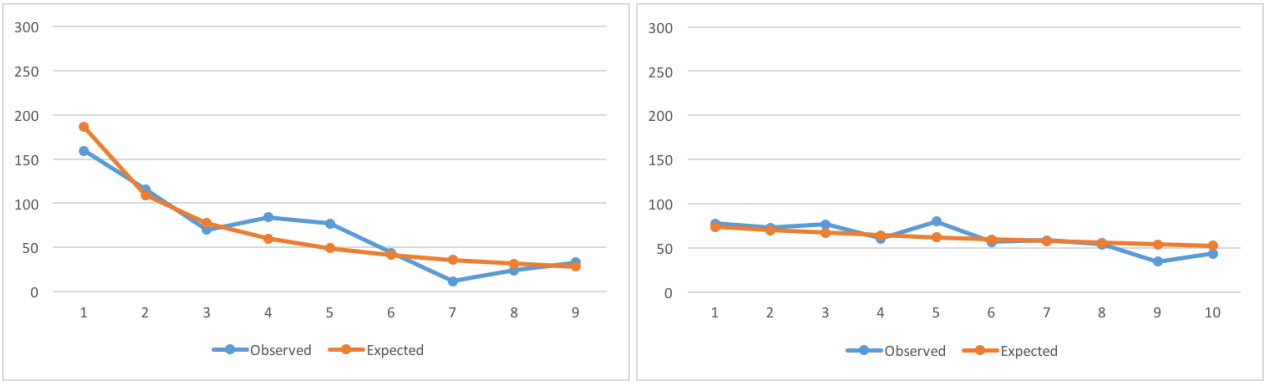
Figures

"I think that it is a relatively good approximation to truth ... that mathematical ideas originate in empirics."

John von Von Neumann (1947)

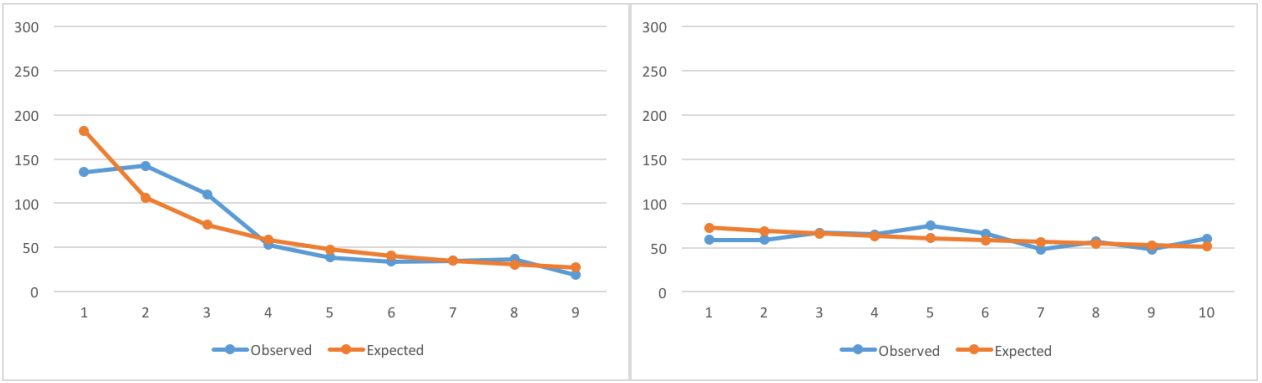
B.1 Observed Counts vs Expected Counts

This section consists in a graphical comparison between each sample's observed leading digits counts and the corresponding BL expected counts. For each dataset, first and second digit occurrences are analysed.



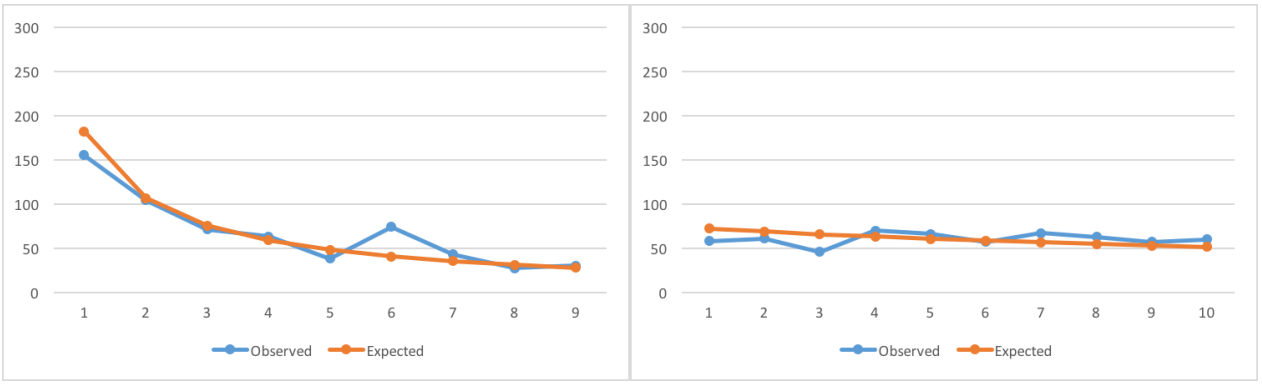
(a) Austria BL1 **(b)** Austria BL2

Figure B.1: Austria – Observed counts vs BL expected counts.



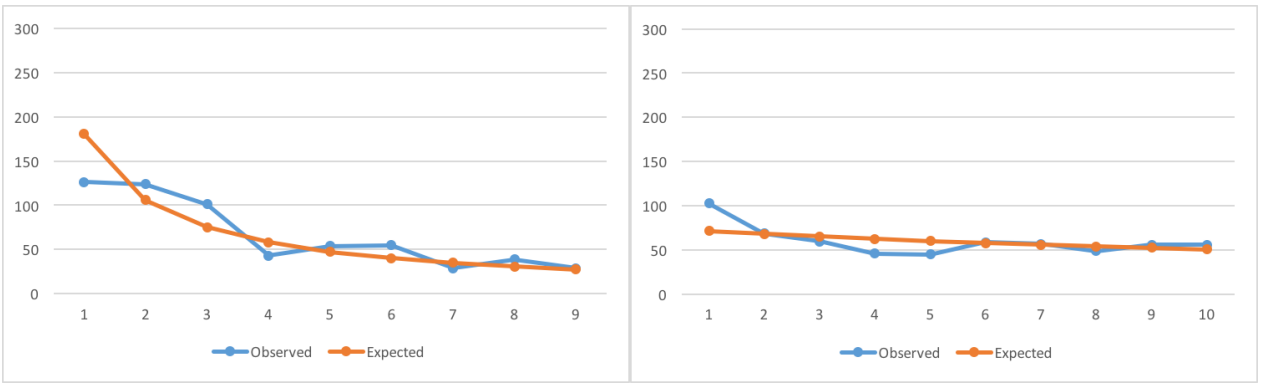
(a) Belgium BL1 **(b)** Belgium BL2

Figure B.2: Belgium – Observed counts vs BL expected counts.



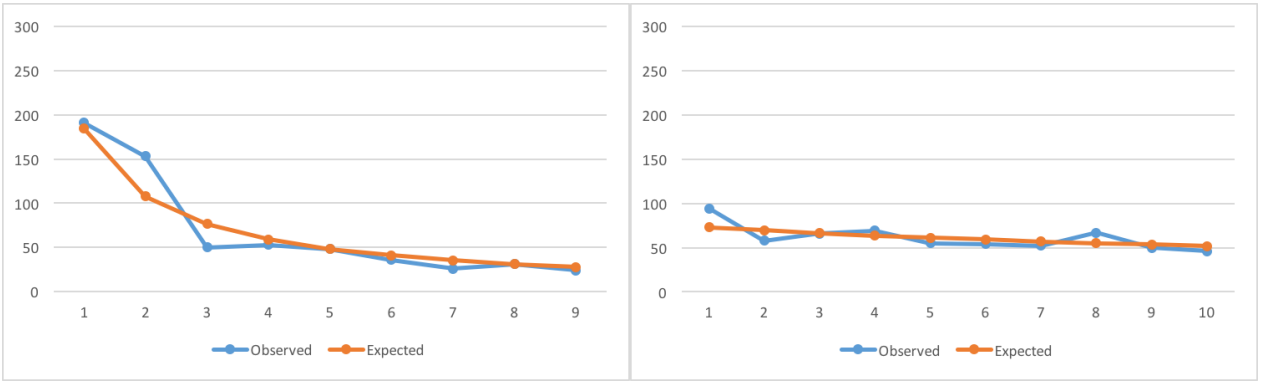
(a) Finland BL1 **(b)** Finland BL2

Figure B.3: Finland – Observed counts vs BL expected counts.



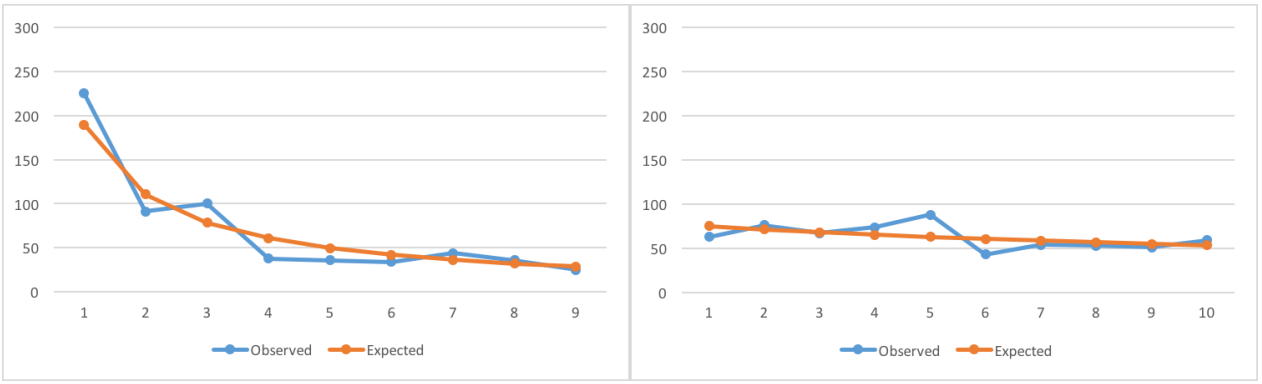
(a) France BL1 **(b)** France BL2

Figure B.4: France – Observed counts vs BL expected counts.



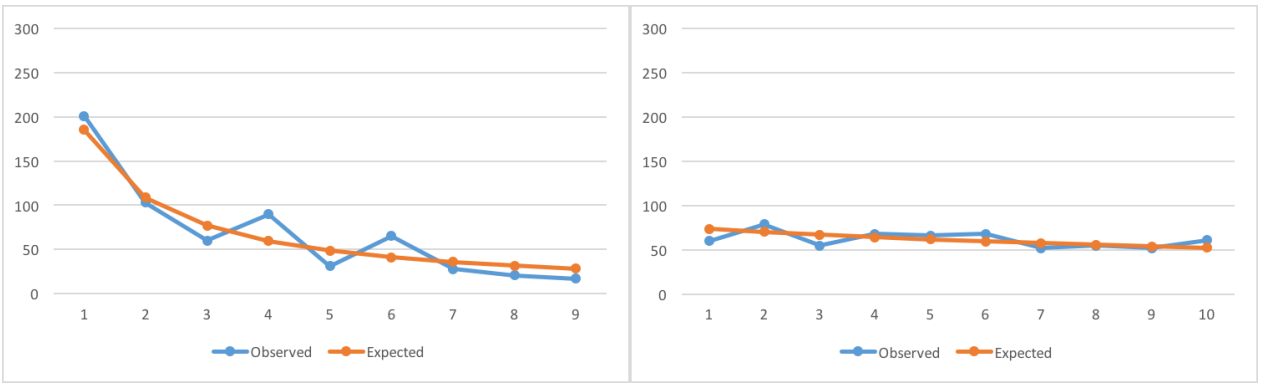
(a) Germany BL1 **(b)** Germany BL2

Figure B.5: Germany – Observed counts vs BL expected counts.



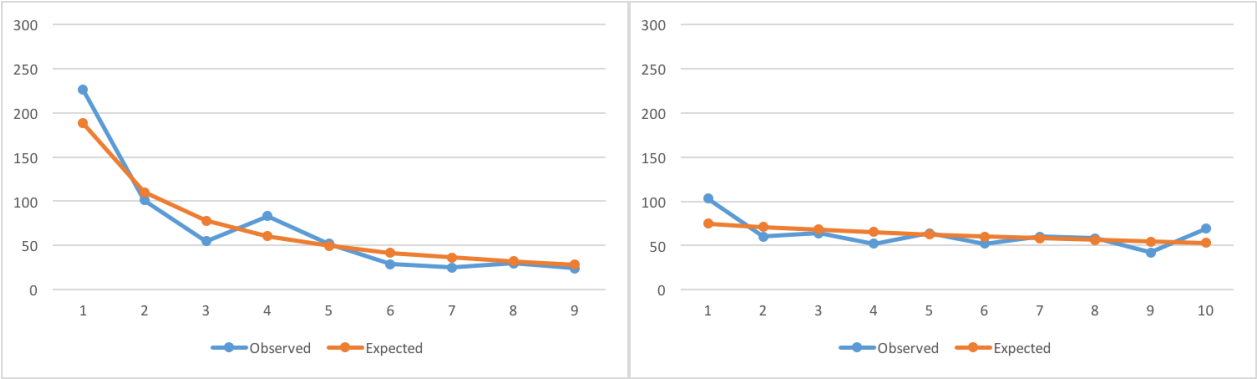
(a) Greece BL1 **(b)** Greece BL2

Figure B.6: Greece – Observed counts vs BL expected counts.



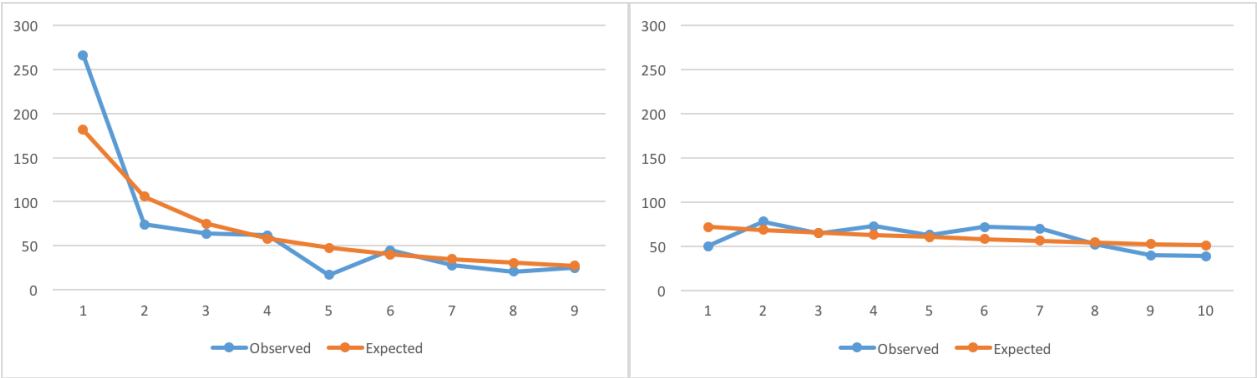
(a) Ireland BL1 **(b)** Ireland BL2

Figure B.7: Ireland – Observed counts vs BL expected counts.



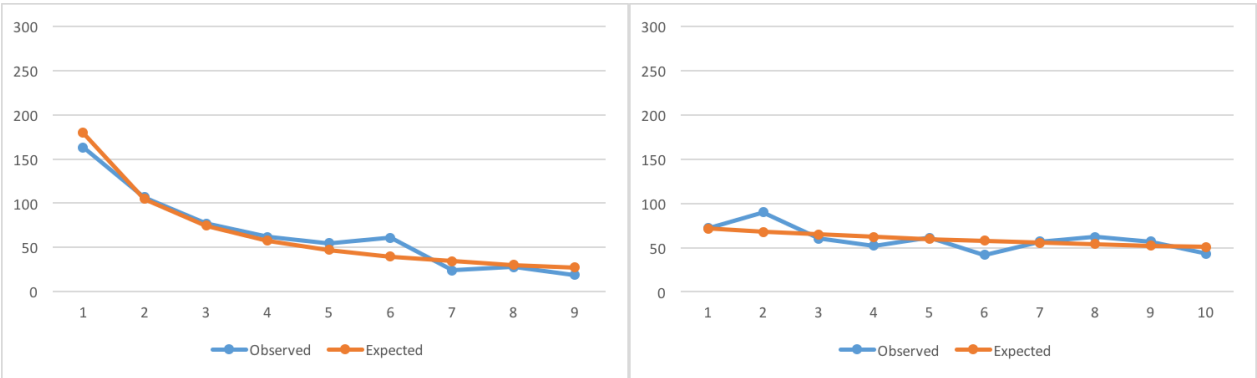
(a) Italy BL1 **(b)** Italy BL2

Figure B.8: Italy – Observed counts vs BL expected counts.



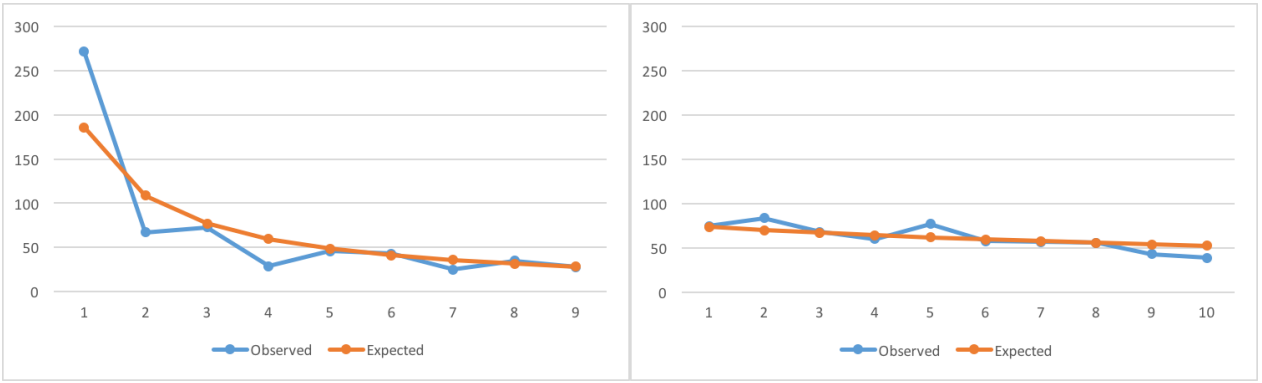
(a) Luxembourg BL1 **(b)** Luxembourg – BL2

Figure B.9: Luxembourg observed counts vs BL expected counts.



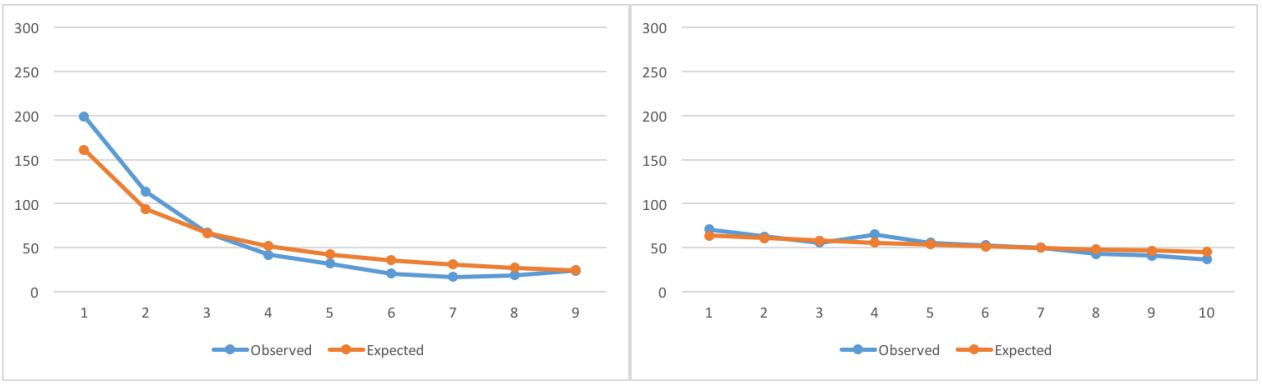
(a) Netherlands BL1 **(b)** Netherlands BL2

Figure B.10: Netherlands – Observed counts vs BL expected counts.



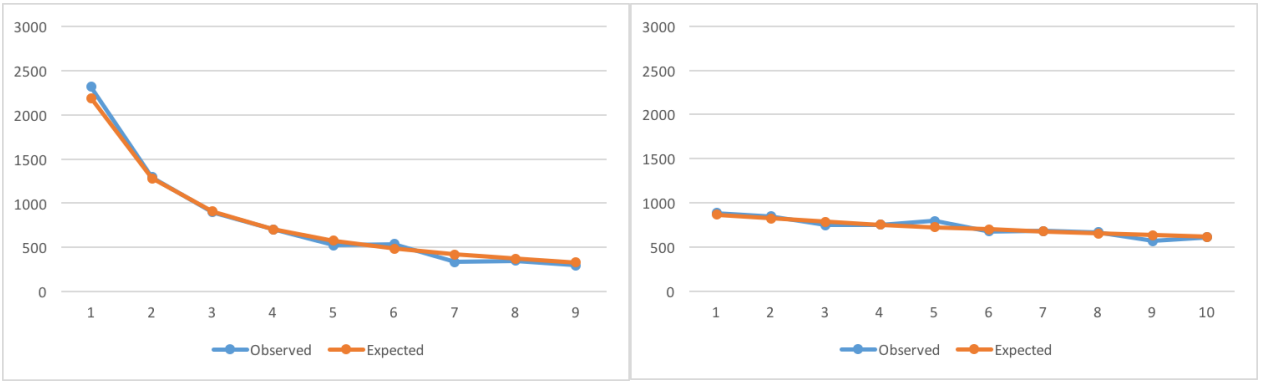
(a) Portugal BL1 **(b)** Portugal BL2

Figure B.11: Portugal – Observed counts vs BL expected counts.



(a) Spain BL1 **(b)** Spain BL2

Figure B.12: Spain – Observed counts vs BL expected counts.



(a) Pooled Sample BL1 **(b)** Pooled Sample BL2

Figure B.13: Pooled Sample – Observed counts vs BL expected counts.

Appendix C

Tables

"The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work."

John von Von Neumann (1955)

C.1 Bayes Factor Interpretation Scale

Table C.1: BF interpretation scale from Kass and Raftery (1995), augmented with one category (the first line) for the case when the data provides more support to H_1 than to H_0. It corresponds to the Jeffreys (1967) original scale, with two of the original categories pooled for simplification.

$\log(B_{01})$	B_{01}	Evidence if favour of H_0
< 0	<1	Negative (Supports H_1)
0 to 0.5	1 to 3.2	Not Worth More Than a Bare Mention
0.5 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
>2	>100	Decisive

C.2 Datasets Summary

Table C.2: Datasets and Descriptive Statistics.

Country	BL1 Sample Size	BL2 Sample Size	Average	Median	Skewness
Austria	619	618	104670	26146	1.53
Belgium	604	604	73711	12911	1.42
Finland	605	605	29658	6782	1.78
France	600	600	170523	29350	1.78
Germany	612	611	52558	8922	1.5
Greece	629	628	29309	4722	1.73
Ireland	616	616	459200	79416	1.52
Italy	625	624	346399	52362	1.67
Luxembourg	602	602	315382	75833	1.67
Netherlands	569	596	6253	875	3.03
Portugal	617	617	31970	6609	1.62
Spain	535	535	43141	6615	1.77
Pooled Sample	7233	7226	138565	11847	3.8

C.3 Multinomial ∧ Dirichlet Model Results

Table C.3: Chi-square goodness-of-fit test and Multinomial ∧ Dirichlet Model results with Dirichlet ($\boldsymbol{\alpha} = \mathbf{1}$) prior distribution, which corresponds to a uniform prior distribution. Strength of evidence measured according to the scale in table C.1.

Country	$\ln(\boldsymbol{B_{01}})$	**Evidence**	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p**-value**
Austria BL1	-3.1	Negative	0.0008	0	0
Austria BL2	5.01	Decisive	0.99999	0.35098	0.07798
Belgium BL1	-1.78	Negative	0.01624	0	0
Belgium BL2	5.9	Decisive	0.99999	0.48083	0.2359
Finland BL1	1.07	Strong	0.9192	0.00036	0
Finland BL2	4.79	Decisive	0.99998	0.33118	0.06762
France BL1	-1.4	Negative	0.03862	0	0
France BL2	3.51	Decisive	0.99969	0.06941	0.00522
Germany BL1	1.21	Strong	0.94159	0.00176	0
Germany BL2	5.67	Decisive	0.99999	0.43638	0.1503
Greece BL1	0.89	Substantial	0.88495	0.00187	0
Greece BL2	4.27	Decisive	0.99995	0.16513	0.01815
Ireland BL1	-2.37	Negative	0.00423	0	0
Ireland BL2	6.34	Decisive	0.99999	0.5	0.38686
Italy BL1	1.23	Strong	0.94487	0.00242	0
Italy BL2	3.43	Decisive	0.99963	0.0545	0.00381
Luxembourg BL1	-8.88	Negative	0	0	0
Luxembourg BL2	3.39	Decisive	0.99959	0.08912	0.00732
Netherlands BL1	3.81	Decisive	0.99985	0.10663	0.00916
Netherlands BL2	4.89	Decisive	0.99999	0.31338	0.0595
Portugal BL1	-8.38	Negative	0	0	0
Portugal BL2	5.76	Decisive	0.99999	0.46127	0.18917
Spain BL1	0.62	Substantial	0.8079	0.00192	0
Spain BL2	7	Decisive	0.99999	0.5	0.79277
Pooled Sample BL1	3.18	Decisive	0.99934	0	0
Pooled Sample BL2	9.63	Decisive	0.99999	0.30444	0.05579

Table C.4: Chi-square goodness-of-fit test and Multinomial ∧ Dirichlet Model results with Dirichlet ($\boldsymbol{\alpha} = \boldsymbol{\theta_0}$) prior distribution. Strength of evidence measured according to the scale in table C.1.

Dataset	$\ln(\boldsymbol{B_{01}})$	**Evidence**	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p-**value**
Austria BL1	1.93	Very Strong	0.98835	0	0
Austria BL2	11.23	Decisive	1	0.35098	0.07798
Belgium BL1	3.35	Decisive	0.99955	0	0
Belgium BL2	12.2	Decisive	1	0.48083	0.2359
Finland BL1	6.45	Decisive	0.99999	0.00036	0
Finland BL2	11.11	Decisive	1	0.33118	0.06762
France BL1	4.02	Decisive	0.99991	0	0
France BL2	9.74	Decisive	1	0.06941	0.00522
Germany BL1	6.14	Decisive	0.99999	0.00176	0
Germany BL2	11.93	Decisive	1	0.43638	0.1503
Greece BL1	5.87	Decisive	0.99999	0.00187	0
Greece BL2	10.53	Decisive	1	0.16513	0.01815
Ireland BL1	2.48	Decisive	0.9967	0	0
Ireland BL2	12.64	Decisive	1	0.5	0.38686
Italy BL1	6.12	Decisive	0.99999	0.00242	0
Italy BL2	9.67	Decisive	1	0.0545	0.00381
Luxembourg BL1	-4.33	Negative	0	0	0
Luxembourg BL2	9.61	Decisive	1	0.08912	0.00732
Netherlands BL1	8.98	Decisive	0.99999	0.10663	0.00916
Netherlands BL2	11.13	Decisive	1	0.31338	0.0595
Portugal BL1	-3.61	Negative	0.00025	0	0
Portugal BL2	11.99	Decisive	1	0.46127	0.18917
Spain BL1	5.19	Decisive	0.99999	0.00192	0
Spain BL2	13.25	Decisive	1	0.5	0.79277
Pooled Sample BL1	8.31	Decisive	0.99999	0	0
Pooled Sample BL2	15.96	Decisive	1	0.30444	0.05579

Table C.5: Chi-square goodness-of-fit test and Multinomial $\wedge$ Dirichlet Model results with Dirichlet ($\boldsymbol{\alpha} = 22\,\boldsymbol{\theta_0}$) prior for BL1 analysis and Dirichlet ($\boldsymbol{\alpha} = 12\,\boldsymbol{\theta_0}$) prior for BL2 analysis. Strength of evidence measured according to the scale in table C.1.

Dataset	$\ln(\boldsymbol{B_{01}})$	**Evidence**	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p-**value**
Austria BL1	-4.92	Negative	0	0	0
Austria BL2	4.56	Decisive	0.99997	0.35098	0.07798
Belgium BL1	-3.59	Negative	0.00026	0	0
Belgium BL2	5.52	Decisive	0.99999	0.48083	0.2359
Finland BL1	-0.6	Negative	0.19978	0.00036	0
Finland BL2	4.46787	Decisive	0.99996	0.33118	0.06762
France BL1	-2.93	Negative	0.00119	0	0
France BL2	3.1	Decisive	0.9992	0.06941	0.00522
Germany BL1	-0.89	Negative	0.11304	0.00176	0
Germany BL2	5.23	Decisive	0.99999	0.43638	0.1503
Greece BL1	-1.14	Negative	0.06684	0.00187	0
Greece BL2	3.87012	Decisive	0.99986	0.16513	0.01815
Ireland BL1	-4.43	Negative	0	0	0
Ireland BL2	5.95	Decisive	0.99999	0.5	0.38686
Italy BL1	-0.91	Negative	0.10967	0.00242	0
Italy BL2	3.02	Decisive	0.99906	0.0545	0.00381
Luxembourg BL1	-10.95	Negative	0	0	0
Luxembourg BL2	2.98	Decisive	0.99896	0.08912	0.00732
Netherlands BL1	1.86	Very Strong	0.98624	0.10663	0.00916
Netherlands BL2	4.47	Decisive	0.99997	0.31338	0.0595
Portugal BL1	-10.29	Negative	0	0	0
Portugal BL2	5.31	Decisive	0.99999	0.46127	0.18917
Spain BL1	-1.75	Negative	0.01752	0.00192	0
Spain BL2	6.55	Decisive	0.99999	0.5	0.79277
Pooled Sample BL1	1.02	Strong	0.9124	0	0
Pooled Sample BL2	9.2	Decisive	0.99999	0.30444	0.05579

C.4 Binomial ∧ Beta Model Results

Table C.6: Z-test and Binomial ∧ Beta Model results for Austria BL1 with a Beta $(1, 1)$ prior distribution, which corresponds to a Uniform prior distribution.

Digit	$\ln(\boldsymbol{B_{01}})$	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p-**value**
1	0.07	0.54162	0.1682	0.01869
2	1.29	0.95114	0.5	0.49281
3	1.32	95419	0.5	0.40591
4	-0.62	0.19392	0.0244	0.0014
5	-1.78	0.01646	0.00117	0
6	1.56	0.97262	0.5	0.74044
7	-3.02	0.00094	0.00152	0
8	1.25	0.94661	0.46235	0.1913
9	1.48	0.96778	0.5	0.42181

Table C.7: Z-test and Binomial ∧ Beta Model results for Ireland BL1 with a Beta $(1, 1)$ prior distribution, which corresponds to a Uniform prior distribution.

Digit	$\ln(\boldsymbol{B_{01}})$	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p-**value**
1	0.93	0.89385	0.45944	0.18573
2	1.35	0.95728	0.5	0.59892
3	0.53	0.77214	0.27458	0.04486
4	-1.8	0.0155	0.00133	0
5	-0.08	0.45297	0.11081	0.00994
6	-1.23	0.05526	0.00413	0.00018
7	1.26	0.94816	0.47244	0.21308
8	0.83	0.87153	0.3302	0.06715
9	0.61	0.80247	0.25707	0.03934

Table C.8: Z-test and Binomial ∧ Beta Model results Luxembourg BL1 with a Beta $(1, 1)$ prior distribution, which corresponds to a Uniform prior distribution.

Digit	$\ln(\boldsymbol{B_{01}})$	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p-**value**
1	-10.21	0	0	0
2	-1.13	0.04634	0.01442	0.00075
3	1.68	0.92122	0.45994	0.18666
4	1.45	0.96602	0.5	0.66329
5	-4.3	0	0.00017	0
6	1.45	0.96555	0.5	0.4936
7	1.33	0.95527	0.48859	0.26358
8	0.93	0.89434	0.363770	0.08555
9	1.63	0.97734	0.5	0.68984

Table C.9: Z-test and Binomial ∧ Beta Model results Portugal BL1 with a Beta $(1, 1)$ prior distribution, which corresponds to a Uniform prior distribution.

Digit	$\ln(\boldsymbol{B_{01}})$	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p**-value**
1	-10.06	0	0	0
2	-3.24	0.00058	0.00042	0
3	1.43	0.96437	0.5	0.66228
4	-2.95	0.00111	0.00104	0
5	1.54	0.97175	0.5	0.72552
6	1.57	0.97403	0.5	0.84751
7	0.87	0.88065	0.34849	0.07658
8	1.55	0.9724	0.5	0.59123
9	1.68	0.97938	0.5	0.95888

Table C.10: Z-test and Binomial ∧ Beta Model results for Austria BL1 with a Beta $(22\,\theta_0, 22-22\,\theta_0)$ prior distribution.

Digit	$\ln(\boldsymbol{B_{01}})$	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p**-value**
1	22.88	1	0.1682	0.01869
2	39.3	1	0.5	0.49281
3	45.9	1	0.5	0.40591
4	47.33	1	0.0244	0.0014
5	48.46	1	0.00117	0
6	53.47	1	0.5	0.74044
7	50.17	1	0.00152	0
8	55.2	1	0.46235	0.1913
9	56.10	1	0.5	0.42181

Table C.11: Z-test and Binomial ∧ Beta Model results for Ireland BL1 with a Beta $(22\,\theta_0, 22-22\,\theta_0)$ prior distribution.

Digit	$\ln(\boldsymbol{B_{01}})$	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p**-value**
1	23.38	1	0.45944	0.18573
2	39.41	1	0.5	0.59892
3	45.18	1	0.27458	0.04486
4	46.08	1	0.00133	0
5	50.37	1	0.11081	0.00994
6	50.57	1	0.00413	0.00018
7	54.32	1	0.47244	0.21308
8	54.76	1	0.3302	0.06715
9	55.2	1	0.25707	0.03934

Table C.12: Z-test and Binomial $\wedge$ Beta Model results for Luxembourg BL1 with a Beta $(22\,\theta_0, 22 - 22\,\theta_0)$ prior distribution.

Digit	$\ln(\boldsymbol{B_{01}})$	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p-**value**
1	11.88	1	0	0
2	36.92	1	0.01442	0.00075
3	45.5	1	0.45994	0.18666
4	49.31	1	0.5	66329
5	46.15	1	0.00017	0
6	53.13	1	0.5	0.4936
7	54.19	1	0.48859	0.26358
8	54.66	1	0.363770	0.08555
9	56.03	1	0.5	0.68984

Table C.13: Z-test and Binomial $\wedge$ Beta Model results for Portugal BL1 with a Beta $(22\,\theta_0, 22 - 22\,\theta_0)$ prior distribution.

Digit	$\ln(\boldsymbol{B_{01}})$	$\boldsymbol{P(H_0\|data)}$	$\underline{\boldsymbol{P}}\boldsymbol{(H_0\|data)}$	p-**value**
1	12.12	1	0	0
2	35.27	1	0.00042	0
3	45.96	1	0.5	0.66228
4	45.43	1	00104	0
5	51.89	1	0.5	0.72552
6	43.47	1	0.5	0.84751
7	53.95	1	0.34849	0.07658
8	55.46	1	0.5	0.59123
9	56.81	1	0.5	0.95888

C.5 Hyperparameter Variances – Multinomial ∧ Dirichlet Model*

Table C.14: Hyperparameter variances and standard deviations in the Multinomial ∧ Dirichlet Model with Dirichlet ($\boldsymbol{\alpha} = \mathbf{1}$) prior. In BL1 analysis the $\text{Dir}_k(\mathbf{1})$ prior for $\boldsymbol{\theta}$ implies a Beta$(1, 8)$ marginal distribution for each parameter in $\boldsymbol{\theta}$. In BL2 the marginal distributions are Beta$(1, 9)$.

	BL1		**BL2**	
i	$\boldsymbol{\sigma^2(\theta_i)}$	$\boldsymbol{\sigma(\theta_i)}$	$\boldsymbol{\sigma^2(\theta_{i+1})}$	$\boldsymbol{\sigma(\theta_{i+1})}$
0	-	-	0.00818	0.09045
1	0.00988	0.09938	0.00818	0.09045
2	0.00988	0.09938	0.00818	0.09045
3	0.00988	0.09938	0.00818	0.09045
4	0.00988	0.09938	0.00818	0.09045
5	0.00988	0.09938	0.00818	0.09045
6	0.00988	0.09938	0.00818	0.09045
7	0.00988	0.09938	0.00818	0.09045
8	0.00988	0.09938	0.00818	0.09045
9	0.00988	0.09938	0.00818	0.09045

Table C.15: Hyperparameter variances and standard deviations in the Multinomial ∧ Dirichlet Model with Dirichlet ($\boldsymbol{\alpha} = \boldsymbol{\theta_0}$) prior. The marginal distributions for each θ_i $(i = 1, \ldots, k)$ are Beta $(\theta_{0i}, 1 - \theta_{0i})$ distributions.

	BL1		**BL2**	
i	$\boldsymbol{\sigma^2(\theta_i)}$	$\boldsymbol{\sigma(\theta_i)}$	$\boldsymbol{\sigma^2(\theta_{i+1})}$	$\boldsymbol{\sigma(\theta_{i+1})}$
0	-	-	0.05268	0.22952
1	0.10521	0.32435	0.05046	0.22463
2	0.07254	0.26934	0.04849	0.22020
3	0.05466	0.23380	0.04672	0.21615
4	0.04376	0.20919	0.04512	0.21242
5	0.03646	0.19093	0.04367	0.20896
6	0.03123	0.17673	0.04233	0.20574
7	0.02731	0.16527	0.04109	0.20272
8	0.02427	0.15578	0.03995	0.19988
9	0.02183	0.14776	0.03889	0.19720

* The results in this section were obtained using the relation: $\theta_i \sim \text{D}_\text{k}(\alpha) \Rightarrow \sigma^2(\theta_i) = \frac{\alpha_i(\alpha_s - \alpha_i)}{\alpha_s^2(\alpha_s+1)}$ where $\alpha_s = \sum_{j=1}^{k} \alpha_k$, and $\sigma(\theta_i) = \sqrt{\sigma^2}$. Because a $\text{Dir}_k(\boldsymbol{\alpha})$ prior distribution for $\boldsymbol{\theta}$ implies a Beta $\left(\alpha_i, \sum_{j=1}^{k+1} \alpha_j - \alpha_i\right)$ marginal distribution for each θ_i in $\boldsymbol{\theta}$, the same variances could be obtained using the formula for the variance of the Beta distribution: $\sigma^2(\theta_i) = \frac{a\,b}{(a+b)^2(a+b+1)}$ where $a = \alpha_i$ and $b = \sum_{j=1}^{k+1} \alpha_j - \alpha_i$.

Table C.16: Hyperparameter variances and standard deviations in the Multinomial $\wedge$ Dirichlet Model with Dirichlet ($\boldsymbol{\alpha} = 22\,\boldsymbol{\theta_0}$) prior. The marginal distributions for each θ_i ($i = 1, \ldots, k$) are Beta $(22\,\theta_{0i}, 22 - 22\,\theta_{0i})$ distributions in BL1 analysis and Beta $(12\,\theta_{0i}, 12 - 12\,\theta_{0i})$ distributions in BL2 analysis.

	BL1		**BL2**	
i	$\boldsymbol{\sigma^2(\theta_i)}$	$\boldsymbol{\sigma(\theta_i)}$	$\boldsymbol{\sigma^2(\theta_{i+1})}$	$\boldsymbol{\sigma(\theta_{i+1})}$
0	-	-	0.00810	0.09002
1	0.00915	0.09565	0.00776	0.08811
2	0.00631	0.07942	0.00746	0.08637
3	0.00475	0.06895	0.00719	0.08478
4	0.00381	0.06169	0.00694	0.08332
5	0.00317	0.05630	0.00672	0.08196
6	0.00272	0.05211	0.00651	0.08070
7	0.00238	0.04874	0.00632	0.07951
8	0.00211	0.04594	0.00615	0.07840
9	0.00190	0.04357	0.00598	0.07735

C.6 Hyperparameter Variances – Binomial $\wedge$ Beta Model*

Table C.17: Hyperparameter variances and standard deviations in the Binomial $\wedge$ Beta Model with Beta $(1, 1)$ prior, resulting in a uniform prior for each parameter.

	BL1		**BL2**	
i	$\boldsymbol{\sigma^2(\theta_i)}$	$\boldsymbol{\sigma(\theta_i)}$	$\boldsymbol{\sigma^2(\theta_{i+1})}$	$\boldsymbol{\sigma(\theta_{i+1})}$
0	-	-	0.08333	0.28868
1	0.08333	0.28868	0.08333	0.28868
2	0.08333	0.28868	0.08333	0.28868
3	0.08333	0.28868	0.08333	0.28868
4	0.08333	0.28868	0.08333	0.28868
5	0.08333	0.28868	0.08333	0.28868
6	0.08333	0.28868	0.08333	0.28868
7	0.08333	0.28868	0.08333	0.28868
8	0.08333	0.28868	0.08333	0.28868
9	0.08333	0.28868	0.08333	0.28868

* The results in this section were obtained using the relations: $\theta_i \sim \text{Beta}(a, b) \Rightarrow \sigma^2(\theta_i) = \frac{a\,b}{(a+b)^2(a+b+1)}$ and $\sigma(\theta_i) = \sqrt{\sigma^2}$.

Table C.18: Prior parameter variances and standard deviations in the Binomial $\wedge$ Beta Model with Beta $(22\,\theta_0, 22 - 22\,\theta_0)$ prior.

	BL1		**BL2**	
i	$\boldsymbol{\sigma^2(\theta_i)}$	$\boldsymbol{\sigma(\theta_i)}$	$\boldsymbol{\sigma^2(\theta_{i+1})}$	$\boldsymbol{\sigma(\theta_{i+1})}$
0	-	-	0.0081	0.09002
1	0.00915	0.09565	0.00776	0.08810
2	0.00631	0.07945	0.00746	0.08637
3	0.00475	0.06895	0.00719	0.08478
4	0.00381	0.06169	0.00694	0.08332
5	0.00317	0.05630	0.00672	0.08196
6	0.00272	0.05211	0.00651	0.0807
7	0.00238	0.04874	0.00632	0.07951
8	0.00211	0.04594	0.00615	0.0784
9	0.0019	0.04357	0.00598	0.07734

Appendix D

VBA Code, Macros and Data

> "*Seeing is not a direct apprehension of reality, as we often like to pretend. Quite the contrary: seeing is inference from incomplete information.*"
>
> Edwin Jaynes (2003)

This link* redirects to a OneDrive shared folder where the data used in this dissertation and the VBA macros developed for the empirical application are stored.

The raw data, consisting in 38 tables extracted from the Eurostat Database can be found in the "Data.xls" file, in the "Data" folder. In that same folder there is a xlsm file with the name of each country being analysed. Each country's xlsm file is where all numbers from all tables of raw data that respect to that country are stored. There is also a xlsm file for the pooled sample, which aggregates the numbers from all countries. The main tool that was used to obtain the study results is the macro in the "Macro_Benford.xlsm" file. By pasting a collection of numbers in the first column of this worksheet, under the cell with the word "Dados", and pressing the "Testar lei de Benford" button, the macro returns: p-values from the chi-square test on the BL1 and BL2, p-values from the z-test on all BL1 and BL2 frequencies, BFs from the Multinomial $\wedge$ Dirichlet Model with $\text{Dir}_k(\mathbf{1})$, $\text{Dir}_k(\boldsymbol{\theta_0})$, $\text{Dir}_k(22\,\boldsymbol{\theta_0})$ and $\text{Dir}_k(\frac{\mathbf{1}}{\boldsymbol{k}})$ prior distributions, BFs from the Binomial $\wedge$ Beta Model with $\text{Beta}(1,1)$, $\text{Beta}(\theta_0, 1-\theta_0)$, $\text{Beta}(22\,\theta_0, 22-22\,\theta_0)$ and $\text{Beta}(\frac{1}{k}, \frac{1}{k})$ prior distributions, the interpretation of each BF in

* https://1drv.ms/f/s!ArrG5X_w8Yihuz036OMFLvva3ZNm

terms of strength of evidence according to the scale in table C.1, the posterior probabilities computed using 3.16 on each BF, the lower bounds on posterior probabilities computed with the p-value calibration on equation 3.20 and the average, median and skewness coefficient of the dataset. For this macro to work properly at least one of the inserted numbers must have more than one digit. It is possible to consult the underlying VBA code by pressing the Visual Basic button on the Developer tab of the Macro_Benford.xlsm file.

The file Variances.xlsx is the excel worksheet where all the hyerparameter variances and standard deviations were computed. The excel formulas that were used can be consulted in this file. The file Macro_Probabilities.xls is an excel worksheet where the BL frequencies for the first digit, second digit, last digit, first two digits, first three digits and last two digits can be consulted.

"The classical theorists resemble Euclidean geometers in a non-Euclidean world who, discovering that in experience straight lines apparently parallel often meet, rebuke the lines for not keeping straight as the only remedy for the unfortunate collisions which are occurring. Yet, in truth, there is no remedy except to throw over the axiom of parallels and to work out a non-Euclidean geometry. Something similar is required today in economics"

John Maynard Keynes (1937)