BEIO

BOLETÍN DE
ESTADÍSTICA E
INVESTIGACIÓN OPERATIVA

2024, Volumen 40, Número 3

Seio
Sociedad de Estadística e Investigación Operativa

Edita **SEIO:**

Facultad de CC. Matemáticas
Universidad Complutense de Madrid
Plaza de Ciencias 3, 28040 Madrid

ISSN: 2387-1725

Ninguna parte de la revista puede ser reproducida, almacenada o trasmitida en cualquier forma o por medios, electrónico, mecánico o cualquier otro sin el permiso previo de la SEIO. Los artículos publicados representan las opiniones del autor y la revista BEIO no tiene por qué estar necesariamente de acuerdo con las opiniones expresadas en los artículos publicados. El hecho de enviar un artículo para la publicación en BEIO implica la transferencia del copyright de éste a la SEIO. Por tanto, el autor(es) firmará(n) la aceptación de las condiciones del copyright una vez que el artículo sea aceptado para su publicación en la revista.

Copyright © 2024 SEIO

BEIO (Boletín de Estadística e Investigación Operativa) es una revista que publica cuatrimestralmente artículos de divulgación científica de Estadística y de Investigación Operativa. Los artículos pretenden abordar tópicos relevantes para una gran mayoría de profesionales de la Estadística y de la Investigación Operativa, primando la intención divulgativa sin olvidar el rigor científico en el tratamiento de la materia en cuestión. Las secciones que incluye la revista son: Estadística, Investigación Operativa, Estadística Oficial, Historia y Enseñanza y Opiniones sobre la Profesión.

BEIO nació en 1985 como Boletín Informativo de la SEIO (Sociedad de Estadística e Investigación Operativa). A lo largo de los años ha experimentado una continua evolución. En 1994, aparece publicado el primer artículo científico y desde entonces el número de artículos científicos publicados ha ido creciendo hasta que en 2008 se segregan del Boletín los contenidos relacionados con la parte informativa y comienza a perfilarse como revista de divulgación de la Estadística y de la Investigación Operativa.

Los artículos publicados en **BEIO** están indexados en Scopus, MathScinet, Biblioteca Digital Española de Matemáticas, Dialnet (Documat), Current Index to Statistics, The Electronic Library of Mathematics (ELibM), COMPLUDOC y Catálogo Cisne Complutense.

La Revista está disponible online en `www.seio.es/BEIO`.

Editores

Jose Ameijeiras Alonso
Universidade de Santiago de Compostela
jose.ameijeiras@usc.es

Consuelo Parreño Torres
Universitat de València
consuelo.parreno@uv.es

Editores Asociados

Estadística
Paula Saavedra Nieves
Universidade de Santiago de Compostela
paula.saavedra@usc.es

Investigación Operativa
Eva Vallada Regalado
Universidad Politécnica de València
evallada@eio.upv.es

Estadística Oficial
Pedro Revilla Novella
Instituto Nacional de Estadística
pedro.revilla.novella@ine.es

Historia y Enseñanza
Mª Carmen Escribano Ródenas
Universidad CEU San Pablo de Madrid
escrod@ceu.es

Resúmenes de tesis Est.
Paula Navarro Esteban
Universidad de Cantabría
pnavarroesteban@gmail.com

Resúmenes de tesis I.O.
Marta Baldomero Naranjo
Universidad de Cádiz
marta.baldomero@uca.es

Editores Técnicos

María Alonso Pena
Universidad de Santiago de Compostela
mariaalonso.pena@usc.es

Miguel Reula Martín
Universidad de Valencia
miguel.reula@uv.es

Normas para el envío de artículos

Los artículos se enviarán por correo electrónico al editor asociado correspondiente o al editor de la Revista. Se escribirán en estilo article de LaTeX. Cada artículo ha de contener el título, el resumen y las palabras clave en inglés sin traducción al castellano. Desde la página web de la revista se pueden descargar las plantillas tanto en español como en inglés, que los autores deben utilizar para la elaboración de sus artículos.

Contents

Editorial
Méndez-Civieta, A., and Morala Miguélez, P. (2024)
Grupo de Tratamiento y Análisis de Big Data (TABiDa) **5**

Estadística
Cuesta-Albertos, J.A., García-Portugués, E., Febrero-Bande, M. and González-Manteiga, W. (2024)
Tests de bondad de ajuste para el modelo lineal funcional basados en proyecciones **7**

Investigación Operativa
Macías Ponce, J. and Olvera - Lopez, W. (2024)
Eligiendo candidatos usando programación lineal y encuestas: Caso MORENA 2024 **21**

Estadística Oficial
Mas Moreno, M. and Ayestarán Arregi, M. (2024)
El sistema integrado de metadatos en Eustat **28**

Historia
Filgueira, C. (2024)
Historia de la Estadística en la Psicología **35**

Opiniones sobre la profesión
Guerrero, V. (2024)
La frontera Pareto académica **48**

Tesis en Estadística
Castillo-Mateo, J. (2024)
Stochastic models for the spatiotemporal analysis of extremes. Applications to the analysis of climate change **52**

Tesis en Investigación Operativa
Segura, P. (2024)
Arc routing problems with drones **57**

Vol. 40 Núm. 3
Editorial

BOLETÍN DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Grupo de Tratamiento y Análisis de Big Data (TABiDa)

A. Méndez-Civieta **P. Morala Miguélez**

Biostatistics department. University of Columbia am5490@cumc.columbia.edu

uc3m - Santander Big Data Institute. Universidad Carlos III de Madrid pmorala@est-econ.uc3m.es

Citar como:
Méndez-Civieta, A., and Morala Miguélez, P. (2024). Grupo de Tratamiento y Análisis de Big Data (TABiDa). Boletín de Estadística e Investigación Operativa. 40(3), 5-6.

El grupo de trabajo **TABiDa**, acrónimo de Tratamiento y Análisis de Big Data, se estableció en 2021 con el objetivo de reunir a investigadores de la Sociedad Española de Estadística e Investigación Operativa interesados en las nuevas tecnologías de Big Data e Inteligencia Artificial. Con una membresía actual de 53 personas, de las cuales 27 son estudiantes, TABiDa refleja el creciente interés y la participación de la comunidad académica joven en estos campos innovadores. La creación de este grupo responde a la necesidad de ofrecer un espacio de reunión, colaboración y desarrollo a todos los investigadores de la SEIO que trabajan en estas áreas emergentes que no encajan con las categorías estadísticas tradicionales, proporcionando así un hogar para aquellas disciplinas que están redefiniendo la estadística y la investigación operativa moderna.

TABiDa se distingue por integrar una diversidad de líneas de investigación que, aunque diversas, comparten un enfoque innovador que no encaja en las estructuras convencionales de los grupos de trabajo existentes hasta el momento. Así, se posiciona como un grupo pionero en la incorporación de tecnologías avanzadas, adoptando un enfoque que refleja la constante evolución del Big Data y la Inteligencia Artificial. Entre las líneas de investigación de los miembros del grupo se incluyen, aunque no de manera exclusiva, el análisis de datos en alta dimensionalidad, el machine learning, el deep learning, la interpretabilidad y la equidad en los algoritmos entre otros temas.

Tras el congreso de la SEIO 2023 en Elche, Álvaro y Pablo hemos asumido la coordinación del grupo. Para reanudar y revitalizar el funcionamiento de este grupo tras un periodo de baja actividad, desde la nueva coordinación queremos presentar un plan de actividades para lograr este objetivo, dando visibilidad al grupo para atraer a más miembros de la SEIO o externos. El primer paso en este plan será la realización del workshop ***AI Horizons 2024*** (https://aihorizons2024.org/), en colaboración con el Proyecto de Prueba de concepto TRANSMODEL (Dept. Estadística, uc3m) del Ministerio de Ciencia e Innovación y Agencia Estatal de Investigación y la Cátedra ENIA: AImpulsa con UNIVERSIA (IBiDat, uc3m) otorgada por el Ministerio de Asuntos Económicos y Transformación Digital. El evento estará dividido en tres temáticas:

1. IA y Salud.
2. IA y Economía del Dato e Interpretabilidad/Sesgos.
3. IA en el ámbito Social.

Cada una de estas temáticas contará con una keynote principal, una mesa redonda con participantes del mundo de la academia y la empresa y una sesión de presentaciones de trabajos (charlas o pósters). Precisamente en esta última modalidad es en la que se busca con más énfasis la participación de miembros del grupo de trabajo TABiDa. Especialmente enfocado a los jóvenes investigadores, contaremos, gracias a la ayuda de la SEIO, con la posibilidad de solicitar financiación para la asistencia de miembros del grupo o doctores con antigüedad no superior a dos años. ¡Os animamos a uniros a TABiDa!

Además, el evento busca revitalizar y fortalecer la comunidad del grupo, facilitando la reconexión y el establecimiento de nuevos lazos entre sus miembros para fomentar la colaboración futura y visibilizar el nuevo enfoque del grupo. Como consecuencia, podrá servir para aumentar la difusión y capacidad para organizar actividades de mayor alcance por parte de TABiDa. Finalmente, el carácter de transferencia con el mundo de la empresa que tendrá el workshop, favorecerá las conexiones entre miembros del grupo y profesionales del sector, pudiendo abrir nuevas oportunidades de colaboración.

Dada la renovación del grupo de trabajo, también queremos remarcar el interés por la creación de sesiones específicas al ámbito de Big Data e Inteligencia Artificial dentro del Congreso de la SEIO, algo que a nuestro entender resultará muy enriquecedor ya que hasta ahora los trabajos presentados en estas áreas no tenían una sesión definida en la que agruparse. Asimismo se retomará la realización de la reunión del grupo de trabajo en este mismo congreso, donde esperamos poder atender todas vuestras sugerencias o ideas para el futuro del grupo.

Agradecimientos

Los coordinadores de este grupo, Álvaro Méndez Civieta y Pablo Morala, quieren agradecer a Rosa Elvira Lillo Rodríguez por la iniciativa de crear este grupo de trabajo. Además, quieren agradecer a todos los miembros del grupo de trabajo por el interés mostrado en los objetivos planteados para el grupo y su apoyo.

Acerca de los autores

Álvaro Méndez Civieta es un investigador postdoctoral en la Universidad de Columbia en la ciudad de Nueva York. Posee un doctorado en Estadística y un máster en Big Data por la Universidad Carlos III de Madrid, además de un grado en Matemáticas por la Universidad de Oviedo. Anteriormente, trabajó como investigador postdoctoral en el Instituto UC3M - Santander de Big Data. Su investigación se centra en problemas de regresión en alta dimensión, técnicas de reducción de dimensionalidad, datos funcionales y regresión cuantílica.

Pablo Morala Pablo es un investigador postdoctoral (FPI) en la Universidad Carlos III de Madrid, donde recientemente ha finalizado su doctorado en Ingeniería Matemática. Previamente, cursó un máster en Estadística y Data Science en esa misma universidad, así como los grados de Matemáticas y Física en la Universidad de Oviedo. Su investigación se centra en técnicas de explicabilidad o interpretabilidad de machine learning, especialmente las enfocadas a redes neuronales.

Vol. 40 Núm. 3
Estadística

Tests de bondad de ajuste para el modelo lineal funcional basados en proyecciones

Juan A. Cuesta-Albertos Eduardo García-Portugués Manuel Febrero-Bande
Wenceslao González-Manteiga

Departamento de Matemáticas, Estadística y Computación. Universidad de Cantabria juan.cuesta@unican.es

Departamento de Estadística. Universidad Carlos III de Madrid eduardo.garcia.portugues@uc3m.es

Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela
manuel.febrero@usc.es

Departamento de Estatística, Análise Matemática e Optimización. Universidade de Santiago de Compostela
wenceslao.gonzalez@usc.es

Citar como:
Cuesta-Albertos, J.A., García-Portugués, E., Febrero-Bande, M. and González-Manteiga, W. (2024). Tests de bondad de ajuste para el modelo lineal funcional basados en proyecciones. Boletín de Estadística e Investigación Operativa. 40(3), 7-20.

Abstract

El modelo lineal funcional con respuesta escalar relaciona una variable aleatoria evaluada en un espacio in nito-dimensional con una variable aleatoria escalar a través de un operador lineal. En el trabajo “Goodness-of- t tests for the functional linear model based on randomly projected empirical processes” (Cuesta-Albertos, García-Portugués, Febrero-Bande y González-Manteiga, 2019) se introduce un contraste de bondad de ajuste de este modelo de regresión un conjunto de datos. Este test está basado en tres ingredientes principales: la caracterización de la hipótesis nula mediante proyecciones aleatorias, la regularización del estimador del modelo lineal, y la consideración de un proceso empírico de los residuos del modelo indexados por las proyecciones aleatorias de la covariable funcional. El propósito de este artículo es, a petición de BEIO, revisar desde una perspectiva lo más sencilla posible el trabajo galardonado con el Premio SEIO-FBBVA 2020 a la Mejor Contribución en Estadística Metodológica. Se proporcionan, además, el análisis de un nuevo conjunto de datos mediante el software existente y una breve revisión de los trabajos relacionados con el test que han aparecido desde su publicación.

Palabras Clave: bootstrap, contraste de hipótesis, datos funcionales, reducción de la dimensión.

1 Introducción

El Análisis de Datos Funcionales (FDA, por sus siglas en inglés) es la disciplina de la estadística que se encarga de analizar variables aleatorias medidas de forma continua sobre un determinado soporte. Este tipo de datos es común en aplicaciones donde las observaciones son curvas, superficies o cualquier otro objeto, de naturaleza posiblemente infinito-dimensional, que pueda ser representado por una función. Algunos ejemplos básicos de datos funcionales son las curvas de crecimiento infantiles, las trayectorias descritas por cotizaciones de acciones o las curvas diarias de variables meteorológicas (véase la Figura 1). Así, el FDA puede ser entendido, intuitivamente, como el "marco límite" del análisis multivariante cuando la dimensión tiende a infinito. A su vez, los procesos estocásticos en tiempo continuo son un pilar fundamental del FDA al formalizar el estudio de la distribución de curvas aleatorias.

Uno de los objetivos habituales en el FDA es analizar la relación entre variables aleatorias funcionales y otras variables de interés mediante modelos de regresión. Al involucrar una variable funcional, en la literatura estadística estos modelos de regresión suelen incorporar el apellido funcional. En el caso más sencillo en el que la variable respuesta Y es escalar y el (único) predictor X es funcional, el modelo

$$Y = m(X) + \varepsilon, \quad m(x) := \mathbb{E}[Y|X = x] \tag{1}$$

recoge la relación entre Y y X, siendo el error ε una variable aleatoria (escalar) tal que $\mathbb{E}[\varepsilon|X] = 0$. En el modelo (1), el operador de regresión m es el objeto de interés y su estimación es crucial para inferir la relación entre Y y X y la predicción de Y a partir de X.

El modelo lineal funcional con respuesta escalar (FLMSR) especifica una relación lineal entre Y y X a través de un operador lineal m. Un operador lineal puede tener formas muy variadas. Por ejemplo, el operador $m(x) = \int x(t)\beta(t)dt$, $\beta \in \{\beta \in \mathcal{C}^2[0,1] : \beta(0) = \beta(1)\}$, siendo $\mathcal{C}^2[0,1] = \{x(t) : [0,1] \to \mathbb{R} \text{ una función diferenciable}\}$, es lineal. En el caso en el que la variable funcional pertenezca a un espacio de Hilbert $\mathcal{H}$ con producto interior $\langle \cdot, \cdot \rangle$, como se explicará en la Sección 2.1, el Teorema de Representación de Riesz garantiza que cualquier operador de regresión lineal es expresable como

$$m(x) = \langle x, \beta \rangle \text{ con } \beta \in \mathcal{H}, \tag{2}$$

asumiendo, sin pérdida de generalidad, que tanto Y como X han sido centradas, es decir, que son tales $\mathbb{E}[Y] = 0$ y $\mathbb{E}[X] = 0$. Cuando $\mathcal{H} = \mathbb{R}^p$ y $\langle x, r \rangle = x^\top r$ es el producto escalar entre dos vectores, el modelo FLMSR se reduce al habitual modelo de regresión lineal múltiple sin término independiente debido al centramiento de las variables predictoras y de la respuesta. Un caso particular de la relación lineal inducida por (2) es $\beta = 0$, en cuyo caso $\mathbb{E}[Y|X] = 0$, para todo x. Por lo tanto, si $\beta = 0$, los modelos (1) y (2) especifican que la dependencia entre Y y X solamente puede existir en momentos condicionales de orden superior, generalmente menos relevantes.

Dada una muestra aleatoria simple $\{(X_i, Y_i)\}_{i=1}^n$ del par $(X, Y) \in \mathcal{H} \times \mathbb{R}$, una pregunta natural relacionada con el FLMSR es si existe una relación lineal entre Y y X. En términos del modelo (2), esta pregunta se puede replantear como ¿ajusta el modelo FLMSR "suficientemente bien" esta muestra o, por el contrario, existen desviaciones significativas entre la muestra y el ajuste dado por el FLMSR? Para responder a esta pregunta formalmente es necesario contrastar la hipótesis nula

$$H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathcal{H}\}. \tag{3}$$

en la que se afirma que el operador de regresión m pertenece a la clase de operadores lineales en $\mathcal{H}$. La hipótesis H_0 es expresable como $H_0 : m(x) = \langle x, \beta_0 \rangle$, para algún $\beta_0 \in \mathcal{H}$ desconocido. Esta hipótesis se puede confrontar a la hipótesis alternativa en la que se afirma que para todo $\beta \in \mathcal{H}$ existe un conjunto $A_1 \subset \mathcal{H}$ de probabilidad positiva, en el que falla la igualdad. Es decir,

$$H_1 : \text{para todo } \beta \in \mathcal{H}, \text{ existe } A_1 \text{ con } P[X \in A_1] > 0, \text{ tal que } m(x) \neq \langle x, \beta \rangle \text{ para } x \in A_1. \tag{4}$$

Obsérvese que esta elección de H_1 es la alternativa más general a la que se podría confrontar H_0, pues no especifica un tipo de estructura no lineal para m (por ejemplo, de tipo cuadrático). Si H_0 no es rechazada, entonces no existen desviaciones significativas entre la muestra y el ajuste dado por el FLMSR.

El contraste de hipótesis en (3)–(4) es un ejemplo de un problema de contraste bondad de ajuste (goodness-of-fit en inglés), un tema clásico en estadística cuyos orígenes se remontan hasta Karl Pearson (véase Barnard, 1992). Así, el conocido test de Kolmogorov–Smirnov (Kolmogoroff, 1933) es un contraste de bondad de ajuste de una distribución conocida, por ejemplo una $N(0,1)$, a la muestra. En cambio, el también conocido test de Lilliefors (1967) evalúa la bondad de ajuste de la familia de distribuciones normales $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$, siendo la media y varianza desconocidas. Estos dos tests son ómnibus, pues son capaces de rechazar H_0 frente a su negación, siempre que H_0 sea falsa y el tamaño muestral sea suficientemente grande. La literatura en contrastes de bondad de ajuste para modelos de regresión con datos multivariantes es extensa, por lo que referimos al lector interesado a la revisión de González-Manteiga y Crujeiras (2013).

En Cuesta-Albertos et al. (2019) (CGFG19 en adelante) se propone un test para evaluar la bondad de ajuste del FLMSR a un

conjunto de datos. Este test está basado en tres ingredientes principales: la caracterización de la hipótesis nula (3) mediante proyecciones aleatorias, la estimación regularizada del parámetro 1 del FLMSR y la consideración de un proceso empírico proyectado de los residuos del FLMSR sobre el que se obtiene el estadístico de contraste. El propósito de este artículo es, a petición de BEIO, revisar desde una perspectiva lo más sencilla posible dicho trabajo, galardonado con el Premio SEIO-FBBVA 2020 a la Mejor Contribución en Estadística Metodológica. Así, este artículo no pretende ser exhaustivo ni completamente detallado, sino que busca centrarse en los elementos principales del trabajo premiado evitando, en lo posible, los detalles más técnicos de CGFG19. El artículo, además, incluye el análisis de un nuevo conjunto de datos mediante el software existente y una breve revisión de las extensiones de CGFG19 que han aparecido desde su publicación.

El resto del artículo se organiza como sigue. En la Sección 2 se introducen en más detalle el análisis de datos funcionales, el modelo lineal funcional con respuesta escalar y un conjunto de datos motivador. La Sección 3 aborda la caracterización de una versión simplificada de la hipótesis nula (3) mediante proyecciones aleatorias. En la Sección 4 se introduce un proceso empírico de los residuos del FLMSR indexados por las proyecciones aleatorias y se presenta la distribución asintótica del estadístico de contraste. La Sección 5 explica el procedimiento que implementa el test del FLMSR en la práctica. Finalmente, en la Sección 6 se proporciona un ejemplo reproducible y en la Sección 7 se revisan trabajos posteriores relacionados con CGFG19.

2 Datos funcionales y regresión lineal

2.1 Datos funcionales

Formalmente, se dice que una variable aleatoria (v.a.) $\mathbf{X}$ es funcional si toma valores en un espacio $\mathcal{F}$ que, siguiendo Ferraty y Vieu (2006), será, como mínimo, un espacio completo métrico posiblemente de dimensión infinita.[1] Esta definición incluye, por supuesto, espacios complejos como los espacios normados (Banach) o los dotados de producto interior (Hilbert). Dos ejemplos notables de espacios funcionales que involucran a funciones reales definidas en un intervalo $[a,b]$ o $\mathbb{R}$ son los siguientes:

- El conjunto de las funciones esencialmente acotadas $L^\infty([a,b]) := \{x : [a,b] \to \mathbb{R} : \text{existe } M \in \mathbb{R}, |x(t)| \leq M \text{ para casi todo } t \in [a,b]\}$ con la métrica $d_\infty(x,y) = \sup_{t\in[a,b]} |x(t) - y(t)|$.
- Los espacios de Lebesgue $L^p([a,b]) := \{x : [a,b] \to \mathbb{R} : \int_a^b |x(t)|^p dt < \infty\}$ con la métrica $d_p(x,y) = \left(\int_a^b |x(t) - y(t)|^p dt\right)^{1/p}$.

Ambos ejemplos son espacios de Banach con la norma definida como $\|x\|_\infty := \sup_{t\in[a,b]} |x(t)|$ y $\|x\|_p = \left(\int_a^b |x(t)|^p dt\right)^{1/p}$, pero solo el espacio $L^2([a,b])$ es Hilbert, esto es, en él podemos definir un producto interior $\langle x, y\rangle = \int_a^b x(t)\overline{y(t)} dt$ coherente con la norma, $\|x\|_2 = \sqrt{\langle x, x\rangle}$ y la métrica $d_2(x,y) = \sqrt{\langle x - y, x - y\rangle}$ definida en el espacio. Disponer de producto interior permite determinar cuando dos elementos son ortogonales ($\langle x, y\rangle = 0$) y también asegurar que existe una base a lo sumo numerable para representar a cualquier elemento, esto es, existe $\{e_j\}_{j=1}^{\infty} \subset L^2([a,b])$ tal que todo elemento puede representarse de manera única como una combinación de elementos de la base: $x = \sum_{j=1}^{\infty} c_j e_j$, donde $c_j = \langle x, e_j\rangle$.

Como espacio de trabajo para el predictor funcional consideraremos un espacio Hilbert separable $\mathcal{H}$ con producto interior $\langle \cdot, \cdot \rangle$ y norma asociada $\|\cdot\|$. Este espacio $\mathcal{H}$ es un espacio de Hilbert real general, pero puede ser interpretado como $L^2([a,b])$.

2.2 Regresión lineal funcional

En el modelo FLMSR la respuesta $Y \in \mathbb{R}$ se explica con el predictor $\mathbf{X} \in \mathcal{H}$ a partir de una muestra $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ independiente e idénticamente distribuida (i.i.d.) del par $(\mathbf{X}, Y)$. Sin pérdida de generalidad y por la conveniencia de eliminar el intercepto del FLMSR, asumimos que las variables están centradas, esto es, $E[\mathbf{X}] = 0$ y $E[Y] = 0$. También trabajaremos con la muestra centrada, de tal forma que $\bar{\mathbf{X}} = 0$ y $\bar{Y} = 0$. El FLMSR utiliza la estructura del producto interior del espacio de Hilbert para establecer la relación

$$Y = \langle \mathbf{X}, \boldsymbol{\beta}\rangle + \varepsilon,$$

donde $\boldsymbol{\beta} \in \mathcal{H}$ y el error ε es tal que $E[\varepsilon \mathbf{X}] = 0$ y $\sigma^2 := \mathrm{Var}[\varepsilon] < \infty$. En el caso en el que $\mathcal{H} = L^2([a,b])$, $\langle \mathbf{X}, \boldsymbol{\beta}\rangle = \int_a^b X(t)\beta(t) dt$. En cierto modo, el FLMSR generaliza el tradicional modelo de regresión lineal, obtenido con $\mathcal{H} = \mathbb{R}^p$, al caso de las infinitas covariables $\{X_j\}_{j=1}^{\infty}$.

La idea clave para estimar el FLMSR es representar las funciones de la muestra y el parámetro en una base adecuada:

$$\mathbf{X} = \sum_{j=1}^{\infty} d_j \mathbf{e}_j \quad \text{y} \quad \boldsymbol{\beta} = \sum_{j=1}^{\infty} c_j \mathbf{e}_j \approx \sum_{j=1}^{p} c_j \mathbf{e}_j$$

[1] Se emplea la negrita para denotar elementos en un espacio funcional y así remarcar la conexión de estos objetos con los vectores columna a los cuales generalizan.

Una de las bases más usadas es la base de componentes principales (PCs), que da lugar a la expansión de Karhunen–Loève:

$$\hat{\alpha} = \sum_{j=1}^{\infty} \sqrt{\hat{\lambda}_j}\, b_j e_j, \tag{5}$$

donde $\{e_j\}_{j=1}^{\infty}$ (la base) es la secuencia de autofunciones ortonormales asociadas con el operador de covarianzas $\hat{\Gamma}$, $\{\hat{\lambda}_j\}_{j=1}^{\infty}$ son sus autovalores y está dado como

$$\alpha \in H \mapsto \langle \alpha, \hat{\Gamma}\alpha \rangle := E\left[\langle \alpha, X - \mu \rangle^2\right] = E\left[\langle \alpha, X - \mu \rangle \langle X - \mu, \alpha \rangle\right], \quad \alpha \in H, \tag{6}$$

para el producto tensorial $x \otimes y$ definido como $\langle x \otimes y, z \rangle := \langle x, z \rangle y$, $x, y, z \in H$. En (5), $\{b_j\}_{j=1}^{\infty}$ es una secuencia de v.a.'s reales de media cero (porque $\hat{\alpha}$ está centrada) e incorreladas dos a dos. A modo de ejemplo, si $H = \mathbb{R}^p$, entonces (6) se convierte en $\alpha \in \mathbb{R}^p \mapsto \alpha^\top \mathrm{Var}[X]\alpha = \alpha^\top \Gamma \alpha$, siendo Γ la matriz de varianzas-covarianzas del vector aleatorio (centrado) X y $\{\lambda_j, e_j\}_{j=1}^{p}$ su sistema de valores y vectores propios. Si $X \in L^2(\Omega)$, entonces $\alpha^\top X = E[\alpha^\top X X^\top \alpha] = \alpha^\top E[X X^\top]\alpha = \alpha^\top \Gamma \alpha$ con $\alpha \in \mathbb{R}^p$).

La estimación de μ mediante mínimos cuadrados requiere, igual que en el caso multivariante, la inversión de la versión empírica del operador Γ, que en base a (6) viene dada por $\hat{\Gamma} := \frac{1}{n}\sum_{i=1}^{n} \hat{\alpha}_i \otimes \hat{\alpha}_i$. Sin embargo, $\hat{\Gamma}$ es, casi seguro, un operador de rango a lo sumo n. Esto no es un impedimento cuando $H = \mathbb{R}^p$ con $p \leq n$, pero cuando H es infinito-dimensional esto implica que $\hat{\Gamma}$ no es un operador invertible. Para evitar este problema, Cardot, Mas y Sarda (2007) proponen una regularización que consiste en definir una secuencia de aproximaciones de μ truncando a un número finito k_n la expansión de Karhunen–Loève:

$$\hat{\alpha}_{k_n} := \sum_{j=1}^{k_n} \frac{1}{\hat{\lambda}_j} \langle \hat{\alpha}, e_j \rangle e_j,$$

siendo $\{\hat{\lambda}_j, e_j\}_{j=1}^{k_n}$ estimadores del sistema propio basados en la muestra.[2] Al elegir k_n de manera que se cumplan ciertas condiciones (véanse las Secciones 2 y 3 de Cardot et al., 2007) que incluyen que $k_n \to \infty$ cuando $n \to \infty$, lo que estamos haciendo es aproximar el modelo infinito-dimensional por uno de dimensión finita donde tanto el parámetro como la covariable funcional se aproximan por k_n sumandos. Más concretamente, tenemos

$$\sum_{j=1}^{k_n} d_j e_j \quad \text{y} \quad \hat{\alpha} \sum_{j=1}^{k_n} G_{ij} e_j, \quad i = 1, \ldots, n, \tag{7}$$

con una precisión creciente según aumenta k_n. Esto proporciona el estimador de μ:[3]

$$\hat{\mu}_{k_n} := \hat{\alpha}_{k_n} = \frac{1}{n}\sum_{i=1}^{n} \hat{\alpha}_i = \frac{1}{n}\sum_{j=1}^{k_n}\sum_{i=1}^{n} \hat{\alpha}_i \frac{\langle \hat{\alpha}_i, e_j \rangle}{\hat{\lambda}_j} e_j. \tag{8}$$

La calidad de las aproximaciones (7) y del estimador (8) vendrán dadas por la capacidad de la base PC para resumir la información de la muestra y de X en k_n sumandos. Aunque en la literatura de FDA hay abundantes referencias sobre el uso de otro tipo de bases en $L^2(\Omega)$ o para el FLMSR (por ejemplo, B-spline en Cardot, Ferraty y Sarda, 2003), la base PC cuenta con tres propiedades importantes: no es específica de $L^2(\Omega)$, es ortogonal y representa la muestra de la forma más sintética posible, esto es, es la base que, a igualdad de número de elementos, es capaz de retener más información de la muestra.

Para determinar k_n, además de las condiciones de Cardot et al. (2007), se debe tomar como referencia el trabajo de Hall y Hosseini-Nasab (2006), que descompone el error cuadrático de estimación de μ con una base PC de k_n elementos:

$$E\left[\|\hat{\mu}_{k_n} - \mu\|^2\right] = \underbrace{\sum_{j=1}^{k_n} \frac{\mathrm{Var}[\langle X, e_j \rangle]}{n}}_{\text{varianza}} + \underbrace{\sum_{j=k_n+1}^{\infty} \langle \mu, e_j \rangle^2}_{\text{sesgo}}. \tag{9}$$

El primer sumando de (9) es un término de varianza que crece con k_n. El segundo sumando es un término de sesgo que disminuye con k_n. Por tanto, la elección de k_n se convierte en un punto crucial tanto para disponer de un estimador útil como para minimizar su error cuadrático. Esta elección suele hacerse con los métodos habituales de criterios de selección de modelos (MSC), que tienen la siguiente formulación general:

$$\begin{aligned} k_n &= \arg\min_{k \geq 1} \mathrm{MSC}(k) \\ &= \arg\min_{k \geq 1} \left\{ -2 \log L(\hat{\mu}_k \mid \hat{\alpha}) + \lambda_n \nu_k \right\}, \end{aligned} \tag{10}$$

donde $L(\hat{\mu}_k \mid \hat{\alpha})$ es la verosimilitud asociada al parámetro y $\lambda_n \nu_k$ es un término de penalización para evitar el sobreajuste.

[2] Si $H = \mathbb{R}^p$, en la práctica $\{\hat{\lambda}_j, e_j\}_{j=1}^{k_n}$ se pueden obtener mediante análisis de componentes principales de la matriz $\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$, $x_i = (x_{i1}, \ldots, x_{ip})$.

[3] Obsérvese que si $H = \mathbb{R}^p$ y $k_n = p$, entonces la primera igualdad de (8) muestra claramente que $\hat{\mu}_{k_n}$ es el conocido estimador de mínimos cuadrados en un modelo lineal sin intercepto y con muestra centrada: $\hat{\mu}_{k_n} = \frac{1}{n}\sum_{i=1}^{n} \hat{\alpha}_i = \frac{1}{n}\sum_{i=1}^{n} \hat{\alpha}_i = 0$. Si $k_n < p$, entonces $\hat{\mu}_{k_n}$ corresponde al estimador del modelo de regresión lineal basado en componentes principales.

2.3 Un ejemplo real

Como ejemplo motivador vamos a usar datos de alquiler ocasional de bicicletas (NBCR) en Washington DC, EEUU durante los años 2011–2012. Estos datos fueron recogidos por el sistema de bicicletas públicas Capital Bikeshare System[4] e inicialmente fueron considerados por Fanaee-T y Gama (2014). En este trabajo hemos considerado solamente los alquileres correspondientes a los sábados para eliminar el efecto del día laborable y no tener que introducir un factor como covariable. También hemos transformado logarítmicamente el agregado de alquileres de todo el día para definir la variable respuesta $\mathcal{Y} = \log^1 \mathrm{NBCR}^{\circ}$). Haremos dos análisis independientes, donde tomaremos como covariables funcionales $\hat{\mathcal{X}}$ dos variables meteorológicas medidas en forma de curvas diarias: la humedad y la sensación térmica. En total, se disponen de $\mathcal{N} = 102$ curvas de cada tipo, estando discretizadas en 24 mediciones horarias. Estas curvas se colorean según la variable respuesta en la Figura 1, poniendo de manifiesto que es muy difícil determinar si la relación de las funciones con la variable escalar sigue un patrón lineal. Esta es una situación habitual en FDA donde, por la propia naturaleza infinito-dimensional de los datos, las herramientas gráficas no suelen ser determinantes o solo proporcionan evidencias indirectas que incluso pueden engañar el ojo del observador. Una aplicación del contraste de bondad de ajuste del FLMSR permitirá evaluar de forma automática la existencia de dicha relación lineal entre estas dos variables.

3 La hipótesis nula proyectada

3.1 Caracterización mediante proyecciones aleatorias

Dado el elemento $h \in \mathcal{H}$, denotaremos por $\hat{\mathcal{X}}^h := \hat{\mathcal{X}} - \hat{\mathcal{X}}_h$ a la proyección de $\hat{\mathcal{X}}$ en la dirección h. Usamos la notación $\mathcal{X} \times \mathcal{Y}! < {}_1{}^1 x^{\circ}$ para denotar el FLMSR y para recalcar que los resultados aquí presentados son también válidos para otra especificación paramétrica; es decir, no utilizan la linealidad del modelo de regresión.

El pilar principal del contraste de bondad de ajuste del FLMSR en CGFG19 es la caracterización casi segura (c.s.) de la hipótesis nula (3), que puede ser reexpresada como

$$\mathcal{H}_0 : \mathcal{E}\left[\mathcal{Y} < {}_{1_0}{}^1 \hat{\mathcal{X}}^{\circ} \mathcal{X}\right] = 0 \text{ c.s., para un } 1_0 \in \mathcal{H} \qquad (11)$$

a través de una hipótesis nula proyectada, definida como

$$\mathcal{H}_0^h : \mathcal{E}\left[\mathcal{Y} < {}_{1_0}{}^1 \hat{\mathcal{X}}^{\circ} \mathcal{X}\right]^h = 0 \text{ c.s., para un } 1_0 \in \mathcal{H} \text{ –}$$

siendo $h \in \mathcal{H}$ una dirección de proyección elegida aleatoriamente. En ambas hipótesis el parámetro verdadero del FLMSR, $1_0 \in \mathcal{H}$, es desconocido. Esta caracterización está recogida en el siguiente resultado, cuya prueba se basa, en particular, en el Teorema 4.1 de Cuesta-Albertos, Fraiman y Ransford (2007).

Teorema 3.1 (Teorema 2.4 y Corolario 2.5 en CGFG19) Sea $\mathcal{Y}$ una medida gaussiana no degenerada en $\mathcal{H}$ y $1_0 \in \mathcal{H}$. Supongamos que $\mathcal{E}_{\mathcal{Y}}^2 \ddot{\mathcal{Y}} \mathcal{Y}$ y que $\hat{\mathcal{X}}$ satisface que

$$\mathcal{Y} < {}_{:} := \mathcal{E}\left[\mathcal{X}^{\circ} \mathcal{X}^{:} \ddot{\mathcal{Y}} \mathcal{Y}\right] \text{ –para todo: } 1 \text{ –y que } \tilde{\mathcal{O}}_{:=1} < {}_{:}^{1 \bullet :} = 1 \bullet$$

Si denotamos $\mathcal{H}_0 := \{h \in \mathcal{H} : \mathcal{E}\left[\mathcal{Y} < {}_{1_0}{}^1 \hat{\mathcal{X}}^{\circ} \mathcal{X}\right]^h = 0 \text{ c.s.}\}$ al subconjunto de $\mathcal{H}$ cuyos elementos verifican $\mathcal{H}_0^h$, entonces

$$\mathcal{E}\left[\mathcal{Y} < {}_{1_0}{}^1 \hat{\mathcal{X}}^{\circ} \mathcal{X}\right] = 0 \text{ c.s.} \Leftrightarrow \mathcal{H}_0 = 1 \bullet$$

De acuerdo con este resultado, si estamos interesados en contrastar la hipótesis nula $\mathcal{H}_0$, entonces podemos hacerlo de la siguiente manera:

(i) Generar una dirección aleatoria h utilizando una medida $\mathcal{Y}$ en $\mathcal{H}$ que sea gaussiana y no degenerada.

(ii) Condicionalmente en h, contrastar la hipótesis nula proyectada $\mathcal{H}_0^h$.

(iii) Para un nivel de significación $\mathcal{U} \in \mathcal{Y}^{10} - 1^{\circ}$, rechazar $\mathcal{H}_0$ si y solo si se rechaza $\mathcal{H}_0^h$.

Este procedimiento es simple pero razonable porque, en virtud del Teorema 3.1:

- Si $\mathcal{H}_0$ se cumple, entonces $\mathcal{H}_0^h$ también se cumple.
- Si $\mathcal{H}_0$ no se cumple, entonces tampoco se cumple $\mathcal{H}_0^h$ de forma $\mathcal{Y}$-c.s. En este caso, con probabilidad uno, habremos elegido una dirección h para la cual $\mathcal{H}_0^h$ no se verifica.

Por supuesto, la principal ventaja de contrastar $\mathcal{H}_0^h$ en lugar de $\mathcal{H}_0$ es que en $\mathcal{H}_0^h$ la v.a. que condiciona es real, lo que simplifica sustancialmente el procedimiento de contraste.

3.2 Contraste de la hipótesis nula simple

El Teorema 3.1 se puede aplicar directamente al contraste de la hipótesis nula simple $\mathcal{H}_0 : \mathcal{E}\left[\mathcal{Y} < {}_{1_0}{}^1 \hat{\mathcal{X}}^{\circ} \mathcal{X}\right] = 0 \text{ c.s.}$, con

[4] Véase https://capitalbikeshare.com/system-data.

Figura 1. Curvas diarias de variables meteorológicas coloreadas en función del logaritmo del número total de alquileres casuales de bicicletas.

1_0 2 H conocido, que consideramos en esta sección. Si bien el objetivo final es (11), donde 1_0 2 H es desconocido, este primer abordaje nos permite introducir la maquinaria de procesos empíricos que se utilizará posteriormente.

Denotemos $/ := . < 1_0 1 \wedge o$. Por el Teorema 3.1, sabemos que $_0$ es `-c.s. equivalente a $E / j \wedge h = 0$ c.s. Por otro lado, como el par $1/ - \wedge h o$ es un par de v.a.'s reales estándar, siguiendo a Stute (1997), se tiene la caracterización

$$E / j \wedge h = 0 \text{ c.s. } () \qquad 1G^o := E / 1_{f \wedge h \; Gg} = 0$$

$$\text{para casi todo } G2 R \bullet \qquad (12)$$

La función es la función de regresión integrada, pues surge de integrar la función de regresión de $/$ dado $\wedge h$, es decir la función $C2 R 7! < / 1C^o := E / j \wedge h = C$, con respecto a h, la distribución de $\wedge h$. En efecto,

$$1 \; G < / 1C^o d \; h 1C^o = E < / 1 \wedge h o 1_{f \wedge h \; Gg}$$

$$= E / 1_{f \wedge h \; Gg} = 1G^o \bullet$$

La principal ventaja del lado derecho (12) es la eliminación del condicionante en la esperanza. Así, dada la muestra i.i.d. $f 1 / 8 - \wedge h o g_{8=1}$, la función se estima directamente mediante

$$G7! = 1G^o := \frac{1}{=} \tilde{G}_{8=1} 1_{f \wedge h \; Gg} / 8$$

$$= \frac{1}{=} \tilde{G}_{8=1} 1_{f \wedge h \; Gg} 1 \cdot 8 < 1_0 1 \wedge 8 oo \bullet \qquad (13)$$

Debido a la concatenación de caracterizaciones, la proximidad de la función $=$ a cero está directamente conectada con la veracidad de h_0 y por tanto de $_0$.

La sucesión de funciones $f = g_{==1}^1$ es lo que se conoce como un proceso empírico, es decir, una sucesión de funciones aleatorias generadas a partir de una muestra de tamaño $=$. Un espacio funcional habitual para analizar estas funciones es el espacio de Skorohod $1R^o$, formado por las funciones cadlag ("continuas por la derecha con límites por la izquierda") de nidas en R (véase Billingsley, 1999, Cápulo 3). Un resultado clásico en la teoría de procesos empíricos es la convergencia débil del proceso empírico asociado a la función de distribución empírica $=$ como estimador de una función de distribución de nida en $\gg 0 - 1\frac{1}{4}$

$$p_{=1} = \; o \quad P \quad \text{en} \quad 1 \gg 0 - 1\frac{1}{4} o$$

donde la notación representa la convergencia débil de la medida asociada al proceso empírico a la medida del proceso límite cuando $n \to \infty$ y P es un puente browniano en $[0,1]$.[5] También utilizaremos para denotar la convergencia débil o en distribución de una secuencia de v.a.'s.

Consideramos el proceso (13) normalizado, de nido como

$$\begin{aligned} \psi_n(\theta) &:= n^{1/2} \psi_n(\theta) \\ &= n^{1/2} \sum_{i=1}^{n} 1_{f(X_i)} g(X_i) - C(\theta), \quad \theta \in \Theta. \end{aligned} \tag{14}$$

El siguiente resultado se sigue a partir del Teorema 1.1 en Stute (1997).

Corolario 3.1 (Corolario 2.6 en CGFG19) Bajo H_0 y asumiendo $f_\theta \in \Psi_1$,

$$\psi_n \to G_1 \text{ en } \ell^\infty(\Theta), \tag{15}$$

siendo G_1 un proceso gaussiano con media cero y función de covarianza dada por $K_1(\theta_1, \theta_2) := \lim_{n} \mathrm{Var}\left[\psi_n(\theta_1), \psi_n(\theta_2) \right]$.

A partir de la convergencia débil de ψ_n bajo H_0 se pueden de nir estadísticos de contraste basados en normas de ψ_n y obtener sus distribuciones nulas bajo H_0. Este paso nal se elabora en la siguiente sección en el caso de la hipótesis nula compuesta.

Concluimos la sección observando que si bien se ha seguido el enfoque de Stute (1997) para contrastar H_0 mediante (12) y posteriormente (15), este enfoque es independiente de la caracterización dada por el Teorema 3.1. Por ejemplo, un abordaje alternativo para contrastar H_0 podría ser confrontar la función de regresión $m(\theta) = E[r(Y, \theta) \mid X]$ frente a un estimador no paramétrico $\hat{m}$ de la misma, siguiendo el enfoque de Härdle y Mammen (1993).

4 El estadístico de contraste y su distribución asintótica

El abordaje descrito en la Sección 3.2 puede ser utilizado también en el caso más realista en el que el parámetro $\theta_0 \in \Theta$ sea desconocido, es decir, cuando nos encontremos frente a la hipótesis nula compuesta (11). La principal dificultad añadida es incorporar los efectos de la estimación del estimador $\hat{\theta}_n$ de nido en (8) en la convergencia del proceso empírico asociado. Al ser los desarrollos de esta sección específicos del FLMSR, este se denotará como $x \mapsto \langle x, \theta \rangle = x^\theta$.

El punto de partida es un proceso empírico análogo a (14), pero sustituyendo el parámetro θ_0 del FLMSR por $\hat{\theta}_n$:

$$\begin{aligned} \beta_n(\theta) &:= \sum_{i=1}^{n} 1_{f(X_i)} g(X_i) \cdot \varepsilon_i^{\hat{\theta}_n} \\ &= \left(\beta_n^{(1)}(\theta), \beta_n^{(2)}(\theta), \beta_n^{(3)}(\theta) \right), \quad \theta \in \Theta, \end{aligned} \tag{16}$$

donde $\{a_n\}_{n=1}^{\infty}$ con $a_n \to 0$ es una secuencia de normalización positiva que se determinará más adelante y los tres procesos empíricos son

$$\beta_n^{(1)}(\theta) := \sum_{i=1}^{n} 1_{f(X_i)} g(X_i) \cdot \varepsilon_i^{\theta_0} - a_n \Gamma_n(\theta, \theta_0) \hat{\theta}_n,$$

$$\beta_n^{(2)}(\theta) := \sum_{i=1}^{n} 1_{f(X_i)} g(X_i) \varepsilon_i - a_n \Gamma_n(\theta, \theta_0) \hat{\theta}_n, \quad \text{y}$$

$$\beta_n^{(3)}(\theta) := a_n \Gamma_n(\theta, \theta_0) \hat{\theta}_n, \quad \text{con} \quad \Gamma_n(\theta, \theta_0) := E\left[1_{f(X)} g(X) X^{\theta} \right].$$

Obsérvese que $\beta_n^{(1)} = n^{1/2} \psi_n$, con lo cual la convergencia débil del primer proceso viene determinada por el Corolario 3.1.

Los procesos $\beta_n^{(2)}$ y $\beta_n^{(3)}$ incluyen los efectos de estimación de θ_0. El siguiente resultado establece que, para cada $\theta \in \Theta$ jo, el efecto de $\beta_n^{(2)}(\theta)$ en (16) es despreciable y que la tasa normalizadora para $\beta_n^{(3)}(\theta)$ depende de la secuencia $\Gamma_n(\theta, \theta_0) \to \Gamma(\theta, \theta_0)$, donde[6]

$$\Gamma(\theta, \theta_0) := E\left[\dot{y}_{\theta_0}(X) X^{\theta} \right] = \sum_{j=1}^{\infty} \frac{\langle h, e_j \rangle \langle \theta_0, e_j \rangle}{\lambda_j}.$$

Esta secuencia se introduce en Cardot et al. (2007) para obtener teoremas centrales del límite involucrando a $\hat{\theta}_n - \theta_0$. La secuencia tiene un orden entre $O(1)$ y $O\left(n^{1/2}\right)$ y, si X es gaussiana y satisface que $E\left[\|X\|^2\right] < \infty$, entonces $\lim_{n} \Gamma_n(\theta, \theta_0) = \Gamma(\theta, \theta_0)$ con $\Gamma(\theta, \theta_0) = \Gamma(\theta) \theta_0$ y $\Gamma(\theta) := \mathrm{Var}\left[X^{\theta} \right] \in \Psi_1$ y Γ siendo la densidad de una distribución normal estándar (Proposición 3.2 en CGFG19).

Teorema 4.1 (Teorema 3.1 en CGFG19) Sea $\theta \in \Theta$ jo. Bajo H_0, $f_\theta \in \Psi_1$, ciertas condiciones[7] y asumiendo independencia entre Y y X,[8] se tiene que:

[5] Es decir, es el proceso de funciones aleatorias que se obtienen al tomar $P(\theta) = B(C(\theta)) - C(1)B(\theta)$ donde B es un movimiento browniano estándar en $[0,1]$.

[6] $\dot{y}_{\theta_0}$ es la versión poblacional de $\dot{y}_{\theta_0,n} := \sum_{j=1}^{p_n} \frac{1}{\lambda_j} \langle \cdot, e_j \rangle e_j$.

[7] Véanse las condiciones A1–A9 en CGFG19.

[8] Esta hipótesis se necesita para probar que $(n^{1/2}) \beta_n^{(2)}(\theta) \to 0$ en probabilidad cuando $n \to \infty$. En el resto del artículo basta con que $E[Y \mid X] = 0$.

(a) $\lim_{n\to\infty} \mathbb{C}_{\mathrm{KS}}^{1}(\mathcal{H}_0) = \mathbb{C}_{\mathrm{KS}}^{2}(\mathcal{H}_0)$ $\mathbb{N}(0, \sigma_{\mathrm{KS}}^2)$.

(b) Si $\lim_{n\to\infty} \mathbb{C}_{\mathrm{KS}}^{1}(\mathcal{H}_0) = 1$, entonces con $\mathcal{H}_0 = \mathcal{H}$ $\mathbb{C}_{\mathrm{KS}}^{1}(\mathcal{H}_0)$ en (16) la distribución asintótica de $\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)$ es la de $\mathbb{C}_{\mathrm{KS}}^{2}(\mathcal{H}_0)\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)$.

(c) Si $\lim_{n\to\infty} \mathbb{C}_{\mathrm{KS}}^{1}(\mathcal{H}_0) \neq 1$, entonces con $\mathcal{H}_0 = \mathcal{H}$ en (16), la distribución asintótica de $\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)$ es la de $\mathbb{C}_{\mathrm{KS}}^{2}(\mathcal{H}_0)\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)$, $\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)$.

El Teorema 4.1 se cumple para todo $G \in \mathcal{R}$. Para el caso (c) del Teorema 4.1 (donde la estimación de 1 no es dominante) y bajo un supuesto adicional, el resultado puede generalizarse a convergencia débil del proceso empírico.

Teorema 4.2 (Teorema 3.3 en CGFG19) Bajo $\mathcal{H}_0$, los supuestos del Teorema 4.1(c) y $\mathcal{H}_0 = \mathcal{H}$, se tiene que:

(a) Para cualquier colección finita $\{G_1, \ldots, G_k\}$ con $k \geq 1$, ocurre que

$$\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)(G_1, \ldots, G_k) \xrightarrow{d} \mathbb{N}(0, \Sigma),$$

con $\Sigma = (\sigma_{ij})_{i,j=1,\ldots,k}$ donde $\sigma_{ij} := \mathbb{C}_{\mathrm{KS}}^{2}(\mathcal{H}_0)\sigma_{G_i,G_j}$, $\sigma_{G_i,G_j} := \mathbb{C}_{\mathrm{KS}}^{1}(\mathcal{H}_0)\sigma_{G_i,G_j} + \mathbb{C}_{\mathrm{KS}}^{1}(\mathcal{H}_0)\sigma_{G_i,G_j}$ con

$$\sigma_{G_i,G_j} = \mathrm{E}\left[\mathrm{Var}\left[\mathbb{1}_{\{\mathcal{X}_n \leq \mathcal{H}_n\}} \mid \mathcal{X}_n = \mathcal{H}_n\right] \mathbb{1}_{\{\mathcal{H}_n \leq G_i\}} \mathbb{1}_{\{\mathcal{H}_n \leq G_j\}}\right]$$

$$\sigma_{G_i,G_j} = \mathrm{E}\left[\mathrm{Var}\left[\mathbb{1}_{\{\mathcal{Y}_m \leq \mathcal{H}_m\}} \mid \mathcal{Y}_m = \mathcal{H}_m\right] \mathbb{1}_{\{\mathcal{H}_m \leq G_i\}} \mathbb{1}_{\{\mathcal{H}_m \leq G_j\}}\right]$$

(b) Si $\mathrm{E}\left[\mathbb{1}_{\{\mathcal{X}_1 \leq \mathcal{H}_0\}}^{4}\right] = O(n^{-2-\varepsilon})$, entonces $\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0) \xrightarrow{d} \mathbb{G}_2$ en $\ell^\infty(\mathcal{R})$, con $\mathbb{G}_2$ un proceso gaussiano con media cero y función de covarianza $\mathcal{H}_0$.

El siguiente resultado proporciona las distribuciones asintóticas bajo $\mathcal{H}_0$ de las normas de Kolmogorov–Smirnov (KS) y Cramér–von Mises (CvM) sobre el proceso empírico $\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)$.

Corolario 4.1 (Corolario 3.4 en CGFG19) Bajo $\mathcal{H}_0$ y los supuestos del Teorema 4.2(b), si $\|\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)\|_{\mathrm{KS}} := \sup_{G\in\mathcal{R}} |\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)(G)|$ y $\|\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)\|_{\mathrm{CvM}} := \int_{\mathcal{R}} \mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)(G)^2 \, d\,\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)(G)$, siendo $\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)$ la distribución empírica de $\mathcal{X}_1, \ldots, \mathcal{X}_n$, entonces

$$\|\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)\|_{\mathrm{KS}} \xrightarrow{d} \|\mathbb{G}_2\|_{\mathrm{KS}} \quad \text{y}$$

$$\|\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)\|_{\mathrm{CvM}} \xrightarrow{d} \|\mathbb{G}_2\|_{\mathrm{CvM}} := \int_{\mathcal{R}} \mathbb{G}_2(G)^2 \, d\,\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)(G).$$

El Corolario 4.1 es interesante desde un punto de vista práctico porque garantiza que un test que rechaza $\mathcal{H}_0$ para $\|\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)\|_{\mathrm{KS}}$ mayor que el cuantil $1-\alpha$ de la v.a. $\|\mathbb{G}_2\|_{\mathrm{KS}}$ asintóticamente respeta un error de Tipo I igual a α (2 [0, 1]), con KS siendo o CvM. Sin embargo, el resultado tiene poca utilidad práctica debido al desconocimiento de la función de covarianza $\mathcal{H}_0$ que, en particular, involucra al inverso del operador de covarianza poblacional $\mathcal{H}_0$.

5 El test en la práctica

La principal ventaja de contrastar $\mathcal{H}_0$ frente a $\mathcal{H}_0$ es que en $\mathcal{H}_0$ la v.a. que condiciona es escalar. Los inconvenientes de este método universal son una posible pérdida de potencia y que el resultado de la prueba puede variar para diferentes proyecciones. Ambos inconvenientes pueden ser mitigados al simular varias direcciones $\mathcal{H}_1, \ldots, \mathcal{H}_n$, contrastar las hipótesis proyectadas $\mathcal{H}_0^{\mathcal{H}_1}, \ldots, \mathcal{H}_0^{\mathcal{H}_n}$, y combinar los p-valores resultantes. Por ejemplo, usando el método FDR propuesto en Benjamini y Yekutieli (2001) (véase la Sección 2.2.2 de Cuesta-Albertos y Febrero-Bande, 2010), es posible controlar la tasa final de rechazo para que sea como máximo α bajo $\mathcal{H}_0$.

La selección de direcciones aleatorias influye claramente en la potencia del test. El compromiso basado en los datos que se propone en CGFG19 para evitar obtener direcciones en subespacios casi ortogonales a los datos es el siguiente: (i) calcular la base de PCs de $\mathcal{X}_1, \ldots, \mathcal{X}_n$, es decir, el sistema propio $\{(\lambda_j, \mathcal{H}_j)\}_{j=1}^{p}$; (ii) elegir $q := \min\left\{k : \sum_{j=1}^{k} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j \geq A\right\}$ para un umbral de varianza A, por ejemplo $A = 0.95$; (iii) generar el proceso gaussiano basado en los datos $\mathcal{Z}_n = \sum_{j=1}^{q} \mathcal{Z}_j \mathcal{H}_j$, con $\mathcal{Z}_j \sim \mathcal{N}(0, \lambda_j)$ y λ_j^2 la varianza muestral de $\{\mathcal{Z}_{ij}\}_{i=1}^{n}$ (véase (7)).[10]

El procedimiento de agregación de p-valores procedentes de las hipótesis proyectadas se describe en el siguiente algoritmo genérico.

Algoritmo 5.1 (Procedimiento de contraste de $\mathcal{H}_0$) Sea $\mathcal{T}_n$ un test para verificar $\mathcal{H}_0$.

(i) Para $k = 1, \ldots, K$, sea p_k el p-valor de la hipótesis proyectada $\mathcal{H}_0^{\mathcal{H}_k}$ obtenido con el test $\mathcal{T}_n$.

(ii) Establecer el p-valor final de $\mathcal{H}_0$ como $\min_{k=1,\ldots,K}\{1 - (1 - p_k)^K, p_k\}$, donde $p_{(1)} \leq \cdots \leq p_{(K)}$.

[9] De acuerdo con el Teorema 1 en Cardot et al. (2007), es imposible que $\mathcal{H}_n$ converja a un elemento aleatorio no degenerado en la topología de $\mathcal{H}$. Esta condición permite obtener la tightness del proceso $\mathbb{T}_{\mathrm{KS}}(\mathcal{H}_0)$.

[10] Obsérvese que la medida gaussiana asociada a λ_j es degenerada. Una medida gaussiana no degenerada puede obtenerse como $\mathcal{Z}_j + \mathcal{G}$, con $\mathcal{G}$ una medida gaussiana fuertemente concentrada alrededor de cero, si bien la consideración o $\mathcal{Z}_j$, $\mathcal{G}$ tiene efectos despreciables en la práctica.

La calibración del estadístico de contraste $k_{N=-h}$ para $\overset{h}{0}$ que permite obtener ?-valores en (i) se realiza mediante el procedimiento wild bootstrap descrito en el Algoritmo 5.2. Este procedimiento de remuestreo perturba los residuos del ajuste del FLMSR, genera muestras bootstrap bajo y reajusta el FLMSR.[11]

Algoritmo 5.2 (Calibración bootstrap del test) Para contrastar $\overset{h}{0}$, con $h \in H$ dada, se procede de la siguiente manera:

(i) Calcular $\hat{1}_{:=}$ según (8) para un $:=$ dado y obtener $\hat{Y}_8 = \cdot_8 \wedge_{8-1}\hat{1}_{:=}$.

(ii) Calcular $k_{N=-h}$ $= \; = \; 1 \cdot 2 \; \acute{\imath} \; {}_{8=1}^{=} 1_{f \wedge \overset{h}{8}} \; G_g \; \hat{Y}_8 \; {}_N$.

(iii) Remuestreo bootstrap. Para $1 = 1 - \cdots -$:

(a) Simular v.a.'s binarias i.i.d. $+_1 - \cdots -_=$ tales que $P \overset{h}{+} = {}^1 1 \; \overset{p}{5} \cdot 2 \; \acute{\imath} = {}^1 5 \; \overset{p}{5} \cdot 10$.

(b) Establecer $\cdot_8 := \wedge_{8-1}\hat{1}_{:=}$, Y_8 usando los errores bootstrap $Y_8 := +_8 \hat{Y}_8$.

(c) Calcular $\hat{1}_{:=}$ según (8) pero a partir de ${}^1 \wedge_{8-} \cdot_8 \; {}^{o}_{8=1}$.

(d) Obtener los residuos bootstrap $\hat{Y}_8 := \cdot_8 \; \overset{D}{\wedge}_{8-1}\hat{1}_{:=}$ $\overset{E}{}$.

(e) Calcular $k_{N=-h}^{1}$ $:= \; = \; 1 \cdot 2 \; \acute{\imath} \; {}_{8=1}^{=} 1_{f \wedge \overset{h}{8}} \; G_g \; \hat{Y}_8 \; {}_N$.

(iv) Aproximar el ?-valor mediante $\overset{1}{-} \; \acute{\imath} \; {}_{1=1} 1_{f k_{N=-h}^{1} \; k_N} \; k_{N=-h} \; g$.

La elección de $:=$ puede realizarse mediante un criterio de selección de modelo, como se avanzaba en (10). En CGFG19 se recomienda usar el SICc (criterio de información de Schwartz corregido de McQuarrie, 1999), donde $\overset{1}{=} - \cdot^{o} = 2 \log^{1} =^{o} : \bullet^{1} = : \; 2^{o}$, para llevar a cabo el Algoritmo 5.2.

En la Sección 5 de CGFG19 se realiza un estudio de simulación para explorar la calibración y la potencia del test bajo diferentes escenarios y variaciones de parámetros, como por ejemplo el número de proyecciones a emplear. Las principales conclusiones que se extraen es que un número relativamente bajo de proyecciones, en torno a $2 \; f \; 3 - 5 \; g$, es un buen compromiso entre mantener el nivel de significación y aumentar la potencia del test. Asimismo, en el estudio de simulación se evidencia una importante mejora computacional frente al contraste de bondad de ajuste del FLMSR propuesto en García-Portugués, González-Manteiga y Febrero-Bande (2014). Esta mejora se atribuye a que el estadístico de contraste en ibid usa una norma CvM que, de forma resumida, integra el proceso $G7! \;)_{=-h} \; {}^1 G^{o}$ con respecto a ${}^1 G - h^{o}$, lo que hace mucho más costoso su cálculo frente a la exploración de direcciones aleatorias que se propone en CGFG19.

6 Ejemplo y software

Volviendo al ejemplo motivador de la Sección 2.3, nuestro objetivo es determinar si la relación de la variable respuesta $\log^{1}$NBCR) sigue o no una relación lineal con alguna de las variables humedad y sensación térmica. Una pista de esta relación se puede obtener estimando ambas regresiones y usando el usual gráfico de diagnóstico de los residuos respecto a los valores ajustados. Si la relación fuese efectivamente lineal, debemos observar un patrón promedio constante (e igual a cero) a lo largo de los valores ajustados. Para los dos modelos planteados, este diagnóstico se muestra en la Figura 2 donde, en ambos casos, la nube de puntos se aleja del patrón constante más acusadamente en el caso de la sensación térmica (derecha).

El resto de análisis de estos dos modelos de regresión no proporcionan mejores pistas sobre el objetivo: los modelos son no triviales usando $:= = 4$ (humedad) y $:= = 6$ (sensación térmica) según determina el SICc, con $'^{2}_{Hum} = 0 - 32$ y $'^{2}_{Sens} = 0 - 70$. Aplicando el test a cada uno de los ejemplos (2000 réplicas bootstrap, hasta $= 5$ proyecciones aleatorias) se obtienen los ?-valores que se pueden ver en la Tabla 1.

	Humedad		Sensación térmica	
Proy.	CvM	KS	CvM	KS
1	0.748	0.852	0.005	0.012
2	0.753	0.852	0.000	0.000
3	0.753	0.852	0.000	0.000
5	0.753	0.852	0.000	0.000

Tabla 1. ?-valores de la aplicación del test para los estadísticos Cramér–von Mises (CvM) y Kolmogorov–Smirnov (KS) a los modelos de regresión con distinto número de proyecciones aleatorias.

Como se puede ver, los tests no rechazan la tesis de linealidad para el modelo de regresión con la humedad como predictor, mientras que la rechazan para el modelo con la sensación térmica.

Los datos y código usados en este ejemplo están disponibles en el repositorio https://github.com/mfebrero/BEIO_bike. El test de bondad de ajuste presentado está implementado en

[11] El caso particular de la hipótesis nula simple corresponde a $1 = 0$, por lo que su calibración corresponde a establecer $\hat{1}_{:=} = \hat{1}_{:=} = 0$ en el algoritmo.

Figura 2. Gráficos de residuos vs valores ajustados para los modelos con humedad (izquierda) y con temperatura (derecha) como predictores.

la función rp.flm.test() de la librería fda.usc (Febrero-Bande y Oviedo de la Fuente, 2024).

7 Algunas extensiones

Como ya se comentaba en la discusión de CGFG19, la metodología allí desarrollada para el contraste de especificación del FLMSR prometía varias extensiones susceptibles de llevarse a cabo en años posteriores. En particular:

(a) Se planteaba la posibilidad de extender el contraste de especificación al modelo parcialmente lineal funcional en el que aparecen covariables funcionales y multivariantes $(X, Z) \in \mathcal{H} \times \mathbb{R}^q$:

$$Y = \langle X, \beta \rangle + m(Z) + \varepsilon,$$

véase Aneiros-Pérez y Vieu (2006) o más recientemente Aneiros-Pérez, Cao, Fraiman, Genest y Vieu (2019) para más detalles sobre este modelo. Desde la publicación de CGFG19 ha habido cierta actividad en esta dirección, destacando los trabajos de Zhang, Sun y Sun (2023) y Yu, Song y Du (2024) que estudian la estimación y la especificación de estos modelos con la presencia de errores en las covariables escalares o con alternativas en base al modelo de regresión cuantil. En Chan, Delsol y Goia (2023) y Xia, Lai y Zhang (2023) se aborda el contraste de especificación del modelo de índice único

$$Y = g(\langle X, \beta \rangle) + \varepsilon,$$

con $X \in \mathcal{H}$, $\|\beta\| = 1$ y $g : \mathbb{R} \to \mathbb{R}$ una función enlace desconocida. En todos estos trabajos las técnicas de contrastes utilizadas se basaron en técnicas de suavizado alternativas a las que se establecieron en CGFG19 con procesos empíricos de regresión.

(b) Se establecía la posibilidad de extender el método de contraste a modelos cuádricos bajo la hipótesis nula. Que sepamos, hasta la actualidad no ha habido avances en esta dirección y el modelo cuadrático de Horváth y Reeder

(2013),

$$\ldots = h^{\wedge} - 1 i \; ¸ \quad {}^{1}B - {}^{o}C \; {}^{1}B^{o} \wedge {}^{1}C^{o} \, dC \, dB -$$

solo se ha considerado en simulaciones comparativas, pero como modelo bajo una hipótesis alternativa (como, por ejemplo, en Patilea y Sánchez-Sellero, 2020).

(c) La tercera línea de extensión que se mencionaba en CGFG19 es sobre la consideración de modelos con respuesta funcional:

$$\_ = < {}^{1}\wedge {}^{o} \; ¸ \; Y -$$

con ${}^{1}\wedge - \_ {}^{o} \; 2 \; H_1 \; H_2$ y $<$ un operador de H_1 en H_2. En esta línea ha habido avances destacables sobre extensiones basadas en proyecciones para el contraste del modelo lineal funcional con respuesta funcional, así como de software y herramientas gráficas (García-Portugués y Álvarez-Liébana, 2023; García-Portugués, Álvarez-Liébana, Álvarez-Pérez y González-Manteiga, 2021; García-Portugués, Álvarez-Liébana, Álvarez-Pérez y González-Manteiga, 2020). Véase también Chen, Jiang, Feng y Zhu (2020), Smaga (2024) y Peng, Wang y Zou (2023) para artículos relacionados.

Además de los avances en las direcciones que se apuntaban en la discusión de CGFG19, en estos cinco años desde la publicación del trabajo se han producido otras contribuciones relacionadas con el trabajo, que destacamos a continuación:

(i) El contraste de especificación de un modelo de regresión cuantil lineal funcional con respuesta escalar corresponde a la hipótesis nula

$${}_{0}: P^{1}. \quad \&. \; {}^{1}g^{o} \; j \wedge {}^{o} = h^{\wedge} - 1 i \quad \text{para un} \; g \; 2 \; {}^{1}0 - 1^{o} \; jo -$$

siendo $\&. \; {}^{1}g^{o}$ el cuantil g (incondicional) de la respuesta $\ldots$. En Feng, Li y Song (2024) se aborda, con un desarrollo muy paralelo a la metodología de CGFG19, el caso de la hipótesis nula simple $1 = 0$ (no asociación cuantílica). En Shi, Du, Sun y Zhang (2021), siguiendo una metodología alternativa con técnicas de suavizado, se trata la hipótesis compuesta. En Liu, Peng, Liu, Lian y Liu (2023) se plantea la estimación del modelo más general

$$P^{1}. \quad \&. \; {}^{1}g^{o} \; j \wedge {}^{o} = \int_{0}^{1} {}^{1}C \triangle {}^{1}C^{o} \, dC$$

para un $g \; 2 \; {}^{1}0 - 1^{o} \; jo -$

que, sin duda, puede representar una base válida para el diseño de nuevos tests en el futuro. Véase también Xia, Zhang y Shi (2024) para una referencia muy reciente.

(ii) En Febrero-Bande, Galeano, García-Portugués y González-Manteiga (2024); Febrero-Bande, Galeano y González-Manteiga (2019) se extiende el estudio de la estimación y de los correspondientes contrastes de especificación para aquellas situaciones en las que la muestra de la variable de respuesta escalar puede presentar datos faltantes que requieren de una adecuada imputación.

(iii) En la línea del estudio de series de tiempo con datos funcionales $f \wedge {}_{C} g_{C} \; {}_{2N} \; H$ (véase Kim, Kokoszka y Rice (2023) para una revisión importante de este contexto moderno), en Álvarez Liébana, López-Pérez, González-Manteiga y Febrero-Bande (2022) se aborda computacionalmente la metodología de CGFG19 trasladada al contraste de especificación de un proceso autorregresivo Hilbertiano:

$${}_{0}: \wedge {}_{C} = < {}^{1}\wedge {}_{C} \; {}_{1}{}^{o} \; ¸ \; Y_{C} -$$

siendo $<$ un operador lineal. Véase González-Manteiga, Ruiz-Medina y Febrero-Bande (2024) para un estudio más teórico de la convergencia del proceso asociado al estadístico de contraste.

(iv) Las técnicas inspiradas en la llamada correlación de distancias también han sido aplicadas para desarrollar nuevos contrastes de especificación de modelos con covariables funcionales. A título de ejemplo mencionamos el trabajo de Lai, Zhang y Wang (2021) y recomendamos, para más detalles sobre otras referencias recientes, la lectura de las revisiones de González-Manteiga, Crujeiras y García-Portugués (2023) y González-Manteiga (2022).

Agradecimientos

Los autores agradecen a los editores de BEIO la invitación para escribir este artículo.

Acerca de los autores

Juan A. Cuesta-Albertos es recientemente jubilado. Fue Catedrático de Universidad del área de Estadística e I.O. en el Departamento de Matemáticas, Estadística y Computación de la Universidad de Cantabria. Es autor de 76 trabajos publicados en revistas científicas centrados en temas como el problema del transporte óptimo, la estadística robusta y el análisis de datos funcionales, entre otros. Ha sido Editor de Test y Editor Asociado del Journal of the American Statistical Association. También ha participado en proyectos de I+D y es un entusiasta de la divulgación científica. Su índice h es 24 en Web of Science y 33 en Google Scholar. Ha recibido el Premio SEIO-Fundación BBVA a la "Mejor Contribución Metodológica en el campo de la Estadística" en las ediciones primera y cuarta de este premio. También es "Honored Fellow" del Institute of Mathematical Statistics.

Eduardo García-Portugués es Profesor Titular en el Departamento de Estadística de la Universidad Carlos III de Madrid y Doctor en Estadística e I.O. por la Universidade de Santiago de Compostela. Su investigación se centra en el desarrollo de metodologías estadísticas relacionadas con datos complejos y no euclidianos (como datos direccionales y funcionales) desde las perspectivas de la estadística matemática y computacional. En particular, ha trabajado en contrastes de bondad de ajuste, inferencia basada en suavizado, modelización de procesos difusivos y desarrollo de software estadístico. Es autor de 28 artículos y 6 capítulos de libros y ha sido IP de 5 proyectos de I+D. Es Editor Asociado en las revistas Journal of Computational and Graphical Statistics, Computational Statistics y Statistics and Probability Letters.

Manuel Febrero-Bande es Catedrático de Universidad del área de Estadística e I.O. en el Departamento de Estadística, Análisis Matemático y Optimización de la Universidade de Santiago de Compostela. Es autor/coautor de 75 trabajos científicos como series de tiempo, estadística espacial, datos funcionales y, en general, en modelos avanzados de predicción y computación estadística. Ha participado en más de 60 proyectos I+D desde 2005 en ámbitos como monitorización para control medioambiental, control de la producción y calidad de procesos industriales, análisis de datos epidemiológicos y sanitarios o reconocimiento de patrones en muestras grandes o complejas. Es Editor Asociado de las revistas Test y Computational Statistics.

Wenceslao González-Manteiga es Catedrático de Universidad en el Departamento de Estadística, Análisis Matemático y Optimización, en la Universidade de Santiago de Compostela. ISI Member, IMS Fellow Member (2017) y Miembro de la Real Academia Galega de Ciencias (2018). En 2020 recibió la Medalla de la SEIO y Premio SEIO-Fundación BBVA a la mejor contribución metodológica en el campo de la Estadística. En el 2021 fue galardonado con el Premio Nacional de Estadística por el Instituto Nacional de Estadística. Recientemente ha sido nombrado Académico Correspondiente de la Real Academia de Ciencias Exactas, Físicas y Naturales de España y de la Academia de Ciencias Matemáticas, Físico-Químicas y Naturales de Granada.

Bibliografía

Álvarez Liébana, J., López-Pérez, A., González-Manteiga, W. y Febrero-Bande, M. (2022). A goodness-of- t test for functional time series with applications to Ornstein–Uhlenbeck processes. arXiv preprint 2206.12821.

Aneiros-Pérez, G., Cao, R., Fraiman, R., Genest, C. y Vieu, P. (2019). Recent advances in functional data analysis and high-dimensional statistics. Journal of Multivariate Analysis, 170, 3–9. doi: 10.1016/j.jmva.2018.11.007

Aneiros-Pérez, G. y Vieu, P. (2006). Semi-functional partial linear regression. Statistics & Probability Letters, 76(11), 1102–1110. doi: 10.1016/j.spl.2005.12.007

Barnard, G. A. (1992). Introduction to Pearson (1900) on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. En S. Kotz y N. L. Johnson (Eds.), Breakthroughs in statistics: Methodology and distribution (pp. 1–10). New York: Springer. doi: 10.1007/978-1-4612-4380-9_1

Benjamini, Y. y Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29(4), 1165–1188. doi: 10.1214/aos/1013699998

Billingsley, P. (1999). Convergence of probability measures. New York: John Wiley & Sons. doi: 10.1002/9780470316962

Cardot, H., Ferraty, F. y Sarda, P. (2003). Spline estimators for the functional linear model. Statistica Sinica, 13(3), 571–592. Retrieved from https://www.jstor.org/stable/24307112

Cardot, H., Mas, A. y Sarda, P. (2007). CLT in functional linear regression models. Probability Theory and Related Fields, 138(3-4), 325–361. doi: 10.1007/s00440-006-0025-2

Chan, L., Delsol, L. y Goia, A. (2023). A link function speci cation test in the single functional index model. Advances in Data Analysis and Classi cation. doi: 10.1007/s11634-023-00545-7

Chen, F., Jiang, Q., Feng, Z. y Zhu, L. (2020). Model checks for functional linear regression models based on projected empirical processes. Computational Statistics & Data Analysis, 144. doi: 10.1016/j.csda.2019.106897

Cuesta-Albertos, J. A. y Febrero-Bande, M. (2010). A simple multiway ANOVA for functional data. Test, 19(3), 537–557. doi: 10.1007/s11749-010-0185-3

Cuesta-Albertos, J. A., Fraiman, R. y Ransford, T. (2007). A sharp form of the Cramér-Wold theorem. Journal of Theoretical Probability, 20(2), 201–209. doi: 10.1007/s10959-007-0060-7

Cuesta-Albertos, J. A., García-Portugués, E., Febrero-Bande, M. y González-Manteiga, W. (2019). Goodness-of- t tests for the functional linear model based on randomly projected empirical processes. The Annals of Statistics, 47(1), 439–467. doi: 10.1214/18-AOS1693

Fanaee-T, H. y Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. Progress in Arti cial Intelligence, 2(2), 113–127. doi: 10.1007/s13748-013-0040-3

Febrero-Bande, M., Galeano, P., García-Portugués, E. y González-Manteiga, W. (2024). Testing for linearity in scalar on function regression with responses missing at random. Computational Statistics, 39(6), 3405–3429. doi: 10.1007/s00180-023-01445-2

Febrero-Bande, M., Galeano, P. y González-Manteiga, W. (2019). Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random. Computational Statistics & Data Analysis, 131, 91–103. doi: 10.1016/j.csda.2018.07.006

Febrero-Bande, M. y Oviedo de la Fuente, M. (2024). fda.usc: Functional data analysis and utilities for statistical computing [Computer software manual]. Retrieved from https://cran.r-project.org/package=fda.usc (R package version 2.2.0)

Feng, Y., Li, J. y Song, X. (2024). Testing conditional quantile independence with functional covariates. Biometrics, 80(2). doi: 10.1093/biomtc/ujae036

Ferraty, F. y Vieu, P. (2006). Nonparametric functional data analysis. New York: Springer. doi: 10.1007/0-387-36620-2

García-Portugués, E. y Álvarez-Liébana, J. (2023). goffda: Goodness-of- t tests for functional data [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=goffda (R package version 0.1.2)

García-Portugués, E., Álvarez-Liébana, J., Álvarez-Pérez, G. y González-Manteiga, W. (2021). A goodness-of- t test for the functional linear model with functional response. Scandinavian Journal of Statistics, 48(2), 502–528. doi: 10.1111/sjos.12486

García-Portugués, E., González-Manteiga, W. y Febrero-Bande, M. (2014). A goodness-of- t test for the functional linear model with scalar response. Journal of Computational and Graphical Statistics, 23(3), 761–778. doi: 10.1080/10618600.2013.812519

García-Portugués, E., Álvarez-Liébana, J., Álvarez-Pérez, G. y González-Manteiga, W. (2020). Goodness-of- t tests for functional linear models based on integrated projections. En G. Aneiros, I. Horová, M. Hušková y P. Vieu (Eds.), Functional and high-dimensional statistics and related fields (pp. 107–114). Cham: Springer. doi: 10.1007/978-3-030-47756-1_16

González-Manteiga, W. (2022). A review on speci cation tests for models with functional data. Spanish Journal of Statistics, 4, 9–40. doi: 10.37830/sjs.2022.1.02

González-Manteiga, W. y Crujeiras, R. M. (2013). An updated review of goodness-of- t tests for regression models. Test, 22(3), 361–411. doi: 10.1007/s11749-013-0327-5

González-Manteiga, W., Crujeiras, R. M. y García-Portugués, E. (2023). A review of goodness-of- t tests for models involving functional data. In N. Balakrishnan, M. A. Gil, N. Martín, D. Morales y M. C. Pardo (Eds.), Trends in mathematical, information and data sciences (Vol. 445, pp. 349–358). Cham: Springer. doi: 10.1007/978-3-031-04137-2
González-Manteiga, W., Ruiz-Medina, M. D. y Febrero-Bande, M. (2024). Linear parametric model checks for functional time series. arXiv preprint 2303.09644.
Hall, P. y Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. Journal of the Royal Statistical Society Series B (Statistical Methodology), 68(1), 109–126. doi: 10.1111/j.1467-9868.2005.00535.x
Härdle, W. y Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. The Annals of Statistics, 21(4), 1926–1947. doi: 10.1214/aos/1176349403
Horváth, L. y Reeder, R. (2013). A test of signi cance in functional quadratic regression. Bernoulli, 19(5A), 2120–2151. doi: 10.3150/12-BEJ446
Kim, M., Kokoszka, P. y Rice, G. (2023). White noise testing for functional time series. Statistics Surveys, 17, 119–168. doi: 10.1214/23-SS143
Kolmogoroff, A. (1933). Sulla determinazione empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari, 4, 83–91. Retrieved from http://digitale.bnc.roma.sbn.it/tecadigitale/giornale/CFI0353791/1933/unico/00000093
Lai, T., Zhang, Z. y Wang, Y. (2021). Testing independence and goodness-of- t jointly for functional linear models. Journal of the Korean Statistical Society, 50(2), 380–402. doi: 10.1007/s42952-020-00083-4
Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62(318), 399–402. doi: 10.1080/01621459.1967.10482916
Liu, Y., Peng, L., Liu, Q., Lian, H. y Liu, X. (2023). Functional additive expectile regression in the reproducing kernel Hilbert space. Journal of Multivariate Analysis, 198. doi: 10.1016/j.jmva.2023.105214
McQuarrie, A. D. (1999). A small-sample correction for the Schwarz SIC model selection criterion. Statistics & Probability Letters, 44(1), 79–86. doi: 10.1016/S0167-7152(98)00294-6
Patilea, V. y Sánchez-Sellero, C. (2020). Testing for lack-of- t in functional regression models against general alternatives. Journal of Statistical Planning and Inference, 209, 229–251. doi: 10.1016/j.jspi.2020.04.002
Peng, L., Wang, G. y Zou, C. (2023). Measuring, testing, and identifying heterogeneity of large parallel datasets. Statistica Sinica, 33(4), 2787–2808. doi: 10.5705/ss.202021.0285
Shi, G., Du, J., Sun, Z. y Zhang, Z. (2021). Checking the adequacy of functional linear quantile regression model. Journal of Statistical Planning and Inference, 210, 64–75. doi: 10.1016/j.jspi.2020.05.003
Smaga, L. (2024). Projection tests for linear hypothesis in the functional response model. Communications in Statistics-Theory and Methods, 53(4), 1402–1419. doi: 10.1080/03610926.2022.2101120
Stute, W. (1997). Nonparametric model checks for regression. The Annals of Statistics, 25(2), 613–641. doi: 10.1214/aos/1031833666
Xia, L., Lai, T. y Zhang, Z. (2023). An adaptive-to-model test for parametric functional single-index models. Mathematics, 11(8). doi: 10.3390/math11081812
Xia, L., Zhang, Z. y Shi, G. (2024). A consistent speci cation test for functional linear quantile regression models. Statistics and Its Interface, 17(4), 649–667. doi: 10.4310/22-SII754
Yu, P., Song, X. y Du, J. (2024). Composite expectile estimation in partial functional linear regression model. Journal of Multivariate Analysis, 203. doi: 10.1016/j.jmva.2024.105343
Zhang, T., Sun, Z. y Sun, L. (2023). Goodness-of- t test for partial functional linear model with errors in scalar covariates. Journal of Statistical Planning and Inference, 227, 91–111. doi: 10.1016/j.jspi.2023.04.001

Vol. 40 Núm. 3
Investigación Operativa

Eligiendo candidatos usando programación lineal y encuestas: Caso MORENA 2024

J. Macías Ponce W. Olvera - Lopez

Departmento de Matemáticas y Física. Universidad Autónoma de Aguascalientes. cesar.macias@edu.uaa.mx

Facultad de Economía. Universidad Autónoma de San Luis Potosí. william.olvera@uaslp.mx

Citar como:
Macías Ponce, J. and Olvera - Lopez, W. (2024). Eligiendo candidatos usando programación lineal y encuestas: Caso MORENA 2024. Boletín de Estadística e Investigación Operativa. 40(3), 21-27.

Abstract

Analizamos la elección de candidatos a las gubernaturas a renovar dentro del proceso electoral del 2024 por parte del partido Movimiento de Regeneración Nacional (MORENA) desde un enfoque basado en problemas de optimización, respetando la cuota de paridad establecida por la autoridad electoral. Planteamos y resolvemos dos problemas de programación lineal que elegían a la combinación de candidatos que maximizan cierta función de utilidad y comparamos los resultados que obtuvimos con lo que realmente ocurrió.

Keywords: elecciones, asignación, optimización, programación lineal, cuota de paridad.

1 Introducción

La participación de las mujeres en México dentro de los procesos de votación democráticos para la elección de puestos populares inicia formalmente en 1947 durante el gobierno de Miguel Alemán Valdés, cuando se reconoce el derecho de las mujeres a votar y ser votadas en procesos de elección municipales únicamente, como aparece publicado en el Diario Oficial de la Federación del 17 de febrero de dicho año en cuestión. Sí que hubo intentos anteriores para la incorporación de las mujeres a la vida política en estados como San Luis Potosí, Yucatán, Tabasco y Chiapas en años anteriores, pero fue hasta 1953 cuando el presidente Adolfo Ruiz Cortines, con base en una iniciativa iniciada por el presidente Lázaro Cárdenas del Río, en que las mujeres obtuvieron el derecho pleno al voto, el cual fue ejercido por vez primera en 1955 de Gobernación (2018); de las Mujeres (2024).

Si bien a partir de esas fechas las mujeres ya contaban con el derecho a votar, el hecho de ser votadas era un proceso que se veía lejano. Los registros históricos nos indican que Rosa Torre González funge, en 1923, como la primera presidenta municipal de México, representando a Mérida, aunque los mismos registros nos indican que tras el asesinato de Felipe Carillo Puerto,

las mujeres en posiciones del gobierno fueron obligadas a renunciar y el voto femenino fue anulado, por lo que se toma este hecho como una victoria parcial de las mujeres en su camino a la obtención de puestos de gobierno de Yucatán (2022). Se reconoce formalmente a Martha Aurora Jiménez Quevedo como la primera Diputada Federal y a María Lavalle Urbina como la primera mujer presidenta del Senado, mientras que la primera gobernadora (por Colima) ocurre en el año 1979 a través de la maestra Griselda Álvarez Ponce de León, a su vez que la primera candidata a la Presidencia de la República fue María del Rosario Ibarra de la Garza en 1982, por el Partido Revolucionario de los Trabajadores Noticias (2023); Olvera (2023). Al ir analizando estas fechas, nos damos cuenta de las desventajas de las mujeres respecto a los hombres en la carrera por puestos de elección popular. Para información más detallada se recomienda consultar a de las Mujeres (2018).

No fue sino hasta 1993 en que se establece en el Código Federal de Instituciones y Procedimientos Electorales (COFIPE) una recomendación hacia los partidos políticos de promover un mayor número de mujeres en actividades políticas de los mismos: fácil de ignorar y por lo tanto seguir con la disparidad a cambio de intereses. En el 1996, el mismo COFIPE, establece que los partidos políticos deben considerar en sus estatutos que las candidaturas por senadores y diputados no excedan del 70 % para un mismo género promoviendo una mayor participación de las mujeres, pero por su ambigüedad, los partidos siguieron ignorando las recomendaciones y la brecha de género no parecía cerrarse. Pero en el 2002, el COFIPE establece con claridad en su artículo 175-A que “De la totalidad de solicitudes de registro, tanto de diputados como de senadores, que presenten los partidos políticos ante el IFE, en ningún caso incluirán más de 70 % de candidatos propietarios de un mismo género”, por lo que es la primera medida real en la legislación mexicana que establece una llamada “cuota de género” o “cuota de paridad” de Gobernación (2002); Reynoso y D'angelo (2006). Y es a partir de entonces en que la brecha de género empieza a reducirse paulatinamente, pero de manera sostenida, y ya podemos ver sin ninguna extrañeza candidatas mujeres a la Presidencia de la República, a gubernaturas y demás puestos de elección. En la Reforma Electoral del 2014 se eleva a rango constitucional la obligatoriedad del principio de paridad en elecciones de legislaturas federales y estatales, pero la Reforma Constitucional de 2019 fue la que estableció la llamada “paridad en todo” Electoral (2024), que establece que en la toma de decisiones debe haber 50 % hombres y 50 % mujeres, con lo cual se pudo ver por primera vez a una mujer como Presidenta de la Suprema Corte de Justicia de la Nación (Norma Lucía Piña Hernández) y hasta 10 mujeres gobernadoras en el año 2023 Navarrete (2023).

Siguiendo esta regla, y buscando aplicar a las elecciones para la gubernatura durante 2024 en nueve entidades federativas, en sesión extraordinaria durante 2023 el INE establece que “los partidos políticos nacionales deberán postular al menos cinco mujeres en las nueve entidades, donde se elige la gubernatura y la Jefatura de Gobierno en la Ciudad de México” de Gobernación (2023); Electoral (2023). La resolución obtuvo apelaciones y rechazos, pero fue finalmente ratificada por el Tribunal Electoral del Poder Judicial de la Federación (TEPJF). Ya quedaba al interior de los partidos discutir y elegir la forma en que se cumpliría la cuota entre sus candidatos propuestos.

El partido político Movimiento de Regeneración Nacional (de aquí en adelante refereido también por el acrónimo MORENA), y partidos de coalición, anunciaron que mediante encuestas en los estados donde se elegía gobernador (Chiapas, Ciudad de México, Guanajuato, Jalisco, Morelos, Puebla, Tabasco, Veracruz y Yucatán) surgirían los llamados “Coordinadores de los Comités de defensa de la Cuarta Transformación”, quienes serán también lo candidatos a dicha gubernaturas. El dirigente de dicho partido, Mario Martín Delgado Carillo, indicó que, si los resultados de las encuestas no daban como ganadoras a cinco mujeres, la Comisión de Elecciones de su partido elegiría las mujeres que tengan los perfiles más competitivos en sus respectivos estados. No quedó claro el criterio formal que se usaría para comparar la competitividad de los perfiles. El 10 de noviembre, en conferencia de prensa, se anunciaron tanto los resultados de las encuestas como los coordinadores/candidatos en cada estado Muñoz (2023); Redacción (2023): Clara Marina Brugada Molina (Ciudad de México), Alma Edwviges Alcaraz Hernández (Guanajuato), Claudia Delgadillo González (Jalisco), Margarita González Saravia Calderón (Morelos) y Norma Rocío Nahle García (Veracruz) serían las mujeres candidatas, mientras que Eduardo Ramírez Aguilar (Chiapas), Alejandro Armenta Mier (Puebla), Javier May Rodríguez (Tabasco) y Joaquín Jesús Díaz Mena (Yucatán) serían los candidatos hombres. Es claro que se tuvo que aplicar algún criterio (presumiblemente, los perfiles más competitivos) para seleccionar a tres de las mujeres ya que los resultados de las encuestas arrojaban como ganadores a siete hombres. Los ganadores hombres de las encuestas que fueron intercambiados por mujeres fueron Omar Hamid Garfuch (Ciudad de México), Francisco Ricardo Sheffield Padilla (Guanajuato) y Carlos Lomelí Bolaños (Jalisco).

Por todo lo anterior, en este trabajo tratamos de encontrar una selección óptima de candidatos a gobernador por el partido MORENA usando criterios de optimización y programación lineal. En la Sección 2 hablamos brevemente de la herramienta utilizada para el análisis y en Sección 3 mostramos su aplicación a los datos, así como las soluciones obtenidas bajo cada uno

de los criterios de maximización que planteamos, comparando los resultados que obtenemos con lo que el partido MORENA eligió. Finalmente, en la Sección 4 mostramos algunas conclusiones del trabajo y recomendaciones para futuros problemas similares.

2 Metodología

Un problema de programación lineal, en su forma más general, consiste de un objetivo a optimizar (que puede ser una maximización o minimización) y un conjunto finito de restricciones, donde tanto el objetivo como las restricciones se representan mediante relaciones lineales. Denotando $x_1, x_2, \dots, x_n$ al conjunto de n variables que deben encontrarse, una forma general de un problema de programación lineal es

$$\max^{1} c_1x_1 + c_2x_2 + \cdots + c_nx_n \text{ sujeto a}$$

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &\leq b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &\leq b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &\leq b_m \end{aligned}$$

y además $x_i \geq 0$ para todo $i \in \{1, 2, \dots, n\}$, donde las constantes c_i, $i \in \{1, 2, \dots, n\}$ y a_{ij}, $i \in \{1, 2, \dots, m\}$, $j \in \{1, 2, \dots, n\}$ no son todas iguales a cero, con b_i, $i \in \{1, 2, \dots, m\}$ cualesquiera números reales. La relación lineal a optimizar (en el caso anterior, a maximizar) se conoce como función objetivo, mientras que al conjunto de m desigualdades se les conoce como restricciones. Para una mayor descripción del problema anterior, los supuestos que involucra y las implicaciones matemáticas del mismo, referimos al lector a Operations Research, an Introduction, de Frederick Hillier Hillier (2015), pero en general se encuentran las bases, al menos, en cualquier libro de investigación de operaciones.

Existen múltiples variaciones al problema original de programación lineal. Por ejemplo, el conjunto de restricciones puede incluir todo tipo de desigualdades e incluso igualdades. También puede darse que, bajo cierto contexto, las variables a encontrar tengan que ser valores enteros, surgiendo lo que se conoce como programación entera. Un sub-caso de mucho interés ocurre cuando además de ser enteras, se requiere que las variables únicamente tomen valores en el conjunto $\{0, 1\}$, y se conoce como programación binaria. La importancia proviene del hecho que una variable, al tomar solamente valores en este conjunto, puede modelar situaciones donde lo que representa la variable es o no está presente. Con todo lo anterior se plantean de manera efectiva (junto con otras restricciones) múltiples situaciones de asignación, que será el caso que ocuparemos en nuestra análisis de elección de candidatos a las gubernaturas.

La programación lineal tuvo un auge importante durante la Segunda Guerra Mundial, ya que se planteó de manera efectiva para resolver problemas de transporte, disponibilidad de recursos y asignación, entre otros. Ello ocasionó que, debido a su estructura, se desarrollaran técnicas y algoritmos especializados para encontrar su solución en un tiempo razonable. En 1947, George B. Dantzig inventó el llamado método símplex Dantzig (1951) que, de manera eficiente, encontraba solución a un gran grupo de problemas de programación lineal. A partir de ese momento, hubo un boom de técnicas y desarrollo de algoritmos particulares para casos particulares de problemas de programación lineal como lo son el método de la M grande, algoritmos de punto interior, ramificación y acotamiento (para programación entera) por mencionar sólo algunos, aunque la investigación sigue muy activa. El segundo gran avance de la investigación de operaciones se da con el desarrollo de la computadora, ya que se permiten realizar los múltiples cálculos que involucra en fracciones de segundo. En la actualidad, es posible encontrar software gratuito y aplicaciones en internet que resuelven de manera eficiente gran cantidad de problemas de investigación de operaciones, por lo que se tiene un acercamiento muy importante de esta disciplina a sectores no especializados en la teoría detrás de los métodos, pero quienes tienen problemas que se resuelven con ellos.

3 Aplicación a la elección de candidatos de MORENA

Definamos el conjunto de los estados en donde se va a elegir candidatos para la gubernatura como

$$E = \{Y, V, T, P, M, J, G, C, D\}$$

donde cada letra indica la inicial del estado correspondiente: Yucatán (Y), Veracruz (V), Tabasco (T), Puebla (P), Morelos (M), Jalisco (J), Guanajuato (G), Chiapas (C) y Ciudad de México (D), y al conjunto de los géneros como $G = \{H, M\}$, que representa a los géneros hombre (H) y Mujer (M). Definamos 18 variables binarias x que tomen valores en $\{0, 1\}$, 9 para los hombres (indizadas por la letra H) y 9 para las mujeres (indizadas por la letra M) e indizadas en una segunda ocasión por el estado al

[1] Se usa "max" para denotar una maximización y "min" para una minimización

que representan. O sea, nuestras variables indicadoras G_{89} serán con $82 \leq y \leq 92$. Si se tiene que $G_{89} = 0$, entonces se interpreta como que el candidato de género 8 del estado 9 no fue elegido como candidato a la gubernatura de ese estado, y si $G_{89} = 1$, el candidato de género 8 sí fue elegido como candidato del partido para la gubernatura del estado 9. Con ello, un conjunto de restricciones necesarias para plantear un problema de optimización será

$$\tilde{O}_{82} G_{89} = 1 - \quad \text{para todo } 92 \bullet$$

Estas nueve restricciones (una para cada entidad) nos indican que en cada estado tiene que elegirse únicamente a un candidato a la gubernatura. De acuerdo a las reglas de paridad de género indicadas por el INE, debe cumplirse que

$$\tilde{O}_{92} G_{9} \quad 5 \quad \text{y} \quad \tilde{O}_{82} \tilde{O}_{92} G_{89} = 9$$

o sea, que el número de mujeres elegidas sea mayor o igual a 5 y que el total de candidatos elegidos sea igual a 9.

Como dato adicional, consideremos los mayores porcentajes de votos obtenidos por un hombre y una mujer en cada uno de los estados, reportados por el propio partido político, y que nosotros mostramos en la Tabla 1.

Por ejemplo, la Tabla 1 indica que, en el estado de Jalisco, la mujer con mayor porcentaje de la votación obtuvo el 18.60 % de los votos mientras que en ese estado el hombre con el mayor porcentaje de votos obtuvo el 23.90 % de los votos. Omitimos los nombres de los candidatos, así como su posición dentro de la votación general considerando todos los candidatos.

El siguiente paso es plantear la función objetivo que se va a optimizar. Nuestra primera propuesta consiste en maximizar la suma de los porcentajes para todos los estados. La intuición detrás de esta propuesta es que la elección de los candidatos sea tal que la mayor proporción de los votantes esté satisfecha con los candidatos elegidos, dado que se elige como representante al candidato por quien votó. Por ejemplo, si se eligiesen a las mujeres de todos los estados como candidatas, la suma de los porcentajes obtenidos será de 155.10, lo que indica que de elegir puras mujeres, solamente el 17.23 % de los votantes se sentirían representados. Pero únicamente en Yucatán se eligiera a un hombre y en todos los demás estados se eligieran mujeres, la suma de los porcentajes obtenidos será ahora de 186.30, indicando una representación del 20.70 %. Esta última asignación tiene a una mayor proporción de gente que en realidad se siente representada por el resultado que a la anterior donde se elegían a puras mujeres. Claramente ninguna de las dos asignaciones cumple con los requerimientos de paridad de género requeridos por el INE, por lo que ninguna de las dos podría ser la óptima bajo esta función objetivo.

Con ello, el primer problema de programación lineal propuesto será

$$\max \quad 8 - 9G_{.} \quad , \quad 40 - 1G_{.} \quad , \quad 16G_{+} \quad , \quad 15 - 2G_{+} \quad , \quad 14 - 5G_{)} \quad ,$$
$$, \quad 51 - 2G_{)} \quad , \quad 11 - 9G_{\%} \quad , \quad 22 - 3G_{\%} \quad , \quad 20 - 4G_{''} \quad ,$$
$$, \quad 18 - 7G_{''} \quad , \quad 18 - 6G_{'} \quad , \quad 23 - 9G_{'} \quad , \quad 23 - 4G_{'} \quad ,$$
$$, \quad 29 - 4G_{'} \quad , \quad 14 - 7G_{'} \quad , \quad 16 - 2G_{'} \quad , \quad 26 - 7G_{'} \quad , \quad 40 - 5G_{'}$$

sujeto a

$$\tilde{O}_{82} G_{89} = 1 - 8 \quad 92 \quad -$$
$$\tilde{O}_{92} G_{9} \quad 5 -$$
$$\tilde{O}_{82} \tilde{O}_{92} G_{89} = 9$$
$$G_{89} 2 \; f \; 0 - 1 \; g \; 8 \; 2 \; (- 8 \quad 92 \bullet$$

Usamos el software QM for Windows para resolver el problema de programación lineal anterior. El resultado obtenido fue el siguiente:

$$G_{.} = G_{+} = G_{)} = G_{\%} = G_{''} = G_{'} = G_{'} = G_{'} = G_{'} = 1$$

$$G_{.} = G_{+} = G_{)} = G_{\%} = G_{''} = G_{'} = G_{'} = G_{'} = G_{'} = 0$$

con valor de la función objetivo igual a 247.2. Con ello, podemos concluir que elegir como candidatas mujeres en los estados de Veracruz, Morelos, Jalisco, Guanajuato, Chiapas y candidatos hombres en Yucatán, Tabasco, Puebla y Ciudad de México es la asignación donde el mayor porcentaje de votantes se encuentra representado, con un total de 27.46 %.

Nótense dos cosas: la asignación que maximiza el porcentaje de los votantes que se encuentran representados consiste en elegir, para cada estado, quien tiene el porcentaje mayor. Esto da un porcentaje máximo de 28.89 % pero no se cumple el criterio de paridad de género requerido por el INE, dado que sólo se elegirían mujeres como candidatas en los estados de Veracruz y Morelos. La otra observación va de la mano con lo que ocurri

en realidad: el partido MORENA eligió candidatas mujeres en los estados de Veracruz, Morelos, Jalisco, Guanajuato y Ciudad de México y candidatos hombre en Yucatán, Tabasco, Puebla y Chiapas. La asignación hecha por el partido obtiene un porcentaje de votantes representado de 26.10 %, el cual es menor bajo nuestra propuesta. La única diferencia entre lo obtenido en nuestra maximización respecto a lo elegido por el partido fueron las elecciones contrarias en Chiapas y Ciudad de México.

Un escenario manejado en varios medios de información impresos y electrónicos era que la designación de la candidata mujer en la Ciudad de México fue más una decisión salomónica que racional. Nuestro segundo caso fue resolver el problema de maximización anterior pero añadiendo la restricción $G_{.} = 1$. Esto es, en nuestro segundo modelo se maximiza nuevamente la proporción de los votantes representados en la asignación sujeto al criterio de paridad de género y que en la Ciudad de México debe elegirse a una mujer. Resolviendo nuevamente el problema asociado, el resultado fue:

$$G_{.} = G_{+} = G_{)} = G_{\%} = G_{"} = G_{} = G_{} = G_{} = G_{} = 1$$

$$G_{.} = G_{+} = G_{)} = G_{\%} = G_{"} = G_{} = G_{} = G_{} = G_{} = 0$$

con un valor máximo de 238.7, lo que representa que bajo esta asignación el 26.52 % de los votantes se encuentran representados por los candidatos elegidos. La observación de este caso es que si se quiere forzar la inclusión de una candidata mujer en la Ciudad de México y además se quiere maximizar el porcentaje de votantes que se encuentra representado por los candidatos elegidos, la asignación debió ser de candidatos hombres en Yucatán, Tabasco, Puebla y Jalisco y candidatas mujeres en Veracruz, Morelos, Guanajuato, Chiapas y Ciudad de México, escenario también diferente a lo que realmente ocurrió.

Nuestro segundo experimento consistió en considerar ya no la proporción de los votantes sino el total de los votantes. Esto porque no es lo mismo el 50 % de un conjunto de 2 votantes que el 20 % de un conjunto de 100 votantes: aunque el porcentaje en el primer conjunto es mayor, el número de votantes es mayor en el segundo caso. Para ello, tomamos en cuenta el número de votantes afiliados a los tres partidos políticos que conforman la coalición, MORENA, PT[2] y VERDE[3] reportados de acuerdo al INE, que se encuentran resumidos en la Tabla 2.

Tomamos los tres partidos políticos MORENA, PT y Partido Verde porque en las elecciones a gobernador el candidato elegido será el candidato de los tres partidos, dado que participan en las elecciones como una coalición.

Convirtiendo la Tabla 2, que se encuentra en porcentajes, a total de personas representadas considerando como total en cada estado al total del padrón de la coalición en cada estado, tenemos como resultado la Tabla 3.

Entonces, nuestro segundo problema de programación lineal es

$$\begin{aligned} \max \; & 3811247G_{.} + 1717202G_{.} + 3803136G_{+} + \\ & 3612979G_{+} + 972123G_{)} + 3432601G_{)} + \\ & 1751549G_{\%} + 3282314G_{\%} + 1294522G_{"} + \\ & 1186645G_{"} + 1749888G_{} + 224851G_{} + \\ & 2528299G_{} + 3176581G_{} + 3366064G_{} + \\ & 3709540G_{} + 977078G_{} + 1482085G_{} \end{aligned}$$

sujeto a

$$\begin{aligned} & \sum_{i=1}^{32} G_{i} = 1 \\ & \sum_{i=1}^{32} G_{i} \geq 5 \\ & \sum_{i=1}^{32} G_{i} = 9 \\ & G_{i} \in \{0,1\} \quad i = 1,2,\ldots,32 \end{aligned}$$

Nuevamente, resolvimos el problema anterior con PuLP en Python for Windows. Los resultados que obtuvimos fueron:

$$G_{.} = G_{+} = G_{)} = G_{\%} = G_{"} = G_{} = G_{} = G_{} = G_{} = 1$$

$$G_{.} = G_{+} = G_{)} = G_{\%} = G_{"} = G_{} = G_{} = G_{} = G_{} = 0$$

con un valor de votantes representados de 35994830. Nótese que la asignación que cumple con el criterio de paridad de género y que además maximiza el número de votantes representados por el resultado es la misma que si se maximizase el porcentaje de votantes representados, la cual, como en el caso anterior, difiere de la asignación que en realidad se decidió.

Como siguiente paso resolvimos el modelo anterior pero también considerando la postura de cierto sector de la prensa que indicaba la designación sin cuestionar de una candidata mujer para la Ciudad de México. Esto es, añadimos nuevamente la

[2] Partido del Trabajo

[3] Partido Verde Ecologista de México

restricción $G_{..} = 1$ al problema anterior. Los resultados que obtuvimos fueron los siguientes:

$$G_{..} = G_{+} = G_{)} = G_{\%} = G_{"} = G_{} = G_{} = G_{} = G_{} = 1$$

$$G_{..} = G_{+} = G_{)} = G_{\%} = G_{"} = G_{} = G_{} = G_{} = G_{} = 0$$

con un valor de la función objetivo de 314434.40. De nuevo, la asignación coincide con la que se encontró cuando se maximizó el porcentaje de votantes representados por la asignación, teniendo nuevamente que la asignación encontrada di ere de lo que ocurrió.

4 Conclusiones

Sin pretender hacer juicios de valor y mucho menos debatir cuál es el mejor candidato o candidata por entidad, este análisis cuantitativo muestra una pérdida de cierta optimalidad al elegir en la Ciudad de México a Clara Marina Brugada Molina como candidata. No obstante que una vez que los procesos electorales se dieron y la coalición o cial (que en lo sucesivo llamaremos coalición) obtuvo triunfos holgados a nivel presidencial, a nivel federal en las cámaras legislativas, en las entidades (incluyendo a la Ciudad de México) y entendiendo la búsqueda de optimalidad como el proceso de encontrar la mejor opción que maximice el rendimiento y no conformarnos con tener una "buena solución" la estrategia de asignación de candidatos por la coalición, exhibe –según nuestro modelo– una pérdida de optimalidad, misma que sugiere que con la asignación que proponemos, se hubieran ganado más representantes populares y en particular en la ciudad de México, donde la coalición ganó 11 de las 16 alcaldías; también se podría decir que se hubieran ganado más alcaldías dado que los registros históricos de votación exhiben que el mexicano no vota diferenciado, es decir, son pocos los mexicanos que votan por un partido para la presidencia y por lo opuesto por las cámaras y municipios (alcaldías). Es decir, algunos votantes que se abstuvieron o votaron por los partidos de oposición, hubieran votado por la coalición del partido o cial por el candidato que sugerimos (Omar Hamid Ga Harfuch) y por su distrito local y alcaldía y por supuesto que en la boleta presidencial. Si nos situamos en los contextos de mayor simple y mayoría cali cada esta decisión que matemáticamente no esóptima, podría ser la diferencia entre obtener o no los objetivos en la cámara de senadores –pues a pesar de todo en la Cámara de diputados se obtuvo la mayoría cali cada: más de dos terceras partes–. Una limitación de nuestro modelo es que no contemplamos el factor "lealtad" a una fuerzapolítica, es decir, que hay ciudadanos que no cambiarían su votación sólo porque no se postuló al que consideraban la primera opción al interior de su partido o fuerza política. De cualquier manera, nuestro análisis reporta que existen mejores asignaciones a la que nalmente decidieron en MORENA y partidos de su coalición.

El hecho que hoy no traiga paraguas y no haya llovido, no signi ca que tomé la decisión óptima.

Acerca de los autores

Julio César Macías Ponce. Licenciado en Matemáticas Aplicadas por la Universidad Autónoma de Aguascalientes (UAA). Maestro y Doctor en Ciencias con orientación en Matemáticas Aplicadas por el Centro de Investigación en Matemáticas (CIMAT). Es profesor investigador de tiempo completo en el Departamento de Matemáticas y Física en la UAA. Sus áreas de interés son: optimización, modelación matemática, teoría de juegos y economía matemática

William Olvera-Lopez Ingeniero en Sistemas Computacionales por la Universidad Autónoma de Campeche. Maestro y Doctor en Ciencias con orientación en Matemáticas Aplicadas por el Centro de Investigación en Matemáticas (CIMAT). Profesor de Tiempo Completo en la Facultad de Economía de la Universidad Autónoma de San Luis Potosí. Sus intereses son la teoría de juegos, investigación de operaciones, probabilidad y estadística y optimización.

Bibliografía

Dantzig, G. B. (1951). Maximization of a linear function of variables subject to linear inequalities. *Activity analysis of production and allocation*, 13, 339–347.

de Gobernación, S. (2002). Diario oficial de la federación: 24 de junio de 2002. Retrieved from https://www.diputados.gob.mx/LeyesBiblio/abro/cofipe/COFIPE_ref10_24jun02.pdf (Consultado el 10 de Enero, 2024)

de Gobernación, S. (2018). Este 3 de julio se cumplen 63 años de la primera vez que la mujer en méxico ejerció su derecho a votar. Retrieved from https://www.gob.mx/segob/articulos/este-3-de-julio-se-cumplen-63-anos-de-la-primera-vez-que-la-mujer-en-mexico-ejercio-su-derecho-a-votar?idiom=es (Consultado el 16 de enero, 2024)

de Gobernación, S. (2023). Diario oficial de la federación: 22 de noviembre de 2023. Retrieved from https://www.dof.gob.mx/nota_detalle.php?codigo=5709199&fecha=22/11/2023#gsc.tab=0 (Consultado el 20 de Diciembre, 2023)

de las Mujeres, I. N. (2018). Las indispensables. recuperando la memoria. Author.

de las Mujeres, I. N. (2024). Aniversario del sufragio femenino en méxico. 17 de octubre. Retrieved from http://cedoc.inmujeres.gob.mx/documentos_download/100698.pdf (Consultado el 04 de Enero, 2024)

de Yucatán, D. (2022). Rosa torres hizo historia elegida a un cargo público. Retrieved from https://www.yucatan.com.mx/imagen/2022/03/09/rosa-torres-hizo-historia-elegida-un-cargo-publico-304811.html (Consultado el 02 de Enero, 2024)

Electoral, I. N. (2023). Determina ine forma en que implementará acciones afirmativas en el registro de candidaturas del pef 2023-2024. Retrieved from https://centralelectoral.ine.mx/2023/11/25/determina-ine-forma-en-que-implementara-acciones-afirmativas-en-el-registro-de-candidaturas-del-pef-2023-2024/ (Consultado el 08 de Enero, 2024)

Electoral, I. N. (2024). Paridad en las candidaturas. Retrieved from https://igualdad.ine.mx/paridad/%20paridad-en-las-candidaturas/ (Consultado el 08 de Enero, 2024)

Hillier, G. J., Frederick S y Lieberman. (2015). Introduction to operations research. McGraw-Hill.

Muñoz, N., A. y Jiménez. (2023). Inicia morena proceso de resultados de encuestas rumbo a gubernaturas. Retrieved from https://www.jornada.com.mx/noticia/2023/11/10/politica/inicia-morena-proceso-de-resultados-de-encuestas-rumbo-a-gubernaturas-5366 (Consultado el 11 de Enero, 2024)

Navarrete, S. (2023). México sumará 10 gobernadoras con el triunfo de del nájmez. Retrieved from https://politica.expansion.mx/mexico/2023/06/09/cuantas-gobernadoras-hay-en-mexico (Consultado el 07 de Enero, 2024)

Noticias, O. (2023). ¿quiénes fueron las primeras mujeres sufragadas en méxico? Retrieved from https://oncenoticias.digital/nacional/quienes-fueron-las-primeras-mujeres-sufragadas-en-mexico/260075/ (Consultado el 04 de Enero, 2024)

Olvera, M. (2023). Las primeras mujeres en la política mexicana. Retrieved from https://zetatijuana.com/2023/03/las-primeras-mujeres-en-la-politica-mexicana/ (Consultado el 03 de Enero, 2024)

Redacción, E. F. (2023). Así fue el maratón de más de 16 horas de morena para elegir a sus candidatos para 2024. Retrieved from https://www.elfinanciero.com.mx/estados/2023/11/10/resultados-encuesta-de-morena-candidatos-cdmx-estado-por-estado-en-vivo/ (Consultado el 09 de Enero, 2024)

Reynoso, D. y D'angelo, N. (2006). Las leyes de cuota y su impacto en la elección de mujeres en méxico. Política y gobierno, 13(2), 279–313.

Vol. 40 Núm. 3
Estadística Oficial

El sistema integrado de metadatos en Eustat

M. Mas Moreno y M. Ayestarán Arregi

Área de metodología, calidad, innovación e I+D. Eustat. m-mas@eustat.eus

Área de metodología, calidad, innovación e I+D. Eustat. m-ayestaranarregi@eustat.eus

Citar como:
Mas Moreno, M., Ayestarán Arregi, M. (2024). El sistema integrado de metadatos en Eustat. Boletín de Estadística e Investigación Operativa. 40(3), 28-34.

Abstract

En Eustat - Instituto Vasco de Estadística, las operaciones estadísticas se documentan mediante ”proyectos técnicos”que actúan como manuales de procedimientos a nivel interno y como herramientas de transparencia para los usuarios de las estadísticas. En los últimos años, estos proyectos técnicos han constituido la base de un sistema de metadatos integrado que incluye definiciones, clasificaciones estándares y fichas metodológicas, conectando aplicaciones internas con el sistema de difusión. Además, se utilizan estándares europeos para la documentación y el intercambio de metadatos. Este sistema continúa desarrollándose, con módulos implementados y otros en fase de implementación o planificación futura.

Palabras Clave: documentación, metadatos, proceso estadístico.

1 Introducción

Durante años, las operaciones estadísticas que elabora el Instituto Vasco de Estadística (en adelante, Eustat), se han documentado mediante los denominados “Proyectos técnicos”. Estos proyectos son auténticos manuales de procedimientos que recopilan las tareas y métodos usados en la elaboración de las estadísticas y sirven, a su vez, como herramienta de transparencia para los usuarios de la estadística oficial. En los últimos años, los proyectos técnicos se han convertido en la base de un sistema integrado de metadatos que incluye diferentes elementos, tales como conceptos y definiciones, clasificaciones y fichas metodológicas. Este sistema (ver Figura 1) conecta diversos subsistemas y aplicaciones internas de Eustat con el entorno de difusión y el sitio web del Instituto. Además, en su implementación se han utilizado estándares europeos de metadatos tales como el Euro SDMX Metadata Structure (ESMS) para la elaboración y difusión de las fichas metodológicas de las operaciones estadísticas, y el formato SDMX de intercambio de datos para la codificación de las clasificaciones.

Figura 1. Esquema del Sistema Integrado de Metadatos de Eustat.

En este artículo describiremos con detalle los diferentes subsistemas y elementos de este sistema integrado de metadatos, así como las fases llevadas a cabo para su implementación. Algunos de los módulos están ya desarrollados y en funcionamiento, como es el caso del módulo de clasificaciones y el de edición y elaboración de proyectos técnicos y fichas metodológicas. El módulo de definiciones está todavía en fase de desarrollo, así como las relaciones de todos los anteriores con el sistema central de metadatos y su conexión con la web de Eustat.

2 Los proyectos técnicos: el origen de los metadatos en Eustat

La documentación y descripción de las tareas y procesos llevados a cabo en las operaciones estadísticas ha sido una de las prioridades y principios de calidad de Eustat. La transmisión interna del conocimiento y la transparencia de cara al usuario final son dos de los objetivos que se pretenden con la elaboración, actualización y mantenimiento de estos proyectos técnicos, además de dar cumplimiento a la normativa que establece su aprobación y gestión por parte de Eustat.

Figura 2. Tabla de contenidos del proyecto técnico.

El artículo 29.a de la Ley 4/1986, de 23 de abril, de Estadística de la Comunidad Autónoma de Euskadi señala como competencia del Instituto Vasco de Estadística la aprobación del proyecto técnico de cada operación estadística. Por su parte, el apartado g) del propio artículo dispone que el Instituto aprobará los cuestionarios integrantes de la Organización Estadística y mantendrá con carácter público, un Registro Central de estos. El Decreto 82/2005 de 12 de abril regula el procedimiento para la aprobación de los proyectos técnicos de las operaciones estadísticas y establecen los contenidos del proyecto que como mínimo deben incluir estos apartados:

a) Definición, objetivos y difusión de la operación.
b) Diseño metodológico y, en su caso, de los cuestionarios.
c) Procedimientos de ejecución.
d) Funciones y responsabilidades.

e) Control de calidad.

f) Informatización.

Tomando como punto de partida la anterior estructura, el contenido de los proyectos técnicos ha ido variando hasta llegar a la actual tabla de contenidos que es la que se ha implementado en el sistema central de metadatos de las operaciones estadísticas (véase Figura 2).

Este sistema contiene los proyectos técnicos, cuestionarios y fichas metodológicas de las operaciones estadísticas y constituye el núcleo central del sistema integrado de metadatos de Eustat.

Figura 3. Listado de proyectos técnicos en el sistema de metadatos de Eustat.

Mediante una aplicación `ad-hoc', el personal estadístico puede acceder a los proyectos técnicos de las operaciones de las que es responsable (véase Figura 3) y, anualmente, editarlos y actualizarlos con los cambios y modificaciones que considere oportunos (véase Figura 4). Si la operación no ha sufrido cambios se puede dar vigencia al proyecto técnico.

Figura 4. Pantalla de edición del proyecto técnico en el sistema de metadatos de Eustat.

Tal y como hemos visto en la Figura 2 el proyecto técnico debe contener, entre otros elementos de los metadatos, las definiciones, conceptos, variables y clasificaciones utilizadas por la operación estadística. Actualmente, estos elementos se definen directamente en el proyecto técnico.

No obstante, la idea del sistema integrado de metadatos contempla la definición y gestión de estos elementos de forma estandarizada e independiente en diferentes subsistemas. Estos subsistemas o módulos estarán conectados con el sistema central de metadatos de forma que desde el proyecto técnico se podrán consultar e insertar elementos de metadatos previamente definidos y que podrán ser reutilizados por una o varias operaciones estadísticas (ver Figura 5).

Figura 5. Relaciones con los subsistemas de clasificaciones y definiciones.

3 Las fichas metodológicas y el ESMS

Otro de los elementos principales del sistema central de metadatos de Eustat, ya implementado en la producción, son las fichas metodológicas. Estas fichas se han definido siguiendo el estándar europeo ESMS que contempla estos 19 apartados:

1. Contacto
2. Actualización de metadatos
3. Presentación estadística
4. Unidad de medida
5. Periodo de referencia
6. Mandato institucional
7. Confidencialidad

8. Política de difusión
9. Frecuencia dela difusión
10. Accesibilidad y claridad
11. Gestión de calidad
12. Relevancia
13. Acuracidad y abilidad
14. Oportunidad y puntualidad
15. Coherencia y compatibilidad
16. Costes y carga
17. Revisión de datos
18. Tratamiento estadístico
19. Observaciones

Además, en la elaboración de estas chas el sistema de metadatos utiliza información presente en otras aplicaciones y subsistemas de Eustat. Concretamente:

- Aplicación del Plan Vasco de Estadística (PVE): Se extrae toda la información o cial sobre la operación estadística publicada en los Planes y Programas de Estadística de la C.A. de Euskadi, tales como objetivos generales y específicos, tipo de operación estadística, periodicidad de recogida y publicación de datos, principales usuarios e interesados en la operación, etc.
- Aplicación de Indicadores de calidad (Hobetu): Se extraen los últimos valores de los indicadores de calidad publicados de cada operación estadística. Estos valores se registran periódicamente en esta aplicación interna de Eustat que sirve como herramienta de evaluación y control de la calidad.
- Sistema de metadatos: Se cargan contenidos de los proyectos técnicos almacenados en el sistema de metadatos de Eustat para completar, principalmente, los apartados que tienen que ver con la presentación y el tratamiento estadístico.

La utilización de todas estas fuentes permite re ejar en la cha metodológica una información coherente, detallada y actualizada sobre la operación estadística. Una vez elaboradas, estas chas se envían a la zona de intercambio de contenidos para nalmente ser publicadas en la Web de Eustat. Las chas metodológicas se pueden consultar en el apartado de Metodología y calidad de cada operación estadística o en el apartado general de documentación en la página principal de la Web.

4 El subsistema de clasi caciones

El subsistema de clasi caciones es otro de los módulos parcialmente implementado dentro del sistema de metadatos de Eustat. Se desarrolló de forma independiente y paralela al sistema central de metadatos, pero teniendo siempre en cuenta su futura integración con éste. Actualmente, contiene las principales clasi caciones (y agregaciones) utilizadas en las operaciones estadísticas del Instituto (ver Figura 6). Un total de 52 clasi caciones y 17 variables. Entre ellas, clasi caciones de actividad económica, productos, bienes y servicios, educación, códigos territoriales y variables demográ cas (sexo, edad, etc.).

Figura 6. Relación de clasi caciones y variables en el módulo de clasi caciones.

El subsistema contempla y almacena la relación de las clasi caciones con las diferentes operaciones estadísticas que la utilizan. Además, aparte de la codi cación con la que se representa cada clasi cación, tiene en consideración el estándar europeo de intercambio de metadatos SDMX. Su enlace con el sistema de difusión de Eustat está completamente desarrollado y ya desde mediados de 2024 las clasi caciones que aparecen en la Web del Eustat se difunden directamente y de forma centralizada desde este módulo de clasi caciones (anteriormente, cada responsable se encargaba de difundir las clasi caciones utilizadas por su operación estadística). Están por desarrollar los enlaces con el módulo de de niciones y con el sistema central de metadatos. Respecto al primero está previsto que las descripciones de cada clasi cación y categoría se almacenen en el módulo de de niciones y se publiquen desde allí en la Web del Instituto. Además, las clasi caciones disponibles se presentarán en el módulo de metadatos para completar los correspondientes apartados de clasi caciones del proyecto técnico y la cha metodológica. Ver Figura 5.

5 El subsistema de definiciones

El módulo de definiciones se encuentra en su primera fase de implementación dentro del sistema integrado de metadatos. Esta fase contempla una carga inicial de definiciones y un proceso de depuración y normalización de las mismas (ver Figura 7).

Figura 7. Consulta de definiciones en el módulo de definiciones.

Respecto a la carga inicial, ésta se ha realizado a partir de las definiciones publicadas en la web de Eustat. El motivo para la realización de esta carga desde la web y no desde otros soportes, como p.ej. los proyectos técnicos, se ha basado en los siguientes factores:

- La existencia de una estructura de base de datos con identificadores únicos para cada definición.
- Existencia de las relaciones:
  - definición $ operación(es) estadística(s)
  - definición $ tema(s) de la web
- Actualidad y vigencia de la definición

Una vez realizada esta primera carga se realizará un proceso de depuración y normalización que pasará por la identificación y eliminación de duplicados y el establecimiento de definiciones estándar cuando existan variantes de un mismo concepto (ver Figura 8).

Figura 8. Utilidades de normalización. Definiciones duplicadas.

Además, se podrá gestionar el histórico de definiciones, indicando una fecha de fin de vigencia, cuando se trate de definiciones que ya no se usan, bien porque el concepto ha quedado obsoleto, o bien porque ha variado notablemente la definición del mismo (ver Figura 9).

Figura 9. Mantenimiento de definiciones. Fecha de vigencia.

Otra de las utilidades fundamentales de este módulo de definiciones es la gestión de propuestas de nuevas definiciones. Estas propuestas se realizarán desde el sistema de metadatos durante la fase de revisión de los proyectos técnicos. Si el responsable de una operación estadística considera la necesidad de definir un nuevo concepto o variable para su operación estadística podrá realizar una propuesta de definición en el momento de la actualización del proyecto técnico. Esta propuesta de definición se mostrará en el módulo de definiciones para su revisión por parte del Área de metodología, calidad, innovación e I+D del Instituto. Esta área podrá aceptar o rechazar la nueva definición y asesorar,

en su caso, en la redacción de la misma o en la reutilización de algún concepto ya existente en el subsistema de definiciones.

Figura 10. Relaciones entre subsistemas. Gestor de definiciones.

Finalmente, cabe destacar que uno de los objetivos primordiales en un sistema de metadatos es la difusión de sus elementos, tanto de forma interna como para el usuario final de la estadística. En la fase final de la implementación del módulo de definiciones, está previsto que la difusión de las definiciones en la Web del Eustat se realice directamente desde este módulo. La decisión sobre qué definiciones deben publicarse de cada operación estadística vendrá dada desde el sistema central de metadatos, concretamente desde el proyecto técnico donde el responsable deberá "marcar" qué definiciones de las incluidas en el proyecto técnico deben difundirse en la web del Instituto.

6 Futuro del sistema integrado de metadatos: hacia la definición de la ontología de Eustat

En un entorno cambiante, como en el que se encuentra actualmente la estadística oficial, cada vez adquiere más relevancia una correcta representación de los procesos, de sus componentes y de las relaciones entre ellos, lo que se denomina comúnmente como "ontología". Estas representaciones pretenden establecer la trazabilidad de los procesos y dar un lenguaje común que pueda ser interpretado por otros sistemas y usuarios. Eustat persigue la definición de esta ontología y desarrolla el sistema de metadatos como parte clave y punto de partida de la misma. Para ello, durante los próximos meses, está previsto completar las relaciones entre los diferentes subsistemas y módulos del sistema integrado de metadatos. Esto es, terminar el desarrollo del módulo de definiciones y clasificaciones con el enlace de éstos con el sistema central de metadatos. Esto permitirá gestionar todos los metadatos desde el módulo central (proyecto técnico y ficha metodológica), realizar la difusión de todos los elementos del sistema y facilitar la identificación de los elementos y procesos a definir en la ontología. Todo ello alineado con los estándares internacionales como el SIMS (Single Integrated Metadata System) que promueve Eurostat, y del que forma parte la estructura de la ficha metodológica (ESMS), y otros como el GSBPM (Generic Statistical Business Process Model) y GSIM (Generic Statistical Information Model) de UNECE, que son los referentes en cuanto a calidad y documentación de los procesos estadísticos.

Acerca de las autoras

Marta Mas Moreno es Diplomada en Estadística por la Universidad de Zaragoza y Licenciada en Investigación y Técnicas de Mercado por la Universitat Oberta de Catalunya (UOC). Ha desarrollado su carrera profesional dentro del ámbito de la estadística oficial del País Vasco, realizando diversos proyectos de investigación y desarrollo en el Instituto Vasco de Estadística (Eustat) y trabajando, principalmente, en el ámbito de las estadísticas de educación y empleo. Actualmente es coordinadora en el Área de metodología, calidad, innovación e I+D de Eustat y es responsable, entre otras, de la operación "Estándares del proceso de producción estadística" en la que se enmarca el sistema integrado de metadatos de Eustat.

Marina Ayestarán Arregui es Licenciada en Matemáticas por la Universidad del País Vasco y en Ciencias Económicas por la UNED. Ha desarrollado su actividad profesional en el ámbito de la estadística oficial del País Vasco desde el año 1997. Inicialmente en el Departamento de Salud (1997-2000) y principalmente en el Instituto Vasco de Estadística (Eustat) en el Área de metodología (2000-2009 y desde 2018) y en el Órgano Estadístico del Departamento de Agricultura y Pesca (2010-2018). Desde 2018 es responsable en el Área de metodología, calidad, innovación e I+D de Eustat.

Bibliografía

BOPV. (2005). Decreto 82/2005, de 12 de abril, por el que se regula el procedimiento para la aprobación de proyectos técnicos de operaciones estadísticas, y se crea el registro de proyectos técnicos y cuestionarios de operaciones estadísticas. (En: Boletín Oficial del País Vasco (BOPV) nº 78 de 26 de abril de 2005)

Eurostat. (2023). RECOMENDACIÓN (UE) 2023/397 DE LA COMISIÓN de 17 de febrero de 2023 sobre los metadatos de referencia y los informes de calidad para el Sistema Estadístico Europeo. (En: Diario Oficial de la Unión Europea de 21 de febrero de 2023)

Eustat. (2023). Sistema de metadatos de operaciones estadísticas. manual de usuario [Computer software manual]. (Besaide 3, Versión 1.6)

Eustat. (2024). Mantenimiento de clasificaciones. manual de usuario [Computer software manual]. (Besaide 84, Versión 1.1)

Ley 4/1986, de 23 de abril, de estadística de la Comunidad Autónoma de Euskadi (1986). (En: Boletín Oficial del País Vasco (BOPV) nº 105 de 30 de mayo de 1986)

Riccio, M. y Scanu, M. (2024). Semantic and ontologies of data sets along a data production process. Conferencia europea de calidad en la estadística oficial. Estoril.

UNECE. (2019). On behalf of the international statistical community. Generic Statistical Business Process Model GSBPM. version 5.1.

UNECE. (2024). On behalf of the international statistical community. Generic Statistical Information model (GSIM). Communication paper for a general statistical audience (Version 2.0)

Vol. 40 Núm. 3
Historia

Historia de la Estadística en la Psicología

Cándida Filgueira Arias

Educación. Universidad San Pablo CEU. candi.bib@ceu.es

Citar como:
Filgueira, C. (2024). Historia de la Estadística en la Psicología. Boletín de Estadística e Investigación Operativa. 40(3), 35-47.

Abstract

The inexcusable con uence of the science of Statistics and the science of Psychology is concretized in this study through the delimitation of the intersection detected and established in approaches, materials and data that are compiled in the historical exposition of both areas of knowledge, with the purpose of channeling and interpreting the historical evolution of both disciplines, thus providing a base of resources to be used for further research.

Psychology turns to Statistics, to the statistical treatment of data, when it begins as a modern science with the objective of satisfying the need to gather parameters in some type of research or investigation with the purpose of clarifying the dependent relationship or not of certain interactions between different dimensions of the studied phenomenon. It is necessary to specify the structure of the functional associations established in both sciences in order to know, identify and clearly distinguish if the interactions recorded are established with qualitative variables or if they are repeated measurements on the same individuals, or if there is certain variability between parametric distributions, etc.

Statistical mathematics intervenes to form and express psychological ideas and hypotheses as precisely as possible; to derive from them, with greater deductive power than verbal reasoning allows and with the maximum guarantee of logical coherence, the observable consequences that follow from these hypotheses; and nally, to strictly and objectively compare the hypothetical predictions with the actual experimental or empirical observations.

In short, we are talking about the emergence of statistics in the history of psychology.

Keywords: History of mathematics, history of psychology, psychology, statistical mathematics, statistics

1 Introducción

La cultura estadística de un país guarda relación con su desarrollo y el grado en que su sistema estadístico es necesario para tomar decisiones acertadas de tipo económico, social y político.

Tanto es así que el conocimiento estadístico juega un papel cada vez más relevante en la sociedad actual debido a la diversidad de uso para explicar hechos de la vida cotidiana y el mundo científico. Un análisis del entorno es suficiente para percatarse que en casi todos los medios impresos y electrónicos existe información con datos estadísticos, gráficas y tablas que hacen referencia, por ejemplo, a tendencias en el consumo de dife-

rentes productos, a la distribución de la población y el uso de recursos naturales, entre otros.

Tal y como nos indica Batanero (2004), constituye un objetivo educativo de gran trascendencia la formación adecuada no solo de los técnicos que producen estas estadísticas, sino también de los profesionales y ciudadanos que deben interpretarlas y tomar a su vez decisiones basadas en esta información, sin olvidar a los que colaboran en la obtención de los datos requeridos.

Y es que tener en consideración la funcionalidad de la Estadística, no solo en la vida cotidiana sino también en el desarrollo de los grandes saberes científicos, nos facilita la comprensión de las cosas y por lo tanto de la ciencia y promueve el entendimiento del origen de los problemas que surgen o que nos planteamos como seres humanos conscientes de un mundo intangible.

Este es el caso de la ciencia de la Psicología que estableció los pilares de dicha conexión para la explicación de la conducta humana en unión con un instrumento[1] tan complejo y delicado junto con la necesidad de elaborar una ineludible vinculación entre ambas. Y es que la "convivencia" entre varios conocimientos científicos comulgan en conjunto sobre el desarrollo de los elementos que lo integran y lo fundamentan de manera ordenada operando, transformando y produciendo resultados a través de estructuras especializadas y todos metodológicos que contribuyen al desarrollo y crecimiento de la ciencia en su campo de de nición e intervención. Estamos aludiendo a la inexcusable con uencia de las ciencias Estadística y Psicología (Amon, 1978a). Para ello, debemos delimitar la intersección establecida en planteamientos, materiales, como los datos que se recopilan y que junto con los objetivos que se propongan se encauzarán hacia la solución o interpretación de las cuestiones y retos planteados y que además proporcione una base de recursos de aprovechamiento para ulteriores nes.

En general, podemos decir que la Psicología acude a la Estadística, al tratamiento estadístico de los datos, cuando se inicia como ciencia moderna y en la necesidad de reunirá patrones en algún tipo de búsqueda o investigación con la nalidad de clari car la relación dependiente o no de ciertas interacciones entre distintas dimensiones del fenómeno estudiado. Se trata en de nitiva de precisar la estructura de las asociaciones funcionales entre diferentes variables y es que la ciencia de la Psicología como tal necesita conocer, distinguir, ver con claridad si dichas interacciones se establecen con variables cualitativas, o si se utilizan medidas repetidas sobre los mismos individuos, o si existe una cierta variabilidad entre distribuciones paramétricas, etc., para tener claramente de nido y delimitado el inicio del estudio que va a constituir la lógica general de indagación. Necesita saber qué es y qué no es "estadísticamente signi cativo" y cómo las características de esta vinculación pueden variar ampliando o restringiendo las muestras. Además, y en distintas circunstancias, obtener un modelo que constituya el paradigma ideal adecuando los datos a los esquemas de dicho modelo (análisis de varianza, correlación, regresión, etc.). Hablamos en de nitiva del surgimiento de la Estadística en la historia de la Psicología

2 Los ancestros

2.1 Breve recorrido histórico de la Psicología

Desde el principio de la historia escrita, los seres humanos han re exionado y discutido sobre la naturaleza humana. Son muchas las cuestiones que se planteaban las antiguas civilizaciones al respecto: ¿Cómo se relacionan el cuerpo y la mente? ¿Cuándo nacemos ya venimos con conocimientos innatos o bien se va registrando a través de la experiencia?

Las respuestas a estas cuestiones las podemos encontrar en los escritos antiguos. Así por ejemplo en la India, Buda se preguntaba cómo se combinan (combinatoria) las sensaciones y las percepciones para formar ideas. En China, Confucio destacaba el poder de las ideas y de una mente instruida. En el Israel antiguo, las escrituras hebreas anticiparon la Psicología actual al relacionar la mente y la emoción con el cuerpo. Se les decía a las personas que tenían que pensar con sus corazones y sentir con sus entrañas.

En Grecia, el lósofo-maestro Sócrates (469-399 a. C.), seguido su discípulo Platón, se planteó una idea muy diferente con relación a la mente y el cuerpo. En efecto, a diferencia de los primeros hebreos, para Sócrates y Platón la mente era separable del cuerpo y seguía viva después de la muerte. Pero consideraba que ya nacemos con el conocimiento. Es curioso como un pensador de la enjundia de Sócrates dedicaba la mayor parte de su tiempo a trabajar en el taller como cincelador de piedra y escultor y curiosamente no cobraba nada por sus clases magistrales, además no dejó escrito ningún texto y hablaba con cualquiera que quisiera confrontar sus ideas. Según Helgunt (1993) p. 21 "Usaba una simple única ráda todo el año y andaba descalzo"

La idea de que la muerte libera a la mente de la prisión que es nuestro cuerpo, se la debemos a Sócrates y curiosamente poco después de su muerte, nació un gran intelectual que fue además alumno de Platón, nos referimos a Aristóteles, nacido de un

[1] Consideramos la comparación de la estadística como instrumento estableciendo los precedentes en Aristóteles que re rió a la lógica como órganono instrumento para el pensamiento, para la ciencia

matrimonio socialmente bien posicionado del norte de Grecia (Filostrato, 1982).

La figura de Aristóteles es muy curiosa porque parece ser que era de carácter cálido y muy amable en el trato, pero también muy sincero, contundente en sus aseveraciones y desarrollaba exposiciones muy directas. Como le había negado la presidencia en la Academia de Platón tras la muerte de éste, Aristóteles fundó una nueva escuela que compartía con ella, el Liceo, y curiosamente duplicó su producción académica delegando la investigación a sus alumnos.

Es indudable que el amor de Aristóteles por los datos constituía una diferencia muy significativa en relación con las disertaciones teóricas y racionales de Sócrates y Platón, quienes dedujeran principios mediante la lógica (?) En este sentido debemos considerar a Aristóteles como el antecesor de los científicos modernos, puesto que inició la formulación de los principios extraídos de las observaciones digamos "cuidadosas" (antecedente de la metodología observacional-sistemática). Sus observaciones le llevaron a indicar que "el alma (De anima) no se puede separar del cuerpo, y lo mismo se debe sostener respecto de partes específicas del alma". Además, Aristóteles decía que el conocimiento no es preexistente, sino que por el contrario se construye a partir de las experiencias acumuladas en nuestros recuerdos. Con esto Aristóteles (1967) se anticipaba a los pensadores posteriores que creían que el ambiente (Nurtura) nos modelaba más que la herencia (Natura).

A partir de este auge de la Psicología digamos precientífica en la antiguo Grecia, los 2.000 años siguientes por increíble que parezca, no consiguieron brindar muchas proposiciones nuevas y perdurables en el tiempo en relación con la explicación y entendimiento de la naturaleza humana y desde luego no fue debido a la ausencia de interés por la indagación e investigación sobre el conocimiento de la realidad humana y natural.

Otro gran pensador y religioso, San Agustín (354-430 d.C.) centró el objeto de estudio hacia la fascinación del conocimiento de los seres humanos. Tal y como nos indica Jaeger (1978), podemos decir que San Agustín fue el pionero de los modernos psicólogos de la salud puesto que elaboró toda una teoría descriptiva que explicaba en él cómo el estado del cuerpo condiciona la mente y cómo ésta, a su vez, influye en el cuerpo.

La ciencia moderna siguió avanzando y 1.200 años después de la época de San Agustín, se inició un florecimiento, un resurgimiento de la ciencia emergiendo las nuevas teorías de la conducta humana. Aparece la brillante figura de René Descartes (1595-1650), nacido en Francia y que expuso sus teorías alineadas con las de Sócrates y Platón. En ellas secundaba la idea de la existencia del carácter innato de las ideas y que la mente, que nada tenía que ver con el cuerpo, podía sobrevivir inexcusablemente a la muerte. No olvidemos que Descartes era filósofo, pero también científico, y supo conjugar la experimentación empírica-estadística con sus disertaciones. La construcción del concepto de mente le obligó a establecer conjeturas en relación con la comunicación de la mente inmaterial y el cuerpo físico hasta tal punto que llegó a la conclusión, un tanto descabellada, que el líquido de las cavidades cerebrales albergaba espíritus en sus animales disecados. Sin embargo, y siguiendo a Laín Entralgo (1987), no por ello debemos desestimar las averiguaciones de este gran pensador que, aunque acumuló el 99% de la sabiduría hasta ahora recopilada de nuestra historia humana, contribuyó a establecer relaciones significativas entre el conocimiento del comportamiento humano (Psicología) y la Estadística.

Avanzando en el devenir de los años, la ciencia fue adquiriendo cada vez más sentido práctico centrándose en el experimento como fuente de obtención de datos y, por lo tanto, de sentido común. En Gran Bretaña, Francis Bacon (1561-1626) centraba sus estudios en la mente humana y sus falencias. No cabe duda de que se anticipó sobre el conocimiento del anhelo de nuestra mente de percibir patrones o pautas establecidas y no solo en esto, sino también en los acontecimientos casuales. En este sentido, Bacon (2000), asevera que la naturaleza tan peculiar del entendimiento humano constituye un nivel superior en cuanto a orden e igualdad en las cosas que realmente encuentra.[2]

La influencia que Francis Bacon ejerció en la ciencia moderna continúa a pesar del tiempo. Hay que reconocer que inició un camino hacia un cambio de paradigma muy importante al asumir que la ciencia tiene un sentido eminentemente práctico y que la fuente de conocimiento emana de los experimentos y del sentido común y no tanto de "creencias", aproximándose así a los experimentos de la Psicología y con ello de la Estadística actual.

Tras la muerte de Bacon surge la figura del filósofo y político británico, John Locke (1632-1704), que, tras una discusión entre amigos escribe un ensayo sobre nuestras propias habilidades. Posteriormente completó este ensayo titulándolo entendimiento humano dando a conocer y exponiendo su elocuente argumento según el cual la mente de los niños al nacer es como una pizarra

[2] Novum organum scientiarum (Nuevos instrumentos de la ciencia), así conocido como Novum organum, es la obra principal del filósofo inglés Francis Bacon, publicada en 1620, quien concebía la ciencia como técnica, capaz de dar al ser humano el dominio sobre la naturaleza. Novum organum trata sobre lógica del procedimiento técnico-científico, una lógica contrapuesta a la aristotélica (cuyos tratados de lógica recibieron, precisamente, el nombre de Organon), y que según Bacon resultaba buena solo para la disputa verbal.

en blanco en la que escriben su experiencia de vida. Con ello se abandonaba la idea original de Platón y Decartes sobre el conocimiento innato puesto que la teoría de John Locke mostraba que la mente actúa a través de lo que percibe, es decir, a través de los sentidos. Después, según las teorías de Bacon y se reforzó el movimiento del empirismo moderno, es decir, el edificio del conocimiento se construye sobre la base de la experiencia y que la ciencia debe establecer sus pilares en la observación y la experimentación.

2.2 Breve recorrido histórico de la Estadística

2.2.1 La Prehistoria

El origen de la Estadística Descriptiva data su origen hace aproximadamente mil o dos mil años antes de Cristo. Fue en Egipto, China y Mesopotamia donde se realizaron censos para la administración de dichos imperios. En el caso de los egipcios se cree un instrumento denominado Nilometro que con el tiempo ha constituida el barómetro económico más antiguo y que medía el caudal del Nilo al mismo tiempo que servía para establecer un índice de fertilidad según el cual se pautaba el monto de los impuestos.

Cuando nos remontamos hacia la "Prehistoria" de la Estadística no debemos olvidar el concepto de "azar" también muy antiguo, tanto como los juegos con la misma denominación y fue óbice y motivo para que los filósofos de entonces le dedicaran sus reflexiones. Así pues, encontramos en las ideas de Aristóteles (384-322) tres tipos de nociones de probabilidad que a siguen vigentes y que de nenás bien actitudes frente al azar y la fortuna, a saber:

- El azar no existe y refleja nuestra ignorancia;

La razón es muy sencilla: apelamos "el concepto de azar" cuando sabemos que el azar carece de concepto. El azar es la instancia paralógica por excelencia, tal y como nos indica Boeri (1993), si escuchamos a Aristóteles en Física II, 5. El azar carece de concepto y el concepto, por su parte, no contiene en sí mismo nada que pueda someter, dominar e identificar eso que, sea lo que sea, aglutinamos bajo el término español de origen árabe azar. El azar no constituye en absoluto un caso de concepto.

- El azar proviene de causas múltiples;

El ámbito de problemas derivados de esta afirmación se vincula, según Brague (1988), a la reivindicación de un espacio reflexivo para el horizonte de lo fortuito, lo casual y lo accidental. En este sentido el esfuerzo conceptual de habilitar filosóficamente los ámbitos coyunturales de la realidad encuentra en Aristóteles uno de sus más firmes representantes .

- El azar es divino y sobrenatural;

Bradie y Miller (1984), afirman que se trata de un factor característico de la condición humana en un nivel que trasciende lo elemental hacia lo divino y sobrenatural. En definitiva, el azar es un índice de la condición humana de carácter indiscutiblemente incierto e inherente a la existencia de esta naturaleza.

Sin embargo, a lo largo de la historia de la humanidad, no pasó mucho tiempo antes de que alguien intentara cuantificar el azar y sus efectos. Durante la Edad Media, en concreto en Oriente, se registró una gran actividad científica y artística y con ello el nombre de azar parece haber venido desde Siria a Europa. No olvidemos que la "flor de azahar" aparece en los dados de la época y podría ser el origen de la palabra.

Fue en el siglo XVII donde surgió la ciencia moderna que, sin lugar a duda, sentó las bases para la futura ciencia en general y, como no, de la ciencia de la Psicología y de la Estadística en particular. Por su parte, hay que resaltar la contribución de las compañías aseguradoras que iniciaron investigaciones matemáticas desde tiempos muy antiguos, y en el siglo XVII aparecieron los primeros famosos problemas de juegos de azar.

En la sociedad francesa, el juego era uno de los entretenimientos más frecuentes. Los juegos se convirtieron en actividades cada vez más complicadas y las apuestas muy elevadas revertieron en la necesidad de calcular las probabilidades de los juegos de manera racional. El caballero de Méré, un jugador apasionado y amante de los juegos, desarrolló una correspondencia epistolar muy activa con Blas Pascal (1623-1662) cuyo tema se centraba, básicamente sobre ciertos juegos de azar. Esta actividad dio origen a su vez a una correspondencia entre algunos matemáticos de la época. Las cuestiones que planteaba De Méré facilitaron, en particular, iniciar una discusión entre Pascal y Pierre Fermat (1601-1665) y así se originó la teoría de las Probabilidades. En el siglo anterior, los italianos Tartaglia (1499-1557), Cardano (1501-1576), e incluso el gran Galileo (1564-1642) abordaron algunos problemas numéricos de combinaciones de dados.

2.2.2 La demografía

Las reglas de cálculo desarrolladas hasta entonces para los juegos de azar vieron sus aplicaciones en otras disciplinas. Los censos demográficos, que se hacían desde la antigüedad, requieren recolectar muchos datos. La demografía y los seguros de vida aprovecharon del desarrollo de la teoría de las probabilidades. Consideramos, por ejemplo, el sexo de una sucesi

de niños recién nacidos. Se puede ver como una repetición de lanzamientos de una moneda, con niño y niña en vez de cara y sello. De la misma manera, podemos considerar un conjunto de hombres mayores de 50 años. Al final del año, una cierta proporción sigue viva.

Durante el siglo XVIII con Pierre Simon, Marqués de Laplace (1749-1827), estos problemas fueron reconocidos como similares a los de un juego, y se encontraron las correspondientes frecuencias relativas, lo que permitió determinar la probabilidad que nazca una niña, o que un hombre mayor que 50 años muera en el año.

Si bien la extensión de los juegos de azar a la demografía y a la matemática actuarial fue extremadamente importante, su planteamiento tiene grandes limitaciones debido a que considera todos los resultados posibles equiprobables. ¿Cuándo una situación real puede expresarse como un juego de azar? Por ejemplo, Daniel Bernoulli, careciendo de datos sobre la mortalidad producida por la viruela a distintas edades, supuso que el riesgo de morir de la enfermedad era el mismo a toda edad. Lo que evidentemente es muy discutible.

2.2.3 La teoría de los errores

Durante los siglos XVIII y XIX la Estadística se expandió sin interrupción mientras la teoría de las probabilidades no mostró progreso. Una de las aplicaciones importante fue desarrollada al mismo tiempo por Gauss (1777-1855), Legendre (1752-1833) y Laplace: el análisis numérico de los errores de mediciones en física y astronomía. ¿Cómo determinar el mejor valor dado por un instrumento que entrega diferentes mediciones del mismo fenómeno?

Esta función cuadrática encuentra su justificación en la distribución normal con Gauss y Laplace, aunque la distribución de los errores fue estudiada mucho antes por Thomas Simpson (1710-1761), que hizo los supuestos que esta distribución debía tener que ser simétrica y que la probabilidad de errores pequeños debía ser más grande que la de los errores grandes. En 1840, Sir Francis Galton (1822-1911) partió de una distribución discreta y la fue refinando hasta llegar en 1857 a una distribución continua muy parecida a la distribución normal. Galton inventó incluso una maquina llamada quincunx, que permite ilustrar la distribución normal. Galton trabajo en meteorología y en herencia. Era el primo de Charles Darwin.

2.2.4 Nacimiento de la Estadística Moderna

Es con la introducción de nuevas aplicaciones que la teoría de las probabilidades del siglo XVIII funda la Estadística Matemática. El término de Estadística se debe posiblemente a G. Achenwall (1719-1772), profesor de la Universidad de Göttingen, tomando del latín la palabra *status*.

Aparte de la demografía y la matemática actuarial, otras disciplinas introdujeron la teoría de las probabilidades. Fue el inicio de la Mecánica Estadística, debido a Maxwell (1831-1879) y Bolzmann, quienes dieron también una justificación de la distribución normal en la teoría cinética de los gases.

La Estadística se empieza a usar de una manera u otra en todas las disciplinas, a pesar de un estancamiento de la teoría de las probabilidades. En particular, muchos vieron la dificultad de aplicar el concepto de simétrico, o de casos igualmente posibles, en todas las aplicaciones. Hubo que esperar a que Andrey Nickolaevich Kolmogorov (1903-1987) separara la determinación de los valores de las probabilidades de sus reglas de cálculo.

Los primeros resultados importantes de la Estadística Matemática se deben al inglés Karl Pearson (1857-1936) y a otros investigadores de la escuela biométrica inglesa.

2.2.5 La segunda mitad del siglo XX: la revolución computacional

Los científicos, especialmente los ingleses, desarrollaron métodos matemáticos para la Estadística, pero en la práctica manipularon cifras durante medio siglo sin disponer de verdaderas herramientas de cálculo. La llegada de los computadores revolucionó el desarrollo de la Estadística. En Francia (Benzécri) y en los Estados unidos (Tuckey) fueron los pioneros en repensar la Estadística en función de los computadores. Mejoraron, adaptaron y crearon nuevos instrumentos para estudiar grandes volúmenes de datos: nuevas técnicas y herramientas gráficas.

De hecho, se pueden encontrar obras como la de J. T. Townsend y F. G. Ashby (1983) que lleva por título Stochastic Modeling of Elementary Psychological Processes. Esta obra, en la misma línea, pero más completa, es la de R. D. Luce sobre tiempos de reacción, publicada en 1984. La obra de Luce, Response Times: Their Role in Inferring Elementary Mental Organization, dedica un primer capítulo al estudio de estas cuestiones: variables aleatorias, funciones de distribución y densidad, independencia y probabilidad condicionada, muestras aleatorias, teoría muestrales, funciones generatrices (convolución, momentos, cumulantes), y un apartado para introducir los conceptos más elementales de los procesos estocásticos.

Con todo, hay que reconocer que la Estadística tiene su origen en el interés de los Estados por conocer los recursos con que contaban, es decir, número de habitantes, edad, tipo de trabajo que realizaban, condiciones de vida, propiedades, etc. Si nos retrotraemos a los tiempos del Imperio romano y egipcio ya aparecen manifestaciones de ese interés por procedimientos que permitieran obtener datos estadísticos (de estado). En cualquier caso y tal y como nos indica Hacking (1990), va a ser en el periodo del dominio napoleónico cuando se produzca el salto más sustancial en ese área por disponer de información estadística. Posteriormente, el interés se ha desplazado a otros niveles de análisis más moleculares (ciudades, barrios, colegios, grupos concretos de personas etc.) o también, en muchas veces, más molares (grupos de naciones, continentes, el mundo, etc.). Por otra parte, Herbert (2019), sugiere que el interés del análisis se ha extendido a todo tipo de variables más allá de las que típicamente cubre el censo, el heredero directo de aquel empeño remoto en que se puede encontrar el origen de la Estadística.

Aunque la Estadística aparezca originalmente asociada al interés por conocer acerca de los habitantes de naciones, regiones u otros tipos de agrupaciones, pronto trascenderá su aplicación a otro tipo de unidades no necesariamente humanas, pero en los que era también habitual la recogida de volúmenes amplios de datos de los que se quería extraer la información esencial –este es el caso, por ejemplo, de la investigación astronómica o, más reciente en el tiempo, la investigación acerca de productos de consumo Ese interés por la Estadística ha dado lugar al desarrollo de una serie de conocimientos y procedimientos orientados a satisfacer dos grandes niveles de competencias: (1) resumir la información recogida, habitualmente cuantiosa, de un modo que resulte más comprensible y permita tomar decisiones útiles; y (2) inferir sobre una población numerosa en su tamaño, a partir de un subconjunto reducido de miembros de esa población

Ambas necesidades, y siguiendo las investigaciones de Amon (1978b), han dado lugar a las dos grandes ramas que tradicionalmente se suelen diferenciar dentro del campo de la Estadística, la Estadística descriptiva y la Estadística inferencial. La aplicación de ambas no es excluyente sino, con frecuencia, complementaria y muy utilizadas ambas en Psicología

Otra diferenciación también tradicional es la de Estadística teórica versus Estadística aplicada, la primera más dirigida al desarrollo y estudio de métodos formalmente válidos para hacer Estadística, mientras que la segunda está orientada a la aplicación de esos métodos a campos de estudio concretos. El análisis de datos es un término que se usa habitualmente con el mismo sentido que el de Estadística aplicada.

3 El método científico y la Estadística

En las investigaciones realizadas por Stiger (2002), nos indica que la Estadística se enmarca dentro del método que utiliza la Psicología en su desarrollo como Ciencia, el conocido como método científico. El método científico representa una estrategia ordenada y sistemática de actuación en la realización de un estudio o investigación. Es precisamente la aplicación del mismo, lo que permite dotar a una investigación del calificativo de científica.

Según Navas (2001), las siguientes son características básicas inherentes al método científico:

(1) se aplica a cuestiones empíricamente contrastables;

(2) busca resultados de aplicación tan general como sea posible;

(3) da pie a la replicabilidad de los estudios basados en su aplicación;

(4) busca explicaciones parsimoniosas.

Los pasos que se pueden diferenciar en la aplicación del método científico (Navas, 2001) se describen en la Figura 1.

Figura 1. Método Científico. Fuente: Navas, (2001)

En síntesis, la Estadística, como eslabón del proceder del método científico, nos va a permitir satisfacer el objetivo de resumir y transmitir de un modo comprensible la información procedente de datos empíricos (Estadística descriptiva) así como, cuando sea oportuno, generalizar a partir de la información recogida de un conjunto reducido de sujetos a una población más amplia a la que éstos representen (Estadística inferencial).

4 La Escuela Experimentalista-Estadística en la Psicología

El siglo XIX se caracterizó por la preocupación de la Psicología por alinearse con las ciencias de la naturaleza y por explicar la conducta del hombre como entendimiento y prolongación de las leyes naturales que la rigen. Sin lugar a duda fue la herencia más directa de la Ilustración.

Las teorías explicativas se centraban en la determinación de asociaciones cuantitativas, en la elaboración de leyes que intervenían de manera similar a las funciones matemáticas, a la recopilación de datos, a la realización de un tratamiento estadístico de los mismos, al desarrollo de hipótesis explicativas (experimentación), etc., contribuyó en definitiva, a la construcción de la metodología estadística adoptada por la Psicología y cuya génesis y desarrollo se ubicaba en las ciencias de la naturaleza. Por lo tanto, el devenir de la Psicología se orientaba hacia el conocimiento positivo descansando sus pilares en dos postulados. El primero circunscrito a la concepción de la verdad entorno al hombre que se agota en su ser natural y el segundo postulado residía en la senda del conocimiento construido a través del ámbito científico estableciendo vínculos cuantitativos, estudios estadísticos, construcción de hipótesis y verificación experimental de los mismos.

Así pues, la ciencia del comportamiento humano se rige, hasta la actualidad, sobre una peana precisa y objetiva en donde la Estadística y las Matemáticas establecen los escenarios adecuados para la experimentación y comprobación empírica que da explicación a los fenómenos de esta naturaleza (Foucault, 1994).

Establezcamos los orígenes de este movimiento digamos "científicista". En el año 1879, el psicólogo, fisiólogo y filósofo, Wilhelm Wundt (1832–1920) fundó el primer laboratorio de Psicología experimental en la Universidad de Leipzig, Alemania. Se establece con ello el referente histórico del que emergerá la Psicología como una disciplina diferenciada de la Filosofía (Brennan, 1999). Por otro lado, no olvidemos que el auge de la ciencia moderna ya estaba consolidado por las investigaciones desarrolladas por Galileo (1564–1642), y en pleno siglo XIX había representado un ingente devenir de descubrimientos científicos y su correspondiente aplicación tecnológica. Y es que, tal y como nos dijo Bertrand Russell (1872–1970), la ciencia moderna comenzó explorando astros y prosiguió con los cuerpos físicos en la Tierra para pasar, a continuación, a los seres vivos para terminar en la explicación del comportamiento del ser humano, es decir, de lo más lejano a lo próximo (la condición humana).

Pero ¿cuándo surge la Psicología como ciencia?, Marx y Hillix (2007), sugieren que todo apunta hacia el positivismo defendido por Auguste Comte (1798–1857) que considera que únicamente las ciencias dignas de tal nombre son las ciencias experimentales como la física, la química, la biología y la astronomía, es decir, aquellas disciplinas que no obedecen a los principios y criterios científicos enmarcados por el positivismo difícilmente podrían asignarle el calificativo de "ciencia". Por su parte Brennan (1999), apunta a que la Psicología tardó resulta menos que el resto de las áreas de conocimiento, argumentándose que su ciente y que justificaba claramente que Wundt estableciera un laboratorio de Psicología experimental en la Universidad de Leipzig.

Entre 1885 y 1890, el psicólogo inglés Conwy Lloyd Morgan (1852-1936), profesor de Zoología y Geología en el Colegio Universitario de Bristol, inicia la Psicología experimental animal en el contexto natural (campo), mientras que E. L. Thorndike (1898) lo desarrollaría en el laboratorio o bien con aparatos (laberintos con libros y la famosa jaula de Thorndike). Leonard Trelawney Hobhouse (1864-1929), después de observaciones cuidadosas del comportamiento de varias especies animales en el zoológico de Manchester, publica en 1901 un interesante libro titulado Mind in Evolution, en el que comenta que los gatos y los perros pueden desarrollar conductas orientadas a solventar problemas, por ejemplo, accionando una palanca para obtener comida.

En este escenario queda evidente la utilización de las herramientas estadísticas y en general matemáticas para el correcto trabajo en laboratorios especialmente diseñados para controlar variables, utilización y manejo de aparataje especializado, cajas problema, laberintos, instrumentos eléctricos en donde presentar y registrar datos procedentes de variables, etc. No olvidemos los trabajos sobre el comportamiento animal procedentes del establecimiento de un paradigma explicativo del comportamiento humano (trabajos de Thorndike y Skinner), en los que los resultados se obtienen a través de un análisis estadístico de los datos registrados como producto de esta experimentación en laboratorio (García Vega, 1985).

W. S. Small (1870-1943), profesor en la Universidad de Clark en 1898, usó por primera vez la rata como objeto de estudio en el análisis del comportamiento de animales en laboratorio, recopilando una serie de datos que posteriormente estudió desde una perspectiva estadística.

Robert Mearns Yerkes (1876-1956) estudiante y profesor en la Universidad de Harvard junto con Segius Margulis (1909), publicaron un artículo sobre el método de estudio de las apreciaciones sensoriales de los animales basado en análisis estadísticos

fruto de la recopilación de datos obtenidos en los estudios experimentales. Durante la primera guerra mundial ejerció como psicólogo del ejército, encargándose de la elaboración de los test de selección de militares. En 1925, abre un laboratorio en New Haven, contando con varios chimpancés y dos años más tarde, en 1929-1930 con la colaboración de la Fundación Rockefeller, crea los famosos laboratorios Yale de Biología de Primates

Yerkes dirigió los laboratorios hasta 1941 sucediéndole Lashley que investigó a los primates desde aspectos de la psicología. En 1955 le sucedió H. W. Nissen, que trabajó con chimpancés centrándose en estudios relacionados con la Psicología Evolutiva, reproducción y comportamiento en general. A éste le sucedieron, primero, A. J. Riopelle y, después, G. H. Bourne, etc. En estos laboratorios se investigó sobre muchos temas, entre otros: reflejos condicionados, programas de refuerzo, conducta social, signos preliminguísticos, solución de problemas, aprendizaje de discriminación, etc., y todo ello, como bien sabemos bajo la operativa instrumental de la Estadística.

Por otro lado, hay que destacar las investigaciones de M. E. Bitterman, director del Instituto de Ciencias Sensoriales en la Universidad de Hawai (Manoa, Honolulu), que realizó multitud de trabajos y publicaciones científicas basadas en resultados estadísticos. En efecto, la obra de Bitterman y sus colaboradores es una auténtica investigación comparada de especies (curvas de aprendizaje, comparación en capacidad discriminativa, solución de problemas, aprendizaje de inversión), sin olvidar los famosos experimentos de aprendizaje de probabilidad, en donde ciertos animales tienden a elegir siempre, entre dos estímulos, el que fue recompensado en primer lugar (y no en el porcentaje de veces que fue reforzado); otros animales, como los peces y las cucarachas, eligen en este caso uno u otro estímulo al azar.

5 Estudios estadísticos del tiempo de reacción

En la década de 1850, Helmholtz, desarrolla experimentos en donde el registro y tratamiento estadístico se manifiestan como imprescindible para la consecución de resultados muy significativos. Helmholtz intentaba medir la velocidad del impulso nervioso. El diseño de su experimento se centraba en el tiempo de reacción fisiológico y para ello aisló un músculo y su correspondiente nervio motor de una rana estimulando distintos puntos del nervio y observando el tiempo exacto que transcurría entre la estimulación y el movimiento muscular.

En la década de 1860, F. C. Donders y De Jaeger, comprueban la importancia de otros factores en sujetos humanos donde no exclusivamente fisiológica, sino también psíquica y de planteamiento experimental acerca de los tiempos de reacción. Por lo tanto, la necesidad de aplicación de la herramienta estadística para la obtención de resultados significativos estrictamente necesaria para el buen fin de estos experimentos.

Otra modalidad de experimento muy frecuente bajo el rigor estadístico fue la denominada de reacción de tipo asociativo. El sujeto, ante una palabra o frase, debe responder verbalmente con otra palabra (sinónima, contraria, homónima, etc.). El aparato más usado para medir el tiempo de reacción entre el estímulo y la respuesta fue, en un principio, el cronoscopio, ideado en 1942 por el famoso relojero y físico Matas Hipp. Este aparato medía el paso del tiempo en centésimas de segundo.

6 La medida de las habilidades

Sir Francis Galton (1822-1911), representa mejor que nadie la actitud evolucionista en Psicología, siendo el primero en aplicar sus principios de variación, selección, adaptación, etc., al estudio de las diferencias individuales entre los hombres. Galton tenía la firme convicción de que las características psíquicas como las físicas eran heredadas y medibles. Por lo tanto y como resultado de esta medición y de su posterior tratamiento estadístico se podían obtener explicaciones científicas válidas y fiables para el estudio del comportamiento humano.

En su obra Hereditary Genius, publicada en el año 1869, Galton investiga las líneas familiares de hombres excepcionales a través del método genealógico, tratando de demostrar a través de la Estadística, que las cualidades eminentes de estos genios se encuentran frecuentemente y con claridad en sus antepasados, y que no sería posible explicar su existencia por la mera influencia del medio ambiente. De los casos estudiados, dedujo que el 48 % de las veces los hombres ilustres habían tenido padres ilustres, por lo que los grandes hombres tienen más posibilidades de tener hijos superdotados que las personas normales. Sin embargo, hay que indicar que estas averiguaciones se habían iniciado sobre presupuestos dudosos puesto que obtuvo conclusiones estadísticas sobre mediciones subjetivas del talento, y valoraba el grado de superioridad de un sujeto por el porcentaje de población que le superaba.

En 1883, Galton publicó la obra Investigaciones sobre el talento humano y su desarrollo. Con esta obra se convierte en el gran precursor de la Psicología científica. Wundt acogió este método de Galton y, mejorándolo, lo usó en su laboratorio de Leipzig. En 1884 se celebra en Londres la Exposición Internacional de la Salud Galton instala en ella un laboratorio Antropométrico donde todo el que lo deseara podía someterse, por tres peniques,

a 17 medidas de diversos aspectos de su persona; altura, peso, vista, ódo, longitud de mano, fuerza (descargando un golpe, apretando, tirando), capacidades respiratorias, etc. Casi todas las medidas correspondían a los procesos elementales que Wundt buscaba en su laboratorio. La poínicial de Galton es la de un antropólogo físico: quiere datos, medidas de las carásticas físicas y metales de las personas, y busca entre ambas las relaciones que estaba seguro de encontrar. En éstico, así en la aplicación estadística, sigue los pasos del matemático belga Adolph Quetelet (1796-1874) que, estudiando la estatura de muestras muy numerosas de personas, llegaba a la conclusión de que los datos se repartían según la ley de Laplace y Gauss, esto es, según la curva normal.

Quetelet supuso entonces que las diferencias se debían "errores accidentales" y que todo parecía indicar que el "propósito inicial" de la Naturaleza era obtener el "hombre medio". Galton creyó que esto era cierto, y no sólo en lo referente a los rasgos físicos, sino también en los psíquicos. Así es que lo primero que hizo con los datos del Laboratorio Antropométrico fue construir las curvas normales correspondientes. Con estos datos obtenidos en 1884 también pudo ofrecer a cada interesado el porcentaje de sujetos que en esa medida era superior e inferior (método de percentiles, tan usado posteriormente en las pruebas psicométricas). También, con los mismos datos, pretendió comprobar si los rasgos entre sí pueden relacionarse mutuamente y si esta covariación es determinable estadísticamente. Así queda esbozado por primera vez en la historia de la Psicología el concepto de correlación.

Charles Spearman (1863-1945), profesor del University College de la Universidad de Londres desde 1907 hasta su muerte, le llama la atención la hipótesis de Galton de que la capacidad intelectual estaba correlacionada con las diferencias de sensibilidad para apreciar tonos, colores, etc. Spearman (1904), trata de comprobar esta hipótesis relacionando ciertos tipos de sensibilidad y las notas obtenidas por alumnos de una escuela elemental. Por este y otros procedimientos llega a la conclusión de que el grado de cada habilidad depende de un factor general (g) y de otro específico de esa habilidad concreta, al que llama factor (s) (specific). A esta teoría se la denomina Teoría de los dos factores, frente a la doctrina de Thurstone, denominada Teoría de los factores múltiples, es decir, apunta la exigencia de más de un factor común, cuando las correlaciones no se disipan, tal y como informa Kelley (1928).

Boring (1978) y Barbado (1974), señalaron que, en la década de 1930 el análisis factorial se desarrolló bajo el liderazgo de tres grandes psicólogos: Cyril Burt, opuesto a la teoría del factor único de Spearman (1927)); Godfrey H. Thompson, cátedra de Psicología de la Educación de la Universidad de Edimburgo, y Louis Leon Thurstone (1887-1955), profesor de Psicología de la Universidad de Chicago desde 1924.

Thurstone (1947), ayudado por su esposa Thelma Gwinn, desarrolla un nuevo método para detectar los factores de la inteligencia: la técnica del análisis factorial múltiple, llegando a aislar ocho habilidades mentales primarias (comprensión verbal, espacial, razonamiento, numérica, fluidez verbal, memoria, velocidad perceptiva y motora), aclarando que según la investigación progresara, se podían ir añadiendo más. Esta afirmación coincide con las investigaciones de Yela (1956).

Desde un punto de vista crítico y seguidor de Galton, irrumpe en escena James McKeen Cattell (1860-1944), que en 1890 puso en circulación un término muy afortunado: los "test". En 1896, Binet y Henry, en un artículo publicado en el Année Psychologique titulado "La Psicología individual", afirmaron que era mejor diferenciar a los sujetos por sus facultades superiores.

Alfred Binet (1857-1911), comenzó su carrera científica en el laboratorio de Psicología fisiológica de la Soborna, fundado en 1889. En 1894, sucede en la dirección de este laboratorio al profesor de fisiología Henri Beaunis (1830-1921). En 1900, su actividad se centra en el laboratorio de pedagogía experimental y como resultado de sus experimentos surge la Escala de inteligencia

En 1912, el alemán W. Stern aporta el valioso concepto de coeficiente de inteligencia.

7 La importancia de la Estadística en Psicología

Como hemos visto, a lo largo de la historia de la Psicología, la Estadística es un instrumento que se emplea en diferentes ramas y/o disciplinas de esta área del conocimientos, es decir, como Psicología experimental, del aprendizaje, educacional, clínica, etc. y es que no debemos olvidar que la Estadística es una ciencia con bases matemáticas que se ocupa de la recolección, análisis e interpretación de los datos obtenidos, permite entender de manera más clara un fenómeno determinado tanto de las ciencias formales como de las empíricas.

Tal y como nos indica Carpintero (1987) citado en Pérez-López (2010), la actividad profesional del psicólogo consiste, entre otras, en el estudio empírico de diversos casos, ya sean individuales o bien grupales, por lo que es necesario y de obligado cumplimiento desarrollar métodos rigurosos de recopilación de datos para que posteriormente sean analizados. De igual manera es necesario tener un diseño de investigación, un método y herramientas para la evaluación del estudio, estamos hablando

en de nitiva de hacer uso de la estadística. El psicólogo debe saber plantear el problema a investigar o bien el caso a tratar, debe saber describir las variables en función de las hipótesis de investigación y conocer el tratamiento estadístico de los datos conducentes a la elaboración de los resultados que proporcionan las pruebas psicométricas.

Y es que el conocimiento de la Estadística interviene en todas las fases del trabajo de investigación, desde la decisión sobre las variables que se investigan y la planificación, hasta la forma en que se han de recoger los datos, para nalmente obtener una interpretación de los resultados obtenidos. En esta fase hay aspectos exclusivamente estadísticos, como serían los relativos al diseño del muestreo o bien la elección del método concreto de análisis y su relación. Otros dependen del problema planteado: elección de las unidades de análisis, variables, escalas de medida; procesos de categorización, etc. Estos puntos no pueden dejarse a la responsabilidad del estadístico puesto que dependen de las preguntas de investigación que solo tienen sentido profundo para la persona que las ha planeado, es decir para el psicólogo. No olvidemos que para la realización de un experimento psicológico son necesarios los siguientes pasos:

1. Planeamiento estadístico de la investigación: Localizar las fuentes de información, escoger el material.
2. Plantear un problema de estudio.
3. Diseñar el experimento. Se valida el modelo comparándolo con lo que sucede en la realidad. Se utilizan métodos estadísticos conocidos como test de hipótesis o prueba de signi cación.
4. Se producen Estadísticas descriptivas.
5. Inferencia estadística. Se llega a un consenso acerca de qué dicen las observaciones acerca del mundo que observamos.

En de nitiva, la Estadística es de suma importancia para la Psicología ya que permite extraer y resumir información útil de las observaciones que se hacen, el psicólogo debe basar sus decisiones en datos limitados y estas serán más faciles de tomar con la ayuda de la estadística, le da mayor claridad y precisión al pensamiento y la investigación empírica psicológica.

8 La enseñanza de la Estadística en Psicología

Cuando se estudia Estadística aplicada a las Ciencias Sociales, la cuestión que se plantea es hasta donde hay que profundizar en estas materias relacionadas con las ciencias humanas puesto que lo que se pretende es desarrollar un buen trabajo y no tanto en ser especialista en Estadística. Hablar del uso de la Estadística facilita las cosas para entender donde veo el problema. Este radica en la conexión del usuario con un instrumento complejo y delicado, análogo al que media en la relación hombre máquina, y en la necesidad de elaborar una adecuada interface entre ambos.

No carece de precedentes esta comparación de la Estadística con un instrumento. Hace ya muchos siglos que Aristóteles se re rió a la lógica como órganon o instrumento para el pensamiento, para la Ciencia. Luego, pasados varios siglos, Condillac (1714-1780) llegaría a decir que la Ciencia es un lenguaje bien hecho (Bono Guardiola, 2002; Cepedello Bolso, 2001). Al n y al cabo un instrumento y una lengua es un conjunto de elementos ordenados, que operan transformando o produciendo resultados mediante ciertos impulsos o esfuerzos indiferenciados recibidos y su propia estructura especializada, que se aplica sobre determinados materiales, lo importante en el instrumento, para quien lo emplea de modo práctico, no están en las variables o pasos intermedios que transforman la acción del usuario en una serie de procesos, sino en el punto de partida -el problema y sus materiales, o datos con los que contaba- y en la meta que se busca -la solución o interpretación, luego aprovechada para ulteriores fines (Matthews, 1994).

En general, el usuario que acude al tratamiento estadístico de los datos reunidos en algún tipo de búsqueda o investigación está interesado principalmente en resolver ésta, y lo que demanda es la clari cación de la dependencia o no de ciertas interacciones entre distintas dimensiones del fenómeno estudiado. Busca, en de nitiva, precisar la estructura de relaciones funcionales entre sus distintas variables. Por supuesto, necesita otras muchas cosas, por ejemplo, conocer y distinguir las variedades de materiales que maneja y las constricciones que estas se deriven. Necesita ver con claridad si se trata de relaciones a establecer con variables cualitativas o no; si emplea medidas repetidas sobre los mismos sujetos, o no; si se mueve entre distribuciones paramétricas, o no; y tantas otras cosas así, que debe tener claras desde el comienzo de su estudio, y que constituyen la lógica general de indagación, tal y como concluye Hothersall (1995).

Así pues, necesita tener muy claro en su pensamiento qué es y qué no es “estadísticamente signi cativo”, y cómo las condiciones de esta relación pueden variar aumentando o disminuyendo las muestras, o encontrando pruebas más potentes que las usadas antes si ello es posible. Y necesita, en muchas ocasiones, tener ante los ojos un modelo, que sirva de paradigma, y cuya lógica pueda ir siguiendo mediante razonamientos de simple analogía, contribuyendo con datos propios en los esquemas del

modelo. Y si esto no es difícil encontrarlo referido al análisis de varianza, a la correlación o a la regresión, es algo menos asequible de hallar referido al análisis discriminante, al path-analysis o al análisis factorial de correspondencias, tan útil en muchos de nuestros trabajos.

Garriga Trillo (1990) y Garriga Trillo y Dorn (1991) coinciden en afirman que cualquier enseñanza de un saber instrumental debe movilizar al sujeto motivándole desde el logro de aquellos fines que espera obtener con el uso de lo aprendido. La Estadística es un lenguaje con que dar sentido a los hechos, y también con el que hablar al especialista para que ayude a refinar los procedimientos de análisis a emplear y que permite, sobre todo, comparar distintas muestras, establecer sus diferencias, y estructurarlas en configuraciones de dependencia, agrupando unas y separándolas de otras; que pondera el peso de ciertas variables dentro de un conjunto, que muestra tendencias y evolución de datos en ciclos.

9 ¿Por qué la Estadística en los estudios de Psicología?

Una divergencia importante es la consideración de la Psicología Matemática como área de conocimiento distinta de la Psicofísica y de la Psicometría. La Psicología Matemática, la Psicometría y la Psicofísica no son áreas distintas de estudio en Psicología, todo lo contrario, su sinergia es fundamental para saber integrar los conocimientos que albergan. Como todos sabemos la Psicofísica es parte de la Psicología Matemática y no un área aparte. Encabeza esta afirmación Fechner (1860/1996), que impulsó la Psicología científica originando la Psicología Matemática. Por lo tanto, podemos decir que ambas disciplinas tienen un comienzo paralelo en el devenir de los tiempos. Posteriormente surge la Psicometría debido al uso de modelos matemáticos aplicados a la Psicología y por lo tanto se va diversificando y agrupando en otras subdisciplinas como es el caso. Tanto es así, que en algunas universidades extranjeras se ha añadido el prefijo Quantitative al de Mathematical Psychology para discernir claramente que se incluye dentro de ella a la Psicometría (Leahey, 1996).

La Estadística nos va a proporcionar un tipo de conocimientos y competencias que favorecen el pensamiento lógico y crítico, nos capacita para realizar estudios (investigaciones) en los que, bien poner a prueba conjeturas (hipótesis) que nos planteemos, bien buscar la respuesta a preguntas que nos surjan. Es crucial tener unos conocimientos básicos de Estadística para poder leer publicaciones (notas de prensa, artículos en revistas especializadas, informes de investigación, etc.) acerca de temas psicológicos, que son los que van a permitir la especialización y la mejora del desempeño profesional del psicólogo.

En efecto, el psicólogo necesita la matemática no sólo para efectuar contrastes estadísticos entre sus modelos teóricos y sus observaciones empíricas, sino también, y quizás preferentemente, para idear modelos matemáticos deterministas y estadísticos que le permitan formular con precisión sus hipótesis y deducir de ellas con exactitud las consecuencias lógicas y experimentalmente comprobables que implican.

10 Conclusiones

Los psicólogos deben ser, sin duda y, ante todo, psicólogos. Su tarea es estudiar los fenómenos psicológicos, pero cuando utilizan la matemática deben hacerlo comprendiéndola y dominarla, al menos en las áreas y en la medida que sus pesquisas psicológicas requieran. Las técnicas estadísticas no son artilugios que pueden conferir significación a cualquier tipo de datos (Freedman, 1999), la Estadística sólo es fecunda si se aplica a ideas psicológicas de rigor, que merezca la pena precisar matemáticamente o contrastar con los datos empíricos. El rigor y la precisión en el registro y ordenación de los hechos y la expresión precisa de las regularidades depende de los procedimientos estadísticos que se utilicen y deben ser elegidos y elaborados previamente, y no posteriori, para ajustarse a los fines de la investigación, a las características de los datos y a las condiciones en que éstos se indagan.

Una de las exigencias principales en la formación estadística del psicólogo es el dominio de técnicas para la recogida, ordenación y análisis de datos de observación. Es un campo muy atendido hoy por la metodología psicomatemática.

La matemática estadística interviene aquí para formar y expresar las ideas e hipótesis psicológicas de la manera más precisa posible; para derivar de ellas, y con mayor poder deductivo que el que permite el razonamiento verbal y con la máxima garantía de coherencia lógica, las consecuencias observables que de esas hipótesis se siguen, y para, finalmente, comparar de forma estricta y objetiva las previsiones hipotéticas con las efectivas observaciones experimentales o empíricas.

Así pues, y siguiendo Salsburg (2002), a la Estadística debe estar siempre al servicio de la Psicología. Al psicólogo le interesa el desarrollo y aplicación de una matemática cualitativa, tanto o más que el dominio de técnicas cuantitativas. La naturaleza de los fenómenos psicológicos, en los que los aspectos y las relaciones cualitativas suelen ser predominantes y decisivos, exige el uso creciente de procedimientos matemáticos fundamentados