Spanish Journal of Statistics

VOLUME 4, NUMBER 1, 2022

INE

NIPO: 096-22-006-0
ISSN: 2695 - 9070
DOI: https://doi.org//10.37830/SJS.2022.1

Spanish Journal of Statistics

Published by the Spanish National Statistical Institute

SPANISH JOURNAL OF STATISTICS

VOLUME 4, NUMBER 1, 2022

Contents

Editorials **2**

Presentation of Volume 4, 1, 2022 **5**
J.M. Sarabia

Research papers **7**

A review on specification tests for models with functional data **9**
W. González-Manteiga

Testing Benford's law: from small to very large data sets **41**
L. Campanelli

The gamma flexible Weibull distribution: Properties and Applications **55**
A.A. Ferreira, G.M. Cordeiro

On moments and entropy of the gamma-Gompertz distribution **73**
F. Castellares, A. J. Lemonte

Official statistics **80**

A first interim assessment of the third round of peer review of the European statistical system **81**
A. Cañada

Use of death statistics according to cause of death in health research **89**
G. Barrio

The Statistics on Causes of Death: characteristics and improvements **99**
M.G. Ferruelo, M.R. González

Mortality statistics **107**
E. Regidor

Acknowledgement to Reviewers **116**

SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1, 2022, Pages 5 – 7
doi: https://doi.org/10.37830/SJS.2022.1.01

EDITORIAL

Presentation of Volume 4, 1, 2022

José María Sarabia

Editor-in-Chief Spanish Journal of Statistics

Dear readers and dear members of the statistical community:

It is a great pleasure for me to present Volume 4, 1, corresponding to the year 2022. This volume is made up of eight papers: one invited article, three articles within the general statistics section and four articles in the section of official statistics.

The invited article has the title: "A review on specification tests for models with functional data", whose author is Professor Wenceslao González-Manteiga, winner of the second National Statistics Prize. Wenceslao is Professor of Statistics and Operations Research at the University of Santiago de Compostela. The award jury highlighted the contribution of Prof. González-Manteiga to non-parametric modeling of dynamics and dependencies in complex systems and to the development of non-parametric statistics over the last 30 years. The winner has been teaching for 42 years, during which he has participated in university management and scientific evaluation at all levels, supervsing more than 30 doctoral theses. Thanks to his work, he has contributed knowledge to society, both in the scientific field (Engineering, Chemistry, Biology, Economics or Medicine) as well as in the industrial sector.

The article presents the most relevant specification tests for models with functional data. Due to the progress in technological advances, massive amounts of data are currently generated and new statistical methodology should be properly deployed to manage this information. The functional data are an example of particular importance. The article reviews the most notable developments in this context, providing some nice illustrations from real data sets.

The next three papers are presented in the general section. The first paper of this section is titled, "Testing Benford's Law: from small to very large data sets", and its author is Leonardo Campanelli. The paper discuss some limitations of the use of generic tests, such as the Pearson's χ^2, for testing Benford's law. The article introduces a new statistic whose sample values are asymptotically independent on the sample size making it a natural candidate for testing Benford's law in very large data sets.

The title of the second paper is, "The gamma flexible Weibull distribution: Properties and Applications", whose authors are Alexsandro A. Ferreira and Gauss M. Cordeiro. The paper proposes a new gamma flexible Weibull distribution, which presents a bathtub-shaped hazard rate, and some of its properties are obtained, including estimation and simulation to examine the consistency of the estimates. The utility of the proposed model is analysed using three real

Published by the Spanish National Statistical Institute

applications.

The last paper of this section is titled, “On moments and entropy of the gamma-Gompertz Distribution", by Fredy Castellares and Artur J. Lemonte. The three-parameter gamma-Gompertz family of distributions was introduced recently in the literature. The analytical expressions provided for the ordinary moments and Shannon entropy are not correct and hence cannot be used for computing such quantities. The authors derive two closed-form expressions for the mean and a closed-form expression for the Shannon entropy in terms of the Whittaker function.

The third section is dedicated to the articles in the the Official Statistics section. We have four interesting papers.

The title of the first paper is, “A first interim assessment of the third round of peer review of the European statistical system", by Agustín Cañada. Peer Reviews are exercises to assess compliance with the principles and indicators of the European Statistics Code of Practice by the members of the European Statistical System: Eurostat and the national statistical systems. Peer Reviews are carried out periodically (every 5/6 years), by agreement of the European Union. To date, three rounds have been carried out: in 2006-2008, in 2013-2015, and a third round is underway between 2021 and 2023. Although the third round is still ongoing at the time of writing, based on the experience of a representative group of the countries already reviewed, a first assessment can already be made of the degree of achievement of the objectives pursued. The aim of this document is to provide a first input for a future comprehensive “lessons learned exercise" and to contribute to the debate on aspects to be taken into account in future peer reviews.

The remaining three articles are dedicated to the study of mortality statistics and causes of death. The first of these articles is titled “Use of death statistics according to cause of death in health research" by Gregorio Barrio. The article discusses several aspects related to the estimation of total and cause-specific mortality rates. On the other hand, the link between socioeconomic indicators and mortality, considered by the Spanish Statistical Office, which makes it possible to study the relationship between socioeconomic factors and mortality and its variation over time is also discussed.

The title of the second paper is, “The Statistics on Causes of Death: characteristics and improvements" by Margarita García Ferruelo and María Rosario González García. This article describes the complex process of the statistics, the advances achieved in recent years, such as the implementation of an international automatic system for coding multiple causes of death and for the selection of the underlying cause or the improvement in obtaining the external causes of death, as well as its usefulness for the epidemiological studies and health research. It is also discussed some of the lessons learned during the worst pandemic period. Finally, it is proposed to collect other variables of interest for the analysis of the causes of death using available administrative sources.

The last paper of this section is titled, “Mortality statistics for assessing population health" by Enrique Regidor. The article deals with several interesting aspects of mortality statistics. From the health system perspective, the adoption of the International Classification of Diseases and Causes of Death was a crucial milestone in population health statistics, shedding light on the diseases responsible for most deaths and the trends in causes of death over time. Morbidity statistics and public health surveillance systems have important objectives, but they do not allow adequate

monitoring of the frequency of diseases and other health problems, nor can they quantify diseases' impact on population health. On the other hand, statistics on cause of death do provide this information thanks to the combination of two features: the exhaustiveness of the data they collect and the objective nature of the phenomenon they quantify.

Finally, I would like to thank again all the authors of this volume for choosing our journal as a means of disseminating their research. I appreciate the work of the editors and reviewers of the papers, who contribute to maintaining a high standard of scientific quality.

SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 9–40
doi:https://doi.org/10.37830/SJS.2022.1.02

INVITED ARTICLE

A review on specification tests for models with functional data

Wenceslao González-Manteiga
Centre for Mathematical Research and Technology Transfer of Galicia (CITMAga).
Department of Statistics, Mathematical Analysis and Optimization.
Universidad de Santiago de Compostela, Santiago de Compostela,
wenceslao.gonzalez@usc.es

Received: Novermber 30, 2022. Accepted: December 14, 2022

Abstract: Nowadays, due to the progress in technological advances, massive amounts of data are generated. As a result, new statistical methodology is needed to properly manage this information. The functional data are an example of special importance. These are mainly obtained by means of high-frequency measurements (spectrometric curves, stock prices recording, etc.). Since the beginning of this century, this type of data has achieved great popularity. This fact has generated new distribution or regression models, among others, appropriate to the functional context. In the last 10 years, novel specification tests are proposed for those models. These are generalizations of methodologies developed for the vectorial framework over the last century. Besides, innovative procedures based on distance correlation ideas have been proposed as well. This article reviews the most notable developments in this context, providing some illustrations from real data sets.

Keywords: distance correlation, functional data, goodness-of-fit, regression models

MSC: 62R10, 62G10

1 Introduction

The invention of computers meant a real change in statistical methodology in the last century. The scientific developments, derived mainly from the first half of the 20th century, were headed to understand existing real data sets information. These were of medium size, obtained wit a great effort in many cases. Other developments in Statistics during the first part of the 20th century were leaded to the design of algorithms for the estimation and testing of different models parameters. In all of these, the computational burden was considerable for the available and quite limited calculus capacity of that moment. Real parameters were estimated, but not curves due to poor graphics resources. The behavior of the statistics distributions were analyzed, under some parametric hypoth-

esis, because of the obvious impossibility of working with large sample sizes in a nonparametric way.

In the 80s, the versatility provided by advances in computer calculus generated new statistical procedures. These are based on simulating artificial data, as is the case of "Bootstrap". Nevertheless, it is not until the next decades, motivated by Internet use, new technologies information, development of distributed as well as parallelized computing, and computational costs reduction for storage and data processing, when the beginning of the "Big Data" age can be established. This phenomenon has a great impact on the development of modern technology in Statistics as well as on all its applications.

Currently, many companies already have continuous and real-time monitoring systems: stock quotes can be measured as high-frequency data, information generated by web pages, social media data or just the credit cards transactions are some examples of massive information generation sources. Other example can be found in the electric market, where high-frequency measures about energy consumption or demand are available as well. In all these cases it is quite relevant to be able to correctly process and control the information.

It is, precisely, in this context of massive, high-frequency or related data, where functional data arise. This kind of data gains an immense popularity with Ramsay and Silverman (2005), Ferraty and Vieu (2006) or more recently with Horváth and Kokoszka (2012), Hsing and Eubank (2015) and Kokoszka and Reimherr (2017), among others. Functional data allows to summary a great amount of information through a curve, surface or, in general, using a "statistical object". This last is typically modeled in a functional space, such as Hilbert spaces.

The management of functional data guide us, naturally, to the consideration of models based on these (distribution models, regression models, etc.), employed for prediction purposes, using interpretation of the results in diverse applications. Thus, the necessity of mechanism for specification testing devoted to models with functional data appears. In this article, the diverse procedures that have emerged during this period are reviewed. These have been mainly developed in the last 10 years, generalizing classic procedures introduced in the first half of the 20th century, essentially based on the empirical distribution, and the more recent advances in the last part of the 20th century making use of nonparametric estimations of the density or regression function.

Although the Goodness-of-Fit (GoF) term is due to Pearson, at the beginning of the 20th century, it is not until the 70s with Durbin (1973) and Bickel and Rosenblatt (1973) where modern specification tests start. These are based on distances between nonparametric estimators of the distribution or density function with respect to hypothetical estimations under the null hypothesis of the model.

In this way, formally, assume that $\{X_1, \ldots, X_n\}$ is an identically and independent distributed (iid) sample of a random variable X with (unknown) distribution F (or density f, if that is the case). If the target function is the distribution F, then the GoF testing problem can be formulated as testing $H_0 : F \in \mathcal{F}_\Theta = \{F_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ vs. $H_1 : F \notin \mathcal{F}_\Theta$, where $\mathcal{F}_\Theta$ stands for a parametric family of distributions indexed in some finite-dimensional set Θ. A general test statistic for this problem can be written as $T_n = T(F_n, F_{\hat{\theta}})$, with the functional T denoting, here and henceforth, some kind of distance between a nonparametric estimate, given in this case by the mentioned empirical cumulative distribution function $F_n(x) = n^{-1} \sum_{i=1}^n \mathbb{I}(X_i \leq x)$, and an estimate obtained under the null hypothesis H_0, $F_{\hat{\theta}}$ in this case. Similarly, for the case of a parametric density model,

the testing problem is formulated as $H_0 : f \in f_\Theta = \{f_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ vs. $H_1 : f \notin f_\Theta$ and can be approached with the general test statistic $T_n = T(f_{nh}, f_{\hat{\theta}})$. In this setting, $f_{\hat{\theta}}$ is the density estimate under H_0 and f_{nh} denotes a general nonparametric density estimate, as for example, the kernel density estimator $f_{nh}(x) = n^{-1}\sum_{i=1}^n K_h(x - X_i)$ introduced by Parzen (1962) and Rosenblatt (1956) where $K_h(\cdot) = K(\cdot/h)/h$, K is the kernel function $\left(K(x) \geq 0 \text{ and } \int K(x)dx = 1\right)$, and h is the smoothing bandwidth.

More recent procedures were generalized to the context of regression models in the 1990s. Consider a nonparametric, random design, regression model such that $Y = m(X) + \varepsilon$, with $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, $m(x) = \mathbb{E}[Y|X = x]$ and $\mathbb{E}[\varepsilon|X = x] = 0$. Denote by $\{(X_i, Y_i)\}_{i=1}^n$ an iid sample of (X, Y) satisfying such a model. In this context, the GoF goal is to test $H_0 : m \in \mathcal{M}_\Theta = \{m_\theta : \theta \in \Theta \subset \mathbb{R}^q\}$ vs. $H_1 : m \notin \mathcal{M}_\Theta$, where $\mathcal{M}_\Theta$ represents a parametric family of regression functions indexed in Θ. Following Durbin (1973) and Bickel and Rosenblatt (1973) ideas, the seminal works of Stute (1997) and Härdle and Mammen (1993), respectively, introduced two types of GoF tests for regression models:

a) Tests based on empirical regression processes, considering distances between estimates of the integrated regression function $I(x) = \int_{-\infty}^x m(t)\,\mathrm{d}F(t)$ (F being the marginal distribution of X under H_0 and H_1). Specifically, the test statistics are constructed as $T_n = T(I_n, I_{\hat{\theta}})$, with $I_n(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(X_i \leq x)Y_i$ and $I_{\hat{\theta}}(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(X_i \leq x)m_{\hat{\theta}}(X_i)$.
b) Smoothing-based tests, using distances between estimated regression functions, $T_n = T(m_{nh}, m_{\hat{\theta}})$, with m_{nh} a smooth regression estimator. As a particular case, $m_{nh}(x) = \sum_{i=1}^n W_{nh,i}(x)Y_i$, with $W_{nh,i}(x)$ some weights depending on a smoothing parameter h. Such an estimator can be obtained, for example, with Nadaraya–Watson or local linear weights (see, e.g., Wand and Jones (1995)).

A complete review of these methodologies, related to specification tests, can be consulted in González-Manteiga and Crujeiras (2013). This is an invited article with discussion for the TEST journal. In this reference, different contributions on this topic since 1990 are reviewed. Resulting statistics for diverse specification tests as well as its distribution calibration, by means of asymptotic techniques or resampling procedures like the Bootstrap, are studied. This review analyzes more than 20 years of developments, being very scarce or almost non-existent the procedures designed for functional data. Very recently, in a chapter of the book González-Manteiga et al. (2022), we perform a review of the existing methodologies for specification tests in the functional data context. These procedures are mainly based on extensions of the methodology introduced in the 90s for specification tests in the vectorial framework to the functional case.

In this paper, an update of the chapter corresponding to the book González-Manteiga et al. (2022) is provided in the next section. Later, in Section 3, the "fundamental case" of the manuscript is presented. This is covered with a detailed review of specification tests based on "distance correlation" ideas and their novel extension to the functional data context. In Section 4 some applications to real data sets for specification testing in the functional framework, applying techniques introduced in previous sections, are displayed. Finally, some conclusions arise in Section 5 and the document finishes with an exhaustive revision of relevant references.

2 Testing specification models for functional data using smoothing or empirical processes

In this section we review the most notable results for specification tests in terms of the distribution function or regression models in the functional data context. For the development of these procedures it is necessary to include functional data in complex structures associated to general spaces (metric or topological ones), as Hilbert spaces. These represent a natural and quite employed way for adequate model description in the functional context.

2.1 GoF for distribution models for functional data

Let $\mathcal{H}$ denote a Hilbert space over $\mathbb{R}$, the norm of which is given by its scalar product as $\|x\| = \sqrt{\langle x, x\rangle}$. Consider $\{X_1, \ldots, X_n\}$ iid copies of the random variable $X : (\Omega, \mathcal{A}) \to (\mathcal{H}, \mathcal{B}(\mathcal{H}))$, with $(\Omega, \mathcal{A}, \mathbb{P})$ the probability space where the random sample is defined and $\mathcal{B}(\mathcal{H})$ the Borel σ-field on $\mathcal{H}$. The general GoF problem for the distribution of X consists on testing $H_0 : \mathrm{P}_X \in \mathcal{P}_\Theta = \{\mathrm{P}_\theta : \theta \in \Theta\}$ vs. $H_0 : \mathrm{P}_X \notin \mathcal{P}_\Theta$, where $\mathcal{P}_\Theta$ is a class of probability measures on $\mathcal{H}$ indexed in a parameter set Θ, now possibly infinite-dimensional, and P_X is the (unknown) probability distribution of X induced over $\mathcal{H}$.

When the goal is to test the simple null hypothesis $H_0 : \mathrm{P}_X \in \{\mathrm{P}_0\}$, a general feasible approach that enables the construction of different test statistics is based on projections $\pi : \mathcal{H} \to \mathbb{R}$, in such a way that the test statistics are defined from the projected sample $\{\pi(X_1), \ldots, \pi(X_n)\}$. Such an approach can be taken on the projected distribution function: $T_{n,\pi} = T(F_{n,\pi}, F_{0,\pi})$ with $F_{n,\pi}(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(\pi(X_i) \leq x)$ and $F_{0,\pi}(x) = \mathbb{P}_{H_0}(\pi(X) \leq x)$. Some specific examples are given by the adaptation to this context of the Kolmogorov–Smirnov, Cramer–von Mises, or Anderson–Darling type tests. As an alternative, based on smoothing techniques tests presented in Section 1, a test statistic can also be built as $T_{n,\pi} = T(f_{nh,\pi}, \mathbb{E}_{H_0}[f_{nh,\pi}])$ with $f_{nh,\pi}(x) = n^{-1}\sum_{i=1}^n K_h(x - \pi(X_i))$ the density estimate of $\pi(x)$. It should be also noted that, when embracing the projection approach, the test statistic may take into account 'all' the projections within a certain space, e.g. by considering $T_n = \int T_{n,\pi}\, \mathrm{d}W(\pi)$ for W a probability measure on the space of the different projections, or take just $T_n = T_{n,\hat{\pi}}$ with $\hat{\pi}$ being a randomly-sampled projection from a certain non-degenerate probability measure W.

Now, when the goal is to test the composite null hypothesis $H_0 : \mathrm{P}_X \in \mathcal{P}_\Theta$, the previous generic approaches are still valid if replacing $\mathrm{P}_{0,\pi}(x)$ with $\mathrm{P}_{\hat{\theta},\pi}(x) = \mathbb{P}_{\mathrm{P}_{\hat{\theta}}}(\pi(X) \leq x)$. Cuesta-Albertos et al. (2006) and Cuesta-Albertos et al. (2007) provide a characterization of the composite null hypothesis by means of random projections, and provide a bootstrap procedure for calibration, see also Bugni et al. (2009) and Ditzhaus and Gaigall (2018). In the space of real square-integrable functions $\mathcal{H} = L^2[0,1]$, one may take $\pi_h(x) = \langle x, h\rangle$, with $h \in \mathcal{H}$. The previous references provide also some approaches for the calibration of the tests under the null hypothesis of the rejection region $\{T_n > c_\alpha\}$, where $\mathbb{P}(T_n > c_\alpha) \leq \alpha$.

A very relevant alternative to the procedures based on projections is the use of the so-called "energy statistics" Székely and Rizzo (2017). Working with $\mathcal{H}$ a general Hilbert separable space (as it can be seen in Lyons (2013)) if $X \sim \mathrm{P}_X$ and $Y \sim \mathrm{P}_Y = \mathrm{P}_0$ (P_0 being the distribution under the null)

then

$$E = E(X,Y) = 2\mathbb{E}[\|X - Y\|] - \mathbb{E}[\|X - X'\|] - \mathbb{E}[\|Y - Y'\|] \geq 0, \tag{1}$$

with $\{X, X'\}$ and $\{Y, Y'\}$ iid copies of the variables with distributions P_X and P_Y, respectively. Importantly, (1) equals 0 if and only if $\mathrm{P}_X = \mathrm{P}_Y$, a characterization that serves as basis for a GoF test. See the nice review of Székely and Rizzo (2017), where a motivation is given for the duality between the expression displayed in (1) and the well-known energy formula of Einstein.

The energy statistic in (1) can be empirically estimated from a sample $\{X_1, \ldots, X_n\}$ as

$$\hat{E}^* = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|X_i - Y_j^*\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|X_i - X_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|Y_i^* - Y_j^*\|,$$

The distribution of $\hat{E}^*$, $\mathbb{P}^*\left\{\hat{E}^* \leq x\right\}$ can be approximated by simulation of the artificial variable $Y^* \sim P_Y$, resulting in $\left\{Y_1^{*b}, \ldots, Y_n^{*b}\right\}$ with $b = 1, \ldots, B$. The critic point for a given α level can be obtained in a natural way from the quantiles of the sorted sample: $\hat{E}^{*(1)} \leq \cdots \leq \hat{E}^{*(B)}$ as a result of the Monte Carlo replicates of the artificial samples.

The most studied case is the Gaussian one, where P_Y follows the distribution of a Gaussian process. This is analyzed in recent literature as in Kellner and Celisse (2019), Kolkiewicz et al. (2021), Górecki and Łukasz (2019), Henze and Jiménez-Gamero (2021) and Bongiorno et al. (2019). In these works, diverse alternatives to the mentioned procedures are provided and reviewed to specification of a functional model of Gaussian distribution or related.

Finally, in the context of tests for distributions, it is worth it to mention the related two-sample problem, a common offspring of the simple-hypothesis one-sample GoF problem. This topic has been extensively studied for scalar random data in the last decades. However, the situation involving functional random data has attracted less attention until now. Three main related approaches have been considered in this setting recently, namely,

a) Comparison of functional means using, e.g., principal component approaches (Horváth and Rice (2015), Ghale-Joogh and E. Hosseini-Nasab (2018)) or adapting the ideas of the F-test to the functional context (Cuevas et al. (2004), González-Rodrígez et al. (2012), Górecki and Łukasz (2019), Lee et al. (2015), Zhang and Liang (2014), Qiu et al. (2021)).

b) Comparison of covariance structures (Boente et al. (2018), Fremdt et al. (2013), Guo et al. (2018), Guo et al. (2019), Guo et al. (2019)).

c) Comparison of the distribution structure in various ways. Tests based on smoothing discrete observed data in potential functional data are developed in Bárcenas et al. (2017) and, similarly, in Estévez-Pérez and Vilar (2013) or Pomann et al. (2016). Empirical processes have been used in Bárcenas et al. (2017). An L^2-type criterion based on empirical distribution functions is used in Jiang et al. (2019). Some Cramér-von Mises-type statistics adapted to the functional case are employed in Bugni and Horowitz (2021).

2.2 GoF for regression models with functional data based on smoothing or empirical processes

We assume in the following, for easier presentation of the different methods, that both the predictor X and response Y are centered, so that the intercepts of the linear functional regression models are null.

A particular case of a regression model with functional predictor and scalar response is the so-called functional linear model. For $\mathcal{H}_X = L^2[0,1]$, this parametric model is given by

$$Y = m_\beta(X) + \varepsilon, \quad m_\beta(x) = \langle x, \beta \rangle = \int_0^1 x(t)\beta(t)\,\mathrm{d}t, \tag{2}$$

for some unknown $\beta \in \mathcal{H}_X$ indexing the functional form of the model and $\mathbb{E}\left[\varepsilon|_X\right] = 0$. This model is the natural generalization of the classical and popular linear (Euclidean) regression models.

In general, there have been two approaches for the inference on (2): (*i*) testing the significance of the trend within the linear model, i.e., testing $H_0 : m \in \{m_{\beta_0}\}$ vs. $H_1 : m \in \{m_\beta : \beta \in \mathcal{H}_X, \beta \neq \beta_0\}$, usually with $\beta_0 = 0$; (*ii*) testing the linearity of m, i.e., testing $H_0 : m \in \mathcal{L} = \{m_\beta : \beta \in \mathcal{H}_X\}$ vs. $H_1 : m \notin \mathcal{L}$.

For the GoF testing problem presented in (*ii*), given an iid sample $\{(X_i, Y_i)\}_{i=1}^n$ and following Härdle and Mammen (1993) ideas in the vectorial case, a test statistic structure can be given by $T_n = T(m_{nh}, m_{\hat{\beta}})$, where $\hat{\beta}$ is a suitable estimator for β and

$$m_{nh}(x) = \sum_{i=1}^n W_{ni}(x) Y_i = \sum_{i=1}^n \frac{K_h(\|x - X_i\|)}{\sum_{j=1}^n K_h(\|x - X_j\|)} Y_i \tag{3}$$

is the Nadaraya–Watson estimator with a functional predictor. In Delsol et al. (2011), a L_2 distance is offered,

$$T_n = \int \left(m_{nh}(x) - m_{nh,\hat{\beta}}(x)\right)^2 \omega(x)\,\mathrm{dP}_X(x),$$

where $m_{nh,\hat{\beta}}$ is a smoothed version of the parametric estimator that follows by replacing Y_i with $m_{\hat{\beta}}(X_i)$ in (3). A crucial problem is the computation of the critical region $\{T_n > c_\alpha\}$, which depends on the selection of h when a class of estimators for β is used under the null. This class of smoothed-based tests, or related, were deeply studied in the Euclidean setting (see González-Manteiga and Crujeiras (2013)). Nevertheless, this is not the case in the functional context, except for this mentioned contribution and others more recent by Maistre and Patilea (2020) and Patilea and Sánchez-Sellero (2020).

As in the vectorial case, it is possible to avoid the bandwidth selection problem using tests based on empirical regression processes. For this purpose, a key element is the empirical counterpart of the integrated regression function $I_n(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(X_i \leq x) Y_i$, where $X_i \leq x$ means that $X_i(t) \leq x(t)$, for all $t \in [0,1]$. In this scenario, the test statistic can be formulated as $T_n(I_n, I_{\hat{\beta}})$, where $I_{\hat{\beta}}(x) = n^{-1}\sum_{i=1}^n \mathbb{I}(X_i \leq x)\hat{Y}_i$, where $\hat{Y}_i = \langle X_i, \hat{\beta}\rangle$. Deriving the theoretical behavior of an empirical regression process indexed by $x \in \mathcal{H}_X$, namely $R_n(x) = \sqrt{n}(I_n(x) - I_{\hat{\beta}}(x))$ is, still today, a challenging task. Yet, as previously presented, the useful projection approach over $\mathcal{H}_X$ can be

considered. The null hypothesis $H_0 : m \in \mathcal{L}$ can be formulated by means of

$$H_0 : \mathbb{E}[(Y - \langle X, \beta\rangle)\mathbb{I}(\langle X, \gamma\rangle \leq u)] = 0, \text{ for a } \beta \in \mathcal{H}_X \text{ and for all } \gamma \in \mathcal{H}_X,$$

which in turn is equivalent to replacing 'for all $\gamma \in \mathcal{H}_X$' with 'for all $\gamma \in \mathcal{S}_{\mathcal{H}_X}$' or 'for all $\gamma \in \mathcal{S}^{p-1}_{\mathcal{H}_X,\{\psi_j\}_{j=1}^{\infty}}$, for all $p \geq 1$', where

$$\mathcal{S}_{\mathcal{H}_X} = \{\rho \in \mathcal{H}_X : \|\rho\| = 1\}, \quad \mathcal{S}^{p-1}_{\mathcal{H}_X,\{\psi_j\}_{j=1}^{\infty}} = \left\{\rho = \sum_{j=1}^{p} r_j\psi_j : \|\rho\| = 1\right\}$$

are infinite- and finite-dimensional spheres on $\mathcal{H}_X$, $\{\psi_j\}_{j=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}_X$, and $\{r_j\}_{j=1}^{p} \subset \mathbb{R}$. As follows from García-Portugués et al. (2014) a general test statistic can be built aggregating all the projections within a certain subspace: $T_n = \int T_{n,\pi}\,\mathrm{d}W(\pi)$ with $T_{n,\pi} = T(I_{n,\pi}, I_{\hat{\beta},\pi})$ based on

$$I_{n,\pi}(u) = n^{-1}\sum_{i=1}^{n}\mathbb{I}(\pi(X_i) \leq u)Y_i \text{ and } I_{\hat{\beta},\pi}(u) = n^{-1}\sum_{i=1}^{n}\mathbb{I}(\pi(X_i) \leq u)\hat{Y}_i, \tag{4}$$

for $\pi(x) = \langle x, \gamma\rangle$. In this case, W is a probability measure defined in $\mathcal{S}_{\mathcal{H}_X}$ or $\mathcal{S}^{p-1}_{\mathcal{H}_X,\{\psi_j\}_{j=1}^{\infty}}$, for a certain $p \geq 1$. Alternatively, the test statistic can be based on only one random projection: $T_n = T_{n,\hat{\pi}}$. More generally, T_n may consider the aggregation of a finite number of random projections, as advocated in the test statistic of Cuesta-Albertos et al. (2019). Both types of tests, all-projections and finite-random-projections, may feature several distances for T, such as Kolmogorov–Smirnov or Cramár–von Mises types.

In the last years, more general procedures for model (2) focus on model specification with scalar response and functional covariate are defined. McLean et al. (2015) consider the functional generalized additive model

$$Y = m_{\mathcal{F}} + \varepsilon = \eta + \int_0^1 \mathcal{F}(X(t), t)dt, \tag{5}$$

being (2) a particular case of (5) taking $\mathcal{F}(x, t) = x\beta(t)$ and $\eta = 0$, whereas Horváth and Reeder (2013) take under consideration the functional quadratic regression model

$$Y = \int_0^1 \beta(t)X(t)dt + \int_0^1\int_0^1 \gamma(s, t)X(t)X(s)dtds + \varepsilon \tag{6}$$

where (2) corresponds with taking $\gamma = 0$ in (6).

Besides, we can highlight some recent alternatives: generalizing the well-known F-test for specification testing or, more generally, the likelihood ratio test. See McLean et al. (2015) or Kong et al. (2016). New ones which establish alternative tests with easy to calibrate distribution as Shi et al. (2022) or devoted to speed computational tasks as in Zhao et al. (2022).

It is also worth mentioning literature comparing the above mentioned procedures. See Tekbudak et al. (2019) for an extensive comparative between procedures based on smoothing techniques, empirical processes and adapted statistics from the likelihood ratio test.

When both the predictor and the response, X and Y, are functional random variables evaluated in $\mathcal{H}_X = L^2[a,b]$ and $\mathcal{H}_Y = L^2[c,d]$, the regression model $Y = m(X) + \varepsilon$ is related with the operator $m : \mathcal{H}_X \to \mathcal{H}_Y$. Perhaps the most popular operator specification is a (linear) Hilbert–Schmidt integral operator, expressible as

$$m_\beta(x)(t) = \langle x, \beta(\cdot,t)\rangle = \int_a^b \beta(s,t)x(s)\,\mathrm{d}s, \quad t \in [c,d], \tag{7}$$

for $\beta \in \mathcal{H}_X \otimes \mathcal{H}_Y$, which is simply referred to as the functional linear model with functional response. The kernel β can be represented as $\beta = \sum_{j=1}^\infty \sum_{k=1}^\infty b_{jk}(\psi_j \otimes \phi_k)$, with $\{\psi_j\}_{j=1}^\infty$ and $\{\phi_k\}_{k=1}^\infty$ being orthonormal bases of $\mathcal{H}_X$ and $\mathcal{H}_Y$, respectively.

Similarly to the case with scalar response, performing inference on (7) have attracted the analogous two mainstream approaches: (*i*) testing $H_0 : m \in \{m_{\beta_0}\}$ vs. $H_1 : m \in \{m_\beta : \beta \in \mathcal{H}_X \otimes \mathcal{H}_Y, \beta \neq \beta_0\}$, usually with $\beta_0 = 0$; (*ii*) testing $H_0 : m \in \mathcal{L} = \{m_\beta : \beta \in \mathcal{H}_X \otimes \mathcal{H}_Y\}$ vs. $H_1 : m \notin \mathcal{L}$. The GoF problem given in (*ii*) can be approached by considering a double-projection mechanism based on $\pi_X : \mathcal{H}_X \to \mathbb{R}$ and $\pi_Y : \mathcal{H}_Y \to \mathbb{R}$. Given an iid sample $\{(X_i, Y_i)\}_{i=1}^n$, a general test statistic follows (see García-Portugués et al. (2021)) as $T_n = \int T_{n,\pi_X,\pi_Y}\,\mathrm{d}W(\pi_X \times \pi_Y)$ with $T_{n,\pi_X,\pi_Y} = T(I_{n,\pi_X,\pi_Y}, I_{\hat\beta,\pi_X,\pi_Y})$, where I_{n,π_1,π_2} and $I_{\hat\beta,\pi_1,\pi_2}$ follows from (4) by replacing π with π_X, and Y_i and $\hat Y_i$ with $\pi_Y(Y_i)$ and $\pi_Y(\hat Y_i)$, respectively. In this case, W is a probability measure is defined in $\mathcal{S}_{\mathcal{H}_X} \times \mathcal{S}_{\mathcal{H}_Y}$ or $\mathcal{S}^{p-1}_{\mathcal{H}_X,\{\psi_j\}_{j=1}^\infty} \times \mathcal{S}^{q-1}_{\mathcal{H}_Y,\{\phi_k\}_{k=1}^\infty}$, for certain $p, q \geq 1$. The projection approach is immediately adaptable to the GoF of (7) with $\mathcal{H}_X = \mathbb{R}$, and allows graphical tools for that can help detecting the deviations from the null, see García-Portugués et al. (2020). An alternative route considering projections just for X is presented by Chen et al. (2020).

The above generalization to the case of functional response is certainly more difficult for the class of tests based on the likelihood ratios. Regarding the smoothing-based tests, Patilea et al. (2016) introduced a kernel-based significance test consistent for nonlinear alternative. Moreover, Smaga (2022) extends the F-test to the context of functional response making use of projections.

3 A new generation of procedures for testing in regression models based on distance correlation

Since the article of Székely et al. (2007), with the first correlation distance methodology development, there has been a huge variety of works using its ideas for independence tests. Some of them focused on the specification testing field. Very recently, in the last five years, new procedures for specification testing have been derived extending correlation distance ideas. These have resulted in novel covariates selection or GoF approaches. In case of covariates selection, this translates in testing if all considered $X_1, \ldots, X_p$ covariates are relevant to explain a variable Y or some can be excluded from the model. For this aim, the covariates selection problem is rewritten as an independence test and distance correlation methodology is used to construct proper statistics. For GoF the model is estimated under the null hypothesis assumptions and then, the independence between the estimation of the model error and covariates is tested. As a result, specification tests result in independence ones which can be performed using distance correlation ideas.

In this section, a first timeline review of classical methods for independence or significance testing in regression models, being special cases of specification tests, is carried out in Section 3.1. We

highlight the most notable procedures and expose their drawbacks. Then, the benefits of the distance correlation based tests, specially in the high-dimensional context of $p > n$, are motivated. Next, a review of the distance correlation and derivatives methodology is introduced in Sections 3.2, 3.3 and 3.4. The distance correlation, the martingale difference divergence and the conditional distance correlation coefficients, as well as their associated independence tests, are described for the vectorial framework in these sections, respectively. Eventually, specific advances for statistics based on distances in the functional data context are detailed in Section 3.5.

3.1 Previous considerations of correlation measures based on distances

During the last decades, covariates selection procedures have received special attention. This study has been specially focused on the big data context, in which the number of covariates (p) is high, even larger than the sample size (n), $p > n$. As a result, several covariates selection techniques have been developed for this framework.

From the beginning, one of the first and well-known dependence measures for random vectors is the correlation coefficient. See, for example, Pearson (1920). This allows to perform covariates selection taking under consideration only covariates with the greatest correlation value with the response. However, this is only able to correctly detect linear relations. As a result, we can only select covariates if we can assume a linear structure in the regression model. With the aim of identifying other types of dependence, other coefficients measures based on ranks were proposed. These are the Spearman's coefficient (Wissler (1905)) or the Kendall's τ (Kendall (1938)). These measures are robust to outliers and detect any type of monotone dependence pattern. Nevertheless, it is not possible to identify non-monotone structures, being unsuitable for some regression models. These techniques only measure the grade of dependence for each covariate separately and do not pay attention to the information provided by the rest of them in the process. Moreover, the computational cost increases in terms of the p size.

If certain structure of the regression model can be assumed, this information can be employed to perform significance tests for covariates selection. For example, under the linearity assumption with Gaussian errors, we can resort to the well-known F-test. Nonetheless, these methodologies are not available in the $p > n$ case and other approaches are needed. In this framework, the most important covariates selection methods are those based on regularizations. These have been specifically proposed for the covariates selection problem in the big data context of $p > n$ to face the problem of the curse of dimensionality. In this way a sparse parameter vector associated with a linear regression model is estimated and those covariates with negligible associated coefficient are excluded. Some examples are the LASSO (Tibshirani (1996)), the SCAD (Fan and Li (2001)), the adaptive LASSO (Zou (2006)) or the Dantzig selector (Candes and Tao (2007)) to name a few. See the review of Freijeiro-González et al. (2022) for in-depth details. However, these procedures and their extensions have some restrictions in practice: it is necessary to assume certain structure in the regression model, which can not always be a reliable assumption, and their behavior is worse when p increases faster than n. Furthermore, some of these techniques require of high computational time and resources for a large number of covariates.

Motivated by these previous limitations, Székely et al. (2007) introduced the concept of distance correlation (DC). This coefficient detects all types of possible dependence relations and, as a result, solve the main drawbacks of the previous correlation coefficients. Besides, no structure assumption

is needed in comparison with regularization techniques. Hence, a covariates selection approach can be performed using the DC coefficient no matter the regression model structure. Consequently, innovative techniques for covariates selection were proposed using DC ideas of Székely et al. (2007). Some examples are the procedure of Székely et al. (2007), the DC-SIS (distance covariance sure independence screening) procedure of Li et al. (2012), using the SIS (sure independence screening) algorithm for linear models of Fan and Lv (2008), or the partial distance correlation methodology introduced in Székely and Rizzo (2014). First and third approaches apply independence tests considering an adequate statistic based on DC ideas. In contrast, the DC-SIS sorts out covariates using the distance correlation values and then applies some cutoff or threshold to consider only the most important ones in model explanation terms, which corresponds with the greatest DC values between covariates and response.

In the last years, two new measures of dependence related with the DC were introduced. The martingale difference divergence (MDD) of Shao and Zhang (2014) and the conditional distance correlation (CDC) of Wang et al. (2015). The MDD is used to test the causality of a vector $Y \in \mathbb{R}^q$ conditioned to a scalar random variable $X \in \mathbb{R}$, whereas the CDC tests the conditional dependence of two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ conditioned to a third one, $Z \in \mathbb{R}^r$. Both coefficients can be employed to derive specification tests and to implement covariates selection procedures. See, for example, the work of Shao and Zhang (2014) and Zhang et al. (2018) for the MDD case and all the details of the procedure proposed in Wang et al. (2015) for the CDC performance.

The necessity of covariates selection and specification testing procedures for the functional data context has motivated the recently development of new procedures for this framework. Here, classic methodologies are not available and thus, new ones are needed. Works as the one developed by Gretton et al. (2005) or Febrero-Bande et al. (2019) in the machine learning context, are examples of novel screening tools and bring out the complexity of the functional data case. In Section 3.5 a review of novel specification tests using DC ideas for the functional context is introduced.

In the following, more details about DC, MDD and CDC are given for a deeper understanding of these three kinds of dependence measures for random vectors in Sections 3.2, 3.3 and 3.4, respectively. Next, recent advances in the functional data context using these ideas are described in Section 3.5.

3.2 Distance correlation

The DC is a measure of dependence which detects all types of relations between two random vectors of different dimensions. This coefficient is introduced for the first time by Székely et al. (2007). The main DC interest is to test if two random vectors, $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with $p, q \geq 1$, are independent. This results in testing

$$H_0 : X \perp Y \quad \text{vs.} \quad H_1 : X \not\perp Y, \tag{8}$$

where $X \perp Y$ denotes independence between X and Y.

Two random vectors are said to be independent if they verify $F_{X,Y} = F_X F_Y$, being F_X, F_Y the distribution functions of X and Y, respectively, and $F_{X,Y}$ their joint distribution. This condition can be rewritten in terms of the characteristic functions and the independence test can be formulated as

$$H_0 : \varphi_{X,Y} = \varphi_X \varphi_Y \quad \text{vs.} \quad H_1 : \varphi_{X,Y} \neq \varphi_X \varphi_Y \tag{9}$$

being $\varphi_{X,Y}$ the joint characteristic function and φ_X, φ_Y the marginal characteristic functions of X, Y.

So, for the testing of the null hypothesis (9) it is needed a statistic measuring if the difference $\varphi_{X,Y} - \varphi_X\varphi_Y$ is significant. This is the main motivation for the introduction of the DC coefficient (Székely et al. (2007), Székely and Rizzo (2017)).

In order to measure the difference between $\varphi_{X,Y}$ and $\varphi_X\varphi_Y$ a weighted L_2 norm ($\|\cdot\|_w^2$) in the $\mathbb{R}^p \times \mathbb{R}^q$ space of complex functions is applied. This is defined as

$$\|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|_w^2 = \int_{\mathbb{R}^p\times\mathbb{R}^q} |\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2 w(t,s)\, dt\, ds \tag{10}$$

where $w(\cdot,\cdot)$ is a weight function properly selected to guarantee the existence of the above integral and $|f| = f\bar{f}$ for $f(\cdot)$, a complex value function with conjugate $\bar{f}(\cdot)$.

Then, once the weight function $w(\cdot,\cdot)$ has been selected, we can take as a measure of dependence $\mathcal{V}^2(X,Y;w) = \|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|_w^2$ satisfying that $\mathcal{V}^2(X,Y;w) = 0$ if and only if X and Y are independent. Particularly, dividing $\mathcal{V}^2(X,Y;w)$ by $\sqrt{\mathcal{V}(X;w)\mathcal{V}(Y;w)}$, where

$$\mathcal{V}^2(X;w) = \int_{\mathbb{R}^{2p}} |\varphi_{X,X}(t,s) - \varphi_X(t)\varphi_X(s)|^2 w(t,s)\, dt\, ds \tag{11}$$

we obtain a type of unsigned correlation $\mathcal{R}_w$.

Following these guidelines, in Székely et al. (2007) it is taken

$$w(t,s) = (c_p c_q |t|_p^{1+p} |t|_q^{1+q})^{-1} dt\, ds \quad \text{for} \quad c_p = \frac{\pi^{(p+1)/2}}{\Gamma((p+1)/2)} \ \text{ and } \ c_q = \frac{\pi^{(q+1)/2}}{\Gamma((q+1)/2)}, \tag{12}$$

denoting by $\|\cdot\|_p$ and $\|\cdot\|_q$ the euclidean norms in $\mathbb{R}^p$ and $\mathbb{R}^q$ and $\Gamma(\cdot)$ the gamma function.

For simplicity, we write $\|\cdot\|^2$ henceforth, instead of $\|\cdot\|_\omega^2$, as the L_2 norm using this weight function. Thus, for finiteness of $\|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|^2$, it is sufficient that $\mathbb{E}[\|X\|_p] < \infty$ and $\mathbb{E}[\|Y\|_q] < \infty$. With this notation, the DC between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{V}^2(X,Y)$ defined by expression (13)

$$\mathcal{V}^2(X,Y) = \|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)\|^2 = \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2}{\|t\|_p^{p+1}\|s\|_q^{q+1}} dt\, ds \tag{13}$$

Similarly, distance variance is given as the square root of

$$\mathcal{V}^2(X) = \mathcal{V}^2(X,X) = \|\varphi_{X,X}(t,s) - \varphi_X(t)\varphi_X(s)\|^2. \tag{14}$$

The DC coefficient between random vectors X and Y with finite first moments is the nonnegative number $\mathcal{R}(X,Y)$ defined by

$$\mathcal{R}(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0, \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases} \tag{15}$$

It is verified that $0 \le \mathcal{R}(X,Y) \le 1$, and $\mathcal{R}(X,Y) = 0$ if and only if X and Y are independent.

Alternative expressions for (13) are

$$\begin{aligned}\mathcal{V}^2(X,Y) =& \mathbb{E}\left[\|X'-X''\|_p\|Y'-Y''\|_q\right] \\ &+\mathbb{E}\left[\|X'-X''\|_p\right]\mathbb{E}\left[\|Y'-Y''\|_q\right]-2\mathbb{E}\left[\|X'-X''\|_p\|Y'-Y'''\|_q\right]\end{aligned} \tag{16}$$

and

$$\begin{aligned}\mathcal{V}^2(X,Y) =& \mathbb{E}_{X'Y'}\left[\mathbb{E}_{X''Y''}\left[\|X'-X''\|_p\|Y'-Y''\|_q\right]\right] \\ &+\mathbb{E}_{X'X''}\left[\|X'-X''\|_p\right]\mathbb{E}_{Y'Y''}\left[\|Y'-Y''\|_q\right] \\ &-2\mathbb{E}_{X'Y'}\left[\mathbb{E}_{X''}\left[\|X'-X''\|_p\right]\mathbb{E}_{Y''}\left[\|Y'-Y''\|_q\right]\right]\end{aligned} \tag{17}$$

being (X',Y'), (X'',Y'') and (X''',Y''') iid copies of (X,Y). See Székely et al. (2007) for more details.

Given $(\mathbf{X_n},\mathbf{Y_n}) = \{(X_i,Y_i), i=1,\ldots,n\}$ an iid sample from the joint distribution function of $(X,Y)\in\mathbb{R}^p\times\mathbb{R}^q$, the empirical sample versions of the estimator of $\mathcal{V}^2(\cdot,\cdot)$ can be obtained as follows. Defining $A_{il} = a_{il}-\bar{a}_{i\cdot}-\bar{a}_{\cdot l}+\bar{a}_{\cdot\cdot}$ by means of quantities

$$a_{il}=\|X_i-X_l\|_p,\quad \bar{a}_{i\cdot}=\frac{1}{n}\sum_{l=1}^n a_{il},\quad \bar{a}_{\cdot l}=\frac{1}{n}\sum_{i=1}^n a_{il}\quad\text{and}\quad \bar{a}_{\cdot\cdot}=\frac{1}{n^2}\sum_{i,l=1}^n a_{il}, \tag{18}$$

similarly $B_{il}=b_{il}-\bar{b}_{i\cdot}-\bar{b}_{\cdot l}+\bar{b}_{\cdot\cdot}$ with $b_{il}=\|Y_i-Y_l\|_q$. The empirical distance covariance $\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})$, based on the empirical estimator of (13), is the nonnegative number given by

$$\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})=\frac{1}{n^2}\sum_{i,l=1}^n A_{il}B_{il}. \tag{19}$$

Respectively, $\mathcal{V}_n^2(\mathbf{X_n})$ is the nonnegative number defined by

$$\mathcal{V}_n^2(\mathbf{X_n})=\mathcal{V}_n^2(\mathbf{X_n},\mathbf{X_n})=\frac{1}{n^2}\sum_{i,l=1}^n A_{il}^2. \tag{20}$$

In summary, the estimation given in (19) is an easier way of obtaining an estimator of $\mathcal{V}^2(X,Y)$ just centering two times the data.

Furthermore, the empirical DC, $\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})$, is the square root of

$$\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})=\begin{cases}\dfrac{\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})}{\sqrt{\mathcal{V}_n^2(\mathbf{X_n})\mathcal{V}_n^2(\mathbf{Y_n})}}, & \mathcal{V}_n^2(\mathbf{X_n})\mathcal{V}_n^2(\mathbf{Y_n})>0,\\ 0, & \mathcal{V}_n^2(\mathbf{X_n})\mathcal{V}_n^2(\mathbf{Y_n})=0.\end{cases} \tag{21}$$

This coefficient takes values $0\le\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})\le 1$, and verifies that, if $\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})=1$, then there exist a vector a, a nonzero real number b and an orthogonal matrix C such that $\mathbf{Y_n}=a+b\mathbf{X_n}C$. Moreover, it is verified almost surely that $\lim_{n\to\infty}\mathcal{R}_n^2(\mathbf{X_n},\mathbf{Y_n})=\mathcal{R}^2(X,Y)$. For more properties about $\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})$, $\mathcal{V}_n(\mathbf{X_n})$ and $\mathcal{R}_n(\mathbf{X_n},\mathbf{Y_n})$ we refer to Székely et al. (2007).

Under the null hypothesis of independence, it is verified that $n\mathcal{V}_n^2(\mathbf{X_n},\mathbf{Y_n})/S_2$ converges in distribution to a quadratic form $Q=\sum_{m=1}^\infty c_m G_m^2$, where S_2 is a normalizing factor defined in Székely et al. (2007), $\{G_m\}_{m=1}^\infty$ are independent standard normal random variables and $\{c_m\}_{m=1}^\infty$

nonnegative constants that depend on the distribution of (X, Y). Moreover, when this hypothesis is violated, $n\mathcal{V}_n^2(\mathbf{X_n}, \mathbf{Y_n}) \to \infty$ in probability as $n \to \infty$. Thus, a test which rejects H_0 for large values of $n\mathcal{V}_n^2(\mathbf{X_n}, \mathbf{Y_n})$ is consistent in an omnibus way against dependence alternatives. In practice, the limiting distribution can be approximated by resampling techniques, as for example using permutation tests.

DC can be also used to perform proper GoF tests. In Xu and He (2021) a procedure based on DC is used to test the null hypothesis $H_0 : X \perp \varepsilon$ and $m \in \mathcal{M}_\beta$ in the regression model $Y = m(X) + \varepsilon$ with $m \in \mathcal{M}_\beta = \{g(x)^\top \beta : \beta \in \mathbb{R}^p\}$ for a given known function $g(\cdot)$. In this context, $\mathbf{Y_n}$ is built with the residuals of the fitted model.

Despite all discussed desirable qualities of the empirical distance covariance coefficient, this is a biased estimator of (13) and its bias increases with dimension of X and Y, i.e. when $p, q \to \infty$. Besides, the DC statistic introduced in (21), and based on these coefficients, exhibits some drawbacks as well. As it is explained in Székely and Rizzo (2013), although distance correlation characterizes independence, interpretation of the size of $\mathcal{R}_n(\mathbf{X_n}, \mathbf{Y_n})$ without a formal test is difficult in high dimensions. An explanation for this is that $\mathcal{R}_n^2(\mathbf{X_n}, \mathbf{Y_n}) \longrightarrow 1$ as $p, q \to \infty$, even though X and Y are independent. Székely and Rizzo (2013) proposed a new unbiased sample estimator for distance covariance and a modified distance correlation statistic based on plug-in these unbiased version in numerator and denominator of expression (21) and verifying that, under the null hypothesis of independence, this converges to a Student t distribution. This new approach solves the inconsistency problem in high dimensions.

An additional problem is the computational cost of the construction of the distance matrices. Some recent works such as Huo and Székely (2016) or Chaudhuri and Hu (2019) propose alternatives to reduce this. However, only the univariate random variables case is considered. New solutions applying for the vectorial framework need to be considered in the future.

Finally, it is remarkable the natural relation between DC and the Hilbert-Schmidt Independence Criterion (HSIC) of Gretton et al. (2005). The HSIC makes use of the cross-covariance operator between two reproducing kernel Hilbert spaces (RKHSs) to measure if there exists some type of dependence between two random vectors defined in two different RKHSs with universal kernel. These vectors will be independent when the HSIC operator will take the null value. The DC is a particular case of HSIC operator where general kernel distances are replaced by Euclidean versions. In some sense, there was a parallel evolution between the HSIC criteria in the machine learning world, related to RHKSs, and the DC ideas in literature. There are really interesting papers published in the last decade giving a unifying framework that links both fields. See Sejdinovic et al. (2013), Hua and Ghosh (2015), Zhu et al. (2020) or Edelmann and Goeman (2022) for examples of this connection under different perspectives. As a result, the HSIC measure can be used to perform independence tests, an example is the work of Song et al. (2012), as well as specification tests, see Sen and Sen (2014) for simultaneous GoF and error-predictor independence tests in linear models.

3.3 Martingale difference divergence

The MDD is a new dependence coefficient introduced by Shao and Zhang (2014). This metric measures the departure from the conditional mean independence hypothesis. This is based on testing if the conditional mean of $Y \in \mathbb{R}$, given $X \in \mathbb{R}^p$, is independent of X. The testing problem is now

given by

$$H_0 : \mathbb{E}[Y|_X] = \mathbb{E}[Y] \text{ almost surely} \quad \text{vs.} \quad H_1 : \mathbb{E}[Y|_X] \neq \mathbb{E}[Y] \text{ almost surely.} \tag{22}$$

Its name comes from the interpretation of martingale difference concept in probability. This means that if H_0 in (22) is verified, then $Y - \mathbb{E}[Y]$ is a martingale difference with respect to X.

As a result, the MDD coefficient is designed to measure the difference between the conditional mean and the unconditional one to perform (22). The MDD of Y given X is the nonnegative number $MDD^2(Y|_X)$ defined by

$$MDD^2(Y|_X) = \frac{1}{c_p} \int_{\mathbb{R}^p} \frac{|\psi_{Y,X}(t) - \psi_Y \psi_X(t)|^2}{\|t\|_p^{p+1}} dt \tag{23}$$

where $\psi_{Y,X}(t) = \mathbb{E}[Ye^{i<t,X>}]$, $\psi_Y = \mathbb{E}[Y]$ and $\psi_X(t) = \varphi_X(t)$.

The MDD coefficient defined in (23) verifies that $MDD^2(Y|_X) \geq 0$ and takes the null value if and only if the null hypothesis (22) holds. This is called divergence and not distance because $MDD^2(Y|_X) \neq MDD^2(X|_Y)$.

Similar to DC, a scale invariant coefficient can be defined. This gives place to the martingale difference correlation (MDC) given by the square root of

$$MDC^2(Y|_X) = \begin{cases} \frac{MDD^2(Y|_X)}{\sqrt{\text{var}^2(Y)\mathcal{V}^2(X)}}, & \text{var}^2(Y)\mathcal{V}^2(X) > 0, \\ 0, & \text{var}^2(Y)\mathcal{V}^2(X) = 0. \end{cases} \tag{24}$$

where $\mathcal{V}^2(X)$ is the distance variance of X defined in (14). It is verified that $0 \leq MDC^2(Y|_X) \leq 1$. Similar properties as DC for $MDD^2(Y|_X)$ and $MDC^2(Y|_X)$ are collected in Shao and Zhang (2014).

For a sample of $i = 1, \ldots, n$ iid observations $(\mathbf{X_n}, \mathbf{Y_n}) = \{(X_i, Y_i), i = 1, \ldots, n\}$ from the joint distribution of $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, it is defined A_{il} as in (18) and $B_{il} = b_{il} - \bar{b}_{i\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}$; being now $b_{il} = |Y_i - Y_l|^2/2$, $\bar{b}_{i\cdot} = \frac{1}{n}\sum_{l=1}^n b_{il}$, $\bar{b}_{\cdot l} = \frac{1}{n}\sum_{i=1}^n b_{il}$ and $\bar{b}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,l=1}^n b_{il}$ for $i, l = 1, \ldots, n$. The empirical estimation of $MDD^2(Y|_X)$, i.e. the sample martingale difference divergence, can be defined as the nonnegative number

$$MDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}}) = \frac{1}{n^2} \sum_{i,l=1}^n A_{il} B_{il} \tag{25}$$

and its associated sample martingale difference correlation coefficient is given by

$$MDC_n^2(\mathbf{Y_n}|_{\mathbf{X_n}}) = \begin{cases} \frac{MDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})}{\sqrt{\text{var}_n^2(\mathbf{Y_n})\mathcal{V}_n^2(\mathbf{X_n})}}, & \text{var}_n^2(\mathbf{Y_n})\mathcal{V}_n^2(\mathbf{X_n}) > 0, \\ 0, & \text{var}_n^2(\mathbf{Y_n})\mathcal{V}_n^2(\mathbf{X_n}) = 0. \end{cases} \tag{26}$$

where $\text{var}_n(\mathbf{Y_n}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})^2$, for $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, and $\mathcal{V}_n^2(\mathbf{X_n})$ is defined in (20).

See the paper of Park et al. (2015) for a nice connection between MDD and DC coefficients.

If $\mathbb{E}\left[\|X\|_p + |Y|^2\right] < \infty$, both estimators, $MDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})$ and $MDC_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})$, converge to their population versions displayed in (23) and (24) almost surely. A prove of this result can be found in Shao and Zhang (2014). Moreover, under the null hypothesis of independence in mean, it is guaranteed that $nMDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}}) \longrightarrow \|\Gamma(t)\|^2$ in distribution when $n \to \infty$, being $\Gamma(\cdot)$ a Gaussian process. In addition, if $\mathbb{E}[Y^2|_X] = \mathbb{E}[Y^2]$ is also guaranteed, $nMDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})/S_n \longrightarrow Q$ in distribution when $n \to \infty$, being Q a nonnegative quadratic form of centered Gaussian random variable with $\mathbb{E}[Q] = 1$ and $S_n = \frac{1}{n^2}\sum_i \sum_l \|X_i - X_l\|_p \frac{1}{n}\sum_i (Y_i - \bar{Y}_n)^2$. Finally, if the null hypothesis is not verified, we have that $nMDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}})/S_n \longrightarrow \infty$ in probability when $n \to \infty$. We refer to Shao and Zhang (2014) for more details. Although the asymptotic distribution under both, H_0 and H_1 hypothesis is known, resampling procedures can be applied in practice to calibrate the distribution of the test statistic, especially for small sample sizes.

Thus, using the estimators of the MDD or MDC, it is possible to perform covariates selection in regression models, specifying which covariates are the relevant ones. Shao and Zhang (2014) propose a screening procedure sorting out the covariates relevance in terms of the regressor function explanation, i.e. based on $\mathbb{E}[Y|_X]$ explanation, and then they establish a proper threshold to detect the important significance covariates. Authors make use of the MDC criteria to measure covariates relevance. A different approach for covariates selection in terms of causality is introduced in Zhang et al. (2018). They propose a statistic based on the MDD ideas to test the null hypothesis of $H_0 : \mathbb{E}\left[Y|_{X_j}\right] = \mathbb{E}[Y]$ almost surely for all $j = 1 \ldots, p$. A wild bootstrap scheme is proposed to approximate the statistics distribution.

All these ideas can be transferred to GoF testing. An example is the work of Su and Zheng (2017). They test the null hypothesis of $H_0 : \mathbb{P}\left(\mathbb{E}[Y|_X] = g(X, \beta)\right) = 1$ for some $\beta \in \mathcal{B}$, being $\mathcal{B}$ the parameter space and assuming $Y = g(X, \beta) + \varepsilon$, with $g(\cdot)$ a known function. The MDD is applied using the residuals calculated under the null hypothesis, and covariates. Calibration of the test is again done by means of wild bootstrap. A similar, but broader approach, using HSIC is also provided by Teran Hidalgo et al. (2018).

3.4 Conditional distance correlation

The CDC was introduced in Wang et al. (2015) to measure the dependence of two random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ conditioned to a third one, $Z \in \mathbb{R}^r$. For this purpose, conditional characteristic functions are employed and ideas of the distance correlation introduced in Section 3.1 are adapted to the conditional framework. The problem to be tested now is

$$H_0 : X \perp_{|Z} Y \text{ almost surely} \quad \text{vs.} \quad H_1 : \mathbb{P}\left(X \not\perp_{|Z} Y\right) > 0 \tag{27}$$

where $X \perp_{|Z} Y$ denotes independence of X and Y conditioned to Z.

Using similar DC arguments, it is possible to rewrite (27) in terms of characteristic functions. The new test is given by

$$H_0 : \varphi_{X,Y|Z} = \varphi_{X|Z}\varphi_{Y|Z} \quad \text{vs.} \quad H_1 : \varphi_{X,Y|Z} \neq \varphi_{X|Z}\varphi_{Y|Z} \tag{28}$$

where $\varphi_{X,Y|Z}$, $\varphi_{X|Z}$ and $\varphi_{Y|Z}$ are the joint and marginal conditional characteristic functions.

Then, the CDC with finite first moments given Z ($\mathbb{E}[|X|_p + |Y|_q|_Z] < \infty$), is defined as the square root of

$$\begin{aligned} CDC^2(X,Y|_Z) =& \|\varphi_{X,Y|_Z}(t,s) - \varphi_{X|_Z}(t)\varphi_{Y|_Z}(s)\|^2 \\ =& \frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{|\varphi_{X,Y|_Z}(t,s) - \varphi_{X|_Z}(t)\varphi_{Y|_Z}(s)|^2}{\|t\|_p^{p+1}\|s\|_q^{q+1}} dt\, ds \end{aligned} \tag{29}$$

where c_p and c_q are the ones defined in (12) and conditional distance variance is the square root of

$$CDC^2(X|_Z) = CDC^2(X,X|_Z) = \|\varphi_{X,X|_Z}(t,s) - \varphi_{X|_Z}(t)\varphi_{X|_Z}(s)\|^2,$$

being $\|\cdot\|$ the weighted norm defined in Section 3.2.

The CDC coefficient defined in (29) has analogues properties to the unconditional version of (13). Particularly, it is verified that $CDC(X,Y|_Z) \geq 0$ if and only if X and Y are conditionally independent given Z.

The conditional distance correlation (CDCor) is the square root of

$$CDCor(X,Y|_Z) = \begin{cases} \frac{CDC^2(X,Y|_Z)}{\sqrt{CDC^2(X|_Z)CDC^2(Y|_Z)}}, & CDC^2(X|_Z)CDC^2(Y|_Z) > 0, \\ 0, & CDC^2(X|_Z)CDC^2(Y|_Z) = 0. \end{cases} \tag{30}$$

and this verifies that $0 \leq CDCor(X,Y|_Z) \leq 1$ and $CDCor(X,Y|_Z) = 0$ if and only if X and Y are conditionally independent given Z.

To construct an estimator of $CDC^2(X,Y|_Z)$ the empirical characteristic functions conditioned to Z are plugged in (29). Note that, for the estimation of conditional characteristic functions, it is needed to resort to some kind of smoothing techniques as for example kernel-type estimators. We refer to Wang et al. (2015) for more details. Denote by $W_i = (X_i, Y_i, Z_i),\ i = 1,\ldots,n$ a sample iid from a random vector $W = (X,Y,Z) \in \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^r$, $\mathbf{X_n} = \{X_1,\ldots,X_n\}$, $\mathbf{Y_n} = \{Y_1,\ldots,Y_n\}$, $\mathbf{Z_n} = \{Z_1,\ldots,Z_n\}$, and $\mathbf{W_n} = (\mathbf{X_n}, \mathbf{Y_n}, \mathbf{Z_n})$. As a result, the sample conditional distance covariance $CDC_n(\mathbf{X_n}, \mathbf{Y_n}|_{\mathbf{Z_n}})$ is the positive quantity defined by

$$\widetilde{CDC}_n^2(\mathbf{X_n}, \mathbf{Y_n}|_{\mathbf{Z_n}}) = \|\varphi^n_{X,Y|_Z}(t,s) - \varphi^n_{X|_Z}(t)\varphi^n_{Y|_Z}(s)\|^2. \tag{31}$$

being $\varphi^n_{X,Y|_Z}$, $\varphi^n_{X|_Z}$ and $\varphi^n_{Y|_Z}$ the corresponding empirical conditional characteristic functions.

Following Wang et al. (2015), letting $d_{ijkl} = \left(a^X_{ij} + a^X_{kl} - a^X_{ik} - a^X_{jl}\right)\left(b^Y_{ij} + b^Y_{kl} - b^Y_{ik} - b^Y_{jl}\right)$ and $d^S_{ijkl} = d_{ijkl} + d_{ijlk} + d_{ilkj}$ for $i,j,k,l = 1,\ldots,n$, where a_{ij} and b_{ij} are defined in (18), and Z_1, Z_2, Z_3 and Z_4 are iid copies of Z, it is verified that

$$CDC^2(X,Y|_{Z=z}) = \frac{1}{12}\mathbb{E}[d^S_{1234}|_{Z_1=z,Z_2=z,Z_3=z,Z_4=z}]$$

As a result, the conditional dependence measures can be estimated by applying kernel regression smoothing ideas to the above expectation estimation. This results in a V-process. The sample conditional distance covariance is defined as the square root of

$$CDC^2_n(\mathbf{W_n}|_Z) = CDC^2_n(\mathbf{X_n}, \mathbf{Y_n}, \mathbf{Z_n}|_Z) = \frac{1}{n^4}\sum_{ijkl} \Psi_n(W_i, W_j, W_k, W_l; Z) \tag{32}$$

where Ψ_n is the symmetric random kernel of degree 4 defined in Schick (1997):

$$\Psi_n(W_i, W_j, W_k, W_l; Z) = \frac{n^4 \Phi_i(Z)\Phi_j(Z)\Phi_k(Z)\Phi_l(Z)}{12\Phi^4(Z)} d^S_{ijkl}$$

for $\Phi_i(Z) = K_H(Z - Z_i)$ and $\Phi(Z) = \sum_{i=1}^n \Phi_i(Z)$, being K a kernel function and H a bandwidth matrix r-dim.

Let $\mathbf{W_{X_n}} = (\mathbf{X_n}, \mathbf{X_n}, \mathbf{Z_n})$ and $\mathbf{W_{Y_n}} = (\mathbf{Y_n}, \mathbf{Y_n}, \mathbf{Z_n})$. Analogously, the sample conditional distance correlation can be defined as the square root of

$$CDCor_n(\mathbf{W_n}|_Z) = \begin{cases} \frac{CDC_n^2(\mathbf{W_n}|_Z)}{\sqrt{CDC_n^2(\mathbf{W_{X_n}}|_Z) CDC_n^2(\mathbf{W_{Y_n}}|_Z)}}, & CDC_n^2(\mathbf{W_{X_n}}|_Z) CDC_n^2(\mathbf{W_{Y_n}}|_Z) > 0, \\ 0, & CDC_n^2(\mathbf{W_{X_n}}|_Z) CDC_n^2(\mathbf{W_{Y_n}}|_Z) = 0. \end{cases}$$

It is verified that $\widetilde{CDC}_n^2(\mathbf{W_n}|_Z) = CDC_n^2(\mathbf{W_n}|_Z)$ given $\mathbf{W_n} = \{W_1, \ldots, W_n\}$ a sample from the joint distribution of (X, Y, Z). Furthermore, if $\mathbb{E}\left[\|X\|_p + \|Y\|_q|_Z\right] < \infty$ and $\Phi(Z)/n$ is a consistent density function estimator of Z, then $CDC_n^2(\mathbf{W_n}|_Z) \longrightarrow CDC^2(X, Y|_Z)$ in probability for each value of Z as $n \to \infty$. See Wang et al. (2015) for more details and properties of $CDC_n^2(\mathbf{W_n}|_Z)$. Analogously, an unbiased version of (32) can be defined with similar properties. For this purpose, U-processes theory is applied.

Wang et al. (2015) make use of these ideas to perform the conditional independence test displayed in (27), applying conditioned covariates selection. In particular, they define a statistic based on the CDC coefficient and implement a test calibrated by means of a local bootstrap. Other procedures related with screening techniques in terms of conditional dependence are the recent works of Song et al. (2020) and Lu and Lin (2020). The first one adapt the ideas of Liu et al. (2014) using the CDCor to specify significant covariates for general varying-coefficient models in regression. Covariates are sorting out based on their CDCor value and then a cutoff is applied. In contrast, Lu and Lin (2020) select an initial set of covariates and measures the importance of remaining ones conditioned to this subset. For this purpose, they use the CDCor, resulting in the CDC-SIS (conditional distance correlation sure independence screening) algorithm.

3.5 A new generation of procedures for testing in regression models based on distance correlation with functional data

In this Section, we assume that both, the explanatory covariate X as well as the output Y of the regression model $Y = m(X) + \varepsilon$, are functions. Here, similar to Section 2, it is assumed that $X \in \mathcal{H}_X$ and $Y \in \mathcal{H}_Y$, being $\mathcal{H}_X$ and $\mathcal{H}_Y$ Hilbert spaces. As it was mentioned in previous sections, results for specification testing in regression models with functional data appear in the last 10 years. However, it is not until very recently when the methodology of the DC is employed, extending procedures of Section 2 to the functional framework.

A first paper is Lee et al. (2020), where it is tested the null hypothesis $H_0 : \mathbb{E}\left[Y|_X\right] = \mathbb{E}[Y]$ as in (22) but now for the functional case. Then, to carry out the test, an statistic based on a generalized version of the MDD coefficient described in Section 3.3 is proposed.

In particular, the vectorial MDD term can be written as (see Shao and Zhang (2014))

$$MDD^2\left(Y|_X\right) = -\mathbb{E}\left[\left(Y - \mathbb{E}[Y]\right)\left(Y' - \mathbb{E}[Y]\right)\|X - X'\|_{\mathcal{H}_X}\right]$$

and this idea is extended to the functional context considering

$$FMDD^2\left(Y|_X\right) = -\mathbb{E}\left[\langle Y-\mathbb{E}[Y], Y'-\mathbb{E}[Y]\rangle_{\mathcal{H}_Y}\|X-X'\|_{\mathcal{H}_X}\right]$$

being, in both cases (vectorial and functional), (X', Y') iid copies of (X, Y).

Hence, based on an iid sample $(\mathbf{X_n}, \mathbf{Y_n}) = \{(X_i, Y_i), i = 1, \ldots, n\}$ of (X, Y), an unbiased estimator of $FMDD^2$ is obtained with the empirical version

$$FMDD_n^2\left(\mathbf{Y_n}|_{\mathbf{X_n}}\right) = \frac{1}{n(n-3)}\sum_{i\neq j}\tilde{A}_{ij}\tilde{B}_{ij} \tag{33}$$

where $\tilde{A}_{ij}$ and $\tilde{B}_{ij}$ are now the corresponding U-centered versions of (18), being the $(i,j)^{th}$ elements of the matrices defined as

$$\tilde{A}_{ij} = \begin{cases} a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, & i \neq j \\ 0, & i = j \end{cases}$$

$$\tilde{B}_{ij} = \begin{cases} b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}, & i \neq j \\ 0, & i = j \end{cases}$$

with $a_{ij} = \|X_i - X_j\| = \sqrt{\langle X_i - X_j, X_i - X_j\rangle_{\mathcal{H}_X}}$, $\bar{a}_{i\cdot} = \frac{\sum_l a_{il}}{(n-2)}$, $\bar{a}_{\cdot j} = \frac{\sum_k a_{kj}}{(n-2)}$, $\bar{a}_{\cdot\cdot} = \frac{\sum a_{kl}}{(n-1)(n-2)}$, $b_{ij} = \|Y_i - Y_j\|^2_{\mathcal{H}_Y}/2$ and $\bar{b}_{i\cdot}$, $\bar{b}_{\cdot j}$ and $\bar{b}_{\cdot\cdot}$ defined in a similar way.

That is, the modified and adapted empirical unbiased version of the estimation given in (25), but now, for the functional context.

Nice results are obtained in Lee et al. (2020) under the assumptions of $\mathbb{E}\left[\|X\|^2_{\mathcal{H}_X} + \|Y\|^2_{\mathcal{H}_Y}\right] < \infty$, $\mathbb{E}\left[\|X-\mathbb{E}[X]\|^2_{\mathcal{H}_X} + \|Y-\mathbb{E}[Y]\|^2_{\mathcal{H}_Y}\right] < \infty$ and the null hypothesis is true: $nFMDD_n^2(\mathbf{Y_n}|_{\mathbf{X_n}}) \longrightarrow \sum_{k=1}^{\infty}\lambda_k(G_k^2-1)$ in distribution, being $\{\lambda_k\}_{k=1}^{\infty}$ the eigenvalues corresponding to the eigenfunctions $\{\Psi_k(\cdot)\}_{k=1}^{\infty}$ such that $J(z,z') = \sum_{k=1}^{\infty}\lambda_k\Psi_k(z)\Psi_k(z')$ with $z = (x,y)$ and $J(z,z') = U(x,x')V(y,y')$, where $U(x,x') = \|x-x'\|_{\mathcal{H}_X} + \mathbb{E}\left[\|X-X'\|_{\mathcal{H}_X}\right] - \mathbb{E}\left[\|x-X'\|_{\mathcal{H}_X}\right] - \mathbb{E}\left[\|X-x'\|_{\mathcal{H}_X}\right]$ and $V(y,y') = -\langle y-\mathbb{E}[Y], y'-\mathbb{E}[Y]\rangle_{\mathcal{H}_Y}$. Here $\{\Psi_k\}$ is an orthogonal sequence in the sense that $\mathbb{E}\left[\Psi_j(z)\Psi_k(z)\right] = \mathbb{I}(j=k)$ and $\{G_k\}_{k=1}^{\infty}$ is a sequence of iid $N(0,1)$ random variables.

This represent the limit distribution of a degenerate U-statistic with kernel $h(\cdot)$, being

$$FMDD_n^2\left(\mathbf{Y_n}|_{\mathbf{X_n}}\right) = \frac{1}{\binom{n}{4}}\sum_{i<j<k<l}h(Z_i, Z_j, Z_k, Z_l)$$

with $h(Z_i, Z_j, Z_k, Z_l) = \frac{1}{4!}\sum_{(s,t,u,v)}^{(i,j,k,l)}(a_{st}b_{uv} + a_{st}b_{st} - a_{st}b_{su} - a_{st}b_{tv})$ the sum over the 24 possible permutations of the indexes (i,j,k,l).

In Lee et al. (2020) it is proposed to reject the null hypothesis of conditional mean independence if and only if $T_n = nFMDD_n^2\left(\mathbf{Y_n}|_{\mathbf{X_n}}\right) > C$, where C is a constant taken based on the α significance level. The power of the test is studied and demonstrated to be consistent under both, local and fixed alternatives, using a consistent wild bootstrap calibration.

A second recent contribution is the paper of Lai et al. (2020), devoted to test a modified null hypothesis: $\widetilde{H}_0$: "X is independent of ε and m satisfies the linear model given by (7) in Section 2". Using the recent results related with the distance covariance (see Székely et al. (2007), Lyons (2013) and Sejdinovic et al. (2013)). Consider now $(\mathcal{X}, \rho_{\widetilde{X}})$ and $(\mathcal{Y}, \rho_{\widetilde{Y}})$ two semimetric spaces of negative type, where $\rho_{\widetilde{X}}$ and $\rho_{\widetilde{Y}}$ are the corresponding semimetrics. Denote by $(\widetilde{X}, \widetilde{Y})$ a random element with joint distribution $\mathrm{P}_{\widetilde{X}\widetilde{Y}}$ and marginals $\mathrm{P}_{\widetilde{X}}$ and $\mathrm{P}_{\widetilde{Y}}$, respectively, and take $(\widetilde{X}', \widetilde{Y}')$ an iid copy of $(\widetilde{X}, \widetilde{Y})$. The generalized distance covariance $(\widetilde{X}, \widetilde{Y})$ is given by

$$\begin{aligned}\theta(\widetilde{X}, \widetilde{Y}) = \mathbb{E}\big[\rho_{\widetilde{X}}(\widetilde{X}, \widetilde{X}')\rho_{\widetilde{Y}}(\widetilde{Y}, \widetilde{Y}')\big] \\ + \mathbb{E}\big[\rho_{\widetilde{X}}(\widetilde{X}, \widetilde{X}')\big]\mathbb{E}\big[\rho_{\widetilde{Y}}(\widetilde{Y}, \widetilde{Y}')\big] \\ - 2\mathbb{E}_{(\widetilde{X},\widetilde{Y})}\big[\mathbb{E}_{\widetilde{X}'}\big[\rho_{\widetilde{X}}(\widetilde{X}, \widetilde{X}')\big]\mathbb{E}_{\widetilde{Y}'}\big[\rho_{\widetilde{Y}}(\widetilde{Y}, \widetilde{Y}')\big]\big].\end{aligned}$$

This corresponds with expression (16) and (17) for the vectorial case.

As noted by Lai et al. (2020) the generalized distance covariance can be alternatively written as

$$\theta(\widetilde{X}, \widetilde{Y}) = \int \rho_{\widetilde{X}}(\widetilde{x}, \widetilde{x}')\rho_{\widetilde{Y}}(\widetilde{y}, \widetilde{y}')\, \mathrm{d}[(\mathrm{P}_{\widetilde{X}\widetilde{Y}} - \mathrm{P}_{\widetilde{X}}\mathrm{P}_{\widetilde{Y}}) \times (\mathrm{P}_{\widetilde{X}\widetilde{Y}} - \mathrm{P}_{\widetilde{X}}\mathrm{P}_{\widetilde{Y}})].$$

where $d[\cdot]$ denotes the differential term of the integral.

Note that $\theta(\widetilde{X}, \widetilde{Y}) = 0$ if and only if $\widetilde{X}$ and $\widetilde{Y}$ are independent. Given an iid sample $\{(\widetilde{X}_i, \widetilde{Y}_i)\}_{i=1}^n$ of $(\widetilde{X}, \widetilde{Y})$, an empirical estimator of θ is given by

$$\theta_n(\widetilde{X}, \widetilde{Y}) = \frac{1}{n^2}\sum_{i,j} k_{ij}\ell_{ij} + \frac{1}{n^4}\sum_{i,j,q,\tau} k_{ij}\ell_{q\tau} - \frac{2}{n^3}\sum_{i,j,q} k_{ij}\ell_{iq}$$

with $k_{ij} = \rho_{\widetilde{X}}(\widetilde{X}_i, \widetilde{X}_j)$ and $\ell_{ij} = \rho_{\widetilde{Y}}(\widetilde{Y}_i, \widetilde{Y}_j)$. Taking $\widetilde{X} = X$ and $\widetilde{Y} = \varepsilon = Y - \langle X, \beta\rangle_{\mathcal{H}_X}$, $\rho_{\widetilde{Y}}$ is the absolute value and $\rho_{\widetilde{X}}$ is the distance associated to the Hilbert space $\mathcal{H}_X$. The test statistic is $T_n = \theta_n(\hat{\varepsilon}, X)$ and is based on $\{(X_i, Y_i - \langle X_i, \hat{\beta}\rangle_{\mathcal{H}_X})\}_{i=1}^n$.

In other recent papers Hu et al. (2020) and Zhao et al. (2022), the null hypothesis about the linearity given in (7) is tested using related approximations based on the MDD adapted to the functional context.

All the tests described in this section have challenging limit distributions and need to be calibrated with resampling techniques.

The references mentioned above are for the extension of DC and MDD coefficients to specification tests in the functional data context. Specification tests, in general, for independence testing between two functional variables X and Y, conditioned to a third one Z, are a really though problem. Some very relevant and recent papers in this topic are the ones of Shah and Peters (2020) or Lundborg et al. (2022). A deep study of the CDC in the functional framework is still an open problem of interest for future research.

4 Applications

In this last section, we illustrate some of the recently developed new methodologies for specification tests in the functional framework introduced along the document. Three real datasets examples with functional nature are employed.

The first application is an illustration of the test of equality of distribution functions. This is devoted to the Medflies data (Carey et al. (1998)). In this example, the Mediterranean fruit flies' lifetime distributions are compared with respect to their fertility (number of eggs). The distinction is done in terms of short-lived or long-lived individuals. As a result, a test of equality of distribution for functional data is performed (Section 4.1).

Secondly, other well-known data set in the functional framework is employed. This is the Tecator database (see Ferraty and Vieu (2006)). In this application, it is wanted to determine if the spectrometric functional variable (absorbance), as well as its first and second derivatives, support relevant information to explain the fat content in a regression model. For this purpose, significance tests are applied over the considered functional covariates (Section 4.2).

Finally, a GoF test based on CD is applied to check if a Ornstein-Uhlenbeck diffusion process explains the evolution of high-frequency financial data. In particular, Johnson & Johnson stock prices from August 2018 to August 2019 are analyzed (Section 4.3).

4.1 Testing equality of distributions in the Medflies data set

Medflies is a functional dataset usually used for classification purposes. See Carey et al. (1998) for more details. This is available in the `ddalpha` (Pokotylo et al. (2019)) package of `R` (R Core Team (2022)). This contains the medflies trajectories for number of eggs laid differentiation between short or long-lived individuals. The goal is to classify a group of Mediterranean flies as short-lived or long-lived (alive after day 50), X and Y populations, respectively, given their fertility up to day 35. The dataset contains 278 trajectories of long-lived and 256 for the short-lived group. This is considered a hard classification problem and the best overall ratio is around 60%. As a result, it makes sense to wonder if it is possible to correctly discriminate between both groups. For this purpose, a test for comparison of populations in the function context is applied.

First, data on flies that have not laid any egg are removed to avoid outliers. We found this type of individuals for both classes: short-lived as well as long-lived flies. This results in a total of 266 long-lived trajectories and 246 of the short-lived ones for the new dataset. As the number of eggs laid is a discrete variable we considered the raw data as well as a logaritmic transformation to avoid heterogeneity problems. Next, functional data is smoothed using nonparametric kernel estimation. This is done using the `optim.np` function of `fda.usc` library (Febrero-Bande and Oviedo de la Fuente (2012)). We have employed a bandwidth parameter value of $h = 1$, other values could be considered as well. However, we have appreciated a suitable smoothing taking this quantity. Results for first 30^{th} resulting samples are displayed in Figure 1 for both groups.

Thus, after smoothing both functional variables X and Y, corresponding to short-lived and long-lived populations, we are interested in determining if there exits significance differences between both groups in terms of their distributions. For this purpose, a test for comparison of the

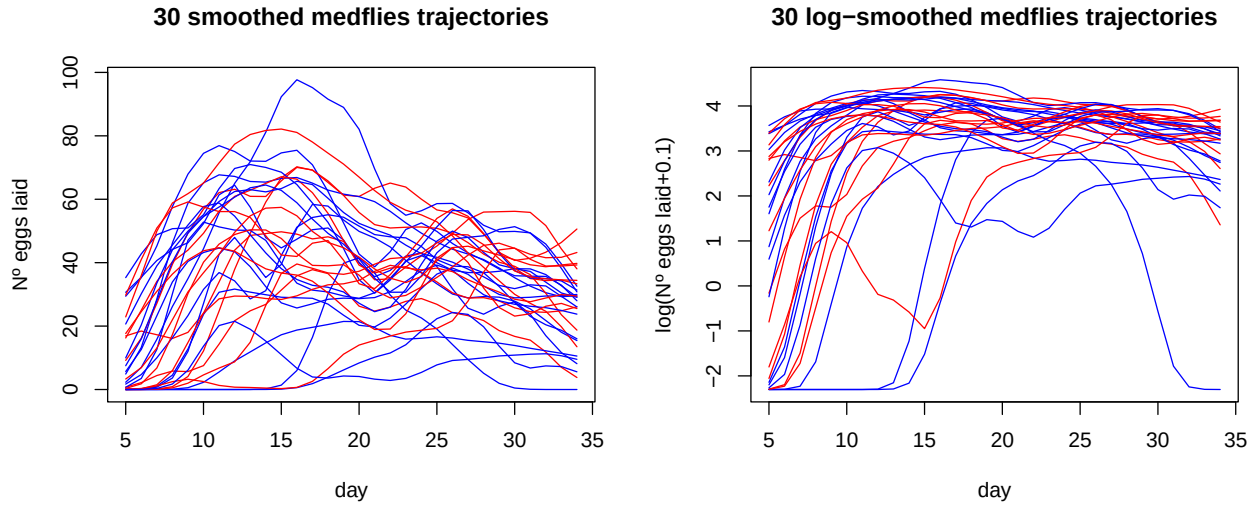


Figure 1: First 30^{th} smoothed medflies trajectories for number of eggs laid (left) and $\log$(number of eggs laid +0.1) (right) for short-lived individuals (—) and long-lived ones (—).

distribution of the populations in the functional framework is needed. This results in testing the null hypothesis of $H_0 : X \sim Y$.

We resort to random projections in functional data (Cuesta-Albertos et al. (2007)) to construct a proper statistic for the test. Once our data is projected, scalar procedures for comparison of populations can be employed. In an illustrative way, we decided to use a total of 10 random projections and then apply Kolmogorov-Smirnov (KS10) and Anderson-Darling (AD10) techniques. For this aim, function `XYRP.test` of the `fda.usc` library (Febrero-Bande and Oviedo de la Fuente (2012)) is applied. Obtained results are collected in Table 1.

	KS10	AD10
smfl	1.3×10^{-4}	6.5×10^{-4}
$\log(\text{smfl} + 0.1)$	0.00804	0.01547

Table 1: Resulting p-values for Kolmogorov-Smirnov (KS10) and Anderson-Darling (AD10) tests using 10 random projections for smoothed medflies trajectories (sfml) and its logarithmic version ($\log(\text{smfl} + 0.1)$).

In view of the results, all p-values< 0.0155, we have evidences to reject the null hypothesis that the number of eggs laid for fruit flies are equally distributed for short and long lived individuals. Thus, we can conclude that there exists difference between the fertility of a fly with long life expectancy compared to one with a lower rate. Therefore, new classification methodologies for the functional data context are needed to correctly discriminate between both groups.

4.2 Significance tests with functional covariates for the Tecator database

The Tecator data set records the content of water, fat and protein percentages jointly with absorbances spectrometric curves, measured in a 100-channel spectrum, of a total of $n = 215$ meat samples. This is available in the `fda.usc` (Febrero-Bande and Oviedo de la Fuente (2012)) package of `R` (R Core Team (2022)). This database is a well-studied real data example in the functional framework. Some examples where this data set is considered are the works of Ferraty and Vieu (2006), García-Portugués et al. (2014), Lee et al. (2020) and Shi et al. (2022) among others. Following previous studies guidelines, we are interested on model the percentage of fat (Y) using the spectrometric curves information (X). In particular, we consider the absorbance (ab) and its first and second derivatives (ab1 and ab2). Representation of the considered covariates is collected in Figure 2.

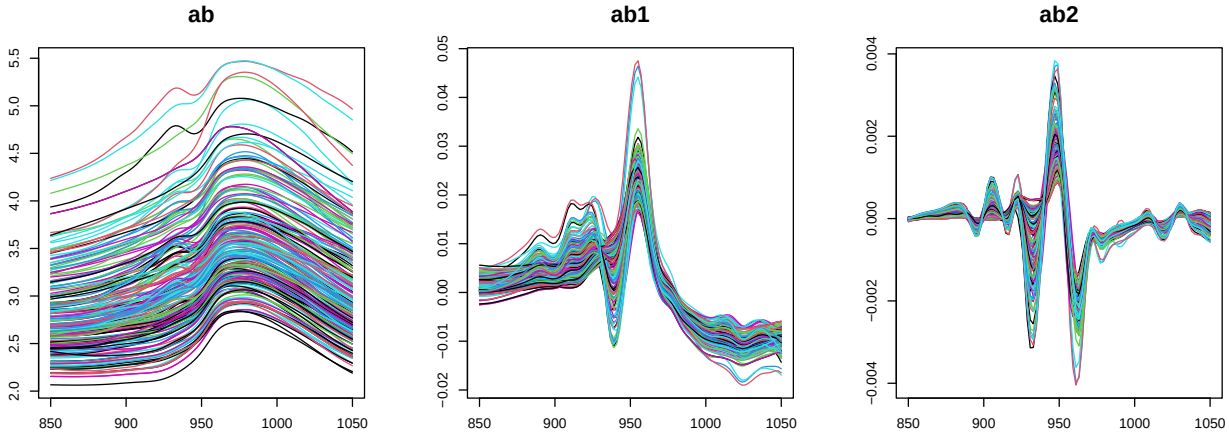


Figure 2: Left: absorbance curves. Middle: first derivative of absorbance curves. Right: second derivative of absorbance curves.

To verify if all considered covariates are relevant in the fat percentage explanation or if some do not support enough information, we can resort to significance tests. In particular, we want to test

$$H_0 : \mathbb{E}\left[Y|_X\right] = \mathbb{E}[Y] \text{ almost surely} \quad \text{vs.} \quad H_1 : \mathbb{P}\left(\mathbb{E}\left[Y|_X\right] \neq \mathbb{E}[Y]\right) > 0,$$

where X can be the absorbance information as well as its first or second derivative. This corresponds with the test displayed in expression (22) of Section 3.3 for the vectorial case.

Following similar ideas to Lee et al. (2020) we can implement the test using the FMDD coefficient introduced previously in Section 3.5. Using a $B = 1000$ resampling wild bootstrap calibration procedure, we obtain null p-values for raw absorbance (ab), first (ab1) and second derivative (ab2). As a result, we have evidences to reject the null hypothesis of conditional mean independence and to claim that these three covariates provide relevant information in the fat percentage explanation. It is interesting to note that no model assumption or structure is needed, as all types of possible dependence in mean are collected in the considered H_0.

Moreover, we can go a step further to detect which covariates are the most and least relevant ones. For this aim, we can define a scale invariant functional martingale difference correlation coefficient (FMDC). This is an extension of the $MDC^2(Y|_X)$ term introduced in (24) for the vectorial context

	ab	ab1	ab2
DC	0.2	0.78	0.91
FMDC	0.45	0.88	0.94

Table 2: Results of distance correlation (DC) and functional martingale difference correlation (FMDC) coefficients for Absorbances (ab), first Absorbances' derivative (ab1) and second one (ab2).

just applying the considered metrics for the functional martingale calculation. Now, we build our functional scale invariant coefficient using the unbiased $FMDD_n^2\left(\mathbf{Y_n}|\mathbf{X_n}\right)$ estimator formula introduced in (33). Results are displayed in Table 2. We see as the ab2 term is the one with the greatest explanation capability, following for ab1 and ab. This highlights the fact that employing the second derivative instead of ab increases the explanatory power of the regression model. Besides, we calculate the DC coefficient and similar results are obtained.

4.3 GoF test for a high-frequency dynamic model example: Johnson & Johnson company stock prices

To illustrate a real-data application for dynamic models, we apply the ideas of DC to test a GoF of the Ornstein-Uhlenbeck process as an autorregresive Hilbertian (ARH) process. We refer the reader to Bosq (2000) for more details about ARH processes.

Let $\{X_t\}_{t\in\mathbb{R}^+}$ be a continuous-time zero-mean stochastic process. Following the ideas in Álvarez-Liébana et al. (2022), we split the path, corresponding to the observed domain of the $t \in \mathbb{R}^+$ term of the stochastic process, as $\mathcal{X}_n(t) = X_{nh+t}$, with $t \in [0,h]$, and $\mathcal{X}_n \in \mathcal{H} = L^2([0,h])$, for each $n \in \mathbb{Z}^+$, constituting an infinite-dimensional discrete-time process. The zero-mean autoregressive Hilbertian process of order one $\mathcal{X} = \{\mathcal{X}_n\}_{n\in\mathbb{Z}^+}$, denoted as ARH(1), satisfies the state equation

$$\mathcal{X}_n(t) = \Lambda\left(\mathcal{X}_{n-1}\right)(t) + \mathcal{E}_n(t), \qquad n \in \mathbb{Z}^+, \quad t \in [0,h],$$

with $\mathcal{X}_n,\ \mathcal{E}_n \in \mathcal{H} = L^2([0,h])$, Λ the linear autocorrelation operator, and $\{\mathcal{E}_n\}_{n\in\mathbb{Z}}$ an independent sequence of Gaussian processes with null mean (strong-white noise) with iid components (see Assumptions considered in Álvarez-Liébana et al., 2022). The Ornstein-Uhlenbeck process,

$$X_t = X_0 e^{-\theta t} + \mu\left(1 - e^{-\theta t}\right) + \sigma\int_0^t e^{-\theta(t-s)}\,\mathrm{d}W_s, \qquad t,s \in \mathbb{R}^+,$$

with X_0 the initial condition at $t_0 = 0$, can be characterized as an ARH(1) process. Let $\mathcal{H}$ be a separable Hilbert space given by $\mathcal{H} = L^2\left([0,h], \mathcal{B}_{[0,h]}, \lambda + \delta_{(h)}\right)$, with $\mathcal{B}_{[0,h]}$ the σ-algebra generated by the subintervals $[0,h]$, λ the Lebesgue measure and $\delta_{(h)}(s) = \delta(s-h)$ the Dirac measure at h. Given a centered process $\{X_t\}_{t\in\mathbb{R}^+}$, the Ornstein-Uhlenbeck process can be characterized as a zero-mean stationary ARH(1) model $\{\mathcal{X}_n(t) := X_{nh+t},\ t \in [0,h]\}_{n\in\mathbb{Z}^+}$, given by

$$\mathcal{X}_n(t) = e^{-\theta t}\mathcal{X}_{n-1}(h) + \sigma\int_{nh}^{nh+t} e^{-\theta(nh+t-s)}\,\mathrm{d}W_s = \Lambda_\theta\left(\mathcal{X}_{n-1}\right)(t) + \mathcal{E}_n(t),$$

with $n \in \mathbb{Z}^+$ and where $\left\{\mathcal{E}_n(t) := \sigma\int_{nh}^{nh+t} e^{-\theta(nh+t-s)}\,\mathrm{d}W_s\right\}_{n\in\mathbb{Z}^+}$ constitutes a $\mathcal{H}$-valued strong white noise and Γ_θ is a bounded linear operator, for each $\theta > 0$.

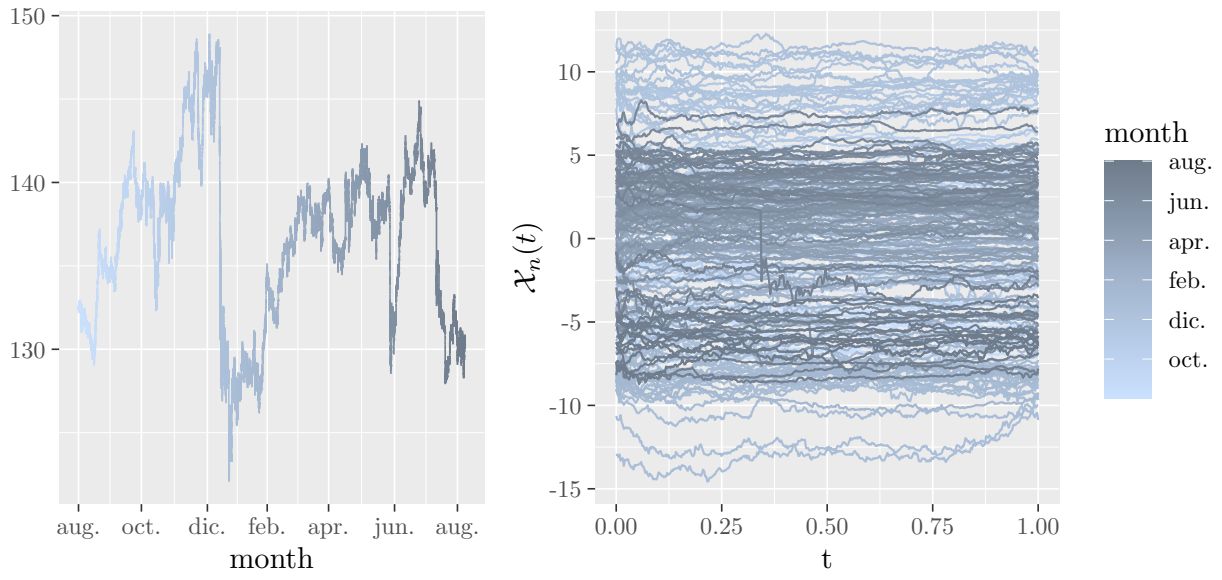


Figure 3: Johnson & Johnson stock prices recorded every minute from August 2018 to August 2019. Observed path (left) and centered daily price curves (right).

The dataset considered consist on Johnson & Johnson stock prices from August 1, 2018 to August 7, 2019, recorded every minute. Figure 3 shows the price path (left) with $98\,280$ observations and the daily curves $\{\mathcal{X}_i(t)\}_{i=1}^n$ with $n = 252$ curves (right) discretized in 390 equispaced grid points, that is, 1-minute data. The daily price curves are evaluated in $\mathcal{H} = L^2\left([0,1], \mathcal{B}_{[0,1]}, \lambda + \delta_{(1)}\right)$, where the $[0,1]$ interval corresponds to a 1-day observation window. We test the parametric form of the Ornstein-Uhlenbeck process, that is, that the daily curves $\{\mathcal{X}_i(t)\}_{i=1}^n$ constitute an ARH(1) process $\mathcal{X}_n(t) = \Lambda(\mathcal{X}_{n-1})(t) + \mathcal{E}_n(t)$ with $\Lambda(\mathcal{X})(t) := \Lambda_\theta(\mathcal{X})(t) = e^{-\theta t}\mathcal{X}(h)$. As in López-Pérez, Febrero-Bande, and González-Manteiga (López-Pérez et al.), to test the specification of the process using the independence test, we have the test

$$H_0 : \mathcal{E}_n(t) \perp \mathcal{X}_n(t) \quad \text{vs.} \quad H_1 : \mathcal{E}_n(t) \not\perp \mathcal{X}_n(t)$$

which is equivalent to test the Ornstein-Uhlenbeck specification. The p-value obtained is 0.0011, therefore the null hypothesis is rejected, as significant evidence is found against the Ornstein-Uhlenbeck as a ARH(1) process for sensible significance levels. Explaining the dynamic of the stock price may require a more intricate model, or coupling the model with jumps, as there was a decline in December due to allegations against the company.

5 Conclusions

In this article, existing procedures about specification tests in the presence of functional data are reviewed. These can be fundamentally differentiated into two types:

a) Extensions of classic procedures developed for the vectorial framework. These are based on distances between a nonparametric universally consistent pilot estimator and another one estimated under the null hypothesis assumptions.

b) Using correlation generalized coefficients. These are employed to measure independence, conditional mean independence and conditional independence. These correspond with the analyzed DC, MDD and CDC coefficients, respectively.

The development of specification tests for the functional context is not an easy task. In fact, most of the references date from the last decade. This field has attracted great interest, resulting in a very fast evolution in the recent years. These novel procedures face some important limitations as the curse of dimensionality in the big data context involving functional data. For this reason, it is currently an interesting line of research for the big data processing.

All the manuscript review is performed for specification tests in static functional models. Nevertheless, an example of specification testing for a functional continuous-time process is given in Section 4.3 to illustrate their possible adaptations.

An open line for future research is the development of specification tests for functional time series. Articles such as the ones of Edelmann et al. (2019), Davis et al. (2018), Dehling et al. (2020), Lee and Shao (2018) or Meintanis et al. (2022) could be a good starting point for construction of new specification tests in dynamic models.

There are several practical problems where functional data are of potential interest. Specially, in the medical context, where a continuous monitoring of patients features can be desirable. An example is the glucose monitoring in diabetes disease. The case of cure models, from the Survival Analysis, is specially relevant. Works as the ones of Zhang et al. (2021) or Edelmann et al. (2022) in the vectorial context based on DC ideas could bridge a gap for specification tests in cure models with functional data.

Eventually, it is important to remark that all the exposition was developed for functional data in Hilbert spaces. There are papers, like Castro-Prado and González-Manteiga (2020) or the excellent review of Jansen (2021), which extend the results to broader spaces. In these last references, a unified version of dependence measures in general metric spaces, being the Hilbertian ones a particular case, is performed. Specification tests for not only the Hilbertian case, but also for general metric spaces, is another open problem for future research.

Acknowledgments

This paper is a consequence of the invitation from Editor of this journal and the President of the Instituto Nacional de Estadística (INE) to produce an article as the second recipient of the Premio Nacional de Estadística. I am really grateful to both of them. The creation of the Spanish National Prize in Statistics was an excellent initiative of professor Juan Manuel Rodríguez Poo (the before mentioned president). This represents a fantastic link between the INE and the developments of Statistics in Spain. This work would not have been possible without the collaboration of all my co-authors: Rosa María Crujeiras Casais, Manuel Febrero Bande, Laura Freijeiro González, Eduardo García Portugués y Alejandra María. López Pérez. My acknowledgment also extends to the rest of my co-authors and all my past students in my academic life. I am an researcher interested in Statistics as a consequence of all I have learned from them.

The research of Wenceslao González-Manteiga is supported by Project PID2020-116587GB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe" and the Competitive Reference Groups 2021 – 2024 (ED431C 2021/24) from the Xunta de Galicia through the ERDF. Besides, I acknowledge the computational resources from the Supercomputing Centre of Galicia (CESGA).

References

Álvarez-Liébana, J., A. López-Pérez, M. Febrero-Bande, and W. González-Manteiga (2022). A goodness-of-fit test for functional time series with applications to Ornstein-Uhlenbeck processes. arXiv Preprint, https://arxiv.org/abs/2206.12821.

Bárcenas, R., J. Ortega, and A. J. Quiroz (2017). Quadratic forms of the empirical processes for the two-sample problem for functional data. *TEST 26*(3), 503–526.

Bickel, P. J. and M. Rosenblatt (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics 1*(6), 1071–1095.

Boente, G., D. Rodríguez, and M. Sued (2018). Testing equality between several populations covariance operators. *Annals of the Institute of Statistical Mathematics 70*(4), 919–950.

Bongiorno, E. G., A. Goia, and P. Vieu (2019). Modeling functional data: a test procedure. *Computational Statistics 34*(2), 451–468.

Bosq, D. (2000). *Linear processes in function spaces: theory and applications*, Volume 149. Springer Science & Business Media.

Bugni, F. A., P. Hall, J. L. Horowitz, and G. R. Neumann (2009). Goodness-of-fit tests for functional data. *The Econometrics Journal 12*(S1), S1–S18.

Bugni, F. A. and J. L. Horowitz (2021). Permutation tests for equality of distributions of functional data. *Journal of Applied Econometrics 36*(7), 861–877.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313 – 2351.

Carey, J. R., P. Liedo, H. G. Müller, J. L. Wang, and J. M. Chiou (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large cohort of mediterranean fruit fly females. *The Journals of Grontology. Series A, Biological Sciences and Medical Sciences 53 4*, B245–51.

Castro-Prado, Fernando and Wenceslao González-Manteiga (2020). Nonparametric independence tests in metric spaces: What is known and what is not. https://arxiv.org/abs/2009.14150.

Chaudhuri, A. and W. Hu (2019). A fast algorithm for computing distance correlation. *Computational Statistics and Data Analysis 135*, 15–24.

Chen, F., Q. Jiang, Z. Feng, and L. Zhu (2020). Model checks for functional linear regression models based on projected empirical processes. *Computational Statistics & Data Analysis 144*, 106897.

Cuesta-Albertos, J. A., E. del Barrio, R. Fraiman, and C. Matrán (2007). The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis 51*(10), 4814–4831.

Cuesta-Albertos, J. A., R. Fraiman, and T. Ransford (2006). Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bulletin of the Brazilian Mathematical Society 37*(4), 477–501.

Cuesta-Albertos, J. A., E. García-Portugués, M. Febrero-Bande, and W. González-Manteiga (2019). Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics 47*(1), 439–467.

Cuevas, A., M. Febrero, and R. Fraiman (2004). An anova test for functional data. *Computational Statistics & Data Analysis 47*(1), 111–122.

Davis, R. A., M. Matsui, T. Mikosch, and P. Wan (2018). Applications of distance correlation to time series. *Bernoulli 24*(4A), 3087 – 3116.

Dehling, H., M. Matsui, T. Mikosch, G. Samorodnitsky, and L. Tafakori (2020). Distance covariance for discretized stochastic processes. *Bernoulli 26*(4), 2758 – 2789.

Delsol, L., F. Ferraty, and P. Vieu (2011). Structural test in regression on functional variables. *Journal of Multivariate Analysis 102*(3), 422–447.

Ditzhaus, M. and D. Gaigall (2018). A consistent goodness-of-fit test for huge dimensional and functional data. *Journal of Nonparametric Statistics 30*(4), 834–859.

Durbin, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *The Annals of Statistics 1*(2), 279–290.

Edelmann, D., K. Fokianos, and M. Pitsillou (2019). An Updated Literature Review of Distance Correlation and Its Applications to Time Series. *International Statistical Review 87*(2), 237–262.

Edelmann, D. and J. Goeman (2022). A Regression Perspective on Generalized Distance Covariance and the Hilbert-Schmidt Independence Criterion. *Statistical Science 37*(4), 562 – 579.

Edelmann, D., T. Welchowski, and A. Benner (2022). A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. *Biometrics 78*(3), 867–879.

Estévez-Pérez, G. and José A. Vilar (2013). Functional anova starting from discrete data: an application to air quality data. *Environmental and Ecological Statistics 20*(3), 495–517.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Febrero-Bande, M., W. González-Manteiga, and M. Oviedo de la Fuente (2019). Variable selection in functional additive regression models. *Computational Statistics 34*(2), 469–487.

Febrero-Bande, M. and M. Oviedo de la Fuente (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software 51*(4), 1–28.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. New York: Springer.

Freijeiro-González, L., M. Febrero-Bande, and W. González-Manteiga (2022). A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *International Statistical Review 90*(1), 118–145.

Fremdt, S., J. G. Steinebach, L. Horváth, and P. Kokoszka (2013). Testing the equality of covariance operators in functional samples. *Scandinavian Journal of Statistics 40*(1), 138–152.

García-Portugués, E., J. Álvarez-Liébana, G. Álvarez-Pérez, and W. González-Manteiga (2020). Goodness-of-fit tests for functional linear models based on integrated projections. In Germán Aneiros, Ivana Horová, Marie Hušková, and Philippe Vieu (Eds.), *Functional and High-Dimensional Statistics and Related Fields*, Cham, pp. 107–114. Springer International Publishing.

García-Portugués, E., J. Álvarez-Liébana, G. Álvarez-Pérez, and W. González-Manteiga (2021). A goodness-of-fit test for the functional linear model with functional response. *Scandinavian Journal of Statistics 48*(2), 502–528.

García-Portugués, E., W. González-Manteiga, and M. Febrero-Bande (2014). A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics 23*(3), 761–778.

Ghale-Joogh, H. S. and S. M. E. Hosseini-Nasab (2018). A two-sample test for mean functions with increasing number of projections. *Statistics 52*(4), 852–873.

González-Manteiga, W. and R. M. Crujeiras (2013). An updated review of goodness-of-fit tests for regression models. *Test 22*(3), 361–411.

González-Manteiga, W., R. M. Crujeiras, and E. García-Portugués (2022). *Trends in Mathematical, Information and Data Sciences*, Volume 445 of *Studies in Systems, Decision and Control*, Chapter A Review of Goodness-of-Fit Tests forÂăModels Involving Functional Data, pp. 349–358. Cham: Springer International Publishing.

González-Rodrígez, G., A. Colubi, and M. A. Gil (2012). Fuzzy data treated as functional data: A one-way anova test approach. *Computational Statistics & Data Analysis 56*(4), 943–955.

Górecki, T. and S. Łukasz (2019). fdanova: an r software package for analysis of variance for univariate and multivariate functional data. *Computational Statistics 34*(2), 571–597.

Gretton, Arthur, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf (2005). Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita (Eds.), *Algorithmic Learning Theory*, Berlin, Heidelberg, pp. 63–77. Springer Berlin Heidelberg.

Guo, J., B. Zhou, J. Chen, and J.-T. Zhang (2019). An L^2-norm-based test for equality of several covariance functions: a further study. *TEST 28*(4), 1092–1112.

Guo, J., B. Zhou, and J.-T. Zhang (2018). Testing the equality of several covariance functions for functional data: A supremum-norm based test. *Computational Statistics & Data Analysis 124*, 15–26.

Guo, J., B. Zhou, and J.-T. Zhang (2019). New tests for equality of several covariance functions for functional data. *Journal of the American Statistical Association 114*(527), 1251–1263.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics 21*(4), 1926–1947.

Henze, N. and M. D. Jiménez-Gamero (2021). A test for Gaussianity in Hilbert spaces via the empirical characteristic functional. *Scandinavian Journal of Statistics 48*(2), 406–428.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. New York: Springer.

Horváth, L. and R. Reeder (2013). A test of significance in functional quadratic regression. *Bernoulli 19*(5A), 2130–2151.

Horváth, L. and G. Rice (2015). An introduction to functional data analysis and a principal component approach for testing the equality of mean curves. *Revista Matemática Complutense 28*(3), 505–548.

Hsing, T. and R. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*, Volume 997. John Wiley & Sons.

Hu, Wenjuan, Nan Lin, and Baoxue Zhang (2020). Nonparametric testing of lack of dependence in functional linear models. *PLOS ONE 15*(6), 1–24.

Hua, W.-Y. and D. Ghosh (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. *Biometrics 71*(3), 812–820.

Huo, X. and G. J. Székely (2016). Fast computing for distance covariance. *Technometrics 58*(4), 435–447.

Jansen, S. (2021). On distance covariance in metric and Hilbert spaces. *ALEA 18*, 1353–1393.

Jiang, Q., M. Hušková, S. G. Meintanis, and L. Zhu (2019). Asymptotics, finite-sample comparisons and applications for two-sample tests with functional data. *Journal of Multivariate Analysis 170*, 202–220.

Kellner, J. and A. Celisse (2019). A one-sample test for normality with kernel methods. *Bernoulli 25*(3), 1816–1837.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*(1/2), 81–93.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. Texts in Statistical Science Series. CRC Press, Chapman & Hall.

Kolkiewicz, A., G. Rice, and Y. Xie (2021). Projection pursuit based tests of normality with functional data. *Journal of Statistical Planning and Inference 211*, 326–339.

Kong, D., A. M. Staicu, and A. Maity (2016). Classical testing in functional linear models. *Journal of Nonparametric Statistics 28*(4), 813–830.

Lai, T., Z. Zhang, and Y. Wang (2020). Testing independence and goodness-of-fit jointly for functional linear models. *Journal of the Korean Statistical Society 50*.

Lee, C. E. and X. Shao (2018). Martingale Difference Divergence Matrix and Its Application to Dimension Reduction for Stationary Multivariate Time Series. *Journal of the American Statistical Association 113*(521), 216–229.

Lee, C. E., X. Zhang, and X. Shao (2020). Testing conditional mean independence for functional data. *Biometrika 107*(2), 331–346.

Lee, J. S., D. D. Cox, and M. Follen (2015). A two sample test for functional data. *Communications for Statistical Applications and Methods 22*(2), 121–135.

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Liu, J., R. Li, and R. Wu (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association 109*(505), 266–274.

López-Pérez, A., M. Febrero-Bande, and W. González-Manteiga. A comparative review of specification tests for diffusion models. arXiv Preprint, https://arxiv.org/abs/2208.08420.

Lu, J. and L. Lin (2020). Model-free conditional screening via conditional distance correlation. *Statistical Papers 61*, 225 – 244.

Lundborg, A. R., R. D. Shah, and J. Peters (2022). Conditional independence testing in hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability 41*(5), 3284–3305.

Maistre, S. and V. Patilea (2020). Testing for the significance of functional covariates. *Journal of Multivariate Analysis 179*, 104648.

McLean, M. W., G. Hooker, and D. Ruppert (2015). Restricted likelihood ratio tests for linearity in scalar-on-function regression. *Statistics and Computing 25*(5), 997–1008.

Meintanis, S. G., M. Hušková, and Z. Hlávka (2022). Fourier-type tests of mutual independence between functional time series. *Journal of Multivariate Analysis 189*, 104873.

Park, T., X. Shao, and S. Yao (2015). Partial martingale difference correlation. *Electronic Journal of Statistics 9*(1), 1492 – 1517.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*(3), 1065–1076.

Patilea, V. and C. Sánchez-Sellero (2020). Testing for lack-of-fit in functional regression models against general alternatives. *Journal of Statistical Planning and Inference 209*, 229–251.

Patilea, V., C. Sánchez-Sellero, and M. Saumard (2016). Testing the predictor effect on a functional response. *Journal of the American Statistical Association 111*(516), 1684–1695.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika 13*(1), 25–45.

Pokotylo, O., P. Mozharovskyi, and R. Dyckerhoff (2019). Depth and depth-based classification with R package ddalpha. *Journal of Statistical Software 91*(5), 1–46.

Pomann, G.-M., A.-M. Staicu, and S. Ghosh (2016). A two-sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 65*(3), 395–414.

Qiu, Z., J. Chen, and J.-T. Zhang (2021). Two-sample tests for multivariate functional data with applications. *Computational Statistics & Data Analysis 157*, 107160.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. New York: Springer.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics 27*(3), 832–837.

Schick, A. (1997). On U-statistics with random kernels. *Statistics & Probability Letters 34*(3), 275–283.

Sejdinovic, D., B. Sriperumbudur, A. Gretton, and K. Fukumizu (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics 41*(5), 2263 – 2291.

Sen, A. and B. Sen (2014). Testing independence and goodness-of-fit in linear models. *Biometrika 101*(4), 927–942.

Shah, R. D. and J. Peters (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics 48*(3), 1514 – 1538.

Shao, X. and J. Zhang (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association 109*(507), 1302–1318.

Shi, E., Y. Liu, K. Sun, L. Li, and L. Kong (2022). An adaptive model checking test for functional linear model. https://arxiv.org/abs/2204.01831.

Smaga, L. (2022). Projection tests for linear hypothesis in the functional response model. *Communications in Statistics - Theory and Methods 0*(0), 1–18.

Song, F., Y. Chen, and P. Lai (2020). Conditional distance correlation screening for sparse ultrahigh-dimensional models. *Applied Mathematical Modelling 81*, 232–252.

Song, L., A. Smola, A. Gretton, J. Bedo, and K. Borgwardt (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research 13*(1), 1393–1434.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics 25*(2), 613–641.

Su, L. and X. Zheng (2017). A martingale-difference-divergence-based test for specification. *Economics Letters 156*, 162–167.

Székely, G. J. and M. L. Rizzo (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis 117*, 193–213.

Székely, G. J. and M. L. Rizzo (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics 42*(6), 2382 – 2412.

Székely, G. J. and M. L. Rizzo (2017). The energy of data. *Annual Review of Statistics and Its Application 4*, 447–479.

Székely, G. J., M. L Rizzo, N. K. Bakirov, et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*(6), 2769–2794.

Tekbudak, M. Y., M. Alfaro-Córdoba, A. Maity, and A. M. Staicu (2019). A comparison of testing methods in scalar-on-function regression. *AStA. Advances in Statistical Analysis 103*(3), 411–436.

Teran Hidalgo, S., M. Wu, S. Engel, and M. Kosorok (2018). Goodness-Of-Fit Test for Nonparametric Regression Models: Smoothing Spline ANOVA Models as Example. *Computational Statistics & Data Analysis 122*.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*, Volume 60 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Wang, X., W. Pan, W. Hu, Y. Tian, and H. Zhang (2015). Conditional distance correlation. *Journal of the American Statistical Association 110*(512), 1726–1734.

Wissler, C. (1905). The spearman correlation formula. *Science 22*(558), 309–311.

Xu, K. and D. He (2021). Omnibus model checks of linear assumptions through distance covariance. *Statistica Sinica 31*, 1055–1079.

Zhang, J., Y. Liu, and H. Cui (2021). Model-free feature screening via distance correlation for ultrahigh dimensional survival data. *Statistical Papers 62*(6), 2711–2738.

Zhang, J.-T. and X. Liang (2014). One-Way ANOVA for Functional Data via Globalizing the Pointwise F-test. *Scandinavian Journal of Statistics 41*(1), 51–71.

Zhang, X., S. Yao, and X. Shao (2018). Conditional mean and quantile dependence testing in high dimension. *The Annals of Statistics 46*(1), 219 – 246.

Zhao, F., N. Lin, W. Hu, and B. Zhang (2022). A faster U-statistic for testing independence in the functional linear models. *Journal of Statistical Planning and Inference 217*, 188–203.

Zhu, C., X. Zhang, S. Yao, and X. Shao (2020). Distance-based and RKHS-based dependence metrics in high dimension. *The Annals of Statistics 48*(6), 3366 – 3394.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 41–54
doi:https://doi.org/10.37830/SJS.2022.1.03

REGULAR ARTICLE

Testing Benford's law: from small to very large data sets

Leonardo Campanelli

All Saints University School of Medicine, leonardo.s.campanelli@gmail.com
Received: May 20, 2022, Accepted: December 30, 2022.

Abstract: We discuss some limitations of the use of generic tests, such as the Pearson's χ^2, for testing Benford's law. Statistics with known distribution and constructed under the specific null hypothesis that Benford's law holds, such as the Euclidean distance, are more appropriate when assessing the goodness-of-fit to Benford's law, and should be preferred over generic tests in quantitative analyses. The rule of thumb proposed by Goodman for compliance checking to Benford's law, instead, is shown to be statistically unfounded. For very large sample sizes ($N > 1000$), all existing statistical tests are inappropriate for testing Benford's law due to its empirical nature. We propose a new statistic whose sample values are asymptotically independent on the sample size making it a natural candidate for testing Benford's law in very large data sets.

Keywords: Benford's law, large data sets, goodness of fit, euclidean distance statistic, Pearson's χ^2

MSC: 62H15, 65C05, 62-07, 62G20

1 Introduction

Benford's law (Benford, 1938) on the distribution of the first significant digit (FSD) of numerical data is an empirical law that has been observed to emerge in disparate data sets, from finance (Nigrini, 1996; Cho and Gaines, 2007) and natural sciences (Sambridge et al., 2010) to COVID 19 data (Sambridge and Jackson, 2020; Campanelli, 2022).

By analyzing the data coming from very different distributions, such as length of rivers, populations of cities, etc., Benford found that the probability of occurrence of the first significant digit d, $P_B(d)$, followed the empirical law

$$\forall\, d \in \{1, ..., 9\}:\ P_B(d) = \log\left(1 + \frac{1}{d}\right). \tag{1}$$

Although we know today that Benford's law holds for some particular distributions (see Morrow (2014) and references therein) and that specific principles lead to the emergence of such a law (Hill,

1995a,b,c), the exact fundamental bases upon which Benford's law reposes are still unknown [for a review of Benford's law, see Miller (2015)].

The most common test in use for testing Benford's law is the Pearson's χ^2. In our case the χ^2 statistic can be written as

$$\chi^2 = N \sum_{d=1}^{9} \frac{[P_B(d) - P(d)]^2}{P_B(d)}, \tag{2}$$

where $P(d)$ is the observed relative frequency of the FSD d, and N is the sample size. However, such a test is based on the null hypothesis of a continuous distribution, and is generally conservative for testing discrete distributions as the Benford's one (Noether, 1963). This problem has been recently solved by Morrow (2014) who has computed asymptotically test values for this statistic under the specific null hypothesis that Benford's law holds.

Another estimator used for checking conformance to Benford's law is the "normalized Euclidean distance", d^*, introduced by Cho and Gaines (2007) and defined by

$$d^* = \frac{1}{D} \sqrt{\sum_{d=1}^{9} [P(d) - P_B(d)]^2}, \tag{3}$$

where $D = \sqrt{\sum_{d=1}^{8} P_B^2(d) + [P(9) - 1]^2}$ is a normalization factor that assures that d^* is bounded by 0 and 1. At the moment of its introduction, however, the properties of this new estimator were not well understood and no test values were reported. These problems have been solved by Morrow (2014), who has provided asymptotically test values for the "Euclidean distance"

$$d_N^* = \sqrt{N \sum_{d=1}^{9} [P(d) - P_B(d)]^2}, \tag{4}$$

and by the author who, recently enough (Campanelli, 2023), has found an empirical expression of its cumulative distribution function. A simple measure of fit to Benford's law, instead, has been proposed by Goodman (2016). His "rule of thumb" for conformance to Benford's law is $d^* \leq 0.25$.

One of the goals of this paper is to show the statistical incorrectness of Goodman's rule of thumb. Also, we will show that the use of p values of the χ^2 statistic for testing Benford's law is only appropriate for "qualitative" analyses, while the use of the Euclidean distance test should be preferred in "quantitative" analyses. Finally, we will discuss some limitations of existing statistical tests to assess the goodness-of-fit to Benford's law for very large number of data points ($N \leq 1000$) and/or small range of data, and we will propose a new statistic that overcome such limitations.

2 Euclidean distance statistic, χ^2 test, and Goodman's rule of thumb

The knowledge of the cumulative distribution function of the Euclidean distance statistic as a function of the sample size N, as derived in Campanelli (2023), makes possible the computation of p values and then allow us to check for the conformance of a set of data to Benford's law in a quantitative way. It is interesting, for example, to reconsider the data that allowed Benford to discover the law that now brings his name. In Table 1, we show the Euclidean distance d_N^* and its corresponding p value for the first-digit distribution of the twenty different groups of counts discussed by Benford in his original paper (Benford, 1938), while in Figure 1, we show the corresponding first-digit frequencies superimposed to Benford's law.

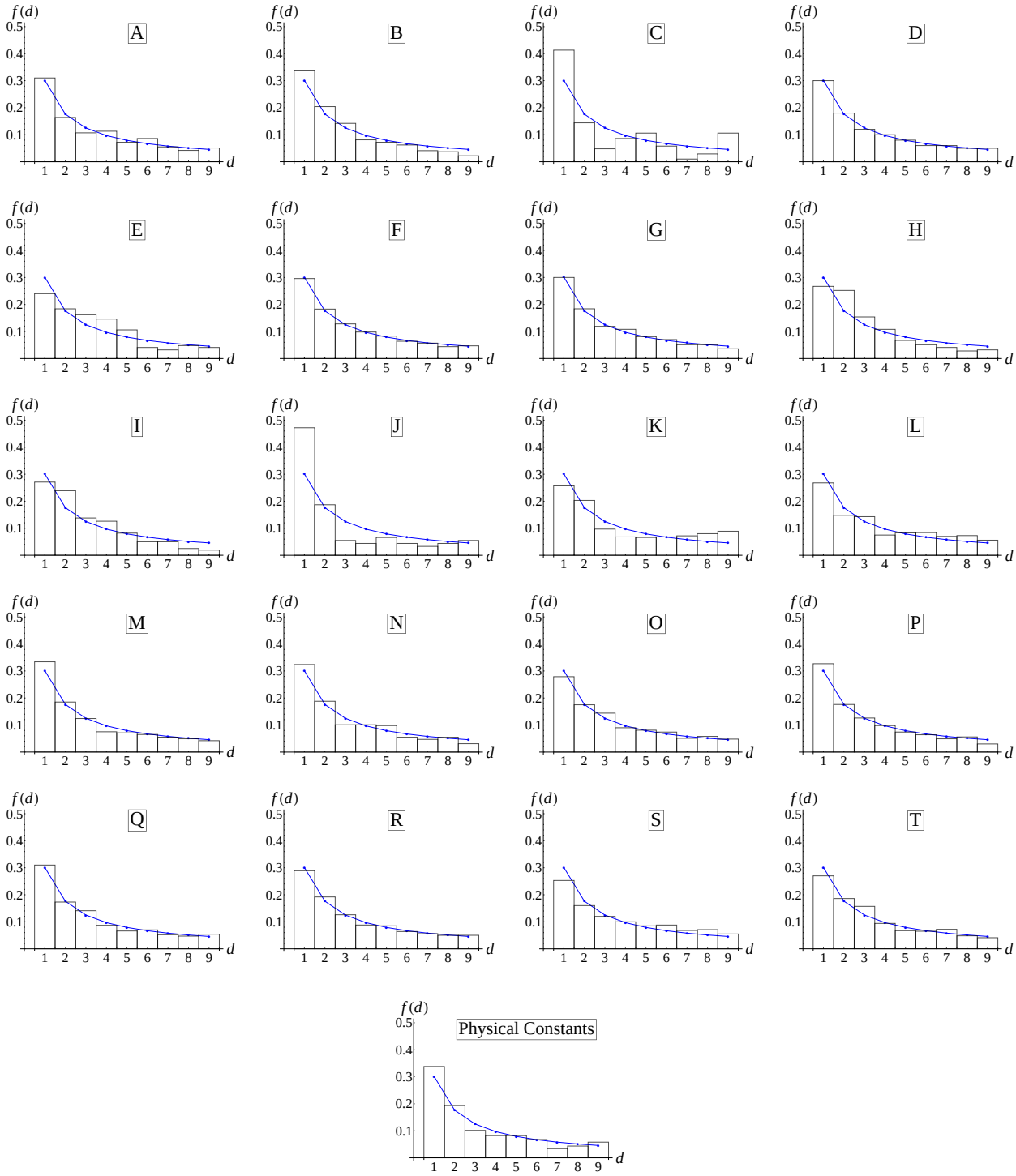


Figure 1: *Panels A to T.* Observed first-digit frequencies for the samples originally considered by Benford (1938) and shown in Table 1. *Bottom panel.* First-digit frequency of the values of the physical constants tabulated in Lide (2002). The (blue) continuous lines represent Benford's law.

Group	Title	N	d^*	d_N^*	p	χ^2	$p(\chi^2)$
A	Rivers, Area	335	0.0354	0.6705	0.78(9)	4.9617	0.7617
B	Population	3259	0.0602	3.5625	0.000(0)	118.63	0.0000
C	Constants	104	0.1581	1.6704	0.00(4)	24.441	0.0019
D	Newspapers	100	0.0107	0.1107	0.999(9)	0.1602	1.0000
E	Spec. Heat	1389	0.0949	3.6652	0.000(0)	111.21	0.0000
F	Pressure	703	0.0122	0.3360	0.99(6)	1.2704	0.9959
G	H.P. Lost	690	0.0188	0.5111	0.94(6)	3.4606	0.9022
H	Mol. Wgt.	1800	0.0931	4.0924	0.000(0)	125.76	0.0000
I	Drainage	159	0.0843	1.1018	0.17(8)	11.142	0.1938
J	Atomic Wgt.	91	0.1893	1.8718	0.000(9)	17.246	0.0277
K	n^{-1}, $\sqrt{n}$,...	5000	0.0827	6.0631	0.000(0)	440.76	0.0000
L	Design	560	0.0588	1.4430	0.02(4)	19.213	0.0138
M	*Digest*	308	0.0403	0.7333	0.69(8)	3.2271	0.9193
N	Cost Data	741	0.0443	1.2503	0.08(0)	15.601	0.0485
O	X-Ray Volts	707	0.0313	0.8622	0.48(9)	5.4256	0.7113
P	Am. League	1458	0.0315	1.2461	0.08(2)	14.595	0.0675
Q	Black Body	1165	0.0264	0.9350	0.37(7)	9.5229	0.3001
R	Addresses	342	0.0225	0.4314	0.98(0)	1.2966	0.9956
S	n^1,n^2,... $n!$	900	0.0583	1.8140	0.00(1)	24.994	0.0016
T	Death Rate	418	0.0480	1.0178	0.26(6)	7.5550	0.4781

Table 1: The Euclidean distance statistic d_N^* and its corresponding p value for the first-digit distribution of twenty different groups of counts discussed by Benford in his original paper (Benford, 1938). Also indicated are the total number of counts for each group, N, and the normalized Euclidean distance d^*. (Digits in parentheses at the third and fourth decimal places indicate an error on those digits of ± 1). The last two columns show the χ^2 score and its corresponding p value, $p(\chi^2)$.

The data considered by Benford were collected from many different and disparate fields, from random numbers appearing within the covers of the same magazine to the values of physical constants [the rows K and S refer to an amalgamation of the observations of the first-digit frequencies of reciprocal and roots (row K) and powers and factorial (row S) of positive natural numbers]. At a first glance, the data suggest a certain regularity in the distribution of the first-digit, as it is evident in Figure 1, and as it was evident to Benford himself to the point that he claimed that "as no definite exceptions have ever been observed among true variables, the logarithmic law for large numbers evidently goes deeper among the roots of primal causes than our number system unaided can explain".

Surprisingly enough, however, half of the cases considered by Benford do not conform to Benford's law at a significance level of 0.10. Moreover, 40% do not conform at a significance level of 0.05 and one quarter do not conform at a significance level of 0.001.

The reasons for a non-conformance to Benford's law can be disparate. As stressed by Benford's himself (Benford, 1938), Benford's law "applies particularly to those outlaw numbers that are without known relationship rather than to those that individually follow an orderly course; and therefore the logarithmic relation is essentially a Law of Anomalous Numbers". Thus, groups E, H, J, and S do not comply to Benford's law probably because the underlying distributions of numbers do not satisfies Benford's requirement of "non-orderliness".

Another possibility is that the range of data is not sufficiently large to ensure conformance to Benford's law, which holds in the limit of an infinite range of data (Benford, 1938). This is probably the case of group C. In fact, if one considers the values of the physical constants as reported in Lide (2002), whose values extend on more than about 68 decades, one finds full conformance to Benford's law (see the bottom panel of Figure 1). In this case, we have $N = 207$, $d^* = 0.0550$, $d^*_{207} = 0.8196$, and $p = 0.55(8)$.

Groups B and K are the groups with the highest number of counts. A possible reason for the non-conformance in the case could be the enormous power of statistical tests for large N, which makes them too rigid to assess the goodness-of-fit well. This problem, and a possible solution, will be discussed in Sec. III.

It is worth observing that the use of the Cho-Gaines' normalized Euclidean distance d^* together with Goodman's rule of thumb for compliance to Benford' law, $d^* < 0.25$, would give a compliance to Benford's law for all groups of data in Table 1. Such a compliance is highly questionable. For example, consider group D and J. Both groups consist of a number of counts of about 100, but while D "seems" to follow Benford's law, J displays a big departure from it (see Figure 1).

Also, group B "seems" to display a high level of "Benfordness", but the use of the Euclidean distance statistic excludes the compliance to Benford's law at a significance level of 0.001.

This sort of "visual Benfordness" is then not reliable. This can be also understood by using the Pearson's χ^2 statistic. In the last two columns of Table 1, we show the value of the χ^2 and its corresponding p value, $p(\chi^2)$, for each group of count discussed by Benford (for the case of the values of the physical constants discussed above, we have $\chi^2 = 5.6983$, and $p(\chi^2) = 0.6810$). As it is clear, the p values of the χ^2 statistic differ, sometimes substantially, from the ones of the Euclidean distance statistic. This strongly indicates that the use of the χ^2 statistic for checking the conformance of a set of data to Benford's law is not completely reliable and should be used only for "qualitative" analyses. This is not surprising since, as it is well known, the χ^2 test has been designed for testing continuous distributions and is generally conservative for testing discrete ones, such as Benford's law (Noether, 1963).

In order to better understand the above two issues, the incorrectness of Goodman's rule of thumb and visual Benfordness, we have prepared six first-digit mock distributions with different total number of counts N. These are presented in Table 2 and visualized in Figure 2.

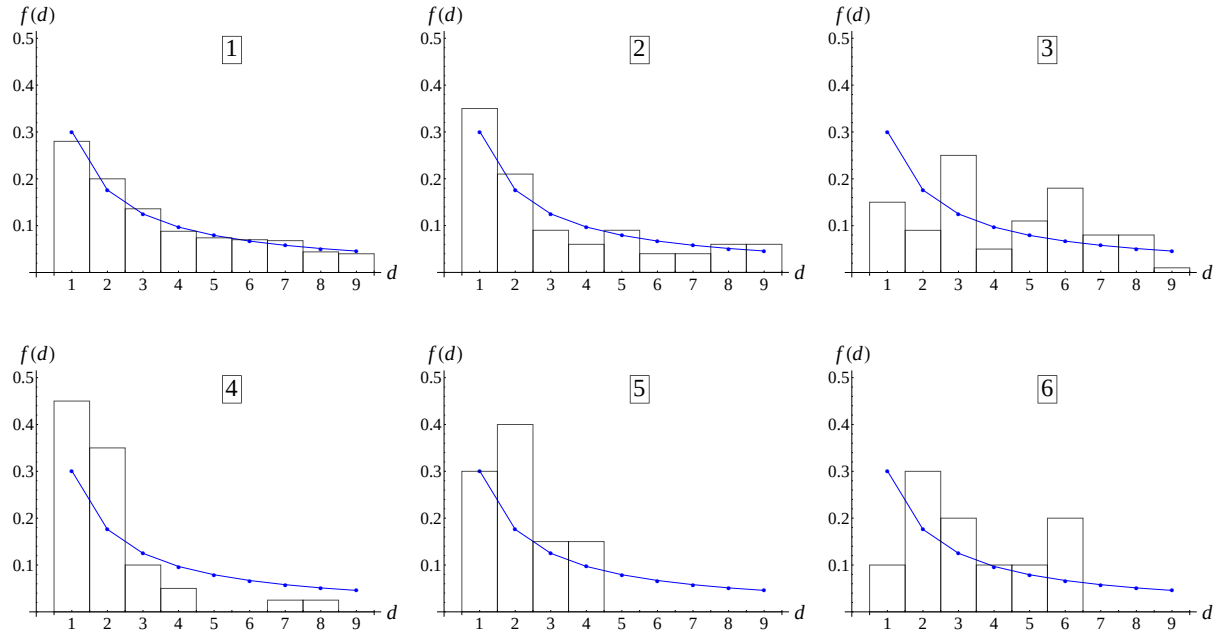


Figure 2: Graphical representations of the first-digit (mock) distributions in Table 2. The (blue) continuous lines represent Benford's law.

Group	N	$f(1)$	$f(2)$	$f(3)$	$f(4)$	$f(5)$	$f(6)$	$f(7)$	$f(8)$	$f(9)$	d^*	d^*_N	p	χ^2	$p(\chi^2)$
1	2500	0.280	0.200	0.136	0.088	0.074	0.070	0.068	0.044	0.040	0.0366	1.8949	0.000(7)	26.110	0.0010
2	500	0.360	0.200	0.100	0.060	0.080	0.040	0.040	0.060	0.060	0.0828	1.9190	0.000(6)	28.117	0.0005
3	100	0.150	0.090	0.250	0.050	0.110	0.180	0.080	0.008	0.010	0.2449	2.5375	0.000(0)	52.123	0.0000
4	40	0.450	0.350	0.100	0.050	0.000	0.000	0.025	0.025	0.000	0.2551	1.6718	0.00(4)	19.887	0.0108
5	20	0.300	0.400	0.150	0.150	0.000	0.000	0.000	0.000	0.000	0.2597	1.2034	0.10(2)	12.397	0.1343
6	10	0.100	0.300	0.200	0.100	0.100	0.200	0.000	0.000	0.000	0.2856	0.9361	$[0.25, 0.50]$	6.9145	0.5459

Table 2: The normalized Euclidean distance d^*, and the Euclidean distance statistic d^*_N and its corresponding p value, for six first-digit mock frequency distributions, $f(d)$, with different total number of counts N. (Digits in parentheses at the third and fourth decimal places indicate an error on those digits of ± 1). The last two columns show the χ^2 score and its corresponding p value, $p(\chi^2)$.

A look at Figure 2 would indicate high Benfordness of group 1, a moderate Benfordness of group 2, a low-level Benfordness of group 4, and non-Benfordness of groups 3, 5, and 6. The Euclidean distance statistic, on the contrary, shows that groups 1 and 2 do not conform to Benford's law at a significance level of 0.001 and group 4 at a significance level of 0.005. Moreover, groups 5 and 6 do conform to Benford's law at significance levels of 0.10 and > 0.25, respectively. Also, although groups 1, 2, and 3 do comply to Benford's law according to Goodman's rule, they do not at a significance level of 0.001 according to the Euclidean distance statistic. Finally, while the Benfordness of groups 5 and 6 should be rejected by Goodman's rule, the Euclidean distance statistic indicates a compliance to Benford's law at very high significance levels. Qualitatively, one can reach similar conclusions by using the χ^2 statistic (see the last two columns in Table 2).

Before considering group 4, it is worth noticing that Goodman's rule of thumb was obtained by the author by considering 40 empirical data sets displaying some visual "degree" of Benfordness (such a degree of Benfordness was not quantified by Goodman). He found that 95% of data sets had a d^* smaller than 0.256. So, he concluded that a value $d^* > 0.25$ is a strong indication of non-compliance to Benford's law, independently on the total number of counts N. We already showed that Goodman's rule generates wrong results when applied to data sets with counts larger or smaller than $N = 40$. But also when considering N around 40, one finds that Goodman's rule is not reliable. Indeed, let's now consider group 4 which contains exactly 40 counts. With a d^* value of 0.2551 and a $d^*_N = 1.6718$, the null hypothesis of conformance to Benford's law cannot be rejected at a significance level of 0.05 according to Goodman results, while it is rejected at a significance level of 0.005 by the Euclidean distance test. In this case, then, Goodman's rule is very conservative in rejecting the null hypothesis. This is probably due to the fact that the 40 empirical data sets used by Goodman had different "levels" of Benfordness.

3 The problem with very large data sets: ε-Benford's law

The Euclidean distance and the χ^2 tests, and in general all other tests used for checking the compliance of a data set to Benford's law, are very sensitive to the sample size N. In particular, they have enormous power for large N making them too rigid to assess the goodness-of-fit well: even a tiny deviation of the first-digit counts from Benford's distribution will be statistically significant. The severity of the existing tests for testing Benford's law for large N, can be traced back to the following reasons:

i) Benford's law does not represent a true law of numbers;

ii) Benford's law emerges in the limit of infinite range of the underlying distribution.

The emergence of Benford's law from a particular sample depends on the properties of the underlying distribution. However, no general criteria has be found that fully explain when and why Benford's law holds for a generic set of data. So, one major problem when testing for Benford's law is that it is not always possible to know in advance if a set of data is expected to follow it or not. This means that the rejection/acceptance of the null can be misleading when the underlying distribution is "close" to but not exactly Benford's, and this regardless of data quality. This problem is exacerbated by the increase of power of statistical tests with the sample size and, for very large sample sizes, say $N \gg 1000$, it makes any statistical test unreliable.

Also, Benford's law, even when it is known to hold exactly Morrow (2014); Hill (1995a,b,c), emerges from underlying distributions that extend on infinite ranges. In real applications, however, the set of data is restricted to a finite range and, typically, to just few decades. For finite ranges,

Year	N	d^*	d_N^*	χ^2	p
1994	9632	0.052	5.3	350	> 0.10
1996	11108	0.081	8.9	510	< 0.001
1998	9694	0.061	6.2	420	$(0.01, 0.05)$
2000	10771	0.072	7.7	670	$(0.001, 0.01)$
2002	10348	0.097	10	1100	< 0.001
2004	8396	0.130	12	2200	< 0.001

Table 3: The normalized Euclidean distance d^*, the Euclidean distance statistic d_N^*, and the χ^2 score for the first-digit distribution of in-kind contributions for six particular election cycles discussed by Cho and Gaines (2007). Also indicated is the total number of counts for each group, N. The last column shows the p values of the Euclidean distance statistic for a ε-Benford's distribution with $\varepsilon = 0.20$.

then, we expect a deviation from Benford's law even if the underlying distribution is exactly Benford. Moreover, we expect that such a deviation becomes statistically significant at large N.

In order to overcome the problem of the enormous power of existing statistical tests for large N, Cho and Gaines (2007) introduced the normalized Euclidean distance statistic in the attempt to quantify the deviation of a data set from Benford's law. However, as pointed out by the authors, the use of this statistic can only identify possible anomalies that deserve further inspection, but does not represent a "quantitative" statistical tool for testing Benford's law.

To better understand this point, let us consider the data analyzed by Cho and Gaines (2007) about the first-digit frequencies of in-kind contributions for six particular election cycles. In Table 3, we show the normalized Euclidean distance d^*, the Euclidean distance statistic d_N^*, and the χ^2 score for such distributions. Due to the extremely large values of both d_N^* and χ^2, the null hypothesis of conformance to Benford's law is rejected at any conceivable significance level for all years. The very large number of counts for each group, of order of 10^5, makes the Euclidean distance and χ^2 tests too powerful to properly assess the goodness-of-fit. However, the values of the normalized Euclidean statistic d^*, as well as those of the Euclidean statistic d_N^*, indicate that the last two elections exhibit a somewhat worse fit than their earlier counterparts (Cho and Gaines, 2007).

In the rest of this Section, we will extend the work of Cho and Gaines by making the identification of anomalies more "quantitative".

We first give the following definition. A random variable X, whose first-digit probability distribution function is $P_\varepsilon(d)$, follows a ε-Benford's distribution iff

$$\forall\, d \in \{1, ..., 9\}: \left| \frac{P_\varepsilon(d) - P_B(d)}{P_B(d)} \right| \leq \varepsilon. \tag{5}$$

Here, the positive parameter $\varepsilon \times 100\%$ quantifies the maximum percentage deviation of the values of the first-digit distribution of X from Benford's law. [Notice that if a random variable X follows a ε-Benford distribution, it automatically ε-satisfies Benford's law in the sense specified byMorrow (2014).]

The first-digit frequencies of in-kind contributions discussed above fail to conform to Benford's law even if their deviations from the law are relatively small, as confirmed by the smallness of the normalized Euclidean distance statistic. Indeed, the underlying random variable could be "intrinsically" ε-Benford, or it became so due to the limitedness of the range of data. Whatever is the case, we may assess the goodness-of-fit of such frequencies to ε-Benford's law after finding the appropriate test values of the Euclidean distance statistic for a ε-Benford distribution.

To this end we performed a Monte Carlo simulation consisting, for each sample size N, of n draws from a Benford's distribution $P_B(d)$, with each value of $P_B(d)$ being multiplied by a (pseudo-)random number in the interval $[1-\varepsilon, 1+\varepsilon]$, thus obtaining the ε-Benford distribution $P_\varepsilon(d)$. In particular, we considered the cases $\varepsilon = 0.05, 0.10, 0.15, 0.20, 0.25$, and we took $n = 10^5$ for $50 \leq N \leq 10000$ and $n = 10^4$ for $10000 < N \leq 100000$. We started with $N = 50$ and $N = 100$, and then we proceeded up to 1000 by steps of 100, up to 10000 by steps of 1000, and up to 100000 by steps of 10000. We then evaluated the Euclidean distance statistic for the ε-Benford distribution, $d_N^{(\varepsilon)}$, as

$$d_N^{(\varepsilon)} = \sqrt{N \sum_{d=1}^{9} [P_\varepsilon(d) - P_B(d)]^2}\,. \tag{6}$$

The observed probability distribution function of the Euclidean distance statistic exhibits a regular dependence of the sample size N. This is apparent in the upper panels of Figure 3, where we show its mean $\overline{d_N^{(\varepsilon)}}$ and its standard deviation $s_N^{(\varepsilon)}$ as a function of the sample size N. In the middle and lower panels of Figure 3, instead, we show the test values $d_{N,1-\alpha}^{(\varepsilon)}$ for $\alpha = 0.10, 0.05, 0.01,$ and 0.001 [the test values $d_{N,1-\alpha}^{(\varepsilon)}$ are defined as $\mathrm{Cdf}\,[d_{N,1-\alpha}^{(\varepsilon)}] = 1-\alpha$, where $\mathrm{Cdf}\,[d_N^{(\varepsilon)}]$ is the (observed) cumulative distribution function of $d_N^{(\varepsilon)}$]. The (blue) continuous lines represent fits of the observed quantities and are divided in two intervals, $50 \leq N \leq 10^3$ and $10^3 \leq N \leq 10^5$. All nonlinear fits can be expressed as

$$\theta_N^{(\varepsilon)} = \left(a + bN^{-1/2} + cN^{-1}\right)\sqrt{N}, \tag{7}$$

where $\theta_N^{(\varepsilon)}$ represents any of the variables $\overline{d_N^{(\varepsilon)}}$, $s_N^{(\varepsilon)}$, and $d_{N,1-\alpha}^{(\varepsilon)}$. The fitting values a, b, and c, for both $50 \leq N \leq 10^3$ and $10^3 \leq N \leq 10^5$, are shown in Table 4.

The choice of the fitting function in Eq. (7) is suggested by the behaviour of the quantities $\theta_N^{(\varepsilon)}/\sqrt{N}$, which numerically are found to be slowly decreasing function of N approaching constant limiting values (see Figure 3). Indeed, assuming that the parameters a, b, and c remain constant for $N > 10^5$, it follows from Eq. (7) that all quantities $\theta_N^{(\varepsilon)}/\sqrt{N}$ approach asymptotic constant values for a given ε,

$$\lim_{N\to\infty} \frac{\theta_N^{(\varepsilon)}}{\sqrt{N}} = \theta^{(\varepsilon)}. \tag{8}$$

A linear fit of these values as a function of ε gives

$$\overline{d^{(\varepsilon)}} = -0.0011 + 0.1960\,\varepsilon, \tag{9}$$

$$s^{(\varepsilon)} = -0.0004 + 0.0596\,\varepsilon, \tag{10}$$

and

$$d_{0.90}^{(\varepsilon)} = -0.0017 + 0.2791\,\varepsilon, \tag{11}$$

$$d_{0.95}^{(\varepsilon)} = -0.0020 + 0.3033\,\varepsilon, \tag{12}$$

$$d_{0.99}^{(\varepsilon)} = -0.0024 + 0.3410\,\varepsilon, \tag{13}$$

$$d_{0.999}^{(\varepsilon)} = -0.0029 + 0.3717\,\varepsilon. \tag{14}$$

We show the limiting values $\theta^{(\varepsilon)}$ and their corresponding linear fits in Figure 4. It is worth noticing that these fits cannot be extrapolated down to $\varepsilon = 0$. Indeed, from the discussion in Campanelli (2023), we expect $\theta^{(\varepsilon)} \to 0$ in the limit $\varepsilon \to 0$.

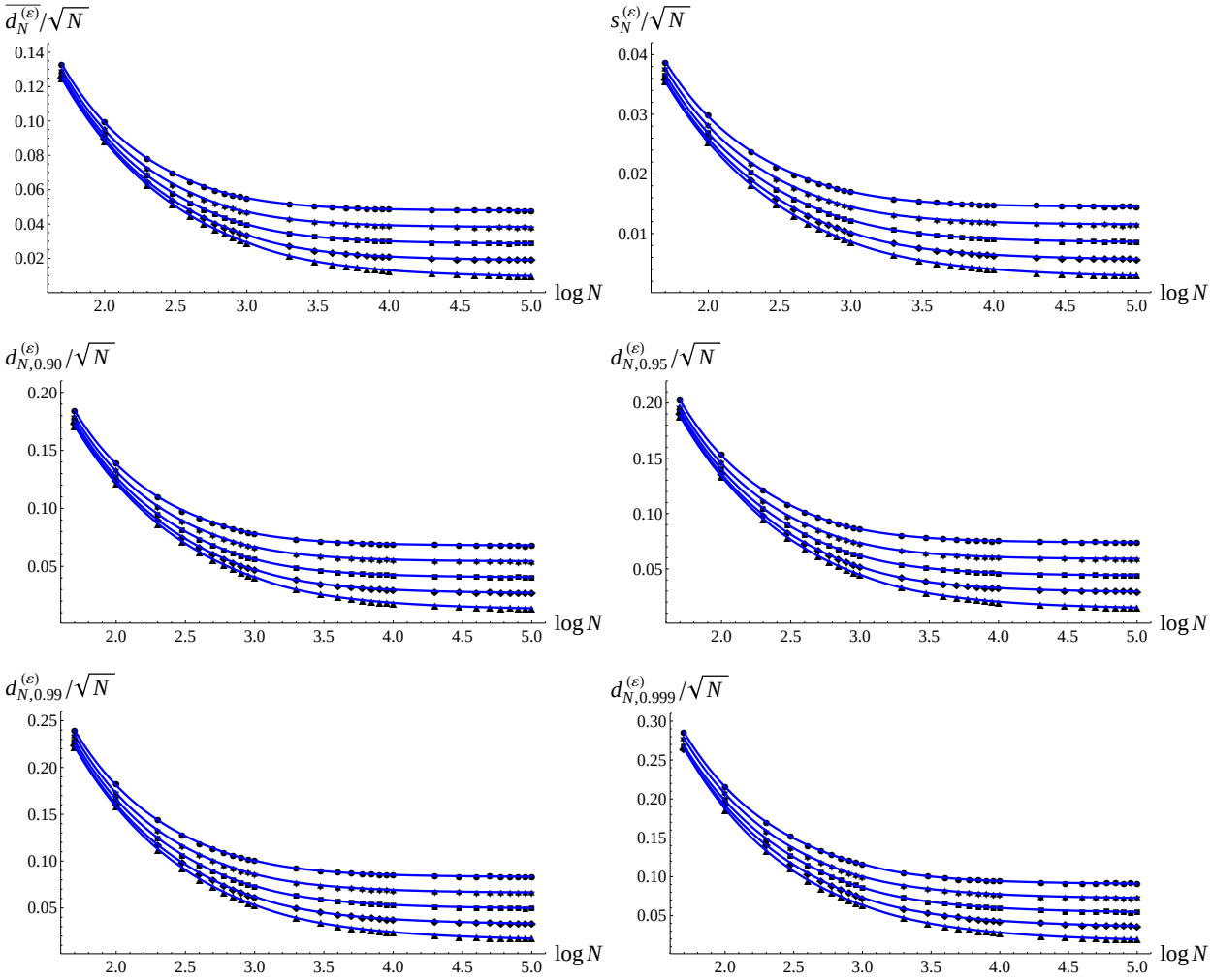


Figure 3: The mean (upper left panel), standard deviation (upper right panel), and test values (middle and lower panels) of the Euclidean distance statistic (6) as a function of the sample size N, together with their nonlinear fits (blue continuous lines), Eq. (7), for different values of ε. From bottom to top: $\varepsilon = 0.05, 0.10, 0.15, 0.20$, and 0.25.

$\overline{d_N^{(\varepsilon)}}$						
	$50 \le N \le 10^3$			$10^3 \le N \le 10^5$		
ε	a	b	c	a	b	c
0.05	0.0024	0.8429	0.1592	0.0085	0.3686	9.329
0.10	0.0090	0.7536	0.5061	0.0187	0.1252	11.10
0.15	0.0180	0.6478	0.8886	0.0285	0.0568	9.570
0.20	0.0287	0.5274	1.3311	0.0382	0.0327	7.999
0.25	0.0393	0.4382	1.5933	0.0478	0.0189	6.811
$s_N^{(\varepsilon)}$						
	$50 \le N \le 10^3$			$10^3 \le N \le 10^5$		
ε	a	b	c	a	b	c
0.05	0.0009	0.2425	0.0171	0.0025	0.1275	2.058
0.10	0.0031	0.2176	0.1027	0.0055	0.0648	2.637
0.15	0.0060	0.1875	0.2099	0.0085	0.0375	2.621
0.20	0.0091	0.1596	0.3028	0.0114	0.0259	2.231
0.25	0.0121	0.1423	0.3310	0.0145	0.0095	2.259
$d_{N,0.90}^{(\varepsilon)}$						
	$50 \le N \le 10^3$			$10^3 \le N \le 10^5$		
ε	a	b	c	a	b	c
0.05	0.0034	1.1701	0.1253	0.0120	0.5272	12.18
0.10	0.0132	1.0400	0.6570	0.0264	0.1943	14.82
0.15	0.0264	0.8842	1.2222	0.0403	0.0922	13.05
0.20	0.0412	0.7373	1.7073	0.0542	0.0480	11.12
0.25	0.0564	0.6109	2.0842	0.0679	0.0331	9.398
$d_{N,0.95}^{(\varepsilon)}$						
	$50 \le N \le 10^3$			$10^3 \le N \le 10^5$		
ε	a	b	c	a	b	c
0.05	0.0040	1.2794	0.1662	0.0130	0.6121	12.63
0.10	0.0151	1.1325	0.7568	0.0286	0.2607	15.38
0.15	0.0291	0.9790	1.2903	0.0437	0.1328	14.28
0.20	0.0452	0.8181	1.8199	0.0588	0.0742	12.41
0.25	0.0610	0.7041	2.1113	0.0737	0.0497	10.73
$d_{N,0.99}^{(\varepsilon)}$						
	$50 \le N \le 10^3$			$10^3 \le N \le 10^5$		
ε	a	b	c	a	b	c
0.05	0.0048	1.5143	0.1805	0.0145	0.8291	12.52
0.10	0.0180	1.3483	0.7872	0.0319	0.4373	16.38
0.15	0.0335	1.1962	1.3040	0.0488	0.2810	15.84
0.20	0.0520	1.0014	2.0509	0.0657	0.2074	13.80
0.25	0.0688	0.9084	2.1364	0.0828	0.0978	14.68
$d_{N,0.999}^{(\varepsilon)}$						
	$50 \le N \le 10^3$			$10^3 \le N \le 10^5$		
ε	a	b	c	a	b	c
0.05	0.0092	1.685	0.9586	0.0158	1.1156	12.73
0.10	0.0189	1.697	0.3014	0.0342	0.7717	14.63
0.15	0.0365	1.534	0.7803	0.0528	0.5464	16.07
0.20	0.0547	1.370	1.4935	0.0712	0.4710	13.99
0.25	0.0770	1.161	2.2538	0.0902	0.2273	18.90

Table 4: The values of the best-fitting parameters a, b, and c in Eq. (7) as a function of ε, for both $50 \le N \le 10^3$ and $10^3 \le N \le 10^5$.

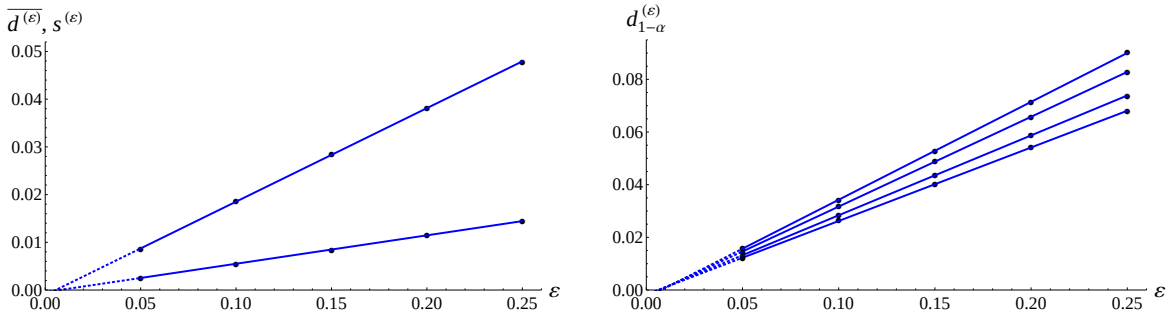


Figure 4: *Left panel.* The limiting values (8) of the mean (upper circles) and standard deviation (lower circles) of the Euclidean distance statistic (6) as a function of ε, together with their corresponding regression lines (blue continuous lines), Eqs. (9) and (10), respectively. *Right panel.* Limiting test-values (8) for the Euclidean distance statistic (6) as a function of ε, together with their corresponding regression lines (blue continuous lines), Eqs. (11)-(14). From bottom to top: $\alpha = 0.1, 0.05, 0.01$ and 0.001.

The mean, standard deviation, and test values of $d_N^{(\varepsilon)}$ grow as $\sqrt{N}$ for large N. Accordingly, the statistic $d_N^{(\varepsilon)}/\sqrt{N}$, whose sample values are by definition independent on N, is asymptotically independent on the sample size making its use a reliable tool for testing Benford's law in samples with large size. This statistic, then, solves the problem of the enormous power for large N of existing statistical tests.

Let us now re-consider the Cho and Gaines data discussed above. These data are divided in "homogeneous" groups, in the sense that the underlying statistical process for each group is the same, namely an in-kind contribution cataloged by the U.S.A. Federal Election Commission. As already noticed, these data sets do not comply to Benford's law (see Table 3). However, we can test the hypothesis that the data comply to a ε-Benford's distribution. We proceed as follows. We fix the value of ε by finding a group conforming to ε-Benford to a large significance level, let's say, bigger than 0.1. This is the case of the first election cycle (1994) for which $d_N^* = 5.3$ and $d_N^{(0.2)} = 5.5$. In other words, the data relative to the 1994 election conform to a 0.20-Benford's distribution at a significance level of 0.1. Accordingly, we can test the null (conformity to a 0.20-Benford's distribution) for the other cycles. The results are shown in the last column of Table 3. While the 1998 in-kind contribution conforms to a significance level of 0.05 and the 2000 one cannot be rejected at a significance level of 0.001, the years 1996, 2002, and 2004 present anomalies: the conformance of the data to a 0.20-Benford's distribution can be rejected at a significance level grater than 0.001. It is interesting to observe that, based on the normalized Euclidean distance statistic, Cho and Gaines (2007) found anomalies only for the years 2002 and 2004. The use of ε-Benford's distribution can be then used not only to identify but also to quantify possible anomalies in homogeneous sets of data.

The use of ε-Benford distributions can be also extended to the case of "non-homogeneous" data sets when the number of counts is large and/or the range of the data is small. An interesting application of Benford's law to physical and mathematical data sets was discussed by Sambridge et al. (2010). Their data are shown in Table 5. More than 50% of the data sets considered in their study has large number counts. In particular, groups S1, S3, S4, S6, S7, S8, S9, and S11 have values of N larger that 10^3 (groups S8 and S11 also have a small range of data, of order of 10^2). Not surprisingly, with the exception of S7 which well conforms to Benford's law, these large-number-count sets have huge values of d_N^* and χ^2 making the goodness-of-fit unreliable. However, they all comply to ε-Benford's

Set	Title	N	Range	d^*	d_N^*	p	χ^2	$p(\chi^2)$	ε
S1	Geomagnetic Field	36512	10^{10}	0.015	2.936	0.	49.90	0.	0.10
S2	Geomagnetic reversals	93	10^3	0.056	0.562	0.91(1)	3.608	0.8907	—
S3	Seis. wavespeeds below SW-Pacific	423776	10^6	0.009	6.041	0.	363.7	0.	0.04
S4	Earth's gravity	25917	10^9	0.035	5.829	0.	188.7	0.	0.15
S5	Exoplanet mass	401	10^5	0.056	1.163	0.13(0)	10.57	0.2274	—
S6	Pulsars rotation freq.	1861	10^4	0.060	2.699	0.	55.03	0.	0.25
S7	Fermi space tel. γ-ray source fluxes	1451	10^5	0.020	0.797	0.59(5)	12.56	0.1280	—
S8	Earthquake depths	248915	10^2	0.027	14.14	0.	1723	0.	0.11
S9	S-A seismogram	24000	10^5	0.030	4.840	0.	191.1	0.	0.15
S10	Green house gas em. by country	184	10^4	0.030	0.421	0.98(3)	2.049	0.9795	—
S11	Glob. Temp. anom. 1880-2008	1527	10^2	0.045	1.828	0.00(1)	34.61	0.	0.15
S12	Fund. Phys. constants	326	10^4	0.058	1.088	0.19(1)	9.615	0.2931	—
S13	Global Infectious disease cases	987	10^6	0.046	1.419	0.02(7)	15.00	0.0592	—
S14	Geometric series	1000	10^{21}	0.007	0.229	0.999(7)	0.417	0.9999	—
S15	Fibbonacci sequence	1000	10^{14}	0.003	0.091	1.	0.122	1.	—

Table 5: The normalized Euclidean distance d^*, the Euclidean distance statistic d_N^* with its corresponding p value, the χ^2 score with its corresponding p value, $p(\chi^2)$, of the first-digit distribution for various physical and mathematical data sets discussed by Sambridge et al. (2010). Also indicated are the total number of counts for each group, N, and the dynamic range of the data (max/min). The last column shows the value of ε such that the first-digit distribution of the counts conform to a ε-Benford's law at a significance level of $\alpha = 0.1$.

law at a significance level of $\alpha = 0.1$, with values of ε ranging from 0.04 to 0.25 (see the last column in Table 5).

4 Conclusions

Benford's law on the distribution of the first digits of numerical data sets has been observed to arise in multifarious classes of data, from natural sciences to finance. Compliance to Benford's law can be tested by using standard test statistics, such as the Pearson χ^2 statistic, the "Goodman's rule of thumb", and/or the recently introduced Euclidean distance statistic.

The main results of our analysis are as follows.

(i) For small and/or large number of data points, $N \leq 1000$, the use of p values of the χ^2 statistic for testing Benford's law is not completely reliable. This is because the χ^2 test, although being a very powerful tool in assessing the goodness-of-fit to any continuous distribution, it is generally conservative for testing discrete ones, like Benford's law. The χ^2 statistic should be then used only for "qualitative" analyses. For quantitative analyses, the Euclidean distance should be used, since the test based on this statistic has been explicitly constructed for testing Benford's law. The Goodman's rule of thumb, instead, should be always avoided when checking for the compliance to Benford's law. Its statistical groundlessness generates results in disagreement with both the χ^2 and the Euclidean distance tests.

(ii) We have discussed some limitations of statistical tests in assessing the goodness-of-fit to Benford's law for very large sample sizes ($N > 1000$) and/or very small ranges of data, and then proposed a possible solution to overcome such limitations. The solution comes from the observation that Benford's law is not in general a limiting distribution nor a fundamental law of numbers and then

real distributions are often ε-Benford, in the sense that they deviate from Benford's law at a relative level of ε. Even a tiny deviation, however, may result in huge values of standard test statistics for large N, making any attempt to quantify the goodness-of-fit unfeasible. We have then considered a new statistic, the Euclidean distance statistic $d_N^{(\varepsilon)}$ for a ε-Benford distribution, and computed appropriate test values. The statistic $d_N^{(\varepsilon)}/\sqrt{N}$, whose sample values are by definition independent on N, is asymptotically independent on the sample size, making it a natural candidate for testing Benford's law in samples with very large size.

References

Benford, Frank (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572.

Campanelli, Leonardo (2022). On the euclidean distance statistic of Benford's law. *Communications in Statistics-Theory and Methods*, 1–24.

Campanelli, Leonardo (2023). Breaking Benford's law: A statistical analysis of Covid-19 data using the euclidean distance statistic. *Statistics in Transition New Series*. In press.

Cho, Wendy K T and Brian J Gaines (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American Statistician 61*(3), 218–223.

Goodman, William (2016). The promises and pitfalls of Benford's law. *Significance 13*(3), 38–41.

Hill, Theodore P (1995a). Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society 123*(3), 887–895.

Hill, Theodore P (1995b). The significant-digit phenomenon. *The American Mathematical Monthly 102*(4), 322–327.

Hill, Theodore P (1995c). A statistical derivation of the significant-digit law. *Statistical Science*, 354–363.

Lide, David R (2002). Handbook of chemistry and physics, eighty. CRC Press LLC, Internet Version of the 84th Edition.

Miller, S. J. (Ed.) (2015). *Benford's Law: Theory and Applications*. Princeton University Press, Princeton.

Morrow, John (2014). Benford's law, families of distributions and a test basis.

Nigrini, Mark J (1996). A taxpayer compliance application of Benford's law. *The Journal of the American Taxation Association 18*(1), 72.

Noether, Gottfried E (1963). Note on the Kolmogorov statistic in the discrete case. *Metrika 7*(1), 115–116.

Sambridge, Malcolm and Andrew Jackson (2020). National COVID numbers–Benford's law looks for errors. *Nature 581*(7809), 384–385.

Sambridge, Malcolm, Hrvoje Tkalčić, and A Jackson (2010). Benford's law in the natural sciences. *Geophysical Research Letters 37*(22).

SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 55–71
doi:https://doi.org/10.37830/SJS.2022.1.04

REGULAR ARTICLE

The gamma flexible Weibull distribution: Properties and Applications

Alexsandro A. Ferreira[1], Gauss M. Cordeiro[2]

[1] Federal University of Pernambuco, alex.ferreira.aaf@gmail.com
[2] Federal University of Pernambuco, gauss.cordeiro@ufpe.br

Received: March 27, 2022, Accepted: December 27, 2022.

Abstract: A new gamma flexible Weibull distribution is introduced, which presents a bathtub-shaped hazard rate, and some of its properties are obtained. The parameters are estimated via maximum likelihood, and a simulation study is performed to examine the consistency of the estimates. The utility of the proposed model is shown using three real applications.

Keywords: Bathtub, Bimodal, COVID-19, Maximum likelihood, Moment, Quantile function

MSC: 33B05, 60E05, 62P99, 65C05

1 Introduction

Various phenomena that occur in the real world can be explained by statistical distributions. For a long time, many of the common distributions (Weibull, gamma, Burr XII, Gumbel) were sufficient for this purpose. However, with computer science development, more flexible distributions have become mandatory. One way to generate new families of distribution is through techniques to generalize existing ones. The main characteristic of these generalizations is the addition of more parameters to their baseline distributions, thus increasing their flexibility.

The Weibull distribution is widely used in many fields, but it is not suitable for bathtub-shaped or unimodal hazard rates. Thus, several models have been developed to extend this distribution and increase the modeling ability, such as those in (Mudholkar and Srivastava, 1993), (Xie and Lai, 1996), (Xie et al., 2002), (Lai et al., 2003), (Famoye et al., 2005), and (Cordeiro et al., 2010), among others.

Of the various modifications made to the Weibull distribution, the one of interest in this article is the flexible Weibull (FW) distribution (Bebbington et al., 2007) with shape parameters $\alpha, \beta > 0$, cumulative distribution function (cdf)

$$G(x;\alpha,\beta) = 1 - \exp\left(-\mathrm{e}^{\alpha x - \frac{\beta}{x}}\right), \quad x > 0,$$

Published by the Spanish National Statistical Institute

and probability density function (pdf)

$$g(x;\alpha,\beta) = \left(\alpha + \frac{\beta}{x^2}\right) \mathrm{e}^{\alpha x - \frac{\beta}{x}} \exp\left(-\mathrm{e}^{\alpha x - \frac{\beta}{x}}\right).$$

For $\beta = 0$ and $\alpha = \log(\lambda)$, the FW model reduces to the exponential, and then it can be regarded as a generalization of the Weibull (Bebbington et al., 2007).

There are several extensions of the FW distribution such as those reported by (El-Gohary et al., 2015), (El-Desouky et al., 2016), (Mustafa et al., 2016), (El-Damcese et al., 2016), (El-Desouky et al., 2017), and (Ahmad and Iqbal, 2017).

Zografos and Balakrishnan (2009) and Ristić and Balakrishnan (2012) defined the cdf of the gamma-G class for any parent cdf $G(x) = G(x;\boldsymbol{\theta})$ with parameter vector $\boldsymbol{\theta}$ of dimension p, by (for $x \in \mathbb{R}$)

$$F(x) = F(x;a,\boldsymbol{\theta}) = \frac{\gamma(a, -\log[1-G(x)])}{\Gamma(a)} = \frac{1}{\Gamma(a)} \int_0^{-\log[1-G(x)]} t^{a-1}\mathrm{e}^{-t}\mathrm{d}t, \tag{1}$$

where $a > 0$ is a shape parameter, and $\Gamma(\cdot)$ is the gamma function. For $a = 1$, Equation (1) reduces to the parent G cdf.

Recently, the gamma-G family has received considerable attention in works by (Nadarajah et al., 2015), (Alzaatreh et al., 2014), (Nadarajah et al., 2015), (Cordeiro et al., 2016), (Bourguignon and Cordeiro, 2016), (Iriarte et al., 2017), (Guerra et al., 2017), and (David et al., 2021), among others.

The article unfolds as follows: Section 2 defines the gamma-flexible Weibull (GFW) distribution and a linear representation for its density. The moments and generating function are reported in Section 3. Section 4 estimates the parameters by the maximum likelihood method and conducts a simulation study. Three real data sets are analyzed in Section 5 to show the utility of the new model. Finally, we draw some conclusions in Section 6.

2 The GFW model and its linear representation

A random variable X follows the GFW distribution, say $X \sim \mathrm{GFW}(a,\alpha,\beta)$, if its cdf and pdf (omitting parameters in the functions) are

$$F(x) = \frac{\gamma\left[a, \exp\left(\alpha x - \frac{\beta}{x}\right)\right]}{\Gamma(a)} = \frac{1}{\Gamma(a)} \int_0^{\exp\left(\alpha x - \frac{\beta}{x}\right)} t^{a-1}\mathrm{e}^{-t}\mathrm{d}t, \quad t > 0, \tag{2}$$

and

$$f(x) = \frac{\left(\alpha + \frac{\beta}{x^2}\right)\left(\mathrm{e}^{\alpha x - \frac{\beta}{x}}\right)^a \exp\left(-\mathrm{e}^{\alpha x - \frac{\beta}{x}}\right)}{\Gamma(a)}, \tag{3}$$

respectively.

The FW distribution was introduced in engineering, but it can be used in several fields. So, the GFW distribution can also be adopted in a similar manner.

The hazard rate function (hrf) of X follows from the last two expressions.

The GFW is identical to the FW distribution when $a = 1$. The calculations in all sections were done using R software (R Core Team, 2020).

Figure 1 displays some plots of the density of X, which can be symmetric, right-symmetric, left-symmetric, or bimodal. Plots of the hrf of X are reported in Figure 2, which has increasing, decreasing, bathtub, and unimodal shapes.

(a) (b)

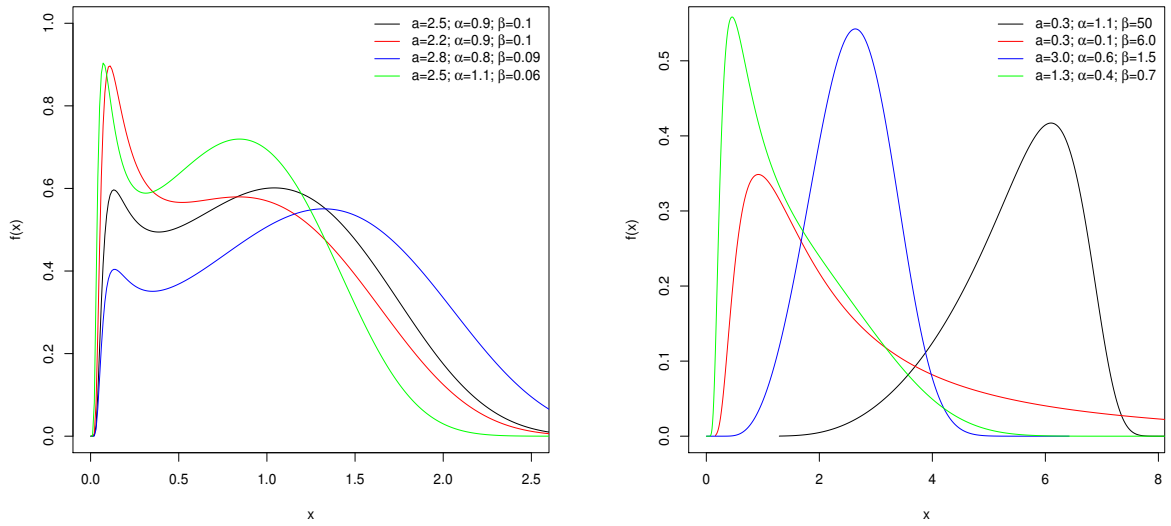


Figure 1: Plots of the density of X.

(a) (b)

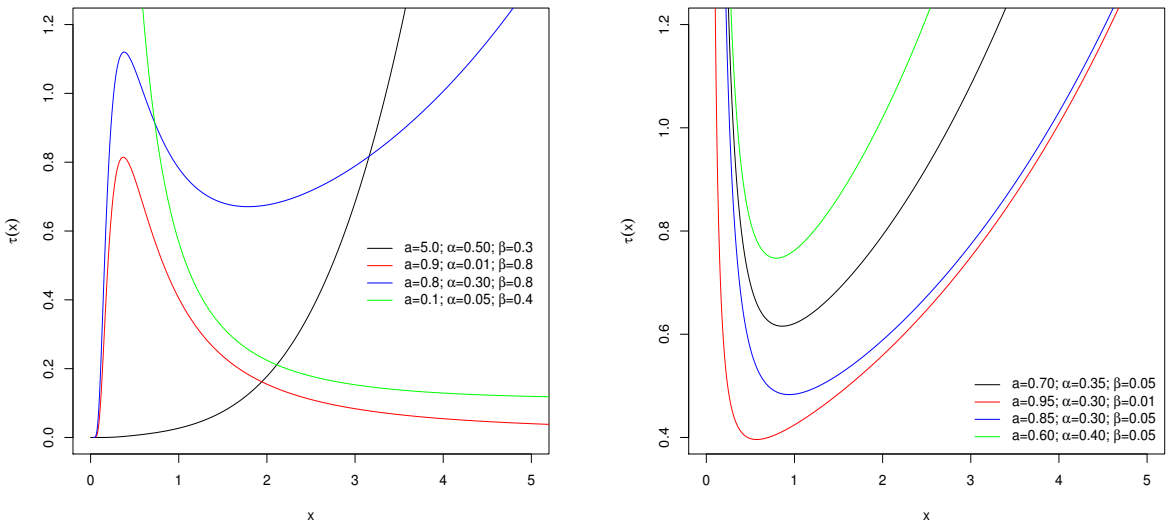


Figure 2: Plots of the hrf of X.

A simple motivation for the GFW distribution follows from Zografos and Balakrishnan (2009), where the GFW density can be approximated by the upper record value density from a sequence of independent and identically distributed FW random variables. Further, we highlight the utility of the proposed distribution in medical data analysis. In fact, the GFW distribution can be selected

as the best model, especially in modeling unimodal and bimodal data of COVID-19 and cancer as illustrated in Section 5.

Following the concept of exponentiated distributions (Cordeiro et al., 2013), the exponentiated FW ("expFW") cdf with power parameter δ, say EFW(α, β, δ) (for $x > 0$), is

$$H_\delta(x; \alpha, \beta) = \left[1 - \exp\left(-\mathrm{e}^{\alpha x - \frac{\beta}{x}}\right)\right]^\delta$$

and the corresponding pdf reduces to

$$h_\delta(x; \alpha, \beta) = \delta \left(\alpha + \frac{\beta}{x^2}\right) \mathrm{e}^{\alpha x - \frac{\beta}{x}} \exp\left(-\mathrm{e}^{\alpha x - \frac{\beta}{x}}\right) \left[1 - \exp\left(-\mathrm{e}^{\alpha x - \frac{\beta}{x}}\right)\right]^{\delta - 1} .$$

From Proposition 2 of Castellares and Lemonte (2015), we can write

$$[-\ln(1 - v)]^c = v^c \sum_{m=0}^{\infty} \rho_m(c)\, v^m, \tag{4}$$

where $c \in \mathbb{R}$, $|v| < 1$, $\rho_0(c) = 1$, $\rho_m(c) = c\,\psi_{m-1}(m + c - 1)$ for $m \geq 1$, and $\psi_m(\cdot)$ are Stirling polynomials, namely

$$\begin{aligned}\psi_{n-1}(w) = &\frac{(-1)^{n-1}}{(n+1)!}\left[T_n^{n-1} - \frac{w+2}{n+2} T_n^{n-2} + \frac{(w+2)(w+3)}{(n+2)(n+3)} T_n^{n-3} - \cdots \right. \\ &\left. + (-1)^{n-1} \frac{(w+2)(w+3)\cdots(w+n)}{n+2)(n+3)\cdots(2n)} T_n^0\right],\end{aligned} \tag{5}$$

where $T_{n+1}^m = (2n + 1 - m)\, T_n^m + (n - m + 1)\, T_n^{m-1}$ are positive integers, $T_0^0 = 1$, $T_{n+1}^0 = 1 \times 3 \times 5 \times \cdots \times (2n + 1)$, and $T_{n+1}^n = 1$.

From Equation (4), we can rewrite Equation (3) as (Castellares and Lemonte, 2015)

$$f(x; a, \alpha, \beta) = \sum_{m=0}^{\infty} p_m\, h_{m+a}(x; \alpha, \beta)\,, \tag{6}$$

where $\varphi_0(a) = \Gamma(a)^{-1}$, $p_m = p_m(a) = \varphi_m(a)/(m + a)$, $\varphi_m(a) = \Gamma(a)^{-1}\rho_m(a - 1) = (a - 1)\,\Gamma(a)^{-1}\psi_{m-1}(m + a - 2)$ (for $m \geq 1$) can be determined from (5), and $h_{m+a}(x; \alpha, \beta)$ denotes the EFW density with power parameter $m + a$.

Equation (6) reveals that the GFW density is a linear combination of EFW densities. So, its properties can follow from those of the EFW distribution.

3 Moments and generating function

We calculate numerically in Table 1 the first four moments, standard deviation (SD), skewness (SK) and kurtosis (KR) of X varying a and β, with $\alpha = 0.04$. The moments increase and the skewness and kurtosis decrease if β increases for a fixed. Note that the same happen when a increases for β fixed.

If $Y_{m+a} \sim$ EFW$(m + a, \alpha, \beta)$, we write from Equation (6)

$$\mu_r' = \mathbb{E}(X^r) = \sum_{m=0}^{\infty} p_m\, \mathbb{E}\left(Y_{m+a}^r\right) . \tag{7}$$

	$a=0.1,\beta=0.5$	$a=0.1,\beta=1.0$	$a=0.1,\beta=1.5$	$a=0.1,\beta=2.0$
μ'_1	0.458	0.6135	0.758	0.897
μ'_2	5.786	6.4989	7.263	8.0748
μ'_3	132.535	143.993	156.166	169.053
μ'_4	3639.751	3919.821	4215.782	4528.010
SD	2.361	2.474	2.586	2.696
SK	9.474	8.745	8.123	7.589
KR	109.462	95.521	84.187	74.898
	$a=0.5,\beta=0.5$	$a=0.5,\beta=1.0$	$a=0.5,\beta=1.5$	$a=0.5,\beta=2.0$
μ'_1	2.644	3.126	3.557	3.956
μ'_2	44.857	49.078	53.432	57.899
μ'_3	1108.423	1189.969	1275.122	1363.795
μ'_4	32176.930	34366.120	36656.030	39047.790
SD	6.153	6.269	6.385	6.500
SK	3.889	3.209	3.052	2.914
KR	15.476	14.290	13.283	12.420
	$a=1.5,\beta=0.5$	$a=1.5,\beta=1.0$	$a=1.5,\beta=1.5$	$a=1.5,\beta=2.0$
μ'_1	10.936	11.709	12.994	13.561
μ'_2	257.408	271.386	298.996	312.685
μ'_3	7526.057	7908.050	8692.924	9095.371
μ'_4	248547.900	261244.500	287697.100	301449.900
SD	11.738	11.588	11.408	11.347
SK	1.049	1.019	0.960	0.932
KR	3.218	3.200	3.143	3.113

Table 1: Numerical results for the GFW model.

Further, the rth moment of the EFW distribution is

$$\mathbb{E}(Y^r_{m+a}) = (m+a)\int_0^\infty x^r \left(\alpha+\frac{\beta}{x^2}\right) \mathrm{e}^{\alpha x-\frac{\beta}{x}} \exp\left(-\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right)\left[1-\exp\left(-\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right)\right]^{m+a-1},$$

where $\left[1-\exp\left(-\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right)\right]^{m+a-1}$ can be written as

$$\left[1-\exp\left(-\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right)\right]^{m+a-1} = \sum_{j=0}^{\infty} \frac{(-1)^j\,\Gamma(m+a)}{j!\,\Gamma(m+a-j)} \exp\left(-j\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right),$$

and then

$$\begin{aligned}\mathbb{E}(Y^r_{m+a}) = &\sum_{j=0}^{\infty} \frac{(-1)^j\,\Gamma(m+a+1)}{j!\,\Gamma(m+a-j)} \int_0^\infty x^r \left(\alpha+\frac{\beta}{x^2}\right) \mathrm{e}^{\alpha x-\frac{\beta}{x}} \\ &\times \exp\left[-(j+1)\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right] \mathrm{d}x\,.\end{aligned}$$

By using power series for $\exp\left[-(j+1)\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right]$ and $\mathrm{e}^{2(k+1)\alpha x}$ gives

$$\mathbb{E}(Y_{m+a}^r) = \sum_{j,k,i=0}^{\infty} \frac{(-1)^{j+k}\,(j+1)^k\,2^i\,(k+1)^i\,\Gamma(m+a+1)\alpha^i}{j!\,k!\,i!\,\Gamma(m+a-j)} \times \int_0^{\infty} x^{r+i}\left(\alpha+\frac{\beta}{x^2}\right)\mathrm{e}^{-(k+1)\alpha x-\frac{(k+1)\beta}{x}}\mathrm{d}x. \tag{8}$$

Based on the result $(\mathbf{3.471}\ 9)$ in Gradshteyn and Ryzhik (2007), we obtain

$$\mathbb{E}(Y_{m+a}^r) = \sum_{j,k,i=0}^{\infty} \frac{(-1)^{j+k}\,(j+1)^k\,2^i\,(k+1)^i\,\Gamma(m+a+1)\alpha^i}{j!\,k!\,i!\,\Gamma(m+a-j)} \times \left[2\alpha\left(\frac{\beta}{\alpha}\right)^{\frac{\nu+1}{2}} K_{\nu+1}\left(2(k+1)\sqrt{\alpha\beta}\right) + 2\beta\left(\frac{\beta}{\alpha}\right)^{\frac{\nu-1}{2}} K_{\nu-1}\left(2(k+1)\sqrt{\alpha\beta}\right)\right], \tag{9}$$

where

$$\nu = r+i,\ K_\nu(z) = \frac{\pi\csc(\pi\nu)}{2}\left[I_{-\nu}(z)-I_\nu(z)\right],\ \text{and}\ I_\nu(z) = \sum_{\ell=0}^{\infty}\frac{1}{\Gamma(\ell+\nu+1)\ell!}\left(\frac{z}{2}\right)^{2\ell+\nu}$$

are the modified Bessel functions of the second and first kind, respectively (for $\nu \notin \mathbb{Z}$).

Substituting (9) into (7) gives the rth moment of the GFW distribution.

In a similar manner, the rth incomplete moment of X, say $m_r(s) = \int_0^s x^r\,f(x)\,\mathrm{d}x$, follows as

$$m_r(s) = \sum_{m,j,k,i=0}^{\infty} \frac{(-1)^{j+k}\,(j+1)^k\,2^i\,(k+1)^i\,p_m\,\Gamma(m+a+1)\alpha^i}{j!\,k!\,i!\,\Gamma(m+a-j)} \times \int_0^s x^{r+i}\left(\alpha+\frac{\beta}{x^2}\right)\mathrm{e}^{-(k+1)\alpha x-\frac{(k+1)\beta}{x}}\mathrm{d}x\,.$$

From Theorem 2 of Chaudhry and Zubair (1994), we obtain (for $r \geq 1$)

$$m_r(s) = \sum_{m,j,k,i=0}^{\infty} \frac{(-1)^{j+k}\,(j+1)^k\,2^i\,p_m\,\Gamma(m+a+1)}{j!\,k!\,i!\,(k+1)^r\,\Gamma(m+a-j)\,\alpha^r} \times \left\{\frac{\gamma\left[(k+1)\alpha s;(r+i+1),(k+1)^2\alpha\beta\right]}{(k+1)} + (k+1)\gamma\left[(k+1)\alpha s;(r+i-1),(k+1)^2\alpha\beta\right]\alpha\beta\right\},$$

where $\gamma(x;a,b) = \int_0^x t^{a-1}\,\mathrm{e}^{-t-b/t}\mathrm{d}t$ is the generalized lower incomplete gamma function.

The generating function (gf) of X can be written from (6) as

$$M(t) = \sum_{m=0}^{\infty} p_m\,M_{m+a}(t)\,, \tag{10}$$

where $M_{m+a}(t)$ is the gf of Y_{m+a}. The gf of the EFW distribution is

$$M_{m+a}(t) = (m+a)\int_0^\infty \mathrm{e}^{tx}\left(\alpha+\frac{\beta}{x^2}\right)\mathrm{e}^{\alpha x-\frac{\beta}{x}}\exp\left(-\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right)\left[1-\exp\left(-\mathrm{e}^{\alpha x-\frac{\beta}{x}}\right)\right]^{m+a-1}.$$

Following a similar algebra as for Equation (8) and again the result $(\mathbf{3.471}\,9)$ (Gradshteyn and Ryzhik, 2007), we obtain (for $t<\alpha$)

$$\begin{aligned} M_{m+a}(t) = \sum_{j,k,i=0}^{\infty} & \frac{(-1)^{j+k}\,(j+1)^k\,2^i\,(k+1)^i\,\alpha^i\,\Gamma(m+a+1)}{j!\,k!\,i!\,\Gamma(m+a-j)} \\ & \times\left\{2\alpha\left[\frac{(k+1)\beta}{(k+1)\alpha-t}\right]^{\frac{i+1}{2}}K_{i+1}\left(2\sqrt{[(k+1)\alpha-t]\,(k+1)\,\beta}\right)\right. \\ & \left.+\,2\beta\left[\frac{(k+1)\beta)}{(k+1)\alpha-t}\right]^{\frac{i-1}{2}}K_{i-1}\left(2\sqrt{[(k+1)\alpha-t](k+1)\beta}\right)\right\}. \end{aligned} \tag{11}$$

Substituting Equation (11) into (10) gives the gf of the GFW distribution.

The quantile function (qf) of the FW distribution is given by (Bebbington et al., 2007)

$$Q_{\mathrm{FW}}(u;\alpha,\beta) = \frac{1}{2\alpha}\left\{\log\left[-\log(1-u)\right]+\sqrt{\left\{\log[-\log(1-u)]\right\}^2+4\alpha\beta}\right\}.$$

By inverting (2) and using results in Nadarajah et al. (2015), the qf of X follows as (for $0<u<1$)

$$Q_{\mathrm{GFW}}(u;a,\alpha,\beta) = \frac{1}{2\alpha}\left\{\log\{Q^{-1}[a,(1-u)]\}+\sqrt{\left\{\log[Q^{-1}(a,1-u)]\right\}^2+4\alpha\beta}\right\}, \tag{12}$$

where $Q^{-1}(a,u)$ is the inverse function of $Q(a,x)=1-\gamma(a,x)/\Gamma(a)$.

Approximations for the skweness and kurtosis of X can be based on quantile measures from (12). Let $Q_{\mathrm{GFW}}(u)=Q_{\mathrm{GFW}}(u;a,\alpha,\beta)$. The Bowley's skewness (Kenney and Keeping, 1962) is

$$\mathcal{B}(a,\alpha,\beta) = \frac{Q_{\mathrm{GFW}}(3/4)+Q_{\mathrm{GFW}}(1/4)-2Q_{\mathrm{GFW}}(1/2)}{Q_{\mathrm{GFW}}(3/4)-Q_{\mathrm{GFW}}(1/4)},$$

whereas the Moors kurtosis (Moors, 1988) is

$$\mathcal{M}(a,\alpha,\beta) = \frac{Q_{\mathrm{GFW}}(7/8)-Q_{\mathrm{GFW}}(5/8)-Q_{\mathrm{GFW}}(3/8)+Q_{\mathrm{GFW}}(1/8)}{Q_{\mathrm{GFW}}(6/8)-Q_{\mathrm{GFW}}(2/8)}.$$

Plots of these quantities for some choices of α and β as functions of a are reported in Figure 3. Note that the skewness increases when a goes to one and decreases from this value. The kurtosis decreases rapidly for small values of a and stabilizes when a increases.

An application of (12) using the first incomplete moment $m_1(s)$ refers to the Bonferroni and Lorenz curves defined by (for a given probability π)

$$B(\pi) = \frac{m_1(q)}{\pi\mu_1'} \quad \text{and} \quad L(\pi) = \frac{m_1(q)}{\mu_1'},$$

respectively, where $q=Q_{\mathrm{GFW}}(\pi)$. Plots of these curves versus π for some choices of a (with $\alpha=0.01$ and $\beta=15$) are displayed in Figure 4.

(a) (b)

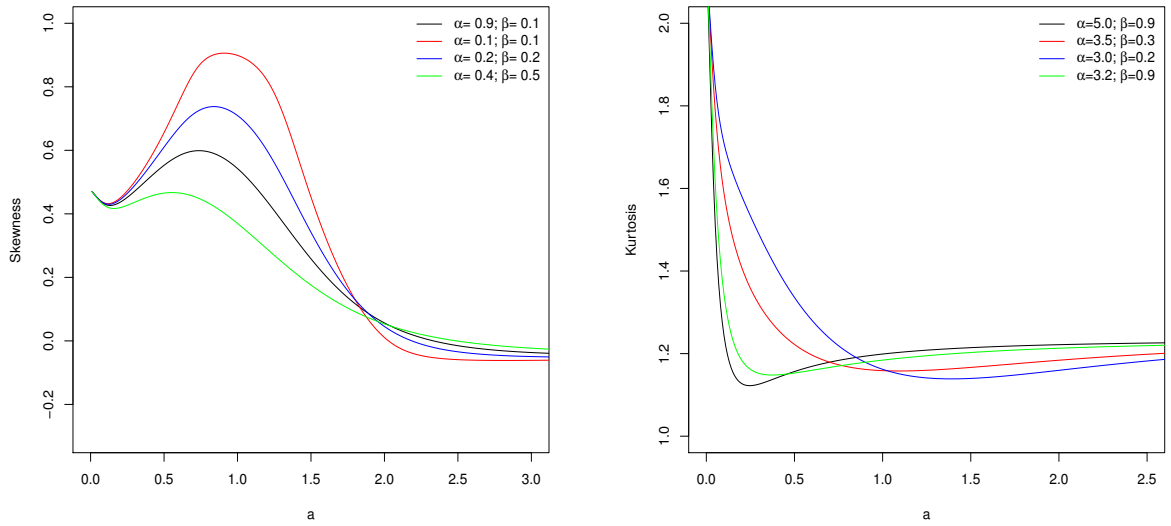


Figure 3: Skewness (a) and kurtosis (b) of X versus a.

(a) (b)

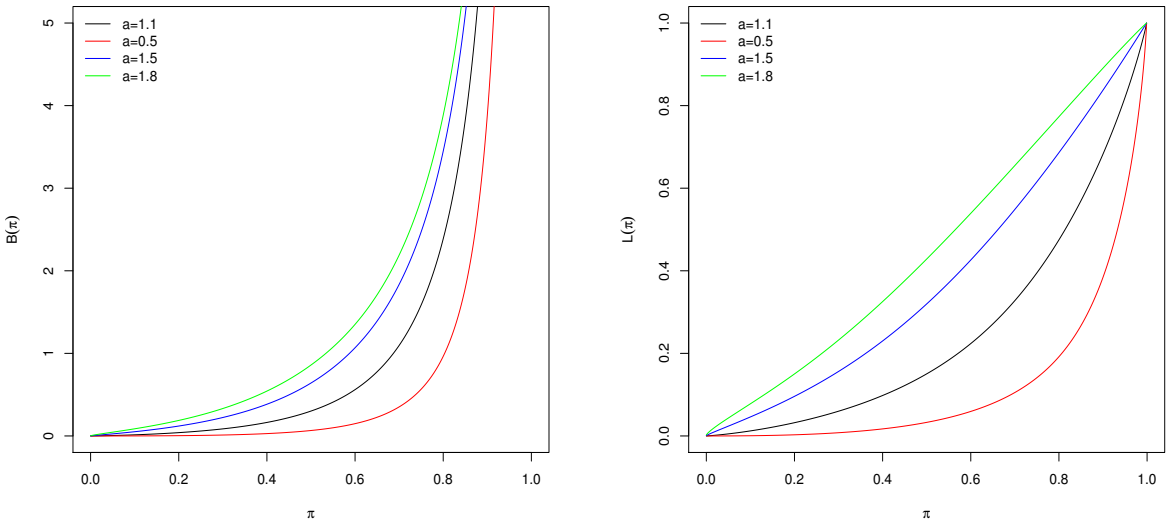


Figure 4: Bonferroni and Lorenz curves of X.

4 Estimation and Simulations

The log-likelihood function for $\boldsymbol{\theta} = (a, \alpha, \beta)^\top$ given the data set $x_1, \ldots, x_n$ from X is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log\left(\alpha + \frac{\beta}{x_i^2}\right) + a\sum_{i=1}^{n}\left(\alpha x_i - \frac{\beta}{x_i}\right) + \sum_{i=1}^{n}\left(-\mathrm{e}^{\alpha x_i - \frac{\beta}{x_i}}\right) - n\log[\Gamma(a)]. \tag{13}$$

The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$, say $\widehat{\boldsymbol{\theta}}$, can be found by maximizing Equation (13) numerically using scripts such as `optim` or `nlm` in `R`, `MaxBFGS` in `Ox`, and `PROC NLMIXED` in `SAS`.

We generate $1,000$ Monte Carlo replicates for the GFW model from Equation (12) with sample sizes $n = 50, 100, 300$, and 500 under three scenarios: $(a, \alpha, \beta) = (0.9, 2, 1.2)$ for scenario I; $(a, \alpha, \beta) = (0.5, 1.5, 3)$ for scenario II; and $(a, \alpha, \beta) = (1.5, 0.5, 0.8)$ for scenario III. We use the `optim` script of `R` to maximize (13). The averages, biases and mean square errors (MSEs) of the estimates are listed in Table 2. The averages tend to the true parameter values and the biases and MSEs converge to zero when n increases, which reveal that the MLEs are consistent.

		Scenario I			Scenario II			Scenario III		
n	parameter	Average	Bias	MSE	Average	Bias	MSE	Average	Bias	MSE
50	a	0.806	-0.094	0.261	0.630	0.130	0.246	1.480	-0.020	0.290
	α	2.815	0.815	4.787	2.099	0.599	2.782	0.514	0.014	0.004
	β	2.312	1.112	8.537	4.623	1.623	26.617	1.044	0.244	0.574
100	a	0.809	-0.090	0.210	0.588	0.088	0.166	1.484	-0.015	0.189
	α	2.430	0.430	0.815	1.826	0.326	0.836	0.509	0.009	0.002
	β	1.840	0.640	1.960	3.894	0.894	9.191	0.962	0.162	0.385
300	a	0.868	-0.031	0.137	0.556	0.056	0.085	1.497	-0.003	0.063
	α	2.156	0.156	0.117	1.583	0.083	0.084	0.502	0.002	0.001
	β	1.450	0.250	0.436	3.203	0.203	1.380	0.840	0.040	0.078
500	a	0.903	0.003	0.110	0.542	0.042	0.049	1.499	-0.001	0.035
	α	2.096	0.096	0.061	1.542	0.042	0.044	0.501	0.001	0.001
	β	1.343	0.143	0.265	3.083	0.083	0.784	0.819	0.019	0.029

Table 2: Simulation results for the GFW model.

5 Applications

We present three applications of the new model and compare it to other distributions: exponentiated Weibull (EW) (Mudholkar and Srivastava, 1993), modified Weibull (MW) (Lai et al., 2003), beta Weibull (BW) (Famoye et al., 2005), FW, Kumaraswamy Weibull (KwW) (Cordeiro et al., 2010), and Kumaraswamy Burr XII (KwBXII) (Paranaíba et al., 2013).

The best model is chosen based on Cramér-von Mises (W^*), Anderson-Darling (A^*), Akaike information criterion (AIC), consistent Akaike information criterion (CAIC), Bayesian information criterion (BIC), and Hannan-Quinn information criterion (HQIC). The MLEs, standard errors (SEs), and the statistics are found using the **AdequacyModel** script (Marinho et al., 2019) of `R` software.

5.1 Failure times

The failure times of 50 components (per 1000h) are (Murthy et al., 2004): 0.036, 0.058, 0.061, 0.074, 0.078, 0.086, 0.102, 0.103, 0.114, 0.116, 0.148, 0.183, 0.192, 0.254, 0.262, 0.379, 0.381, 0.538, 0.570, 0.574,0.590, 0.618, 0.645, 0.961, 1.228, 1.600, 2.006, 2.054, 2.804, 3.058, 3.076, 3.147, 3.625, 3.704, 3.931, 4.073, 4.393, 4.534, 4.893, 6.274, 6.816,7.896, 7.904, 8.022, 9.337, 10.940, 11.020, 13.880, 14.730, 15.080.

Table 3 gives some descriptive statistics. The mean is greater than the median, and then the data are right-skewed and leptokurtic.

Mean	Median	SD	Variance	Skewness	Kurtosis	Min.	Max.
3.3430	1.4140	4.1395	17.1350	1.4167	4.0846	0.0360	15.0800

Table 3: Descriptive statistics for failure times.

Table 4 reports the MLEs and their SEs (in parentheses). The MW, EW, KwW, and BW models have higher SEs related to their estimates, whereas the GFW, FW, and KwBXII models have accurate estimates.

Table 5 indicates that the GFW model gives the best fit to the data since it has the lowest statistics among all models. The generalized likelihood ratio (GLR) test (Vuong, 1989) is used to compare the GFW model against the FW ($GLR = 3.911$), MW ($GLR = 3.503$), EW ($GLR = 3.455$), KwW ($GLR = 3.372$), KwBXII ($GLR = 3.452$), and BW ($GLR = 3.450$) models for a significance level of 5%. These results show that the GFW distribution provides the best fit to the current data.

The plots of the estimated densities and estimated survival functions for the most competitive models are shown in Figure 5. The GFW distribution provides the closest approximations to the histogram and empirical survival function, which shows its utility for real-life applications.

Model	MLEs (SEs)				
GFW (a, α, β)	1.362	0.109	0.126		
	(0.190)	(0.013)	(0.033)		
FW (α, β)	0.099	0.183			
	(0.012)	(0.034)			
MW (α, λ, β)	0.496	0.034	0.562		
	(0.099)	(0.025)	(0.098)		
EW (α, λ, β)	0.290	0.770	0.785		
	(0.681)	(0.990)	(1.546)		
KwW (a, b, α, β)	0.118	2.368	4.551	0.046	
	(0.024)	(1.555)	(0.099)	(0.025)	
KwBXII (a, b, c, k, s)	0.121	2.199	4.381	1.193	21.015
	(0.019)	(0.477)	(0.147)	(0.217)	(0.125)
BW (a, b, α, β)	0.708	0.703	0.412	0.819	
	(1.392)	(1.460)	(1.575)	(1.057)	

Table 4: Findings from the fitted models to failure times.

5.2 COVID-19

The numbers of deaths from COVID-19 in 83 Illinois counties in the United States through December 2021 are: 169, 13, 28, 91, 13, 107, 4, 41, 31, 89, 46, 57, 108, 146, 35, 30, 32, 156, 38, 52, 21, 113, 73, 59, 130, 93, 10, 40, 101, 25, 36, 16, 15, 95, 90, 101, 21, 150, 57, 32, 34, 127, 184, 38, 69, 115, 78, 121, 165, 24, 53, 58, 72, 15, 38, 108, 85, 104, 39, 110, 82, 16, 58, 7, 15, 7, 107, 67, 74, 14, 8, 56, 29, 124, 52, 19, 72, 30, 66, 34, 196, 201, 98. See https://data.world/associatedpress/johns-hopkins-coronavirus-case-tracker.

Model	W^*	A^*	AIC	CAIC	BIC	HQIC
GFW	**0.042**	**0.257**	**193.850**	**194.372**	**199.586**	**196.035**
FW	0.079	0.414	195.846	196.101	199.670	197.302
MW	0.130	0.850	208.727	209.249	214.463	210.912
EW	0.150	0.946	210.713	211.234	216.449	212.897
KwW	0.131	0.861	210.706	211.595	218.355	213.619
BW	0.149	0.942	212.696	213.585	220.344	215.608
KwBXII	1.132	0.870	213.086	214.450	222.646	216.726

Table 5: Adequacy measures for the models fitted to failure times.

(a) (b)

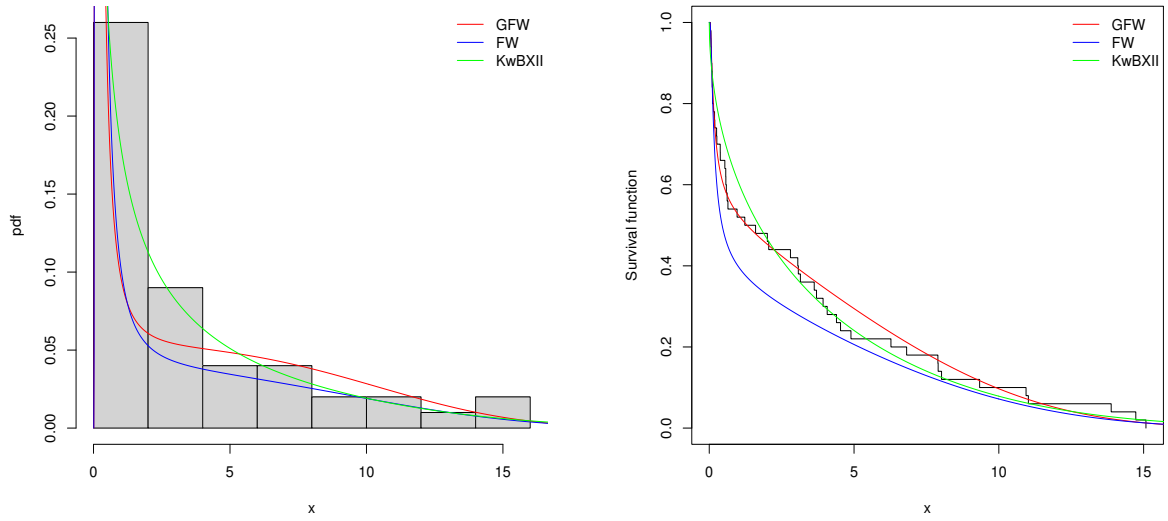


Figure 5: (a) Estimated densities of three models; (b) empirical and estimated survival functions of the models.

Table 6 shows some descriptive statistics for these data. The skewness is positive, and the kurtosis indicates mesokurtic distribution. The MLEs and their SEs (in parentheses) reported in Table 7 reveal

Mean	Median	SD	Variance	Skewness	Kurtosis	Min.	Max.
67.8670	57.0000	48.3850	2341.1	0.8198	2.9783	4	201

Table 6: Descriptive statistics for COVID-19 data.

that the GFW, FW, and KwW distributions have accurate estimates, and the other ones have high SEs relative to their estimates. The results in Table 8 indicate that the GFW model has the lowest values of the criteria, so it can be chosen as the best model. Additionally, the GLR test also reveals that the GFW model is better than the FW ($GLR = 3.383$), MW ($GLR = 3.961$), EW ($GLR = 2.925$), KwW

($GLR = 2.473$), KwBXII ($GLR = 3.502$), and BW ($GLR = 4.698$) models for a significance level of 5%.

Figure 6 reports plots of the estimated densities and estimated cumulative functions for the most adequate models. The fit of the new distribution is closer to the histogram and empirical cumulative function than those of the other distributions. So, these results support that the GFW distribution is better suited to the current data.

Model	MLEs (SEs)				
GFW (a, α, β)	1.702	0.010	14.005		
	(0.253)	(0.001)	(4.458)		
FW (α, β)	0.008	32.812			
	(0.001)	(4.290)			
MW (α, β, λ)	0.005	0.003	1.161		
	(0.002)	(0.002)	(0.116)		
EW (α, β, λ)	0.013	1.418	0.986		
	(0.005)	(0.503)	(0.574)		
KwW (a, b, α, β)	1.333	0.117	1.336	0.069	
	(0.083)	(0.055)	(0.039)	(0.019)	
KwBXII (a, b, c, k, s)	10.526	72.271	0.327	1.393	40.836
	(25.224)	(95.623)	(0.401)	(2.142)	(122.510)
BW (a, b, α, β)	3.697	3.665	0.011	0.615	
	(1.303)	(1.943)	(0.006)	(0.120)	

Table 7: Findings from the fitted models to COVID-19 data.

Model	W^*	A^*	AIC	CAIC	BIC	HQIC
GFW	**0.034**	**0.241**	**855.333**	**855.637**	**862.590**	**858.248**
FW	0.095	0.596	859.516	859.666	864.353	861.459
MW	0.057	0.348	858.767	859.071	866.024	861.682
EW	0.059	0.351	858.690	858.994	865.946	861.605
KwW	0.056	0.335	860.250	860.763	869.925	864.137
BW	0.088	0.544	863.876	864.389	873.551	867.763
KwBXII	0.083	0.508	864.870	865.649	876.964	869.728

Table 8: Adequacy measures for the models fitted to COVID-19 data.

5.3 Laryngeal cancer

The data set corresponds to the lifetime (in months) of 90 male patients with laryngeal cancer. The data are (Colosimo and Giolo, 2006): 0.6, 1.3, 2.4, 3.2, 3.3, 3.5, 3.5, 4.0, 4.0, 4.3, 5.3, 6.0, 6.4, 6.5, 7.4, 2.5, 3.2, 3.3, 4.5, 4.5, 5.5, 5.9, 5.9, 6.1, 6.2, 6.5, 6.7, 7.0, 7.4, 8.1, 8.1, 9.6, 10.7, 0.2, 1.8, 2.0, 3.6, 4.0, 6.2, 7.0, 2.2, 2.6, 3.3, 3.6, 4.3, 4.3, 5.0, 7.5, 7.6, 9.3, 0.3, 0.3, 0.5, 0.7, 0.8, 1.0, 1.3, 1.6, 1.8, 1.9, 1.9, 3.2, 3.5, 5.0, 6.3, 6.4, 7.8, 3.7, 4.5, 4.8, 4.8, 5.0, 5.1, 6.5, 8.0, 9.3, 10.1, 0.1, 0.3, 0.4, 0.8, 0.8, 1.0, 1.5, 2.0, 2.3, 3.6, 3.8, 2.9, 4.3.

Some descriptive statistics in Table 9 reveal that the data are right-skewed and platykurtic. For these data, we compare the GFW distribution with other models that also have the bimodal shape,

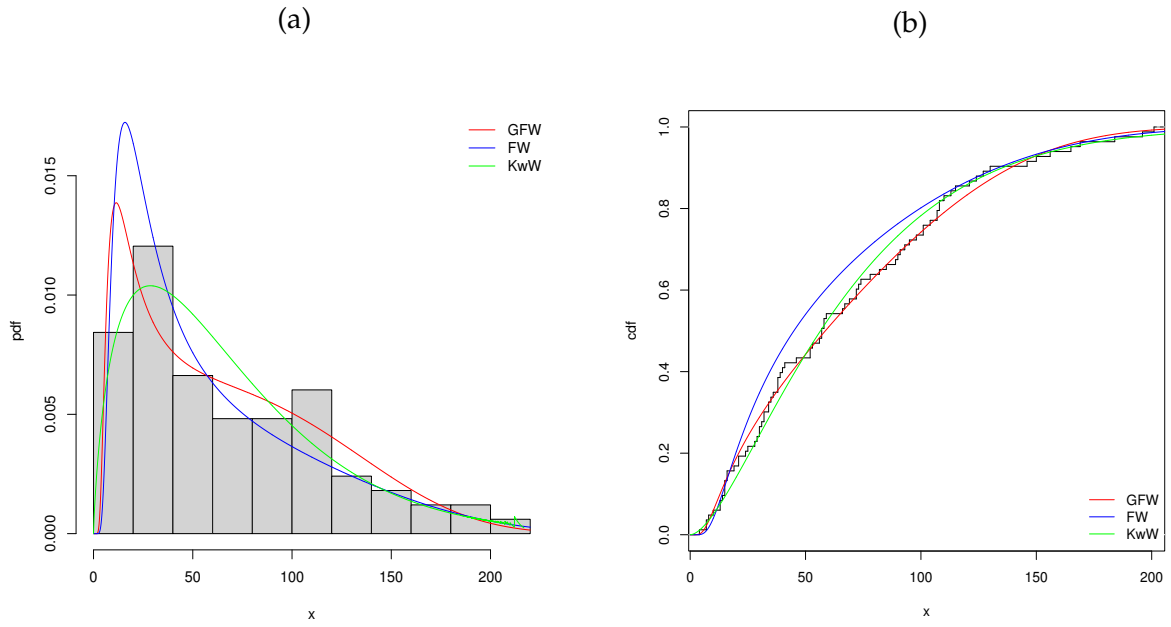


Figure 6: (a) Estimated densities of three models; (b) empirical and estimated cumulative functions of the models.

namely, the Odd log-logistic flexible Weibull (OLLFW) (Prataviera et al., 2018), extended Weibull log-logistic (EWLL) (Abouelmagd et al., 2019), Marshall-Olkin flexible Weibull (MOFW) (Mustafa et al., 2016), and FW.

Mean	Median	SD	Variance	Skewness	Kurtosis	Min.	Max.
4.197	4	2.612	6.901	0.343	2.367	0.1	10.700

Table 9: Descriptive statistics for laryngeal cancer data.

The MLEs and their corresponding SEs (in parentheses) in Table (10) show that the GFW, OLLFW, MOFW, and FW distributions have accurate estimates. The GFW distribution has the lowest values of the adequacy measures in Table (11) and can provide a better fit than the other distributions. The GLR test confirms that the GFW distribution fits the current data better than the OLLFW ($GLR = 4.657$), EWLL ($GLR = 3.741$), MOFW ($GLR = 11.556$), and FW ($GLR = 3.603$) distributions for a significance level of 5%. The plots in Figure 7 also support our claim.

6 Conclusions

We introduced a new versatile distribution called the gamma flexible Weibull and provided some of its properties. A simulation study demonstrated that the maximum likelihood estimates of the parameters are consistent. Three real applications showed that the new distribution is extremely competitive to other lifetime models for unimodal and bimodal medical data.

Model	MLEs (SEs)		
GFW (a, α, β)	2.527	0.201	0.197
	(0.223)	(0.012)	(0.070)
OLLFW (a, α, β)	0.359	0.288	2.869
	(0.053)	(0.023)	(0.106)
EWLL (λ, α, β)	0.072	0.925	21.689
	(0.459)	(0.457)	(135.256)
MOFW (a, α, β)	6.118	0.188	0.453
	(1.528)	(0.011)	(0.142)
FW (α, β)	0.142	1.286	
	(0.012)	(0.198)	

Table 10: Findings from the fitted models to laryngeal cancer data.

Model	W^*	A^*	AIC	CAIC	BIC	HQIC
GFW	**0.035**	**0.211**	**414.251**	**414.530**	**421.751**	**417.275**
OLLFW	0.310	1.744	440.411	440.690	447.910	443.435
EWLL	0.148	0.913	425.014	425.293	432.514	428.038
MOFW	0.048	0.278	415.136	415.415	422.635	418.160
FW	0.564	3.218	462.138	462.275	467.137	464.154

Table 11: Adequacy measures for the models fitted to laryngeal cancer data.

(a) (b)

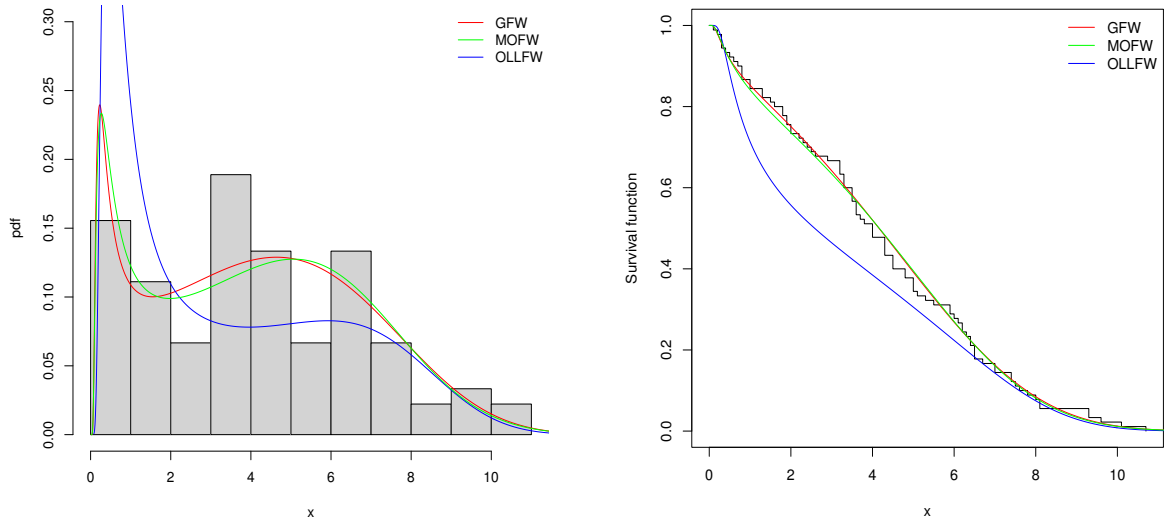


Figure 7: (a) Estimated densities of three models; (b) empirical and estimated survival functions of the models.

Acknowledgments

This work was supported by the Fundação de Amparo á Ciência e Tecnologia do Estado de Pernambuco (FACEPE) [IBPG-1448-1.02/20]. Thanks to the Associate Editor and the reviewers.

References

Abouelmagd, T.H.M., M.S. Hamed, J.A. Almamy, M.M. Ali, H.M. Yousof, M.C. Korkmaz, et al. (2019). Extended Weibull log-logistic distribution. *Journal of Nonlinear Sciences and Applications 12*, 523–534.

Ahmad, Z. and B. Iqbal (2017, 05). Generalized flexible Weibull extension distribution. *Circulation in Computer Science 2*, 68–75.

Alzaatreh, A., F. Famoye, and C. Lee (2014, 01). The gamma-normal distribution: Properties and applications. *Computational Statistics & Data Analysis 69*, 67–80.

Bebbington, M., C.-D. Lai, and R. Zitikis (2007). A flexible Weibull extension. *Reliability Engineering & System Safety 92*, 719–726.

Bourguignon, M. and G.M. Cordeiro (2016, 01). New results on the Ristić-Balakrishnan family of distributions. *Communication in Statistics- Theory and Methods 45*, 6969–6988.

Castellares, F. and A.J. Lemonte (2015). A new generalized Weibull distribution generated by gamma random variables. *Journal of the Egyptian Mathematical Society 23*, 382–390.

Chaudhry, M.A. and S.M. Zubair (1994). Generalized incomplete gamma functions with applications. *Journal of Computational and Applied Mathematics 55*, 99–123.

Colosimo, E.A. and S.R. Giolo (2006). *Análise de sobrevivência aplicada*. Editora Blucher.

Cordeiro, G.M., E.M.M. Ortega, and D.C.C. Cunha (2013, 01). The exponentiated generalized class of distributions. *Journal of Data Science 11*, 1–27.

Cordeiro, G.M., E.M.M. Ortega, and S. Nadarajah (2010). The Kumaraswamy Weibull distribution with application to failure data. *Journal of the Franklin Institute 347*, 1399–1429.

Cordeiro, G. M., M.D.C.S. Lima, A.H.M.A. Cysneiros, M.A.R. Pascoa, R.R. Pescim, and E.M.M. Ortega (2016). An extended Birnbaum-Saunders distribution: Theory, estimation, and applications. *Communications in Statistics - Theory and Methods 45*, 2268–2297.

David, L.R.R., G.M. Cordeiro, and M.D.C.S. Lima (2021, 04). The gamma-Chen distribution: a new family of distributions with applications. *Spanish journal of statistics 2*, 23–40.

El-Damcese, M.A., A. Mustafa, B.S. El-Desouky, and M.E. Mustafa (2016, 12). The Kumaraswamy flexible Weibull extension. *International Journal of Mathematics And its Applications 4*, 1–14.

El-Desouky, B.S., A. Mustafa, and S. Al garash (2016, 05). The exponential flexible Weibull extension distribution. *Open Journal of Modelling and Simulation 05*, 83–97.

El-Desouky, B. S., A. Mustafa, and S. Al-Garash (2017). The beta flexible Weibull distribution. *preprint arXiv: Statistics Theory*.

El-Gohary, A., A. El-Bassiouny, and M. Elmorshedy (2015, 07). Exponentiated flexible Weibull extension distribution. *International Journal of Mathematics And its Applications Volume 3*, 1–12.

Famoye, F., C. Lee, and O. Olumolade (2005, 01). The beta-Weibull distribution. *Journal of Statistical Theory and Applications 4*, 121–136.

Gradshteyn, I. S. and I. M. Ryzhik (2007). *Table of integrals, series, and products* (Seventh ed.). Elsevier/Academic Press, Amsterdam. Translated from the Russian, Translation edited and with a preface by Alan Jeffrey and Daniel Zwillinger, With one CD-ROM (Windows, Macintosh and UNIX).

Guerra, R.R., F.A. Pe na Ramírez, and G.M. Cordeiro (2017). The gamma Burr XII Distributions: Theory and Applications. *Journal of Data Science 15*, 467–494.

Iriarte, Y.A., J.M. Astorga, H. Bolfarine, and H.W. Gómez (2017). Gamma-Maxwell distribution. *Communications in Statistics - Theory and Methods 46*, 4264–4274.

Kenney, J. and E. Keeping (1962). *Moving averages*. 3 edn. NJ: Van Nostrand.

Lai, C.D., M. Xie, and D.N.P. Murthy (2003). A modified Weibull distribution. *IEEE Transactions on Reliability 52*, 33–37.

Marinho, P.R.D., R.B. Silva, M. Bourguignon, G.M. Cordeiro, and S. Nadarajah (2019, 08). AdequacyModel: An R package for probability distributions and general purpose optimization. *PLOS ONE 14*, 1–30.

Moors, J. J. A. (1988). A quantile alternative for kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician) 37*, 25–32.

Mudholkar, G.S. and D.K. Srivastava (1993). Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability 42*, 299–302.

Murthy, D. P., M. Xie, and R. Jiang (2004). *Weibull models*, Volume 505. John Wiley and Sons.

Mustafa, A., B.S. El-Desouky, and S. Al-Garash (2016, 10). The exponetiated generalized flexible Weibull extension distribution. *Fundamental Journal of Mathematics and Mathematical Sciences 6*, 75–98.

Mustafa, A., B.S. El-Desouky, and S. AL-Garash (2016). The Marshall-Olkin Flexible Weibull Extension Distribution. *arXiv preprint arXiv:1609.08997*.

Nadarajah, S., G.M. Cordeiro, and E.M.M. Ortega (2015). The ZografosâĂŞBalakrishnan-G family of distributions: Mathematical properties and applications. *Communications in Statistics - Theory and Methods 44*, 186–215.

Paranaíba, P.F., E.M.M. Ortega, G.M. Cordeiro, and M.A.R. de Pascoa (2013). The Kumaraswamy Burr XII distribution: theory and practice. *Journal of Statistical Computation and Simulation 83*, 2117–2143.

Prataviera, F., E. M. M. Ortega, G.M. Cordeiro, R. R. Pescim, and B.A.W. Verssani (2018). A new generalized odd log-logistic flexible Weibull regression model with applications in repairable systems. *Reliability Engineering & System Safety 176*, 13–26.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ristić, M.M. and N. Balakrishnan (2012). The gamma-exponentiated exponential distribution. *Journal of Statistical Computation and Simulation 82*, 1191–1206.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*, 307–333.

Xie, M. and C.D. Lai (1996). Reliability analysis using an additive Weibull model with bathtub-shaped failure rate function. *Reliability Engineering & System Safety 52*, 87–93.

Xie, M., Y. Tang, and T.N. Goh (2002). A modified Weibull extension with bathtub-shaped failure rate function. *Reliability Engineering & System Safety 76*, 279–285.

Zografos, K. and N. Balakrishnan (2009). On families of beta- and generalized gamma-generated distributions and associated inference. *Statistical Methodology 6*, 344–362.

SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 73–79
doi:https://doi.org/10.37830/SJS.2022.1.05

REGULAR ARTICLE

On moments and entropy of the gamma-Gompertz distribution

Fredy Castellares[1], Artur J. Lemonte[2]

[1] Universidade Federal de Minas Gerais, fwcc29@gmail.com
[2] Universidade Federal do Rio Grande do Norte, arturlemonte@gmail.com

Received: June 22, 2022. Accepted: November 17, 2022.

Abstract: The three-parameter gamma-Gompertz family of distributions was introduced recently in the literature. We verify that the analytical expressions provided for the ordinary moments and Shannon entropy are not correct and hence cannot be used for computing such quantities. We derive two closed-form expressions for the mean and a closed-form expression for the Shannon entropy in terms of the Whittaker function.

Keywords: Whittaker function, moments, entropy

MSC: 60E05, 60E10

1 Gamma-Gompertz distribution

The gamma-Gompertz ("GGo" for short) family of distributions was defined by Shama et al. (2022), and its cumulative distribution function (CDF) and probability density function (PDF) are given, respectively, by (Shama et al., 2022, Definition 1)

$$G(x) = \frac{\gamma\left(\theta, \lambda(\mathrm{e}^{\alpha x}-1)/\alpha\right)}{\Gamma(\theta)}, \quad x > 0,$$

$$g(x) = \frac{\lambda}{\Gamma(\theta)} \exp\left\{\alpha x - \frac{\lambda}{\alpha}\left(\mathrm{e}^{\alpha x}-1\right)\right\}\left[\frac{\lambda}{\alpha}(\mathrm{e}^{\alpha x}-1)\right]^{\theta-1}, \quad x > 0,$$

where $\lambda > 0$, $\theta > 0$, $\alpha > 0$, $\Gamma(\cdot)$ is the complete gamma function, and $\gamma(a,z) = \int_0^z t^{a-1}\,\mathrm{e}^{-t}\,dt$ is the lower incomplete gamma function for $a > 0$ and $z > 0$. We can also express the PDF of the GGo distribution in the form

$$g(x) = \frac{\lambda^\theta\,\mathrm{e}^{\lambda/\alpha}}{\alpha^{\theta-1}\,\Gamma(\theta)} \exp\left(\alpha\,\theta\,x - \frac{\lambda}{\alpha}\,\mathrm{e}^{\alpha x}\right)(1-\mathrm{e}^{-\alpha x})^{\theta-1}, \quad x > 0. \tag{1}$$

Published by the Spanish National Statistical Institute

The GGo distribution reduces to the Gompertz distribution (see, for example,c Garg et al. (1970)) when $\theta = 1$.

It is worth stressing that the failure rate (FR) function plays a substantial role in the lifetime data analysis, mainly in survival and reliability studies. Indeed, the mathematical characterization of a lifetime distribution for a certain life phenomena can be made on the basis of its failure rate pattern. In particular, many real-life data, particularly in reliability engineering, exhibit bathtub-shaped FR, which contains the three main regions: early FR region followed by constant FR region and, then, the wear-out region when the FR growths significantly. However, the assumption that the FR increases rapidly with time is not always true. In particular, Bartley (2003) provides an example from electric power industry where some high voltage transformers that survive before the mean life tend to have extremely long lives and the FR is eventually constant. Many distributions have bathtub-shaped FR, but the vast majority are V-shaped, and so these distributions may not fit appropriately bathtub-shaped data with a flat region. However, this region may be very important in real applications and, hence, the correct modeling of a flat region becomes very important. From Shama et al. (2022, Eq. (10)), we have that the FR function of the GGo distribution has the form

$$h(x) = \frac{\lambda \exp\{\alpha x - \lambda(\mathrm{e}^{\alpha x} - 1)/\alpha\}}{\Gamma(\theta, \lambda(\mathrm{e}^{\alpha x - 1} - 1)/\alpha)} \left[\frac{\lambda}{\alpha}(\mathrm{e}^{\alpha x - 1} - 1))\right]^{\theta - 1}, \quad x > 0,$$

where $\Gamma(\cdot, \cdot)$ is the upper incomplete gamma function defined by

$$\Gamma(a, z) = \int_z^\infty t^{a-1}\mathrm{e}^{-t}\, dt, \quad a \in \mathbb{C}, \quad z \in \mathbb{C}.$$

A closer look at Figure 2 in Shama et al. (2022, p. 694) reveals that the FR function of the GGo distribution can present bathtub-shaped FR with a flat region, and so this distribution can be useful in practice to fit real data with a long flat region.

Shama et al. (2022) have derived various distributional properties of the GGo distribution, and have provided an extensive Monte Carlo simulation study to assess the effectiveness of some classical estimation approaches to estimate the GGo distribution parameters. They have also considered a re-parametrized log-GGo distribution and, based on this re-parametrized distribution, a log-GGo regression model was introduced. However, we note that closed-form expressions of some mathematical properties provided by these authors do not appear correct, and so cannot be recommended to users.

2 Moments and entropy

It is well-known that some important statistical measures as, for example, variance, skewness and kurtosis can be obtained in terms of moments. Thus, it is quite important to have a valid expression for the moments in order to compute such quantities. Unfortunately, the analytical expression for the rth ordinary moment of the GGo distribution provided by Shama et al. (2022) does not appear correct. Shama et al. (2022, Theorem 5) have derived a closed-form expression for the moments of the GGo distribution, which is given by

$$\mu'_r = \lambda \left(\frac{\lambda}{\alpha}\right)^{\theta - 1} \mathrm{e}^{\lambda/\alpha} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(-1)^{i+j+r+1}\,(\theta + j - i)^{-r-1}\,\lambda^j}{\Gamma(i+1)\,\Gamma(\theta - i)\,\Gamma(j+1)\,\alpha^{j+r+1}} \Gamma(r+1). \tag{2}$$

From (2), note that the quantity $(\theta + j - i)^{-r-1}$ may be undefined, and so is impossible to compute the moments from the above expression. For example, let $\theta = 1$, $j = 1$ and $i = 2$. In this case, it

follows that

$$(\theta + j - i)^{-r-1} = \frac{1}{(1+1-2)^{r+1}} = \frac{1}{0^{r+1}},$$

which is obviously undefined for all r. In addition, a closer look at the proof of Theorem 5 in Shama et al. (2022, p. 695) reveals that closed-form expression (2) comes from an integral which is not convergent; that is, after a change of variable, Shama et al. (2022, p. 695) have provided an analytical expression for the following integral

$$\int_0^\infty x^r \mathrm{e}^{\alpha(\theta+j-i)x}\, dx.$$

However, if there exists at least a pair (i, j) such that $j-i > 0$, then it is evident that the above integral diverges, since $\alpha > 0$ and $\theta > 0$. In short, the above integral diverges for infinite pairs (i, j), and so the moments from expression (2) do not exist.

The entropy of a random variable is a measure of variation of the uncertainty. Entropy has been used in various situations in science and engineering, and numerous measures of entropy have been studied and compared in the literature. Let N be a random variable with PDF v. The Shannon entropy of N is defined by $\mathbb{E}[-\log(v(N))]$. Shama et al. (2022, Theorem 7) have derived a closed-form expression for the Shannon entropy of the GGo distribution, which is given by

$$H(g) = -\tau - \ln(\lambda) + \theta + \ln(\Gamma(\theta)) + (1-\theta)\Psi(\theta), \tag{3}$$

where $\Psi(\cdot)$ denotes the digamma function, and

$$\tau = \left(\frac{\lambda}{\alpha}\right)^\theta \mathrm{e}^{\lambda/\alpha} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{(-1)^{i+j}(\lambda/\alpha)^j\,(\theta+j-i)^{-2}}{\Gamma(i+1)\,\Gamma(\theta-i)\,\Gamma(j+1)\,\alpha^{j+r+1}}.$$

From (3) (i.e., specifically from τ), the quantity $(\theta + j - i)^{-2}$ may be undefined, and so is not possible to compute the Shannon entropy from the analytical expression $H(g)$. For example, let $\theta = 1$, $j = 1$ and $i = 2$. Hence, it follows that $(\theta + j - i)^{-2} = (1 + 1 - 2)^{-2} = 0^{-2}$, which is obviously undefined. Therefore, the closed-form expression for the Shannon entropy obtained in Shama et al. (2022, Theorem 7) is not a valid analytical expression and, consequently, cannot be recommended to users.

3 Statistical properties

In the following, we provide explicit closed-form expressions of some statistical properties of the GGo distribution.

3.1 Moment generation function

From Gradshteyn and Ryzhik (2015, p. 340), we have that

$$\int_0^\infty (1-\mathrm{e}^{-z})^{\nu-1} \exp\left(-\mu z - \beta \mathrm{e}^z\right) dz = \Gamma(\nu)\,\beta^{\frac{\mu-1}{2}}\,\mathrm{e}^{-\beta/2}\,W\left(\frac{1-\mu-2\nu}{2}, \frac{-\mu}{2}; \beta\right), \tag{4}$$

where $\mu \in \mathbb{C}$, $\beta \in \mathbb{C}$ such that the real part of β is positive, $\nu \in \mathbb{C}$ such that the real part of ν is positive, and $W(a, b; v)$ denotes the Whittaker function with $a \in \mathbb{C}$, $b \in \mathbb{C}$ and $v \in \mathbb{C}$ (Whittaker, 1903). We have the following proposition.

Proposition 1. *The moment generation function of the GGo distribution is given by*

$$M(t) = \mathrm{e}^{\lambda/2\alpha}\left(\frac{\lambda}{\alpha}\right)^{\theta/2-1/2-t/2\alpha} W\left(\frac{1}{2}+\frac{t}{2\alpha}-\frac{\theta}{2}, \frac{\theta}{2}+\frac{t}{2\alpha}; \frac{\lambda}{\alpha}\right). \tag{5}$$

Proof. We have that $M(t) = \int_0^\infty \mathrm{e}^{tx} g(x) dx$, for $t \in \mathbb{R}$. Hence, from the PDF in (1), we have that

$$M(t) = \frac{\lambda^\theta \, \mathrm{e}^{\lambda/\alpha}}{\alpha^{\theta-1}\,\Gamma(\theta)} \int_0^\infty \mathrm{e}^{tx} \, \exp\left\{\alpha\,\theta\,x - \frac{\lambda}{\alpha}\,\mathrm{e}^{\alpha\,x}\right\} (1-\mathrm{e}^{-\alpha\,x})^{\theta-1}\, dx.$$

Let $z = \alpha x$, and so

$$M(t) = \frac{\lambda^\theta \, \mathrm{e}^{\lambda/\alpha}}{\alpha^{\theta}\,\Gamma(\theta)} \int_0^\infty \exp\left\{(t/\alpha+\theta)\,z - \frac{\lambda}{\alpha}\,\mathrm{e}^{z}\right\} (1-\mathrm{e}^{-z})^{\theta-1}\, dz.$$

From (4), the result follows. □

Corollary 1. *If $\theta = 1$, the moment generating function of the GGo distribution reduces to moment generating function of the Gompertz distribution given by*

$$M(t) = \mathrm{e}^{\lambda/\alpha}\left(\frac{\lambda}{\alpha}\right)^{-t/\alpha} \Gamma\left(1+\frac{t}{\alpha}, \frac{\lambda}{\alpha}\right), \quad t \in \mathbb{R}.$$

Proof. From (5) and when $\theta = 1$, we have that

$$M(t) = \mathrm{e}^{\lambda/2\alpha}\left(\frac{\lambda}{\alpha}\right)^{-t/2\alpha} W\left(\frac{t}{2\alpha}, \frac{1}{2}+\frac{t}{2\alpha}; \frac{\lambda}{\alpha}\right).$$

From Olver et al. (2010, p. 177), we have that $\Gamma(\xi+1; v) = \mathrm{e}^{-v/2}\, v^{\xi/2}\, W(\xi/2, (\xi+1)/2; v)$, where $\xi > 0$ and $v > 0$. The result follows by considering $\xi = t/\alpha$ and $v = \lambda/\alpha$ in $\Gamma(\xi+1; v)$. □

Proposition 2. *The characteristic function of the GGo distribution is given by $\varphi(s) = M(is)$, where $s \in \mathbb{R}$, and $i = \sqrt{-1}$ is the imaginary unit.*

3.2 Moments

We have the following proposition.

Proposition 3. *The rth ordinary moment of the GGo distribution is given by*

$$\mu'_r = \frac{d^r}{dt^r} M(t)\bigg|_{t=0}.$$

Remark 1. *proposition 3 relies on the fact that the moments of a distribution can be obtained from the moment generating function.*

Remark 2. *It is worth stressing that the computation of the ordinary moments of the GGo distribution from proposition 3 is not a trivial problem, since the analytical derivatives of the Whittaker function are not easy to obtain.*

The next propositions present the mean of the GGo distribution.

Proposition 4. *If $\theta = n \in \mathbb{N}$, the first moment (mean) of the GGo distribution reduces to*

$$\mathbb{E}[X] = \frac{\mathrm{e}^{\lambda/2\alpha}}{\alpha} \sum_{k=0}^{n} \left(\frac{\lambda}{\alpha}\right)^{(k-1)/2} W\left(\frac{-k-1}{2}, \frac{k}{2}; \frac{\lambda}{\alpha}\right). \tag{6}$$

Proof. The survival function of the GGo distribution when $\theta = n \in \mathbb{N}$ can be expressed as

$$\bar{G}(x) = \mathrm{e}^{\lambda/\alpha} \sum_{k=0}^{n} \frac{(\lambda/\alpha)^k}{k!} (1 - \mathrm{e}^{-\alpha\, x})^k \, \exp\left\{\alpha\, k\, x - \frac{\lambda}{\alpha}\, \mathrm{e}^{\alpha\, x}\right\}, \quad x > 0.$$

We have that $\mathbb{E}[X] = \int_0^\infty \bar{G}(x)dx$, and so

$$\mathbb{E}[X] = \mathrm{e}^{\lambda/\alpha} \sum_{k=0}^{n} \frac{(\lambda/\alpha)^k}{k!} \int_0^\infty (1 - \mathrm{e}^{-\alpha\, x})^k \, \exp\left\{\alpha\, k\, x - \frac{\lambda}{\alpha}\mathrm{e}^{\alpha\, x}\right\} dx.$$

Let $z = \alpha x$. We have that

$$\mathbb{E}[X] = \frac{\mathrm{e}^{\lambda/\alpha}}{\alpha} \sum_{k=0}^{n} \frac{(\lambda/\alpha)^k}{k!} \int_0^\infty (1 - \mathrm{e}^{-z})^k \, \exp\left\{k\, z - \frac{\lambda}{\alpha}\mathrm{e}^{z}\right\} dz.$$

From (4) with $\nu = k + 1$, $\mu = -k$ and $\beta = \lambda/\alpha$, the result follows. □

Proposition 5. *If $\theta > 0$, the first moment (mean) of the GGo distribution reduces to*

$$\begin{aligned}\mathbb{E}[X] = {} & \frac{\mathrm{e}^{\lambda/\alpha}}{\alpha\Gamma(\theta)} \sum_{m=0}^{\infty} \frac{(\lambda/\alpha)^m\, (1-\theta)_m}{m!} \left[\Psi(\theta - m) - \log(\lambda/\alpha)\right] \Gamma(\theta - m) \\ & + \frac{\mathrm{e}^{\lambda/\alpha}(\lambda/\alpha)^\theta}{\alpha\Gamma(\theta)} \sum_{m=0}^{\infty} \frac{(1-\theta)_m}{m!(\theta-m)^2}\, {}_2F_2(\theta - m, \theta - m; \theta - m + 1, \theta - m + 1; -\lambda/\alpha),\end{aligned}$$

where $(a)_n := a(a+1)\cdots(a+n-1)$ is the rising factorial with $a \in \mathbb{R}$, $(a)_0 := 1$ and $n \geq 1$, and ${}_2F_2(a, b; c, d; z)$ is the generalized hypergeometric function defined by

$${}_2F_2(a, b; c, d; z) = 1 + \sum_{n=1}^{\infty} \frac{(a)_n\, (b)_n}{(c)_n\, (d)_n} \frac{z^n}{n!}, \quad z \in \mathbb{R},$$

and $\Psi(\cdot)$ is the digamma function defined by

$$\Psi(z) = \frac{d}{dz} \log(\Gamma(z)).$$

Proof. We have that $0 < \mathrm{e}^{-\alpha\, x} < 1$ for all $x > 0$, and so $(1 - \mathrm{e}^{-\alpha\, x})^{\theta-1} = \sum_{m=0}^{\infty} \frac{(1-\theta)_m}{m!} \mathrm{e}^{-m\, \alpha\, x}$. Hence, we can express the PDF in (1) of the form

$$g(x) = \frac{\alpha(\lambda/\alpha)^\theta\, \mathrm{e}^{\lambda/\alpha}}{\Gamma(\theta)} \sum_{m=0}^{\infty} \frac{(1-\theta)_m}{m!} \exp\left\{(\theta - m)\alpha\, x - \frac{\lambda}{\alpha}\, \mathrm{e}^{\alpha\, x}\right\}, \quad x > 0.$$

Using the above PDF, it follows that

$$M(t) = \frac{\alpha(\lambda/\alpha)^\theta\, \mathrm{e}^{\lambda/\alpha}}{\Gamma(\theta)} \sum_{m=0}^{\infty} \frac{(1-\theta)_m}{m!} \int_0^\infty \exp\left\{(\theta - m + t/\alpha)\alpha\, x - \frac{\lambda}{\alpha}\, \mathrm{e}^{\alpha\, x}\right\} dx.$$

Let $z = \alpha x$, and so

$$M(t) = \frac{(\lambda/\alpha)^{\theta}\, \mathrm{e}^{\lambda/\alpha}}{\Gamma(\theta)} \sum_{m=0}^{\infty} \frac{(1-\theta)_m}{m!} \int_0^{\infty} \exp\left\{(\theta - m + t/\alpha)z - \frac{\lambda}{\alpha}\,\mathrm{e}^{z}\right\} dz.$$

From Gradshteyn and Ryzhik (2015, p. 340), we have that $\int_0^{\infty} \exp\left(-pz - q\mathrm{e}^{z}\right) dz = q^{p}\,\Gamma(-p, q)$, where $p \in \mathbb{C}$, and $q \in \mathbb{C}$ such that the real part of q is positive. Hence,

$$M(t) = \frac{\mathrm{e}^{\lambda/\alpha}}{\Gamma(\theta)} \sum_{m=0}^{\infty} \frac{(\lambda/\alpha)^m\,(1-\theta)_m}{m!} \left(\frac{\alpha}{\lambda}\right)^{t/\alpha} \Gamma\left(t/\alpha + \theta - m, \frac{\lambda}{\alpha}\right).$$

From Brychkov (2008, p. 22), we have that

$$\frac{d}{da}\Gamma(a,z) = [\psi(a) - \log(z)]\Gamma(a) + \Gamma(a,z)\,\log(z) + \frac{z^a}{a^2}\,{}_2F_2(a,a;a+1,a+1;-z).$$

Now, using the above derivative and that $\mathbb{E}[X] = (d/dt)M(t)|_{t=0}$, the result follows. □

Remark 3. *The algebraic developments considered in this section reveal that is not easy to obtain a general closed-form expression for the ordinary moments of the GGo distribution. This is still an open problem regarding the three-parameter GGo family of distributions introduced by Shama et al. (2022).*

3.3 Entropy

We have the following proposition

Proposition 6. *The Shannon entropy of the GGo distribution is given by*

$$H(g) = -\log\left[\frac{\lambda^{\theta}\,\mathrm{e}^{\lambda/\alpha}}{\alpha^{\theta-1}\,\Gamma(\theta)}\right] - \alpha\,\theta\,\mathbb{E}[X] + \frac{\lambda\,M(\alpha)}{\alpha} + (\theta-1)\sum_{n=1}^{\infty}\frac{M(-\alpha\,n)}{n}, \tag{7}$$

where $\mathbb{E}[X]$ is the mean of the GGo distribution provided in proposition 5, and $M(\cdot)$ is the moment generating function of the GGo distribution provided in proposition 1.

Proof. The Shannon entropy of the GGo distribution is given by $H(g) = -\mathbb{E}[\log g(X)]$, where $g(x)$ is the PDF of the GGo distribution. Note that

$$\ln(g(x)) = \ln\left[\frac{\lambda^{\theta}\,\mathrm{e}^{\lambda/\alpha}}{\alpha^{\theta-1}\,\Gamma(\theta)}\right] + \alpha\,\theta\,x - \frac{\lambda}{\alpha}\,\mathrm{e}^{\alpha\,x} + (\theta-1)\,\ln(1-\mathrm{e}^{-\alpha\,x}), \quad x > 0.$$

Using the expansion $-\ln(1-z) = \sum_{n=1}^{\infty}\frac{z^n}{n}$, for $|z| < 1$, and taking the expected value, the result follows. □

4 Numerical study

In what follows, we provide some numerical values for the mean and Shannon entropy of the GGo distribution. We use the proposed closed-form expression in proposition 5 to obtain numerical values for $\mathbb{E}[X]$, and the proposed closed-form expression in (7) to obtain numerical values for the entropy. Table 1 lists the values of $\mathbb{E}[X]$ and $H(g)$ for $\lambda = 0.8$, $\alpha = 1.0$, and different values of θ. In this table, $n_{\max}$ means the number of terms considered in the expansion for $H(g)$ in (7) , while the last column shows the corresponding values of the entropy by numerical integration. Note that the numerical values delivered by the expression (7) and numerical integration for the entropy are near, mainly when $n_{\max} = 1000$.

		$H(g)$				
θ	$\mathbb{E}[X]$	$n_{\max} = 10$	$n_{\max} = 60$	$n_{\max} = 150$	$n_{\max} = 1000$	numerical integration
0.6	0.454810	0.292393	0.200059	0.179110	0.159559	0.150320
0.7	0.518039	0.370634	0.320716	0.310929	0.302876	0.299965
0.8	0.578384	0.435732	0.411674	0.407606	0.404650	0.403818
0.9	0.636056	0.488960	0.480238	0.478969	0.478153	0.477971
1.0	0.691245	0.531898	0.531898	0.531898	0.531898	0.531898
1.1	0.744126	0.566114	0.570739	0.571234	0.571485	0.571521
1.2	0.794856	0.593027	0.599791	0.600410	0.600690	0.600721
1.3	0.843579	0.613869	0.621310	0.621891	0.622125	0.622147
1.4	0.890428	0.629689	0.636984	0.637469	0.637644	0.637657
1.5	0.935522	0.641361	0.648084	0.648465	0.648589	0.648594
1.6	0.978973	0.649613	0.655578	0.655864	0.655950	0.655951
1.7	1.020881	0.655048	0.660206	0.660416	0.660472	0.660472
1.8	1.061341	0.658159	0.662540	0.662690	0.662726	0.662726
1.9	1.100437	0.659356	0.663028	0.663134	0.663157	0.663157
2.0	1.138249	0.658975	0.662021	0.662096	0.662110	0.662110

Table 1: Mean and entropy.

Acknowledgments

The authors would like to thank the Editor and an anonymous reviewer for their insightful comments and suggestions. Fredy Castellares gratefully acknowledges the financial support from FAPEMIG (Belo Horizonte/MG, Brazil). Artur Lemonte gratefully acknowledges the financial support of the Brazilian agency CNPq (grant 304776/2019-0).

References

Bartley, William H (2003). Analysis of transformer failures, wgp 33 (03). In *Proceedings of the 36th annual conference on international association of engineering*.

Brychkov, Yury A (2008). *Handbook of special functions: derivatives, integrals, series and other formulas*. Chapman and Hall/CRC.

Garg, Mohan L, B Raja Rao, and Carol K Redmond (1970). Maximum-likelihood estimation of the parameters of the gompertz survival function. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 19*(2), 152–159.

Gradshteyn, Izrail Solomonovich and Iosif Moiseevich Ryzhik (2015). *Table of integrals, series, and products* (8th Edition ed.). Academic press.

Olver, Frank WJ, Daniel W Lozier, Ronald F Boisvert, and Charles W Clark (2010). *Handbook of mathematical functions*. Cambridge University Press, Cambridge, UK.

Shama, MS, Sanku Dey, Emrah Altun, and Ahmed Z Afify (2022). The gamma–gompertz distribution: Theory and applications. *Mathematics and Computers in Simulation 193*, 689–712.

Whittaker, Edmund T (1903). An expression of certain known functions as generalized hypergeometric functions. *Bulletin of the American Mathematical Society 10*(3), 125–134.

Spanish Journal of Statistics
Vol. **4** No. 1 2022, Pages 81–88
doi:https://doi.org/10.37830/SJS.2022.1.06

Spanish Journal of Statistics

Official Statistics

A first interim assessment of the third round of peer review of the European statistical system

Agustín Cañada

Subdirección General de Difusión, Instituto Nacional de Estadística, agustin.canada.martinez@ine.es
Received: November 3, 2022. Accepted: December 26, 2022.

Abstract: Peer Reviews are exercises to assess compliance with the principles and indicators of the European Statistics Code of Practice by the members of the European Statistical System: Eurostat and the national statistical systems (composed of statistical offices and other institutions). Peer Reviews are carried out periodically (every 5/6 years), by agreement of the European Union. To date, three rounds have been carried out: in 2006-2008, in 2013-2015, and a third round is underway between 2021 and 2023. Although the third round is still ongoing at the time of writing (December 2022), based on the experience of a representative group (14) of the countries already reviewed, a first assessment can already be made of the degree of achievement of the objectives pursued. The aim of this document is to provide a first input for a future comprehensive "lessons learned exercise" and to contribute to the debate on aspects to be taken into account in future peer reviews.

Keywords: Code of Practices, Peer review, Quality of official statistics

MSC: 62-03, 62-04, 62P99

1 Introduction

Peer reviews (PR) are exercises to assess compliance with the principles and indicators of the European Statistics Code of Practices (CoP) by members of the European Statistical System: Eurostat and national statistical systems (NSI, other institutions...). The ultimate objective of these assessments is to strengthen the statistical systems at national and European level, thus increasing the reliability and credibility of users and other stakeholders in European official statistics (Cañada, 2019).
The PR are carried out periodically (every 5/6 years), by agreement of the European Union: There have been three exercises ("rounds"): The first "round" of PR took place between 2006 and 2008; the second, more far-reaching one, was conducted between 2013-2015 and a third round is underway between 2021 and 2023.
At the time of drafting this report (November 2022), the halfway point of the third round of peer reviews has been reached, both in terms of time and number of countries evaluated, with 19 of the 31

planned evaluation visits having been completed. Of these, Eurostat has published 14 final reports on its website. In the case of Spain, the experts' visit to INE took place from May 31 to June 3, 2022, and the final report is not yet available. However, based on published reports and Spain's experience, it is possible to make an initial assessment of the degree of achievement of the objectives pursued. The purpose of this document is to provide decision-makers with a first input for a future and more complete "lessons learned exercise" and to contribute to the debate by the countries on the aspects to be considered in future peer reviews.

2 Peer review of the European Statistical System: An overview

Table 1 summarizes the key features of the PR, highlighting the differences between the three rounds conducted to date.

The first "round" of reviews focused only on some CoP principles and was conducted under a proper peer review approach: statistical institutes in each EU country were assessed by experts from statistical institutes of other countries.

The second round, carried out between 2013-2015, was broader in scope: all the principles of the Code of Practice were assessed; (a sample of) other institutions producing European statistics in each country - the so-called Other National (Statistical) Authorities ONA) (ministries, agencies, etc.)- was brought in; the process was applied not only to the EU countries, but also to the four "European Free Trade Association" (EFTA) countries; an attempt was made to move closer to an audit approach, using teams of expert reviewers from outside the NSIs, to avoid the impression of less independence and objectivity linked to the internal reviewers as in the first round.
From the point of view of its audit approach, conventional audit principles are followed: information on the national statistical system is prepared by the NSI, based on a self-assessment questionnaire (SAQ), supporting the answers with documents that serve as evidence; the team of reviewers analyses this information in depth; there is an "audit" visit to each NSI in which the reviewers complete their on-site knowledge of the system assessed; the last stage is the preparation by the reviewers of a report on the degree of compliance with the CoP by the NSI (countries) that includes "recommendations" on areas for improvement; and then, in response to these recommendations, the NSI would prepare a multi-year action plan for improvement. Implementation will be monitored annually by the European authorities.
The third round, which runs from 2021 to 2023, shares common features with the second round, as it covers the same scope of countries (all EU countries plus the four EFTA countries), the same content, as it covers all CoP principles, and the same reviewed institutions, Eurostat, the NSIs and the ONA. However, it presents differences that seek to improve or correct more controversial aspects of the second round, which can be grouped into six fundamental ones (Cañada and Muñoz, 2016):

1) Methodological approach. Although like the 2013 PR, the new phase is still close to the audit methods, it is nevertheless intended to be a combination of audit and PR approaches. This is perfectly illustrated by the change in the composition of the teams explained below.
There are also changes in some of the tools used: in particular, the NSI self-assessment questionnaire has been simplified so that, without losing its global character (it covers all the principles of the Code), it allows for greater agility and speed in the review process. The questionnaire is much simpler than that of the 2nd round: the questions are adapted to the 84 CoP indicators, which in fact

Topic	2006-2008	2013-2015	2021-2023
1. Territorial scope	EU countries	28 EU + 4 EFTA countries	27 EU + 4 EFTA countries
2. Conceptual scope	Principles: 1 to 6 and 15 of CoP (2005).	- All CoP principles (2011) + coordination + EU cooperation	- All CoP principles (2017). Emphasis on certain principles
3. Institutional Scope	NSI	NSI + ONA + Eurostat	NSI + ONA + Eurostat
4. Methodology 4.1. Reviewers	Peer review approach Internal to the ESS: NSIs active staff	Quasi-audit Experts from outside the system (+ Eurostat observer)	Combination PR + audit Internal + External + Eurostat Member + Eurostat Harmonization Observer (in some PR)
4.2. Procedure	PR approach: self-assessment; review; NSI visit; reporting	"Audit-type" approach. self-assessment; review; NSI visit; report	"Audit-type" approach.
4.3. Self-assessment questionnaire(s)	CoP indicators only for the selected principles	a) NSI: 3 SAQ: - CoP SAQ: defined by principles/ indicators (broken down at QAF detail) - 2 Ancillary SAQ on: Coordination and Cooperation. [Total number of questions: 400] (b) ONA: simplified questionnaire	a) NSI (only one SAQ) - 84 questions on each indicator of the CoP. - 60 SWOT questions. - 9 questions by area of the CoP. - 29 additional questions (innovations, COVID-19 pandemic) [Total number of questions: 182] (b) ONA: simplified questionnaire
4.4. Evaluation report	Country-specific reports: NSIs	Country-specific report, focusing on NSIs (with some reference to ONA)	a) Specific report for each country, covering both the NSI and the ONA. b) Recommendations differentiated between Compliance-relevant and improvement-related
5. Others			Eurostat + countries communication campaign

Table 1: Peer review of the ESS: key features of the three rounds

reveal compliance with the 16 principles of the Code; some other questions are included (see Table 1) to reach a total of 182 questions. We can remain that in the 2nd round, the SAQ was structured according to QAF (quality assurance framework, Eurostat (2019)) methods linked to the CoP indicators, which involved completing a questionnaire with more than 400 questions.
2) Seek greater harmonization of reviews across countries. In the 2013 round, there was great heterogeneity and lack of harmonization in the country reviews and in the PR reports: there were no well-

established general criteria on what the most relevant and accessory points could be when assessing countries; the heterogeneity of the reports and of the recommendations to the countries also translated into heterogeneity of the improvement action plans.... This heterogeneity greatly conditioned one of the objectives of the PR, which is to contribute to the improvement of the European statistical system, by developing actions that would have led the countries to advance along common lines.
Therefore, more "harmonization of the review along the different countries" is a Priority goal of the current 3rd round. Several elements are used to achieve this harmonisation (Eurostat, 2021a):
- On the one hand, an effort towards greater standardisation of the methodologies applied by the different teams: Although the report and the recommendations will obviously depend on the outcome of the evaluation, and the review teams have autonomy of decision, priority themes have been recommended for the review (according to European agreements). A guide of suggested types and categories of recommendations, has ben published in one of the annexes of the methodology (Eurostat, 2020a).
- Furthermore, a differentiation into two types of recommendations was introduced: (...) "The recommendations issued by the peer review team should be split into "Compliance-relevant" - recommendations fundamental/important to ensure compliance/alignment with the ES CoP - (...), and "Improvement-related" - less critical/technical supporting improvements recommendations-.
- An additional element is the role of Eurostat technicians in the process, which is discussed in a later section.
3) Composition of the review teams. In the third round, the composition of the teams of reviewers uses a mixed formula from previous experiences: each team of reviewers includes both "internal" evaluators from the statistical system (-PR approach-) and experts external to official statistics (-audit approach-). This is an attempt to solve a problem detected in the second round: having reviewers external to the system guarantees greater objectivity but has the counterpart that they may not have a sufficiently updated knowledge of the situation and trends of the statistical system.
For this reason, the third round has opted for this compromise between the two previous ones, combining the independent and objective vision provided by the external reviewers with the updated knowledge of the practices and criteria of European official statistics that can be provided by the NSI and Eurostat staff.
Finally, the teams of reviewers are made up of four members: a serving member of an NSI ("internal", "peer review" approach); an expert external to the ESS ("audit" approach); an expert from Eurostat (participating as a reviewer); and, as Chairman of each team of experts, a senior statistician, with experience in NSI management.
The first group, people who are currently working in the NSIs, is proposed by the countries. Within this group, some coordinators of the PR Process (and/or responsible of the Quality management) in their own countries, are simultaneously reviewers for other countries, which reveals the difficulty in finding people specialised in quality topics. An implicit issue is whether this direct involment of quality managers as reviewers may give an impression of less objectivity of the process, as already raised in the first round of the PR.
4) Role of Eurostat. One of the aspects of the previous round questioned by the countries was the lack of active participation of Eurostat in the reviews (only a Eurostat technician participated as a "mere observer" in the country visits). In response to this criticism, Eurostat has taken a more active role in the 3rd round:
- On the one hand, a Eurostat technician participates as a member of the expert team. This active role, besides serving to improve the quality of the process, is guided by the objective of improving harmonization in the assessment and reporting of countries.
- Moreover, in pursuit of this harmonization objective, Eurostat proposed to incorporate as an ad-

ditional "observer" to the process an expert, specialized in quality issues, who would support the expert team during the country visit and in the drafting of the report and, especially, of the recommendations to the countries. "The role of Eurostat observers would be to support the expert team in formulating more harmonized recommendations, especially in terms of scope and magnitude, during the PR visit and, in particular, on the last day of the visit." (Eurostat 2020b). However, due to lack of resources, Eurostat observers are only involved in some of the reviews. It is Eurostat itself that chooses the visits in which it wishes to participate, with the prior approval of the country under review.

5) Strengthening the role of ONA. This round of the PR aims to strengthen the participation and role of the ONA. Although they had already been included in the previous round, their role in the process and in the reports was considered very marginal. Therefore, the new round aims to give a greater role to the ONA and, in short, to have a more complete view of the situation of the National Statistical Systems as a whole. To this end, a specific procedure and objective criteria agreed upon by Eurostat and the NSIs have been established for greater participation of the ONA in the process.

6) Improved communication on the PR to stakeholders. One of the novelties of the third round is the attention given to communication aspects, through the design and implementation of a communication campaign by Eurostat and the countries on the process, its objectives and results. This is a reaction to one of the most questioned aspects of the second round, which was the limited impact of the process outside the statistical world. There was a critical view of the (limited) dissemination of the process.

But communication is essential to achieve one of the fundamental objectives of the PR: to contribute to the external image of quality and credibility of official statistics. That is, to demonstrate to the institutions most closely linked to statistics (the "stakeholders": informants, policy makers, users), but also to society at large, that the European Statistical System "operates within a sound quality framework". In short, to contribute to the credibility and confidence of users in statistical institutions. At the same time, it will also promote that governmental institutions support the improvement actions derived from the PR.

To this end, a communication strategy to accompany the third round of ESS peer reviews was defined based on the design of common instruments and means of dissemination for all countries.

3 Some drawbacks of the third round: Lessons learned and alternatives for future PR

In November 2022, the halfway point of the third round of peer reviews has been reached, both in time and in the number of countries evaluated, with 19 of the 31 peer review visits having been completed.

In the case of Spain, the experts' visit to INE took place from May 31 to June 3, 2022. Part of the meetings were held with the statistical production units of INE (35 managers and technicians of the institution participated in the meetings), but also with 30 representatives of the main stakeholders of the Spanish Statistical System: other producing institutions, such as Ministries and Bank of Spain; qualified users -business federations, trade unions, non-governmental organizations-; managers of administrative registers; representatives of the scientific community: universities and researchers; media. At the time of writing (November 2022) only a provisional version of the report is known.

Based on the data known so far from the reports published by Eurostat on its website (www), a first assessment can already be made of the degree to which the objectives pursued are being achieved.

1) Harmonization as a main challenge. A greater harmonisation of the countries' processes was

one of the basic objectives of the third round the ESS peer reviews. A challenge of this objective is how to achieve more harmonisation of the reports while there is flexibility for the reviewers' team in the choice of principles/indicators to review. Therefore, harmonization has been focused on the recommendations: "The aim of harmonisation is about the outcome/ the final results of the peer review meaning the scope, magnitude and number of recommendations" (Eurostat, 2021b).
Concerning scope and magnitude, to ensure greater harmonization, several elements already mentioned, were introduced: a guide of the types and categories of recommendations; the above-mentioned distinction between "Compliance-relevant" (CR) and "Improvement related" (IR); the inclusion of a Eurostat observer...
As a reference of the outcome of the process, we can summarise figures for the 14 countries: a total of 251 recommendatiosn were made, most of them (221) belong to the "Improvement" category, and only 30 are "Compliance-relevant". This is undoubtedly a very positive result, as it reveals the high degree of compliance with the CoP by European countries. However, an in-depth analysis of the reports reveals some questionable aspects: on the one hand, there are some countries without any "CR" recommendations. Although this is possible, as it is reflecting a very high level of compliance with the CoP, in practice it is questionable, if we remember how the PR self-assessment questionnaire is designed: For those countries where no CR recommendations have been identified, this means that the level of compliance with the 84 CoP indicators should be almost total or perfect.
The second doubt that arises from the differences in interpretation among the different teams of reviewers as to what is included in one category or another, even with different evaluations for similar recommendations. Of course, the recommendations and their classification in one category or another is responsibility of the Expert teams; and they are autonomous to guarantee objectivity of the process. But more homogeneous criteria would be useful.
Another aspect linked to harmonization is the aforementioned issue of Eurostat observers, who, due to limited resources, only participate in some of the visits (chosen by Eurostat). Thus, this initiative would have more scope if observers could participate in all reviews.
2) Recommendations and the communication issue. One of the aims of the 3rd round is the effort to a more intensive dissemination and communication campaign of the process to reassure stakeholders about the quality of European Statistics. On the other hand, being one of the main objectives of the PR assessing whether NSI are fully compliant with the CoP, that means to identify aspects for improvements and/or where there is not compliance with the CoP (as in any auditing process). That is, the final and more evident outcome of the PR is the list of the recommendations stated to the countries. The obvious problem is that if the report places excessive emphasis on the recommendations (by their number, by their nature) and/or points to be improved by the country under evaluation, the final view portrayed to the stakeholders about the country and its statistical system is debatable. This may have the opposite effect of what was intended by the review process: by causing stakeholders to question the quality of official statistics.
Back to the figures for the 14 countries available, the most frequent number of recommendations is 22. This number of recommendations necessarily implies devoting a good part of the country report to their justification (moreover, the recommendations appear twice in the report: in the executive summary and in a specific chapter). Thus, for a non-expert reader or a reader outside the statistical world, it is uncertain what impression can be obtained of the statistical situation of a country from a report of these characteristics.
3) The scope of the PR. In terms of scope, the PR aims to assess the overall status of countries' statistical systems, through the level of achievement with the CoP as a whole. This objective is clear, but probably too broad or ambitious, given the current complexity of statistical systems and the obvious resource and time constraints faced by the reviews. In practice, it is a difficult challenge,

firstly for countries to summarize the status of the statistical system in a simple questionnaire and supporting documentation; secondly, it is also difficult for the team of reviewers, a small group of people working in a necessarily small amount of time, to analyze the documentation in detail and to adequately understand and assess a country's statistical system.
For example, the official self-assessment questionnaire (SAQ) of the third round contains a total of 182 questions: 84 for the CoP indicators; 60 for the typical SWOT questions; 9 for a general self-assessment by groups of principles (institutional environment, processes, and products); 29 for additional questions on other topics.
Moreover, in the case of Spain, the SAQ has been designed under a structure like that of the 2nd round questionnaire; that is, describing not only compliance with the 84 CoP indicators, but also compliance with the methods recommended by the QAF 2019. And, in support of this questionnaire, INE prepared more than ninety documents, in some cases written or updated (and translated into English) especially for the PR. The investment made by the Spanish NSIs in the preparation of the PR was very considerable.
After the analysis of all this documentation, the second step of the PR is the visit to the country assessed by the reviewers. Over the course of five days, the reviewers try to complement their vision of the statistical system through meetings with INE staff and different stakeholder groups.
Realistically, despite the undeniable effort and professionalism of the experts, it is still a system that faces obvious limits (in terms of time and resources) to adequately capture the complex reality of current European statistical systems. And the question remains whether the final reports could not reflect the great efforts made by the NSIs in this field.
Returning to one of the recurring questions throughout the different rounds, it is worth asking whether, instead of a global approach aiming to assess the full coverage of the Code, a more in-depth but narrower scope analysis, focused on groups of principles, would be more appropriate. In addition, a more concentrated and narrower scope of the PR could also contribute to the recurrent objective of harmonization.

4 Final comment

There is no doubt that the PR assessments are contributing to the improvement of the quality of the European statistical system. They constitute a balanced mechanism between simple internal evaluations and audit exercises, adapted to the peculiarities of official statistics. The third round represents a new step forward, overcoming some of the limitations of the previous rounds.
Although the third round is still in progress at the time of reviewing this document (December 2022), based on the final reports for 14 of the countries already reviewed, (and the practical experience of the author in the case of Spain) a first assessment can already be made of the degree of achievement of the objectives pursued. As a result of this analysis, some suggestions are made on areas where there is still room for further efforts: in relation to the scope of application, as in previous rounds, the question arises once again as to whether the PR can be approached from a perspective more focused on specific topics, rather than attempting to cover the entire CoP; it is also pointed out that further progress is needed in the harmonization of results between countries, with greater homogenization of criteria among the review teams; aspects that could contribute to the objective of better communication of the nature and value of these exercises aimed at the key players in the statistical system.

References

Cañada, Agustín (2019). Peer reviews of the european statistics and the involvement of users in the quality assurance of official statistics. *Spanish Journal of Statistics 1*(1), 33–40.

Cañada, Agustín and Luisa Muñoz (2016). Peer review 2013-2015. lessons learned, challenges and opportunities. In *European Conference in Official Statistics 2016, Madrid*.

Eurostat (2019). Quality assurance framework. Technical report, Eurostat.

Eurostat (2020a). Guides' annex vi: Formulation of issues and recommendations. Technical report, Eurostat.

Eurostat (2020b). Overall methodology for the third round of peer reviews. Technical report, Eurostat.

Eurostat (2021a). Annex on the harmonisation of results of the peer reviews. third round: 2021-2023. Technical report, Eurostat.

Eurostat (2021b). Frequently asked questions on the 3rd round of ess peer reviews (last update: 3.6.2021). Technical report, Eurostat.

SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 89–98
doi:https://doi.org/10.37830/SJS.2022.1.07

OFFICIAL STATISTICS

Use of death statistics according to cause of death in health research

Gregorio Barrio

Biomedical Research Center Network for Epidemiology and Public Health (CIBERESP)
National School of Public Health, Carlos III Health Institute, gbarrio@isciii.es
Received: November 3, 2022. Accepted: December 26, 2022.

Abstract: Estimates of total and cause-specific mortality rates require information on the number of deaths (numerator) and the population at risk (denominator). In unlinked mortality studies, the numerator and denominator come from different sources, so there may be a numerator/denominator bias when estimating mortality rates according to certain individual attributes. This bias does not occur in linked mortality studies, in which data from the census or general population surveys are linked to vital records, and in the case of death, to the date and cause of death. However, regulations to protect individuals' confidentiality greatly limit the use of linked and unlinked mortality statistics for scientific research, whether due to the regulations themselves or because of the restrictive interpretations thereof by some statistical offices not always sufficiently argued. On the other hand, some methodological developments by these offices are of enormous relevance, for example, the linkage between socioeconomic indicators and mortality by the National Statistics Institute of Spain, which enables the study of the relationship between socioeconomic factors and mortality and its variation over time.

Keywords: Cause of death, Confidentiality, Health research, Mortality registers, Numerator-denominator bias, Record linkage

MSC: 60E05, 62P99

1 Introduction

The creation of civil registries in the 19th century to collect data on deaths and other sociodemographic characteristics was an important milestone for health research. The information contained in the death certificates, necessary for registering the deceased in these registries, began to be used in different studies on mortality according to factors such as occupation and place of residence. William Farr was the first to use these vital records to calculate the mortality rate across various occupations and in different geographical areas, in the mid-19th century in England and Wales

(Drever and Whitehead, 1997). The denominator for calculating this rate came from the population census. Farr passionately advocated for the development of a standard international nomenclature for collecting statistics on cause of death, coming to regard it as even more important to research than establishing a standard system of weights and measures in the physical sciences.

In the early 20th century, researchers in England and Wales also began to use information on occupation, as provided by the census and by death registries, to calculate mortality rates in different social strata. They used a social class scheme developed in 1911, which categorized occupations based on their social prestige. Known as the Registrar General's social class scheme, it was used throughout the 20th century in countless studies by British authors. Researchers from other countries have also developed similar classifications to study socioeconomic differences in mortality.

On the other hand, scientists in numerous countries have used census and other population-based data to characterize geographic or political-administrative areas based on demographic, socioeconomic, or environmental variables, using this information to assess the relationship between different geographic areas and the mortality of the population residing in them. On other occasions, investigators have evaluated the impact of an unexpected event or a health intervention on population health by comparing the mortality rates between areas or over time.

Thus, despite their limitations, cause-specific mortality statistics have helped define the main public health problems and carry out innumerable studies on the epidemiology and natural history of diseases. In fact, Hill considered that vital statistics laid the groundwork for the birth of epidemiology, and indeed, Snow used London's vital statistics, provided by the Registrar General in the mid-nineteenth century, in his landmark study of cholera transmission in that city (Hill et al., 1955).

2 Mortality studies with unlinked information

One of the earliest insights in occupational medicine was the recognition of the healthy worker effect (Checkoway et al., 2004). Occupational studies of mortality at the end of the 19th century described this bias after observing lower mortality in people who were employed relative to the general population. This phenomenon is due to the fact that people with a chronic disease are less likely to enter or remain in the labor market.

Likewise, for most of the 20th century, studies of socioeconomic inequalities in mortality have used the death registry and census data. In countries where these data were available (usually because occupation was recorded on the death certificate), researchers were able to describe the evolution of socioeconomic differences in overall and cause-specific mortality. For example, comparative studies in several European countries found that in the 1990s, socioeconomic differences in mortality were smallest in southern European countries like Italy and Spain (Mackenbach et al., 1997).

Since the 19th century, studies using a certain characteristic of the geographic or political-administrative area as the main independent variable have also generated knowledge of interest to public health. In the 19th century, rural populations had lower mortality than urban ones (Cosby et al., 2008) — a relationship that has been inverted in high-income countries. Other studies in these settings have investigated the relationship between various characteristics of the neighborhood of

residence and mortality, showing that the population residing in more deprived neighborhoods have higher mortality (Meijer et al., 2012).

Similarly, the availability of data on deaths and on populations in cities, municipalities, regions, and countries has made it possible to assess the impact of periods of high air pollution, heat waves, macroeconomic fluctuations, public health regulations, or access to medicines on overall and cause-specific mortality. For example, one study showed a sharp acceleration in the decline of mortality due to hepatitis C and other related causes, such as liver cancer and HIV infection, after the implementation of the Hepatitis C Strategic Plan in Spain, in April 2015, whose main component was providing universal and free access to direct-acting antivirals against this disease (Table 1) (Politi et al., 2022).

	Annual percent change in mortality rate(*)	
Cause of death	**Pre-intervention period**	**Post-intervention period**
Hepatitis C	-3.2	-18.4
Hepatocarcionoma	-0.9	-2.7
Cirrhosis	-3.7	-3.7
HIV disease	-8.3	-15.6
Non-C hepatitis	-5.8	-1.4
Other live diseases	-3.1	-1.6
All non-hepatic causes	-2.2	0.1

(*) 1. The pre-intervention period included from the first quarter 2001 to the first quarter 2015, inclusive. The post-intervention period included from the second quarter 2015 to the last quarter 2018.

Table 1: Comparison of mortality trends from hepatitis C and other hepatic and non-hepatic causes of death in the general population, before and after the implementation of the Strategic Plan for Tackling Hepatitis C: Spain, 2001 – 2018

Scientists' access to the data needed for such studies is uneven, which helps explain the absence of these types of investigations in some countries or regions. Restrictions are rooted in the fact that some statistical offices consider that removing the personal identification of the deceased —first name, last name, personal identification number— is not enough to protect individuals' confidentiality. If the population is small, they include characteristics like occupation, day of death, neighborhood, or municipality of residence within the scope of statistical confidentiality. There are even statistical offices that consider individual age as an object of special protection, so instead of providing the age of each deceased, they share only the five-year age group to which the deceased belongs. Others consider that it is the combination of variables that should remain secret, so they do not provide both age and cause of death for the deceased. There are endless ways of thinking on this matter.

Such obstacles make it difficult, if not impossible, to perform spatiotemporal analyses of great epidemiological and public health interest. For example, restrictions on data access preclude the study of mortality and daily ecological variables (e.g., temperature, air pollution) or socioeconomic indicators about the neighborhood or municipality of residence. It is also difficult to develop and

evaluate highly relevant public health interventions. In some cases, researchers can access these data once they formalize and fulfill certain administrative requirements imposed by the statistical offices. But many give up or do not even attempt it in the face of the heavy bureaucratic burden entailed.

Some scientists are surprised at this limitation, first of all, because individual observations are never disseminated in the findings of medical research, except in some clinical studies based on a very few patients. Second, the characteristics subject to special statistical protection are instrumental to generating results of interest. Third, the data that some offices treat as a statistical secret are not considered as such by others. These scientists probably forget that, in statistical offices, as in many other places (including research centers), ethics committees establish these limitations according to different criteria, either due to variations in national regulations on data protection or in the interpretation of these laws by different offices.

These heterogeneous interpretations can generate paradoxical situations, as in Spain, where for reasons of confidentiality, the National Statistics Institute does not routinely provide some attributes of the deceased person's microdata file, while some regional statistics offices do.

3 Mortality studies with linked information

In classic mortality studies by occupation or social class, researchers have calculated the number of deaths occurring among people of a given occupation during a given time period, divided by the number of people in that occupation for half the period. As noted, the data source for the numerator was the death certificate, and for the denominator the population census. As the numerator and denominator come from different sources, these are unlinked cross-sectional studies, which are at risk of a numerator/denominator bias, since a person's occupation in the death registry may not match that person's occupation in the census (Lynge, 2011).

This bias can also occur with other variables, such as sex or age, since both come from different sources; some people may appear as men in the death registry and as women in the population census, or vice versa, and the age may different. However, the scientific literature has never made any reference to these errors because they are probably infrequent and have a negligible impact on the study results.

Starting in the second half of the 20th century, the central statistics offices of several countries began to link data from the census and the death registry in individual entries, making it possible to avoid this numerator/denominator bias in research. The central statistics offices provided researchers with the linked data set, with census variables along with the date and cause of death, after removing personal identifiers to protect privacy. The first countries to implement this methodology were the USA, in the 1960s, followed by Denmark, Finland, Norway, England, Wales, and France in the 1970s. Some countries established this linkage in the entire population, while others did so only with a representative census sample (Fox, 1989). Subsequently, some central statistics offices began to link data from general population surveys to that from the death registry (Duleep, 1989). Since then, and especially from the 1990s, the statistical offices of other countries have followed suit, implementing the methodologies needed to create a linked data set. In Spain and Italy, some regional statistics offices have been ahead of the national statistical offices in that regard. In Spain, this change was of great importance due to the decrease in the proportion of

death certificates containing the occupation of the deceased, which made it impossible to perform mortality studies according to this variable. The creation of a linked data set enabled researchers to continue investigating the relationship between socioeconomic status and mortality.

Population and educational level(*)	Nineties	First half of' the 2000s'	Nineties	First half of the 2000s'
Finland				
Low	1.97	2,08	1,59	1,84
Medium	1.6	1.61	1.22	1.35
High	1	1	1	1
Sweden				
Low	1.78	1.9	1.88	1.88
Medium	1.42	1.47	1.47	1.42
High	1	1	1	1
Norway				
Low	1.88	2.35	1.76	2.12
Medium	1.43	1.63	1.3	1.45
High	1	1	1	1
Denmark				
Low	1.77	2	1.62	1.85
Medium	1.46	1.53	1.24	1.34
High	1	1	1	1
France				
Low	2.23	2.37	1.64	1.8
Medium	1.6	1.7	1.17	1.38
High	1	1	1	1
Switzerland				
Low	1.95	2.22	1.43	1.54
Medium	1.41	1.52	1.11	1.13
High	1	1	1	1
Region of Madrid				
Low	1.55	1.56	1.37	1.3
Medium	1.3	1.27	1.18	1.23
High	1	1	1	1
Basque Country				
Low	1.49	1.51	1.25	1.39
Medium	1.2	1.16	1.12	1.22
High	1	1	1	1

(*) Low level: lower than primary studies, primary studies and the first cycle of secondary education. Intermediate level: second cycle of secondary education and studies after secondary education. High level: university studies.

Table 2: Mortality rate ratio according to educational level in various European populations. Subjects from 30 to 74 years. Nineties and first half of the two thousand years

Comparative studies with this type of data from various Western European countries have confirmed the smaller socioeconomic differences in mortality in southern compared to northwestern European countries (Mackenbach et al., 2015). These results are consistent regardless of whether

the measure of socioeconomic status is based on occupation or educational attainment. Table 2 shows these findings according to education. Data for Spain and Italy are generally collected and reported at a regional level, but this pattern is similar in analyses of data in the entire population.

Apart from avoiding the numerator/denominator bias, linked data sets are fertile grounds for testing hypotheses because of the large number of variables they contain. For example, based on a linkage between the 2001 population census and the death registry over the following ten years, implemented by the National Statistics Institute in Spain, a study found an acceleration in mortality decline during the 2008 economic crisis with respect to the previous period, which was more pronounced in people with low compared to high socioeconomic status (Table 3). In that study, the size of the dwelling (m^2) and the number of cars in the household, as recorded in the 2001 population census, were used as indicators of socioeconomic status (Regidor et al., 2016).

	APC in mortality rate		
Indicators of wealth	**(1) Precrisis (2004-2007)**	**(2) Crisis (2008-2011)**	**Effect size (2)-(1)**
Household floor space(m^2)			
Low (<72)	-1.7	-3.0	-1.4
Medium (72-104)	-1.7	-2.8	-1.1
High (>104)	-2.0	-2.1	-0.1
Household car ownership			
Low (no car)	-0.3	-2.3	-2.0
Medium (1 car)	-1.6	-2.4	-0.8
High (2 or more cars)	-2.2	-2.5	-0.3

Table 3: Table 3. Trends in premature mortality (lower than 75 years old) in Spain. Annual percent change (APC) in mortality rate before and during the 2008 economic crisis, and effect size according to indicators of wealth

Statistics offices apply the same interpretation of statistical confidentiality to linked and unlinked data, withholding information on variables they consider should not be disclosed. Furthermore, when population samples are linked to the death registry, statistics offices do not provide detailed information on some variables in the resulting data set. For example, they do not release the specific cause of death (at the level of the fourth digit of the International Classification of Diseases), but only the large disease group to which those causes belong.

Statistics offices consider that a low number of deaths from a specific cause cannot be used for statistical analysis. This paternalistic criterion is inadequate, since it is impossible to know the hundreds of possibilities that the availability of the specific cause of death would provide in answer to an infinite number of research questions. Statistics offices have absolute legitimacy when determining what data are subject to statistical confidentiality, based on whatever law, moral reasoning, or ethical criteria they deem appropriate to apply. But they lack legitimacy when they resort to supposedly technical criteria without adequate theoretical and empirical arguments to support them.

A common feature of linked mortality studies in many countries is that investigators often have little or no control over the linkage processes or the techniques used to build the database. To make

matters worse, they may also have little information about the processes and techniques used. This knowledge is essential because linkage errors, materialized in the impossibility of linking some individual records (missing links) or falsely linking records, could generate bias when they do not occur randomly. In fact, scientific journals often ask researchers for details about the process before publishing the results. It is understandable that statistics offices prevent access to individualized records with personal identifiers, but it is highly questionable that they rarely provide basic information on linkage techniques or the validity of the linkage made. It is also problematic that researchers do not participate in the planning of these processes, since it could allow them to assess the quality of the linkage, reducing errors and enabling a better interpretation of the study results drawn from these linked databases (Harron et al., 2017; Harron, 2022).

4 Other mortality studies with linked information

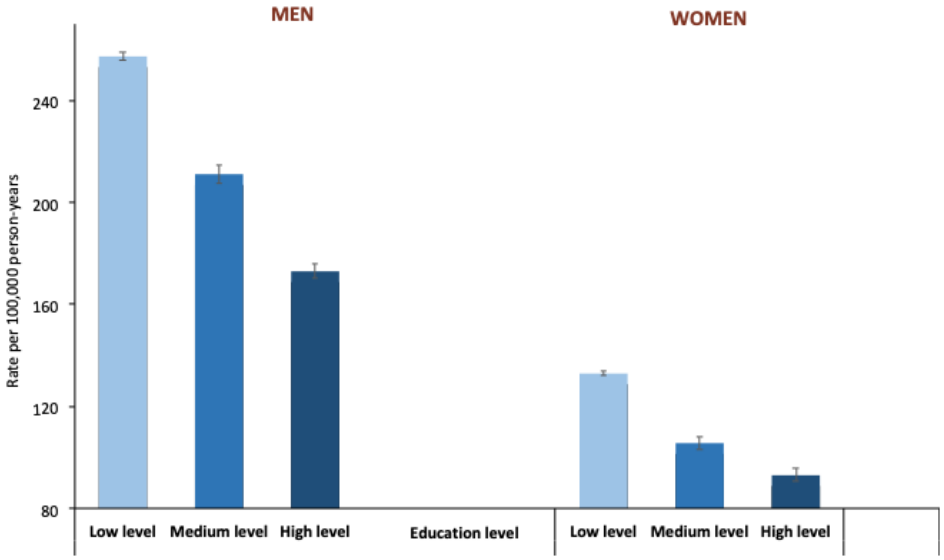
Another methodological option that avoids numerator/denominator bias in studies of socioeconomic characteristics and mortality is that developed by the National Statistics Institute of Spain in the continuous updating of population figures. By crossing numerous sources of administrative information, all the socioeconomic indicators collected in any of those sources are made available for each citizen. Subsequently, the National Statistics Institute links these variables to each deceased person in the death registry, thus avoiding the numerator/denominator bias when calculating mortality rates based on these attributes.

These population and mortality data, plus data on sex, age, and level of studies, were used to calculate the estimates that appear in Figure 1. In mortality from causes of death strongly related to alcohol, there is an inverse gradient according to the level of studies in both men and women, a result that is similar to those from other high-income countries. Likewise, in Spain and other countries, there is an inverse gradient in alcohol intake according to the level of studies in men, while in women, alcohol intake is most frequent in those with a high level of studies (Boyd et al., 2021; Donat et al., 2022). This discordance between the findings on alcohol-related mortality and alcohol intake in women has been called the alcohol harm paradox, the reasons for which are still unclear among the international scientific community.

5 Epidemiological follow-up studies

In epidemiological follow-up studies, a large amount of information is obtained over time from a large sample of research subjects. The data collection questionnaires include numerous variables obtained through personal interviews, physical tests, and blood and urine analyses. In this way it is possible to estimate the relationship between a wide variety of factors and the appearance of diseases and cause-specific mortality.

Researchers turn to statistics offices to find out the vital status of research subjects and, if deceased, the cause of death. Some offices make this information contingent on different agreements and protocols with researchers, so that investigators can obtain the information they need from the death registry. In Spain, to determine an individuals's vital status, researchers can request the data from either the National Statistics Institute or from the National Death Index, a database



1.Alcohol-induced pseudo-Cushing's syndrome (E24.4), alcohol use disorders (F10), alcoholic nervous system degeneration (G31.2), alcoholic polyneuropathy (G62.1), alcoholic myopathy (G72.1), alcoholic cardiomyopathy (I42.6), alcoholic gastritis (K29.2), alcoholic liver disease (K70), alcohol-induced acute pancreatitis (K85.2), alcohol-induced chronic pancreatitis (K86.0), maternal complication of foetal alcohol injury (O35. 4), foetus and newborn affected by maternal alcoholism (P04.3), foetal alcohol dysmorphic syndrome (Q86.0), blood alcohol finding (R78.0), accidental alcohol exposure poisoning (X45), intentional self-inflicted alcohol exposure poisoning (X65), alcohol exposure poisoning, undetermined intent (Y15) and evidence of alcohol involvement (Y90-Y91).Chronic hepatitis, not elsewhere classified (K73) and fibrosis and cirrhosis of liver (K74, except biliary cirrhosis -K74.4 to K74.5-). Cancers of lip, oral cavity and pharynx (C00-C13), oesophagus (C15), larynx (C32) and liver (C22). Tuberculosis (A15-A19, B90, K67.3, P37.0) and lower respiratory infection/pneumonia (A48.1, A70, J09-J15.8, J16, J20-J21, P23.0-P23.4), pancreatitis (K85-K86, except alcohol-induced pancreatitis ?K85.2 and K86.0-) and epilepsy (G40-G41).

Figure 1: Age-standardized mortality rate from causes closely related to alcohol in people aged 25 years, by educational attainment. Spain, 2016-2019

managed by the Ministry of Health; however, this index does not provide access to the cause of death.

Based on the results obtained in some of these investigations, it is possible to quantify the impact that certain circumstances or factors may have on the burden of disease and death in the population. For example, using estimates on the relationship that tobacco and alcohol consumption have with mortality from various causes of death, together with information on the prevalence of these behaviors, it is possible to estimate deaths potentially attributable to tobacco and alcohol from various causes of death, together with information on the prevalence of these behaviors, it is possible to estimate deaths potentially attributable to tobacco and alcohol. According to two studies, 14.0 and 4.0 of deaths were attributable to tobacco and alcohol use, respectively, over the first two decades of the 21st century in Spain (Ministerio de Sanidad, Servicios Sociales e Igualdad (MSSSI), 2016; Donat et al., 2022).

6 Conclusions

Mortality registries and cause-of-death statistics are of immense importance to clinical and public health research, far outweighing the value of existing morbidity records. In Spain, for example, the quality of the mortality data and their value have increased considerably by including new sociodemographic variables, such as educational level or occupation. In addition, the validity of registered causes of death has increased by better incorporating judicial and forensic information. However, fully capitalizing on the potential of these data is still not possible due to limitations in accessing some variables, such as the day or municipality of death. These restrictions derive from a very conservative interpretation of personal data protections by the ethics committees in some statistics offices. This rigid position should be reconsidered, and efforts made to design simple procedures so that scientists can access this information —with privacy guarantees but without tedious bureaucratic procedures. After all, the greatest benefits for the population derive, surely, from achieving an adequate balance between protecting people's right to privacy and carrying out research that improves their health and quality of life.

On the other hand, the usefulness of mortality registries for health research would be greatly improved if investigators requesting linkages to other registries could participate in some way in planning the linking procedures or at least receive detailed information about them. In this way, they could more adequately interpret the findings of their research and respond with confidence to the editors of the journals that disseminate their work.

References

Boyd, Jennifer, Clare Bambra, Robin C Purshouse, and John Holmes (2021). Beyond behaviour: How health inequality theory can enhance our understanding of the 'alcohol-harm paradox'. *International Journal of Environmental Research and Public Health 18*(11), 6025.

Checkoway, Harvey, Neil Pearce, and David Kriebel (2004). *Research methods in occupational epidemiology*, Volume 34. Monographs in Epidemiology.

Cosby, Arthur G, Tonya T Neaves, Ronald E Cossman, Jeralynn S Cossman, Wesley L James, Neal Feierabend, David M Mirvis, Carol A Jones, and Tracey Farrigan (2008). Preliminary evidence for an emerging nonmetropolitan mortality penalty in the united states. *American Journal of Public Health 98*(8), 1470–1472.

Donat, Marta, Gregorio Barrio, Juan-Miguel Guerras, Lidia Herrero, José Pulido, María-José Belza, and Enrique Regidor (2022). Educational gradients in drinking amount and heavy episodic drinking among working-age men and women in spain. *International Journal of Environmental Research and Public Health 19*(7), 4371.

Drever, Frances and Margaret Whitehead (1997). *Health inequalities: decennial supplement*. Decennial supplement. Series DS No. 15. London: The Stationery Office, London.

Duleep, Harriet Orcutt (1989). Measuring socioeconomic mortality differentials over time. *Demography 26*(2), 345–351.

Fox, AJ (1989). Longitudinal studies based on vital registration records. *Revue D'épidémiologie et de Santé Publique 37*(5-6), 443–448.

Harron, Katie (2022). Data linkage in medical research. *BMJ Medicine 1*(1).

Harron, Katie L, James C Doidge, Hannah E Knight, Ruth E Gilbert, Harvey Goldstein, David A Cromwell, and Jan H van der Meulen (2017). A guide to evaluating linkage quality for the analysis of linked data. *International Journal of Epidemiology 46*(5), 1699–1710.

Hill, AB et al. (1955). Snow-an appreciation. *Proceedings of the Royal Society of Medicine 48*(12), 1008–12.

Lynge, Elsebeth (2011). Occupational mortality. *Scandinavian Journal of Public Health 39*(7_suppl), 153–157.

Mackenbach, Johan P, Ivana Kulhánová, Gwenn Menvielle, Matthias Bopp, Carme Borrell, Giuseppe Costa, Patrick Deboosere, Santiago Esnaola, Ramune Kalediene, Katalin Kovacs, et al. (2015). Trends in inequalities in premature mortality: a study of 3.2 million deaths in 13 european countries. *Journal of Epidemiology and Community Health 69*(3), 207–217.

Mackenbach, Johan P, Anton E Kunst, Adriënne EJM Cavelaars, Feikje Groenhof, and Jose JM Geurts (1997). Socioeconomic inequalities in morbidity and mortality in western europe. *The lancet 349*(9066), 1655–1659.

Meijer, Mathias, Jeannette Röhl, Kim Bloomfield, and Ulrike Grittner (2012). Do neighborhoods affect individual mortality? a systematic review and meta-analysis of multilevel studies. *Social Science & Medicine 74*(8), 1204–1212.

Ministerio de Sanidad, Servicios Sociales e Igualdad (MSSSI) (2016). Muertes atribuibles al consumo de tabaco en españa, 2000-2014. Technical report, Ministerio de Sanidad, Servicios Sociales e Igualdad, Madrid.

Politi, Julieta, Juan-Miguel Guerras, Marta Donat, María J Belza, Elena Ronda, Gregorio Barrio, and Enrique Regidor (2022). Favorable impact in hepatitis c–related mortality following free access to direct-acting antivirals in spain. *Hepatology 75*(5), 1247–1256.

Regidor, Enrique, Fernando Vallejo, José A Tapia Granados, Francisco J Viciana-Fernández, Luis de la Fuente, and Gregorio Barrio (2016). Mortality decrease according to socioeconomic groups during the economic crisis in spain: a cohort study of 36 million people. *The Lancet 388*(10060), 2642–2652.

Spanish Journal of Statistics
Vol. 4 No. 1 2022, Pages 99–105
doi:https://doi.org/10.37830/SJS.2022.1.08

Official Statistics

The Statistics on Causes of Death: characteristics and improvements

Margarita García Ferruelo[1], M. Rosario González García[2]

[1] Instituto Nacional de Estadística, margarita.garcia.ferruelo@ine.es
[2] Instituto Nacional de Estadística, rosario.gonzalez.garcia@ine.es

Received: November 3, 2022. Accepted: December 26, 2022.

Abstract: The Statistics on Causes of Death is a key tool for the Public Health. This article describes the complex process of the statistics, the advances achieved in recent years, such as the implementation of an international automatic system for coding multiple causes of death and for the selection of the underlying cause (IRIS) or the improvement in obtaining the external causes of death, as well as its usefulness for the epidemiological studies and health research. It is also discussed some of the lessons learned during the worst pandemic period, that, without any doubt, have highlighted the need of a more efficient method to get information through the implementation of an Electronic Death Certificate. And, finally, it is proposed to collect other variables of interest for the analysis of the causes of death using available administrative sources.

Keywords: underlying cause of death, multiple cause of death, International Classification of Diseases, coding, IRIS System, IMLweb, administrative data

MSC: 62P10, 62P25, 91B82

1 Introduction

This work is part of the review process within the strategy of the European Statistical System (ESS) to implement one of the recommendations of the Code of Good Practices (CoP) through the evaluation of key statistics by external experts. For this purpose, a session has been included within the framework of the XXXIX Spanish conferences on Statistics and Operation Research and of the XIII Conference of Public Statistics to present the Statistics on causes of death methodology, with two speakers from Complutense University of Madrid and the National School of Health of the Carlos III Health Institute.

The Statistics on Cause of Death is an annual operation legally supported by Regulation (EU)

328/2011 of the Commission, of April 5, 2011, which develops Regulation (EC) 1338/2008 of the European Parliament and of the Council regarding community statistics on public health and safety and health at work in the field of statistics on causes of death, which establishes the commitments acquired by the Member States and Eurostat in relation to the statistics on Causes of Death, Commission Regulation (EC) 328/2011 specifies the scope, the definitions, the list of variables, the reference period, the deadline to send the data and the set of variables and metadata to be provided to Eurostat.

The main objective of the statistics is to know the pattern of mortality. Information on causes of death is the traditional study that has allowed us to know better the health situation since the 19th century and continues to have great potential as a tool for decision-making in public health. This statistic began based on a list of 5 diseases. Later, the cause of death was classified according to a list of 99 diagnoses, origin of the International Classification of Diseases (ICD) of the World Health Organization (WHO) and nowadays there is the tenth revision of this classification (ICD- 10), containing more than 12,000 diagnostic codes, which allows classifying the mortality causes with a high degree of clinical specification (OPS 1995).

The purpose of the causes of death data is to know the pattern of mortality associated with each sex, group of age and geographical areas, as well as its evolution over the time. The key variable is the underlying cause of death, which is selected following the ICD-10 criteria from the diseases informed by the physician in the Medical Death Certificate. (INE 2020)

The study of mortality based on the underlying cause of death constitutes a great value tool; in fact, it has allowed to detect changes in the trend of certain diseases. Four decades ago, cardiovascular diseases were the main cause of almost half of deaths, but their relative weight has been decreasing since then (26% in 2021) and, although they remain the leading cause of death, they have been displaced for other causes such as tumours, dementia, or Alzheimer's disease. These last two diseases mentioned have been the main cause of more than 30,000 deaths in 2021, 66% more than in 2005. And with the outbreak of the pandemic, this pattern has undergone some new changes.

Another important contribution of the underlying cause of death, with relevant social and economic impact, is to identify the "guilty" causes of premature deaths and quantify the years of life lost (INE 2021) due to these causes. Violent deaths —that is, traffic accidents, poisonings, falls or other accidents, suicides, or homicides— are the causes that deprive of more years of life. People who die of a violent cause, live, on average, 27 years less than the years they would theoretically have to live. And if we refer to tumours, we would be talking about 12 years of life lost. However, if Statistics on causes of death were limited to the study of the underlying cause, the opportunity to identify the associations of the most frequent pathologies and to know the true dimension of mortality would be lost. An aging population, like the case of Spain, have a significant increase of chronic diseases in which several pathologies converge, although they do not lead to death, they can contribute to hastening it. This is the case, for example, of diabetes or hypertension.

When one of these diseases is informed on the certificate, its selection as underlying cause will depend on the rest of the diseases that have also been mentioned. For example, if hypertension and COVID-19 appear together, depending on the order in which both diseases are informed, the ICD-10 selection rules may penalize hypertension by displacing it in favour of COVID-19. Thus, the opportunity to know and analyse the true dimension of hypertension in mortality would be lost, apart from to identify the most frequent associations with other pathologies and, therefore, to adopt more effective prevention action measures. Another recent example of the importance of the multiple causes of death availability (INE 2021) has been their effect on respiratory diseases with the irruption of the COVID-19. Respiratory system diseases have been displaced as underlying cause by the COVID-19 when both were informed in the same certificate. Thanks to the multiple causes, the

impact of respiratory diseases on mortality was not lost and it was possible to obtain information on the complications derived from COVID-19 and the comorbidities presented by the people who died. Apart from that and without any doubt, the usefulness of the causes of death data is also linked to the continuous quality improvement to be able to respond to new demands.
In recent years, great efforts have been made to improve the quality of these statistics. An international automatic system for coding (IRIS) has been implemented and, the Institutes of Legal Medicine and Forensic Sciences have been incorporated as a source of information of the causes of deaths in case of deaths with judicial intervention.
In terms of responding new demands, the Causes of Death Statistics offer very precise and high-quality information, however it does not provide information in the short term as evidenced during the COVID-19 crisis. This fact is due to the administrative steps required by the Civil Registers before sending the Death Certificate (paper document with the needed information) to the NSI and its complex process of mass scanning and OCR review. The solution is found in the Electronic Death Certificate, a project framed in the digitization of the Civil Register as established by Law 20/2011 and in which the NSI collaborates with the Civil Register. Nevertheless, Causes of Death Statistics is the source of numerous epidemiological studies and health research. The pandemic, that has given visibility to epidemiology, has also made this Statistics more visible, proof of this is the huge number of accesses to its results. In 2019 there were around 260,000 queries and since 2020 they are approaching two million.

2 Statistical processes in the pandemic context

When the COVID-19 pandemic emerged, the information on mortality that was becoming known referred to estimated mortality data (Daily Mortality Monitoring System - MoMo - and Experimental Statistics on weekly estimations of deaths) and mortality declared to the National Epidemiological Surveillance Network (RENAVE). On the one hand, the mortality declared by COVID-19, provided by the Ministry of Health from the registers of the Autonomous Communities, met the criterion of deaths with a positive COVID test without distinguishing the direct cause of these deaths and, on the other hand, the estimated mortality made it possible to know the excess mortality, but without having information on the causes of death.

This daily information was essential for the epidemiological surveillance of the pandemic; however it was not enough to know the real impact of COVID-19 on mortality. The source that would provide the best estimation of mortality attributable to the pandemic would be the Statistics on Causes of Death.
The Death Certificate, document in which the physician informs the sequence of diseases that finally lead to death, is the source of information for the Causes of Death Statistics.

The Civil Registers send the death certificates to the Provincial Delegations of NSI to be scanned. These certificates are designed for optical recognition. Taking into account that the terms to be recognized are diseases, their recognition is quite complicated, for that reason it is necessary to develop a diseases dictionary from the information provided by physicians in the death certificates over the years. Currently, this dictionary has around 170,000 different terms and the recognition success level is around 85-90Some difficulties had to be faced during the health crisis, such as the collapse in the Civil Registers during the first wave of the pandemic, a greater volume of deaths

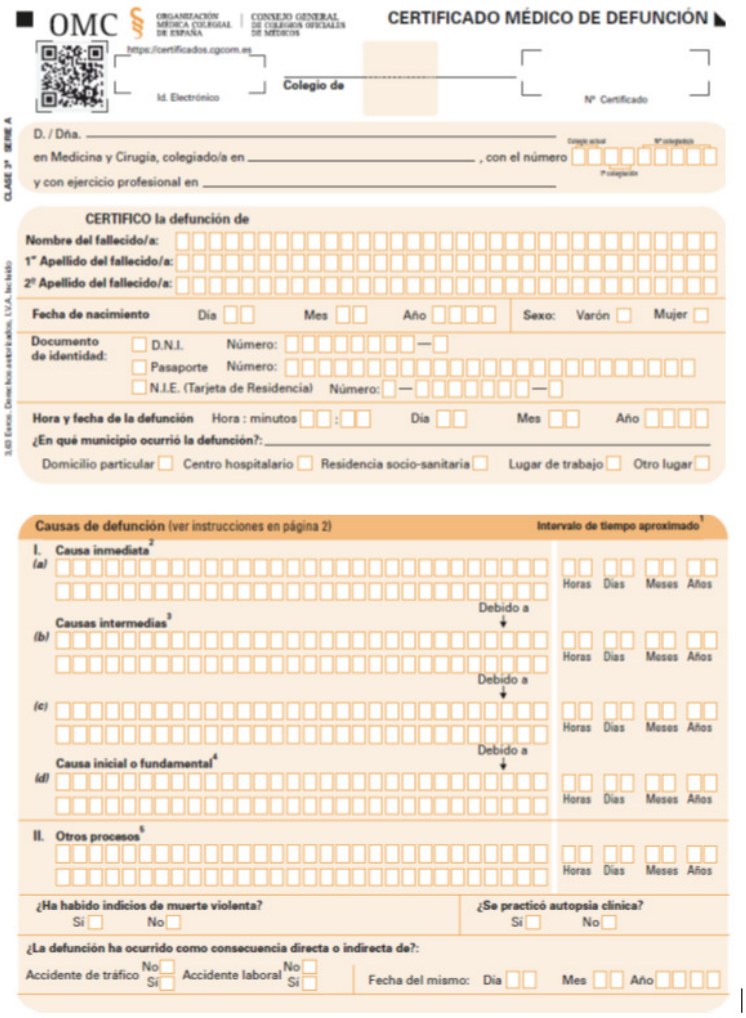


Figure 1: Medical death certificate

and, the most important, dealing with a new disease in the process of codification. In March 2020, when the first deaths due to covid occur, the Optical Recognition System did not recognize the terms referring to this new disease because of not being included in the dictionary. For that reason, coders from the Autonomous Communities, within the collaboration agreements with the NSI for the work of the Statistics, had to compare exhaustively the recognition result with the certificate images and correct any errors. As soon as possible, the terms used in the first certificates referring to COVID were included in the dictionary (around 145 different expressions).

Once the information on the certificates has been scanned and reviewed by coders, the next step is to identify the initial cause of death, applying the ICD-10 rules in the sequence of diseases informed by the physician in the Death Certificate, and finally determine the underlying cause of death. The Volume 2 of ICD-10 describes these rules according to medical criteria and applying medical logical relations between the diseases informed in the certificate. This is undoubtedly the most complicated part of the process and requires continuous medical supervision and advice.

Until 2013 the coding process was manual, and the emerging doubts were solved through a forum created for this purpose between the NSI and the Regional Mortality Registers. However, the manual

coding of causes of death, apart from being affected by the same problems as other manual coding, that is, it requires time, requires numerous human and financial resources, and is very sensitive to systematic errors by coders, presents another specific problem: the selection of the underlying cause of death must be based on the guidelines described in the ICD and these are characterized by their complexity and, above all, their numerous exceptions. In the context of improving the quality of this statistics, the IRIS automatic system was implemented in Spain with the 2014 data, as a consequence, a fundamental advance has been achieved in terms of punctuality and comparability of the information, both at national and international level (Carrillo and González, 2016). IRIS is an intelligent system prepared to work with the WHO medical death certificate model, which is in force in Spain. It works using algorithms based on codes assigned to medical terms, on causal medical relations and on the application of selection rules in accordance with the guidelines of Volume 2 of the ICD-10. To get an idea of how valuable this tool, it is enough to mention that the number of relations between diseases programmed exceeds 29 million.

Although Iris is linked to the names of its two co-founders, Lars Age Johansson and Gérard Pavillon, as well as the rest of the Core Group members, it is important to highlight that the automatic system success is also a consequence of the involvement of all countries that, to a greater or lesser extent, are part of the project. The development of a tool that guarantees the comparability of causes of death statistics at a world level could not be understood without a coordinated international cooperation.

IRIS is a language-independent international software because it works with codes and this implies the development of a dictionary in the national language that associates an ICD-10 code with each disease. Currently, the Spanish dictionary consist of 167,000 standardized terms associated with their ICD-10 code. The COVID-19 pandemic brought new challenges for IRIS since their decision tables had to be modified to include the new disease.

The inclusion of COVID-19 in the statistical process began from the moment that WHO incorporated the new disease into the ICD-10, assigning two different codes to distinguish between COVID-19 virus identified (with positive test) and COVID-19 virus not identified (suspected), thereby making it possible to determine the mortality directly caused "by" confirmed COVID and "by" suspected COVID.

Apart from that, as mentioned before, in addition to the underlying cause, the Statistics also provides information on the diseases that have contributed to the death and have been reported by the physician in the Death Certificate (multiple causes). In this way, deaths with the presence of COVID-19 without being this disease the underlying cause, that is, mortality "with" confirmed COVID and "with" suspected COVID could be identified.

In the same way, multiple causes gave information about complications due to COVID-19 and the comorbidities of people who died of this disease. Respiratory failure and pneumonia were the most frequent complications reported on the Death Certificates of people who died âĂIJofâĂİ COVID-19, both identified and not identified virus. In terms of comorbidities, hypertensive disease and renal failure were the main comorbidities in COVID-19 virus identified mortality and dementia in the case of COVID-19 virus not identified mortality. The results showed that there were 60,358 deaths due to COVID-19 virus identified in 2020 and another 14,481 deaths due to suspected COVID-19 due to having symptoms compatible with the disease (COVID-19 virus not identified). In addition, physician certified 8,275 deaths due to other causes, but having COVID-19 as comorbidity contributing to the death. In 3,770 cases the physicians identified the virus and in 4,505 cases they suspected its presence due to having symptoms compatible with the disease.

The other action to improve quality that deserves to be highlighted is the incorporation of the Institutes of Legal Medicine and Forensic Sciences as a source of information on deaths with judicial intervention. Although quantitatively judicial deaths have little relative weight (between 5-6% of

mortality), they are very relevant qualitatively because most of them are premature and avoidable deaths.
Up to 2019, the main source of this information was the judicial authorities and around 40% of reported judicial deaths had incomplete information on causes of death. This implied that the Mortality Register of Autonomous Communities, depending on their available human resources, had to improve this information by contacting the coroners, creating comparability problems between regions and time series.
The NSI has developed a software in order the coroner to provide the information on deaths with judicial intervention, the tool is in line with the statistical needs, that is, focused on obtaining the ICD-10 code and guarantees coverage and comparability. The incorporation of the Institutes of Legal Medicine and Forensic Sciences in the circuit of deaths with judicial intervention information, as recommended by numerous studies, has led to a very significant improvement in quality. In addition, the collaboration of the Institutes of Legal Medicine and Forensic Sciences positions them as a key source of information in the circuit of official statistics on causes of death, reinforcing their social and health projection. The software also provides reports with the same format and common international health language that facilitates comparability between Institutes of Legal Medicine and Forensic Sciences and the preparation of its annual reports.

3 Some examples of epidemiological studies and health research based on the Statistics on Causes of Death

As mentioned in the introduction, the Statistics on Causes of Death is the source of information for numerous epidemiological studies and health research. In order to show the potential offered by the Statistics, it is presented below some of the numerous research projects based on monitoring the mortality of specific cohorts' members:
- Monitoring of the mortality of the cohort of patients affected by the Toxic Oil Syndrome. The Carlos III Health Institute has been monitoring for 37 years the Toxic Oil Syndrome epidemiological cohort with 20,643 affected since the beginning of the outbreak. The NSI annually provides the vital status of those affected and the cause of death in case of decease to identify a mortality pattern of this epidemic. Up to now, more than 5,000 deaths have been identified.
- EPI-CT project on potential health effects of exposure of children and adolescents to ionizing radiation during TAC scans carried out by the Centre for Research in Environmental Epidemiology (CREAL) of Catalonia.This project came to light through a collaboration agreement between the NSI and the Centre for Research in Environmental Epidemiology (CREAL) of Catalonia. The objective was to determine the potential effects of the ionizing radiation doses applied during the TAC scans, to see the possible effects and, according to the results, to reduce and optimize these doses. There was first a pilot study with 10,000 patients and later with more than 200,000 patients. The cohort was crossed with 27 mortality data files.
- BIFAP Project: Database for Pharmacoepidemiologic Research in Primary Care. The Spanish Agency for Medicines and Health Products requires information on mortality by cause to evaluate the safety and effectiveness of new medicines.
- Prospective study of the Health Research Institute of the Hospital La Paz on the follow-up during 12 months of a 2,000 individual cohort with a suicide attempt. This study aims at assessing the incidence of re-attempted suicide and identify risk factors. For this purpose, this cohort has been crossed with the mortality data files. - Mortality study among medical professionals. The Council

of Official Colleges of Physicians has prepared a mortality study among the medical professionals. This initiative analyses, for the first time in Spain, the expectancy life and causes of death of Spanish physicians, based on the data from its register and mortality data from the NSI. The study analyses the evolution of the number of medical professional deaths in the period 2005-2014 and the main causes of death in this group.

4 Other user demands

The availability of additional variables of interest for the mortality analysis is a reiterated demand from researchers.
Based on 2016 data, variables such as educational level and activity status have been obtained from administrative sources and incorporated into the microdata file. Occupation, at the one-digit of the National Classification of Occupations (CNO 11) level, has been assigned in case the deceased over 16 years of age was working at the time of death.
For this purpose, the information from the pre-census population files that were prepared for the elaboration of the 2021 Population and Housing Census has been used (INE 2014).
The sources of information have been the Mutual insurance companies of civil servants (MUFACE, MUGEJU and ISFAS), the file of current contracts from the Public Employment Service (SEPE) and the 2011 and 2001 Censuses.
Also from 2010 data, the geographical coordinates of the census section corresponding to the residence of each deceased are made available to the researchers in the microdata file.
For the future, other information of interest, such as the average income of the census section, it will be incorporated into the microdata files for researchers as soon it is available.

References

Carrillo, Jesús and María del Rosario González (2016). Iris: Codificador automático internacional de causas de muerte. Technical report, Instituto Nacional de Estadística.

Instituto Nacional de Estadística (INE) (2014). Método de asignación de nivel educativo, relación con la actividad laboral y ocupación. Technical report, Instituto Nacional de Estadística (INE).

Instituto Nacional de Estadística (INE) (2020). Certificado médico de defunción/ boletín estadístico de defunción. Technical report, Instituto Nacional de Estadística (INE).

Instituto Nacional de Estadística (INE) (2021). Metodología de los resultados detallados. Technical report, Instituto Nacional de Estadística (INE).

Organización Panamericana de la Salud (OPS) (1995). Clasificación estadística internacional de enfermedades y problemas relacionados con la salud – 10 revisión. Technical report, Washington, D.C.

SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 107–115
doi:https://doi.org/10.37830/SJS.2022.1.09

OFFICIAL STATISTICS

Mortality statistics

Enrique Regidor

Subdirección General de Información Sanitaria, Ministerio de Sanidad, eregidor@sanidad.gob.es
Received: November 3, 2022. Accepted: December 26, 2022.

Abstract: The creation of civil registries, together with the obligation to report information on the deceased from the death certificate, have enabled the monitoring of various population health indicators. Data from death certificates, as compiled and disseminated by central statistics offices, are used to estimate different measures, most classically infant mortality and life expectancy. However, in high-income countries, infant mortality is no longer considered an appropriate measure of population health due to its low magnitude. From the health system perspective, the adoption of the International Classification of Diseases and Causes of Death was a crucial milestone in population health statistics, shedding light on the diseases responsible for most deaths and the trends in causes of death over time. Morbidity statistics and public health surveillance systems have important objectives, but they do not allow adequate monitoring of the frequency of diseases and other health problems, nor can they quantify diseases' impact on population health. On the other hand, statistics on cause of death do provide this information thanks to the combination of two features: the exhaustiveness of the data they collect and the objective nature of the phenomenon they quantify.

Keywords: Civil registries, death certificate, central statistics offices, infant mortality, life expectancy, causes of death, public health surveillance

MSC: 62-03, 62B86, 62P10, 62Q05

1 Introduction

The emergence of the State entailed tremendous political, economic, and administrative advances in human societies. From an administrative point of view, the obligation to record vital events —births, deaths, stillbirths, marriages, and divorces— was established in order to know at all times what was the situation in which its citizens were in terms of their vital and marital status. Most European countries created such systems over the 19th century, although in some, such as Sweden, they were

introduced as early as the 18th century (Mackenbach, 2020).

Prior to the 19th century, information on these phenomena was often available from parish records of baptisms, burials, and marriages. The information contained in these registries has been used to estimate changes in the structure of the population, such as birth, death, and marriage rates. But its uneven implementation does not provide a comprehensive vision of the evolution of these phenomena during this period. And a sufficiently valid comparison of these estimates according to different sociodemographic characteristics of citizens is not possible either.

Since the creation of civil registries, central statistics offices of the countries have been in charge of managing the data obtained in these registries for their compilation and subsequent dissemination. This has made it possible to rigorously assess trends in births, deaths, and marriage rates and to assess variations in the magnitude of these rates according to sociodemographic and geographic characteristics. Data on deaths in the civil registries come from the death certificates, and it is this information that central statistics offices compile and disseminate for their mortality statistics.Estimates of total and cause-specific mortality rates require information on the number of deaths (numerator) and the population at risk (denominator). In unlinked mortality studies, the numerator and denominator come from different sources, so there may be a numerator/denominator bias when estimating mortality rates according to certain individual attributes. This bias does not occur in linked mortality studies, in which data from the census or general population surveys are linked to vital records, and in the case of death, to the date and cause of death. However, regulations to protect individuals' confidentiality greatly limit the use of linked and unlinked mortality statistics for scientific research, whether due to the regulations themselves or because of the restrictive interpretations thereof by some statistical offices not always sufficiently argued. On the other hand, some methodological developments by these offices are of enormous relevance, for example, the linkage between socioeconomic indicators and mortality by the National Statistics Institute of Spain, which enables the study of the relationship between socioeconomic factors and mortality and its variation over time.

2 Classical population health indicators

In high-income countries, information on deaths, compiled and disseminated for central statistics offices, has been used to document the enormous reduction in infant mortality rate and the rise in life expectancy over the 20th century. For example, it is known that the infant mortality rate in different Western European countries ranged from 80 to 210 deaths per 1,000 live births in 1990. By the end of the century, these rates had converged and dropped precipitously across the region, standing at 3 to 6 deaths per 1,000 live births. In Spain, infant mortality fell from 204 to 4 per 1,000 live births between 1900 and 2000 (Gómez, 1991; Viciana, 2003). This reduction, together with parallel advances in medical treatments and living conditions, led to a dramatic increase in life expectancy, from 35.7 years in 1900 to 79.3 years in 2000 (Goerlich and Pinilla, 2006). Similar trends were observed in surrounding countries with a similar socioeconomic situation (MSC 2005).

Traditionally, infant mortality and life expectancy have been the classic indicators of population health status. However, in high-income countries, the low rates of infant mortality have reduced their usefulness as a sentinel indicator to reflect population health. On the other hand, life expectancy

continues to be an ideal measure in that regard. Life expectancy at age x is the average number of years that a person of that age is expected to continue to live. This is a hypothetical measure since it does not measure the actual chances of survival. Its calculation is based on current mortality rates, which logically are subject to changes over time. Its fundamental advantage is that it can be used to compare different regions or countries and to observe their evolution over time, since it is not influenced by differences in the age structure of the populations being compared. In fact, the estimation of life expectancy has revealed the reversal of a key trend in Western Europe over the 20th century. People born in southern European countries in the late 19th and early 20th centuries could expect to live to around 40 years of age in Spain, Portugal, Italy, and Greece, lagging far behind their northwestern neighbors. However, the generation born in southern Europe the 1960s had a similar life expectancy as their northern peers, and by the 1980s life expectancy in southern European countries had surpassed that of some countries of higher latitudes (Mackenbach, 2020).

In the 21st century, life expectancy continues to be an ideal indicator for monitoring population health worldwide, since it reflects the impact that numerous health problems have on population mortality (WHO, 2022b). Figure 1 shows the evolution of life expectancy at birth in Spain from 2001 to 2020.

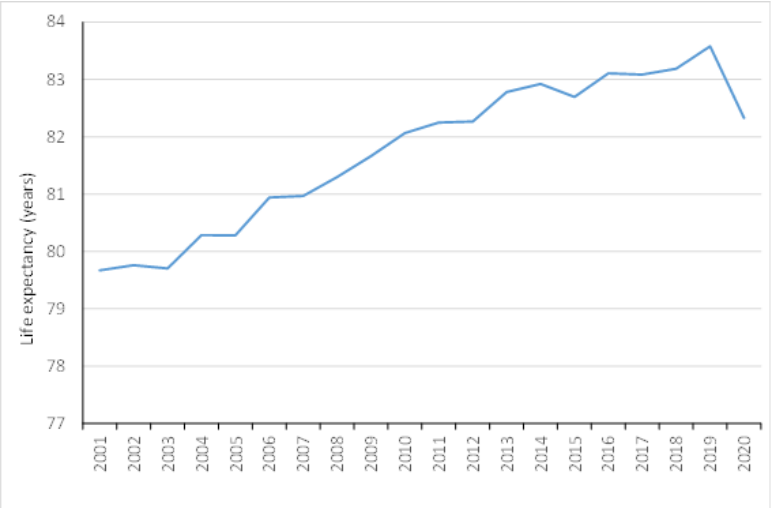


Figure 1: Life expectancy at birth. Spain, 2001-2020

Several years showed year-on-year decreases in life expectancy at birth, specifically 2003, 2005, 2015, and 2020, while this indicator hardly changed in 2007 and 2012. The increase in recorded deaths accounts for the decrease or maintenance of life expectancy each year compared to the previous one. The increase in deaths in the second half of 2003 was due to the heat wave that occurred that summer. In the rest of the years, except for 2020, the spikes in deaths were most likely the result of the increased intensity and/or duration of the influenza virus. The decrease in life expectancy from 2014 to 2015 —not only in Spain but also in most countries in the northern hemisphere— is notable (Ho and Hendi, 2018). In the 50 previous years, there had been no comparable reductions in life expectancy from one year to another; this decline is probably attributable to the combined effects of the two phenomena mentioned above: a particularly virulent flu season and a heat wave. As for

2020, the striking decline was due to the increase in deaths as a result of the COVID-19 pandemic.

The year-on-year changes in the number of deaths are also evident in the crude death rate, as shown in Figure 2. After all, life expectancy at birth can be considered a snapshot of the mortality of the population in a given period.

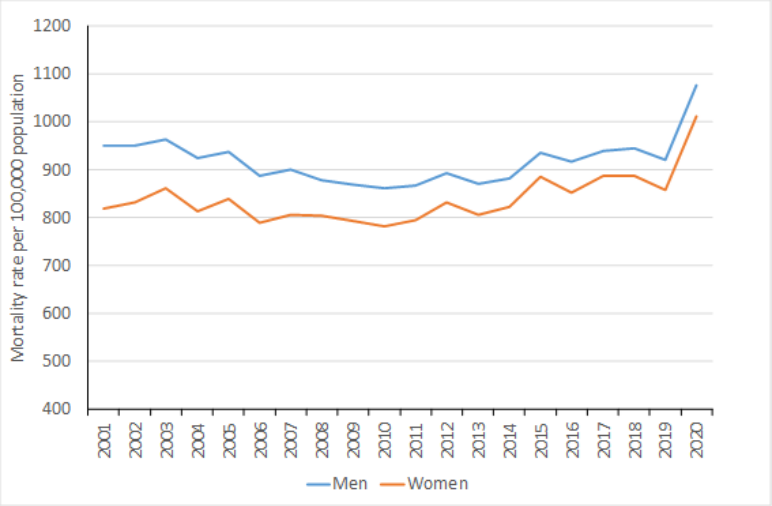


Figure 2: Crude death rate per 100,000 population. Spain, 2001-2020

3 International Classification of Diseases and Causes of Death

From a public health perspective, standardized recording of the cause of death in the civil registry was a milestone, providing valuable insight on the diseases responsible for the most deaths. A second great achievement was the establishment of an international classification of causes of death, which provided standard criteria for central statistics offices to use in the collection, processing, and classification of information, and for the presentation of death statistics.

The need for a common system to classify causes of death was recognized at the first International Statistical Congress, held in Brussels in 1853. The Congress asked William Farr and Marc d'Espine to prepare a standard classification of causes of death applicable to all countries. At the next Congress, in Paris in 1855, Farr and d'Espine presented two separate lists based on very different principles; the Congress adopted a version representing a compromise between the two. In successive years, that list was revised according to Farr's criteria, although it never gained universal acceptance. The first International List of Causes of Death would have to wait until Jacques Bertillon's proposal was approved by the International Institute of Statistics in 1893. This classification adopted Farr's criteria for distinguishing systemic diseases from those with a precise anatomical location. In 1899, the International Statistical Institute agreed to revise the list every 10 years in order to reflect advances in medical science and statistical procedures. Numerous revisions were undertaken throughout the 20th century. The most recent, known as the International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10), was released in 1999. Since the sixth revision in 1948, when the World Health Organization (WHO, 2022a) took over its develop-

ment, the Classification has also contained a list of causes of morbidity (Alderson 1988, WHO, 2022a).

The list of causes of morbidity has given rise to various classifications in different medical specialties and areas of the health system. Possibly the best known and most used is the International Classification of Diseases, Ninth Revision, Clinical Modification, which was published in the USA to classify diagnoses and reasons for health care contacts. It is a classification of both diseases and procedures, used to code clinical information derived from health care, mainly in hospitals and other specialized medical care settings.

Mortality statistics remain the most suitable indicator for monitoring trends in the burden of disease over time and from one place to another. However, this is not, nor can it be, the objective of the morbidity statistics derived from health services data. These statistics reflect the burden of disease that the health system has to deal with at any given time, but this information is not suitable for reflecting the trend in the frequency of diseases. This important distinction can be explained by the increase in healthcare resources, the growing size of the population covered by all types of healthcare insurance, improvements in diagnosis, and new indications for care. Moreover, some patients suffering from a disease may not use any health services, while others could be registered several times in the information systems of different health services. Likewise, patients could present to different health centers whose information systems are not interconnected. These circumstances, inherent to the health system of any country, make it impossible to ascertain whether the variations observed in morbidity reflect changes in the frequency of disease or only variations in the use of health services. Morbidity statistics have similar limitations in terms of their aptness for estimating geographic variations in disease frequency.

These imprecisions do not exist in death statistics, because a person dies only once. In addition to the objective nature of the phenomenon, the exhaustive nature of vital statistics makes mortality indicators very useful for monitoring diseases and other health problems and for establishing public health priorities. It is true that morbidity statistics from population-based disease registries enable an adequate comparison of the incidence —new cases— of diseases. But various limitations related to the clinical characteristics and diagnostic criteria for different pathologies, together with the inefficiencies inherent to these registries, mean that relatively few population-based disease registries exist.

4 Main causes of death

The available data on cause of death have revealed important changes in the percentage of deaths attributable to specific diseases. In high-income countries, the proportion of deaths from infectious diseases —tuberculosis, intestinal infectious diseases, influenza, and pneumonia— declined in the first half of the 20th century, while deaths from cardiovascular diseases and accidents rose. During the second half of the century, the contribution of these causes to total mortality decreased, while cancer deaths increased. Recent decades have also seen a rise in deaths from mental and neurological diseases, namely Alzheimer's disease and Parkinson's disease. However, since mortality statistics cannot reflect the burden of diseases and health problems that are not lethal, mental illnesses impose a much larger burden to the population than that reflected in these statistics.

Adequate estimation of mortality from cause of death requires high-quality information on this variable in death certificates. This quality has gradually improved. The quality of information on cause of death in national mortality registries has gradually improved. In Spain, for example, through the better inclusion of judicial and forensic information in cases where professionals from these fields are involved. This was not always the case: into the 1980s and 1990s, specific registries of mortality from AIDS or from acute reactions to drugs were necessary to adequately monitor the impact of these problems on population health (De la Fuente et al., 1995; Brugal et al., 1999). Another indicator of the quality of mortality statistics by cause of death is through the proportion of deaths that cannot be assigned to a specific cause of death. In Spain, the figure is around 2.

Currently, the leading causes of death in in high-income countries are cancer, heart disease and cerebrovascular disease. These three causes of death represent half of the deaths, as can be seen in Table 1 referring to Spain. The exception was the year 2020, since deaths from COVID-19 represented the third leading cause of death.

Cause of death	**Codes**	**2017**	**2018**	**2019**	**2020**
All causes		424,523	427,721	418,703	493,776
Malignant neoplasms (cancer)	C00-C97	109,073	108,526	108,867	108,533
Diseases of heart (heart disease)	I00-I09,I11, I13, I20-I51	85,143	83,744	80,444	82,309
COVID-19	U07.1, U07.2	..	..	..	74,839
Cerebrovascular diseases (stroke)	I60-I69	26,937	26,420	25,712	25,817
Alzheimer disease	G30	15,201	14,929	14,634	15,571
Chronic lower respiratory diseases	J40-J47	15,486	14,607	13,808	12,734
Influenza and pneumonia	J10-J18	11,397	12,267	10,843	11,676
Accidents (unintentional injuries)	V01-X59,Y85-Y86	11,502	11,530	11,827	11,297
Diabetes mellitus	E10-E14	9,773	9,921	9,644	9,662
Nephritis, nephrotic syndrome and nephrosis	N00-N07, N17-N19, N25-N27	6,757	7,269	7,369	7,517
Hypertensive disease	I10, I12, I15	4,787	4,998	4,912	6,239
Parkinson disease	G20-G21	4,656	4,583	4,615	5,008
Chronic liver disease and cirrhosis	K70, K73-K74	4,236	4,001	4,021	3,976
Suicide	X60-X84+Y87.0	3,680	3,541	3,673	3,941
Septicemia	A40-A41	3,800	3,040	2,885	2,745

Table 1: Number of deaths for the 15 leading causes of death for the total population. Spain, 2017-2020. (Cause of death and codes based on International Classification of Diseases, 10th Revision)

Many criteria are considered in the development of the ICD, for instance disease etiology, anatomical location, and clinical manifestations. Some diseases, including several infectious diseases, are even encompassed under other groups because one criterion is prioritized over another. Therefore, there is no single way of presenting the information. Table 1 shows the method used by the US National Center for Health Statistics and by the Spanish Ministry of Health to determine the main causes of death. But other institutions and central statistics offices use other techniques. For example, based on the large groups of the ICD, diseases of the circulatory system constitute the main cause of death in Spain. This statement is correct, but so is the more granular information in Table

1. Thus, when presenting such statistics, it is important to report the ICD codes used to group the selected causes of death.

The cause of death tabulated is the underlying cause of death. In accordance with WHO recommendations, medical death certificates include three causes of death: the immediate cause (the disease or condition that directly led to death), the intermediate cause of that disease or condition, and finally, the initial or basic cause (the disease or injury that initiated the aforementioned events that led to death). This last cause is the one used to disseminate information on causes of death, monitor the main diseases and health problems, and conduct research. However, on occasion, some countries' central statistics offices and some research groups disseminate information on the number of deaths according to multiple causes in order to consider various combinations of causes appearing on medical death certificates.

These calculations show that certain causes appear on the death certificates of many deceased people, even though they are not the basic cause of death. The utility of this information, however, is unknown. It can confound conclusions about mortality patterns by cause of death. A physician can include in the death certificate certain intermediate and immediate causes, but another physician can include other different intermediate and immediate causes. Furthermore, from a statistical point of view it makes no sense. If, instead of including the three causes of death mentioned above, numerous other causes could be included in the medical death certificate, the presence of some causes would increase even more than in the combination of multiple causes. The only way for an analysis of multiple causes of death to be logically rigorous would be to include all the causes of death from the ICD in a death certificate, and for the physician certifying the death to explicitly indicate whether each one was or not relevant to the case at hand —something completely crazy.

5 Public health surveillance

One objective of public health surveillance is the early detection of epidemic outbreaks for purposes of disease control. This endeavor requires a continuous collection of health data for analysis and interpretation, but not exhaustive information on each and every disease case (de Mateo and Regidor, 2003). Indeed, not all sick people use health services. Furthermore, the collection criteria may change to increase the validity of the measurement of the phenomenon under surveillance, as information emerges around the transmissibility of the infectious agent, its clinical manifestations, and the prognosis of the patients. Therefore, the real impact of an epidemic outbreak on the mortality of the population does not necessarily correspond to the deaths from cases that have been detected.

In most epidemic outbreaks, numerous media sources and large swaths of the scientific community criticize the mortality figures produced by public health surveillance systems. These criticisms were highly publicized during the COVID-19 pandemic. Most likely, such judgments are due to the poor understanding of this public health practice, together with the anxiety generated by epidemic situations. A very rough estimate of the impact of epidemic outbreaks on mortality is made using the daily mortality monitoring systems of public health surveillance institutions, or through weekly death statistics from central statistics offices. The true estimate can only be made much later, when the mortality statistics by cause of death have been consolidated, since these statistics are the ones

that offer exhaustive data on deaths.

Notably, before the end of 2021, Spain was one of a very few countries worldwide that had exhaustive information on cause-specific mortality from 2020, when the first two waves of the COVID-19 pandemic occurred. This achievement was only possible due to the diligent efforts of the National Statistics Institute and the regional statistics agencies and death registries to expedite the compilation and dissemination of the cause-specific death statistics for 2020.

6 Conclusions

Mortality statistics remain the most suitable indicator for monitoring trends in the burden of disease over time and from one place to another. Information on deaths, compiled and disseminated for central statistics offices from the civil registries, has been used to document the enormous reduction in infant mortality rate and the rise in life expectancy. Likewise, from a public health perspective, standardized recording of the cause of death in the civil registry was a milestone, providing valuable insight on the diseases responsible for the most deaths. A second great achievement was the establishment of an international classification of causes of death.

The tabulation on basic cause of death is used monitor the main diseases and health problems, and conduct research. Sometimes, some countries' central statistics offices and some research groups disseminate information on the number of deaths according to multiple causes in order to consider various combinations of causes appearing on death certificates. The utility of this information, however, is unknown and it can confound conclusions about mortality patterns by cause of death.

Whereas the objective of public health surveillance is the early detection of epidemic outbreaks for purposes of disease control, the true estimate of impact of epidemic outbreaks can only be known much later, when the mortality statistics by cause of death have been consolidated, since these statistics are the ones that offer exhaustive data on deaths.

References

Alderson, Michael (1988). *Mortality, morbidity and health statistics*. Springer.

Brugal, MT, G Barrio, E Regidor, M Mestres, JA Caylà, and L De la Fuente (1999). Discrepancias en el número de muertes por reacción aguda a sustancias psicoactivas registradas en España. *Gaceta Sanitaria 13*(2), 82–87.

De la Fuente, Luis, Gregorio Barrio, Julián Vicente, María J Bravo, and José Santacreu (1995). The impact of drug-related deaths on mortality among young adults in Madrid. *American Journal of Public Health 85*(1), 102–105.

de Mateo, S and E Regidor (2003). Public health surveillance systems: let's not ask for the impossible. *Gaceta Sanitaria 17*(4), 327–331.

Goerlich, Francisco José and Rafael Pinilla (2006). Esperanza de vida en España a lo largo del siglo XX. Technical report, Fundación BBVA, Bilbao.

Gómez, Rosa (1991). *La mortalidad infantil española en el siglo XX*.

Ho, Jessica Y and Arun S Hendi (2018). Recent trends in life expectancy across high income countries: retrospective observational study. *British Medical Journal 362*.

Mackenbach, Johan P (2020). *A history of population health: rise and fall of disease in Europe*. Brill.

Ministerio de Sanidad y Consumo (MSC) (2005). La salud de la población española en el contexto europeo y del sistema nacional de salud. indicadores de salud. Technical report, Madrid: Ministerio de Sanidad y Consumo.

Viciana, Francisco (2003). Tendencias demográficas durante el siglo xx en españa. Technical report, FundaciÃşn Dialnet, Universidad de La Rioja, Logroño. En: Arroyo Pérez A (coord.).

World Health Organization (WHO) (2022a). History of the development of the ICD. Technical report, World Health Organization. License: CC BY-NC-SA 3.0 IGO.

World Health Organization (WHO) (2022b). World health statistics 2022: monitoring health for the SDGs, sustainable development goals. Technical report, World Health Organization. Accessed October 31, 2022.

5 Acknowledgement to Reviewers

The Editors of Spanish Journal of Statistics gratefully acknowledge the assistance of the following people, who reviewed manuscripts:

Enrique Calderín-Ojeda, Centre for Actuarial Studies, Department of Economics, The University of Melbourne, Australia.

Agustín Cañada, Instituto Nacional de Estadística, INE, Spain.

Emilio Gómez-Déniz, University of Las Palmas de Gran Canaria, Spain.

Jorge Navarro, University of Murcia, Spain.

Spanish Journal of Statistics

General Information

The Spanish Journal of Statistics (SJS) is the official journal of the National Statistics Institute of Spain (INE). The journal replaces Estadística Española, edited and published in Spanish by the INE for more than 60 years, which has long been highly influential in the Spanish scientific community. The journal seeks papers containing original theoretical contributions of direct or potential value in applications, but the practical implications of methodological aspects are also welcome. The levels of innovation and impact are crucial in the papers published in SJS.

SJS aims to publish original sound papers on either well-established or emerging areas in the scope of the journal. The objective of papers should be to contribute to the understanding of official statistics and statistical methodology and/or to develop and improve statistical methods; any mathematical theory should be directed towards these aims. Within these parameters, the kinds of contribution considered include:

- Official Statistics.
- Theory and methods.
- Computation and simulation studies that develop an original methodology.
- Critical evaluations and new applications
- Development, evaluation, review, and validation of statistical software and algorithms.
- Reviews of methodological techniques.
- Letters to the editor.

One volume is published annually in two issues, but special issues covering up-to-date challenging topics may occasionally be published.

Author guidelines

The Spanish Journal of Statistics publishes original papers in the theory and applications of statistics. A PDF electronic version of the manuscript should be submitted to José María Sarabia, Editor in chief of SJS via email to sjs@ine.es. Submissions will only be considered in English.

Manuscripts must be original contributions which are not under consideration for publication anywhere else. Its contents have been approved by all authors and. A single-blind refereeing system is used, so the identity of the referees is not communicated to the authors. Manuscripts that exceed 30 journal pages are unlikely to be considered for publication. More detailed information can be found at https://www.ine.es/sjs.

Published by the Spanish National Statistical Institute