OPOSICIONES

A LOS CUERPOS DE ESTADÍSTICOS DEL ESTADO

EXÁMENES PRÁCTICOS RESUELTOS

CDEE CSEE

PARTE COMÚN
ESPECIALIDAD CIENCIA DE DATOS

JULIO CÉSAR HERNÁNDEZ SÁNCHEZ • MIGUEL GARCÍA DÍAZ



OPOSICIONES A LOS CUERPOS DE ESTADÍSTICA DEL ESTADO

EXÁMENES PRÁCTICOS RESUELTOS

© Publicaciones Instituto Nacional de Estadística, Madrid 2025

Julio César Hernández Sánchez

Miguel García Díaz

ISBN: 9788409557516

Edición: 2ª

Nº de páginas: 506

Actualizado a 28 de Marzo de 2025

Ilustración de la portada y diseño gráfico:

Tomás Hernández Sánchez

A Isa, Julia y Teresa, que son mi vida.

A mi hermano Tomy, un genio.

A mis padres que me enseñaron la humildad.

A mis padres, mi hermana y Mario por estar ahí siempre.

A Silvia, Bego y Dulce por todo lo compartido, la luz y el optimismo.

Y a esos buenos profesores.



Julio César Hernández Sánchez

Licenciado en Matemáticas

Ingeniero Técnico en Informática de Sistemas

Estadístico Superior del Estado desde 2005

Delegado Provincial del INE en Zamora



Miguel García Díaz

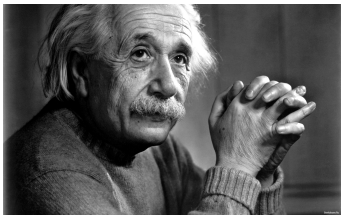
Licenciado en Matemáticas
(Ciencias de la Computación)

Máster en Finanzas Cuantitativas (AFI)

Mathematics is the most beautiful and
most powerful creation of the human spirit.
Stefan Banach

I have had my results for a long time: but I do not yet know
how I am to arrive at them.
Carl Friedrich Gauss

Agradecimientos



Try not to become a man of success but a man of value.

Albert Einstein

Nos gustaría agradecer el esfuerzo de todas las personas que han hecho posible que este documento pueda ver la luz y pueda ayudar a aquellos opositores ilusionados con este camino. En particular queremos hacer una mención especial a Cristobal Rojas Montoya, María de Blas Portero, Román Mínguez Salido, Mariano Matilla García, María Dolores Lorca y Carlos Valero, así como a Francisco Romero Cerrato y a Joana Morillo Martín, que con sus ideas y su experiencia han ayudado a que la versión de este libro haya mejorado, además de aquellos compañeros del INE que han prestado su disposición y colaboración en todo aquello que hemos necesitado.

Queremos agradecer también a Miguel Ángel de Castro, a Miguel Ángel Martinez y a Adolfo Gálvez toda su ayuda y apoyo en este proyecto, continuación del que surgió en 2021, pero que con un cambio de temario muy significativo en los procesos selectivos había que actualizar, y que ellos reci-

bieron con entusiasmo y han impulsado para que pudiera ser una realidad.

En particular nos acordamos de nuestras familias, a las que hemos robado mucho tiempo para que esta iniciativa saliera adelante, y que siempre nos apoyan, y a tantos amigos que nos han animado con esta aventura.

Prólogo

Este libro ha nacido para que puedan utilizarlo todos aquellos opositores que estén preparando el acceso a los cuerpos de estadística del estado, aunque por su carácter multidisciplinar puede ser de utilidad en otros ámbitos.

Esperamos que tengáis la bondad de perdonar los errores que pudiéramos haber cometido al elaborarlo.

También nos gustaría agradecer las correcciones, aportaciones e ideas de quien en el uso de este libro pudiera detectar alguna cuestión que fuera susceptible de cambio con el objetivo de mejorar este material. Todo aquello que recibamos será considerado para publicar futuras versiones mejoradas y que puedan seguir siendo útiles para toda la comunidad estadística. Nuestro contacto es el siguiente:

juhersan.zz@gmail.com

zeycus@gmail.com

Este libro se irá actualizando con los exámenes de problemas que cada año vayan publicándose de las sucesivas convocatorias del cuerpo de diplomados en estadística del estado y con los exámenes de la parte común del cuerpo superior y de la especialidad de ciencia de datos, con la idea de crear una herramienta útil y actual para el opositor/a.

Índice general

Agradecimientos **VII**

Prólogo **IX**

1. Introducción **1**

2. Año 2022. Turno libre. **11**

2.1. Soluciones del segundo examen del Cuerpo Superior de Estadísticos del Estado de 2022. Parte Común. 12

2.2. Soluciones del cuarto examen del Cuerpo Superior de Estadísticos del Estado de 2022. Especialidad ciencia de datos. 43

2.3. Soluciones del segundo examen del Cuerpo de Diplomados en Estadística del Estado de 2022. 86

3. Año 2022. Promoción interna. **117**

3.1. Soluciones del segundo examen del Cuerpo Superior de Estadísticos del Estado de 2022. Parte común. Promoción interna. 118

3.2. Soluciones del cuarto examen del Cuerpo Superior de Estadísticos del Estado de 2022. Ciencia de datos. Promoción interna. . 149

3.3. Soluciones del segundo examen del Cuerpo de Diplomados en Estadística del Estado de 2022. Promoción interna del grupo C1. 183

4. Año 2023. Turno libre. 213

4.1. Soluciones del segundo examen del Cuerpo de Diplomados en Estadística del Estado de 2023. 214

4.2. Soluciones del segundo examen del Cuerpo Superior de Estadísticos del Estado de 2023. Parte Común. 249

4.3. Soluciones del cuarto examen del Cuerpo Superior de Estadísticos del Estado de 2023. Especialidad ciencia de datos. Turno libre. 286

5. Año 2023. Promoción interna. 353

5.1. Soluciones del segundo examen del Cuerpo Superior de Estadísticos del Estado de 2023. Promoción interna. 354

5.2. Soluciones del cuarto examen del Cuerpo Superior de Estadísticos del Estado de 2023. Especialidad ciencia de datos. Promoción interna. 361

Bibliografía 363

APÉNDICES

Apéndice A. Exámenes de problemas de 2022. Turno libre. 369

A.1. Segundo examen del Cuerpo Superior de Estadísticos del Estado. Parte común. Turno libre. 369

A.2. Cuarto examen del Cuerpo Superior de Estadísticos del Estado. Especialidad ciencia de datos. Turno libre. 378

A.3. Segundo examen del Cuerpo de Diplomados en Estadística del Estado. Turno libre. 387

Apéndice B. Exámenes de problemas de 2022. Promoción interna. 399

B.1. Segundo examen del Cuerpo Superior de Estadísticos del Estado. Parte común. Promoción interna. 399

B.2. Cuarto examen del Cuerpo Superior de Estadísticos del Estado. Especialidad ciencia de datos. Promoción interna. 406

B.3. Segundo examen del Cuerpo de Diplomados en Estadística del Estado. Promoción interna grupo C1. 415

B.4. Segundo examen del Cuerpo de Diplomados en Estadística del Estado. Promoción interna grupo C1 especialidad estadística. . 426

Apéndice C. Exámenes de problemas de 2023. Turno libre. 435

C.1. Segundo examen del Cuerpo Superior de Estadísticos del Estado. Parte común. Turno libre. 435

C.2. Cuarto examen del Cuerpo Superior de Estadísticos del Estado. Especialidad ciencia de datos. Turno libre. 447

C.3. Segundo examen del Cuerpo de Diplomados en Estadística del Estado. Turno libre. 457

Apéndice D. Exámenes de problemas de 2023. Promoción interna. 469

D.1. Segundo examen del Cuerpo Superior de Estadísticos del Estado. Parte común. Promoción interna. 469

D.2. Cuarto examen del Cuerpo Superior de Estadísticos del Estado. Especialidad ciencia de datos. Promoción interna. 477

Apéndice E. Galeria de imágenes **487**

E.1. Autoría y referencias bibliográficas de las imágenes portada de cada capítulo 487

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA



Capítulo 1

Introducción



"True science teaches, above all, to doubt and to be ignorant."

Miguel de Unamuno

Actualmente los procesos selectivos a los Cuerpos de Estadísticos del Estado se ajustan a lo dispuesto en un Real Decreto de 2015, de 30 de octubre, por el que se aprueba el texto refundido de la Ley del Estatuto Básico del Empleado Público, y constan de una fase de oposición, tanto para los aspirantes que se presenten por el sistema general de acceso libre, como para los aspirantes que se presenten por la convocatoria de promoción interna. Además el proceso selectivo,



para aquellos que superen la fase de oposición, incluye un curso selectivo que se valora y que influye en la puntuación final de los candidatos.

En la convocatoria de 2022 del cuerpo superior de estadísticos del estado el temario ha sufrido un cambio muy significativo respecto de los procesos anteriores que se venían desarrollando. A partir de este año hay dos especialidades, una relativa a Estadística-Ciencia de datos y otra relativa a Estadística aplicada a las ciencias económicas y sociales, compartiendo ambas una parte común de temario que tendrá que ser preparada por todos los candidatos. Relativa a esa parte de materias comunes versa el primer examen, que es una prueba de contestar por escrito a veinticinco preguntas cortas, a elegir de entre treinta propuestas por el Tribunal relacionadas con los temas de dicho bloque común. Este conjunto de conocimientos comunes a ambas especialidades se concentra en temas de derecho, demografía, cuentas nacionales, inferencia y modelización, producción estadística oficial y almacenamiento y modelos de datos. El segundo examen es un examen de problemas también de este conjunto de temas comunes en el cual de 8 problemas se eligen 4 (excepto el apartado de derecho que no se incluye en este examen). El tercer examen es un examen de idiomas en el que se establecen dos modalidades excluyentes de evaluación: una presencial y, otra, sustitutiva de la anterior, mediante la cual los candidatos podrán acreditar su conocimiento de inglés equivalente a B2 o superior, de acuerdo con los niveles del Marco Común Europeo de Referencia de las Lenguas (MCER). La prueba presencial consistirá en la realización de una prueba, escrita y otra oral. Por último, el cuarto examen consistirá en contestar por escrito a cuatro cuestiones prácticas a elegir entre ocho propuestas por el Tribunal, de los temas del grupo de materias específicas según la especialidad consignada en la solicitud de cada candidato. Un aspecto muy significativo en este cambio de paradigma



es que se han eliminado exámenes escritos de desarrollar diversos temas y se ha eliminado el examen oral, que tradicionalmente era el más duro con diferencia y el que requería del mayor esfuerzo para poder asimilar y exponer ante el tribunal los temas requeridos.

El contenido del temario ha supuesto un punto de inflexión en los conocimientos que se demandan que tengan los opositores, de acuerdo al conjunto de exigencias cada vez más novedosas que se van implantando en la estadística oficial y que requieren de unos perfiles diferentes a los que habitualmente estaban entrando en estos cuerpos de estadística. En la especialidad de ciencia de datos se incluyen bloques de bases de datos, programación, tecnologías web, simulación, análisis de algoritmos, entre otros que no aparecían anteriormente. Igualmente, en la especialidad más orientada a las ciencias sociales aparecen como necesarios tener conocimientos en contabilidad financiera o el sistema fiscal español, que son novedosos también. Todos los exámenes son eliminatorios de forma que sólo aquellos aspirantes que superen todos y cada uno de ellos podrán acceder a dicho cuerpo superior. Es decir, parece que esta nueva estructura confiere más importancia a la resolución de problemas que a la acumulación memorística de conceptos y exposición de los mismos de una u otra forma, por lo que todavía es más importante para los opositores poder contar con ayudas como la que este libro intenta ofrecer.

Por el contrario, en las oposiciones al Cuerpo de Diplomados en Estadística del Estado el temario no ha sufrido grandes variaciones, con un planteamiento basado en dos exámenes, uno de preguntas cortas sobre diversos bloques de contenido y uno de problemas que engloba a todas las áreas de conocimiento del temario. La incorporación de exámenes de preguntas cortas ha sustituido a pruebas consistentes en redactar uno o varios temas de diversas áreas temáticas, pero lo que ha permanecido constante a lo largo de



muchos años, y sigue vigente actualmente son los exámenes de problemas en una u otra fase de cada proceso selectivo.

Para preparar las oposiciones es necesaria la utilización de un gran número de textos de diversa índole, tanto de teoría como de problemas, pero dada la diversidad del temario no existen referencias que cubran conjuntamente aspectos de campos tan dispares como demografía, econometría, economía, estadística y muestreo. Es decir, que para profundizar en el conocimiento de cada parte se necesita bibliografía especializada. Y esta situación se traslada de igual forma a la preparación de los exámenes de problemas, teniendo que trabajar para cada apartado en la búsqueda de problemas que pudieran ser objeto selección en estas pruebas. No hay ningún material que aglutine de alguna manera el conocimiento necesario de lo que ha sido susceptible de ser preguntado y tampoco hay material que explique y resuelva los problemas en los diversos procesos de oposición a lo largo de los años. De esta forma, al opositor/a le resulta muy complicado poder tener cierta certeza de cómo se resuelven aquellas cuestiones de ramas tan diversas.

Por este motivo, este libro se enfoca en dichas pruebas, que en el acceso a ambos cuerpos corresponden a varios de los exámenes eliminatorios. En el caso del Cuerpo de Diplomados es el segundo y último examen de la oposición, en el cual se decide el proceso y cobra una importancia vital, y en el caso del Cuerpo Superior es el segundo examen de la parte común, que es la puerta al examen de idiomas y el cuarto y último examen de problemas de la parte específica.

El objetivo del libro no es otro que el de intentar que los opositores cuenten con un material de ayuda que les acompañe en todo el proceso y en su camino de preparación y que de alguna forma les genere seguridad sobre cómo abordar esta parte de la oposición, puesto que un examen de problemas



no está exento de cierta incertidumbre.

Además, se han incluido en el texto los exámenes de promoción interna, tanto del cuerpo de diplomados como del cuerpo superior para poder contar con todos los exámenes disponibles de problemas, puesto que todas las modalidades de acceso a los diferentes cuerpos aportan valor al conocimiento de la resolución de diversas cuestiones y son vitales en un momento de cambio como el que hemos comentado y del que no se disponen de problemas orientativos de diversos bloques que pueden ser susceptibles de ser preguntados en futuras convocatorias. La decisión de añadir estos exámenes se ha hecho para que el libro sea lo más compacto posible y ayude a cada aspirante desde cualquier situación administrativa que pudiera tener.

Los cuerpos de estadística del estado para los que este material pretende ser de ayuda han sufrido muchos cambios a lo largo de los últimos tres siglos. Los antecedentes se remontan a 1856 cuando se creó la Comisión Estadística General del Reino, que estaba compuesta por un grupo de especialistas en estadística que se dedicaron a supervisar, coordinar y asesorar a los diversos Departamentos de la Administración en el ámbito estadístico.

Los Cuerpos de Estadística en sus orígenes estuvieron muy vinculados a lo que hoy conocemos como Catastro, de forma que había una interrelación entre la topografía catastral, la geografía y la estadística. En 1870 se creó una Dirección General de Estadística Autónoma y se publicó un listado con los funcionarios que llevaron a cabo trabajos estadísticos y topográficos. Pero es en 1873 cuando se vio conveniente crear un Cuerpo de Estadística con diversas categorías administrativas, cómo había que acceder, así como las retribuciones y otros aspectos de relevancia.

Fue en 1902 cuando un Real Decreto del entonces Ministerio de Instrucción Pública y Bellas Artes establecía una separación del Cuerpo de Estadísti-



ca en dos partes: el Cuerpo Facultativo de Estadística y el Cuerpo Auxiliar de Estadística, fijando las plantillas de ambos.

En los años treinta la estadística tuvo una gran explosión con los avances de la estadística matemática, así como de otros campos como la probabilidad y la inferencia, pero la guerra civil afectó a todo el engranaje estadístico de forma considerable. Se incorporaron, no obstante, máquinas para el procesamiento de información que hasta entonces se hacía de forma manual, contribuyendo a la formación del Censo de Población en 1930.

El Real Decreto de 29 de enero de 1930 del Ministerio de Trabajo y Previsión aprobó un reglamento que establecía las funciones de los Cuerpos Facultativo y Técnico de Ayudantes de Estadística, incluyendo aspectos sobre el ingreso, vacantes, tribunales de oposición etc...

El ingreso en el Cuerpo Facultativo de Estadística se llevaría a cabo por oposición con ciertos requisitos sobre la edad y la formación de los candidatos, teniendo que superar una serie de pruebas sobre álgebra, cálculo, estadística, economía, derecho y geografía. Para el Cuerpo de Ayudantes de Estadística también se estableció un sistema de oposición, pero no se requería titulación académica.

En 1942 se creó el Cuerpo Técnico de Estadística y Colocación que requería para poder formar parte de él también un proceso de concurso-oposición para el que era necesario tener un título universitario. Es en la Ley Estadística de 1945 cuando queda reflejado que los trabajos del INE deberían ser llevados a cabo por dos cuerpos de nueva creación, que fueron el Cuerpo de Estadísticos Facultativos y el Cuerpo de Estadísticos Técnicos. Al primero le correspondería el trabajo de dirección y de investigación y el acceso al mismo sería mediante oposición entre titulados de grado superior. Respecto al segundo, se encargaría de la ejecución técnica de los diversos



trabajos estadísticos.

En los años 60 se produjo un gran cambio en los cuerpos de estadística puesto que había que acometer nuevos retos profesionales e institucionales, a la vez que una evolución en la legislación del funcionariado también requería diversas adaptaciones. El INE inicia operaciones importantes en esos años, como la Encuesta de Población Activa(EPA), pero se producen situaciones que restan importancia al papel que la institución tenía y trataba de adquirir. Otros organismos e incluso los propios Ministerios comienzan a elaborar estadísticas de diversa índole al margen del INE, de forma que poco a poco esa labor que se supone debía tener la institución como responsable del sistema estadístico se va debilitando en cierto modo.

Para el acceso al Cuerpo de Estadísticos Facultativos, en los años 60, se requería un título universitario de grado superior, pero a partir del 68 se limitaba fundamentalmente, salvo alguna excepción en alguna convocatoria, a licenciados en matemáticas y en ciencias políticas, económicas y comerciales. El temario fue cambiando con el paso del tiempo, y se introdujeron materias como informática y econometría a las ya existentes de derecho, demografía, economía y estadística. Esta circunstancia hizo que la dificultad en las mismas fuera en aumento puesto que no se adecuaba el temario a ningún plan de las carreras universitarias existentes. La plantilla del Cuerpo era de 200 funcionarios en 1945 y se amplió en 40 plazas en 1973, con la idea a partir de 1974 de dotar de 10 por año al cuerpo, pero muchas quedaron sin cubrir por la dificultad de los conocimientos necesarios para el acceso, así como de las pruebas selectivas.

En cuanto al Cuerpo de Estadísticos Técnicos, inicialmente se requería un título de enseñanza media o asimilado, pero a partir de 1968 había que poseer un título de Bachiller Superior, Perito Mercantil o Maestro de Prima-



ria, a la vez que se actualizaron los contenidos y exámenes de la oposición, que se centrarían en Matemáticas, Estadística, Legislación y Economía fundamentalmente. La plantilla inicial fue de 300 funcionarios y la Ley del 73 la incrementó en 120 plazas, que se convocarían a partir de 1974 de forma gradual.

El inicio del uso de los ordenadores en el INE se remonta al año 1964, en el que se utilizó un equipo IBM con tarjetas perforadoras y 16Kb de memoria y para el que fue necesario formar a un equipo de 14 personas, entre facultativos y técnicos. En la década de los 70 se van sustituyendo esos equipos por otros más modernos y con más capacidad, que van dotando al organismo de experiencia en la evolución de las posibilidades técnicas que van ofreciento estos avances.

En el año 77, mediante Real Decreto, se creó el Cuerpo de Estadísticos Técnicos Diplomados, extinguiéndose así el Cuerpo de Estadísticos Técnicos. Una orden en ese mismo año regularía la integración de un cuerpo en el otro.

A finales de esta década se empezó a gestar la elaboración de una nueva ley de estadística que dejase atrás la ley de 1945. Ya la Constitución del 78 establecía las competencias del Estado y entre ellas incluía la estadística para fines estatales, aunque este último término no estaba suficientemente definido. Por este motivo era necesario elaborar una nueva ley de Estadística que pudiera aclarar y desarrollar esta problemática.

Ya en el año 80 se configuraba en los distintos documentos previos a la nueva ley al INE como punto central y pilar fundamental del Sistema Estadístico Nacional, de forma que sería el Plan Nacional de Estadística el que reflejara las estadísticas con fines estatales, así como aquellas que fueran de interés para las CCAA.

En la década de los 80 se exigió ser Licenciado, Ingeniero, Arquitecto



o estudios equivalentes para poder acceder al Cuerpo de Estadísticos Facultativos y al igual que ocurrió anteriormente el temario se fue adaptando progresivamente a los avances en las distintas disciplinas que lo componían, como informática y econometría.

La publicación de la Ley de la Función Estadística Pública tuvo lugar en el año 89, en la cual se elevaba el rango administrativo del INE, que pasaba a ser Subsecretaría de Estado, y establecía la posibilidad de tener autonomía dentro de la Administración, de forma que pasaba a ser un Organismo Autónomo adscrito al Ministerio de Economía y Hacienda.

En el año 2001, la Ley de Medidas Fiscales Administrativas y de Orden Social cambia la denominación del Cuerpo de Estadísticos Facultativos, que a partir de ese momento se llamaría Cuerpo Superior de Estadísticos del Estado, e igualmente cambia la del Cuerpo de Estadísticos Técnicos Diplomados por el nombre Cuerpo de Diplomados en Estadística del Estado, que son los nombres actualmente vigentes de los cuerpos sobre los que este libro de problemas hace referencia. Puede consultarse una exposición mucho más extensa y detallada de la historia de los cuerpos de estadística de la Administración General del Estado en referencias como [Celestino, 2010] y [Merediz, 2004].

Por último, hemos considerado importante añadir en los anexos de este ejemplar, para cada año, el enunciado de los exámenes de problemas (uno del cuerpo de diplomados y los dos, segundo y cuarto, del cuerpo superior, enfocado en la especialidad de ciencia de datos) tal y como se han presentado a los opositores, teniendo cada capítulo del libro el objetivo de resolver todas o la mayoría de las cuestiones propuestas con el mayor detalle posible. Además iremos actualizando el contenido de esta versión con la resolución de los exámenes que se realicen en futuras convocatorias para que este material se

vaya consolidando y sea útil para tantos estudiantes cuyo objetivo es poder formar parte de estos colectivos de la administración.

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA



Capítulo 2

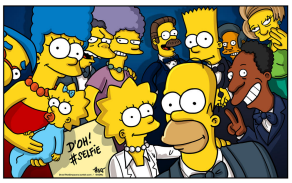
Año 2022. Turno libre.

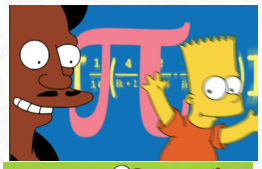
"Exploring pi is like exploring the universe."

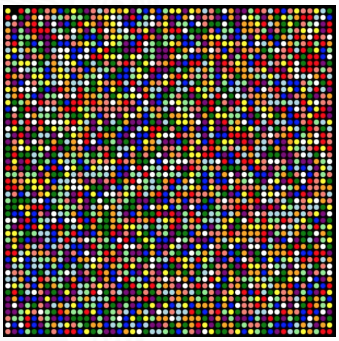
—David Chudnovsky

We'll never know the precise area or circumference of a circle.

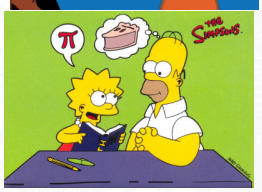
Why? Because π exact value is impossible to be calculated.











The ratio between the circumference and diameter of a circle is one of the most famous numbers in all of mathematics, π. Irrational, transcendental, and never-ending, it has frustrated and fascinated mathematicians across cultures for centuries, and it has become a universally recognized symbol for the field in popular culture.

One of the earliest formulae for calculating pi was proposed by English mathematician John Wallis in 1656. While calculating an integral in an attempt to find the area of a circle with a radius of one, he established a formula involving the multiplication of an infinite series of fractions that was based on the value of one-half of π. Later that century, Sir Isaac Newton used his binomial theorem to quickly calculate the value of pi up to 16 decimal places. Next came the Gregory-Leibniz series, which made use of both infinite series and trigonometric functions to develop formulae for values of pi divided by four and six. in 1761, Johann Lambert proved that pi was irrational and Ferdinand von Lindemann concluded that it was transcendental soon after. It was also during this period that the use of the Greek letter pi was introduced by William Jones, in 1706, and that it was later popularized by Leonhard Euler, beginning in 1737. In the early twentieth century, indian mathematician Srinivasa Ramanujan developed an incredibly efficient formula for calculating π based on its reciprocal fraction, which was later incorporated into computer algorithms.

$2022 = 2 \cdot 3 \cdot 337$

Adding to 2022 its reverse (2202), we get a palindrome (4224).



2.1. Soluciones del segundo examen del Cuerpo Superior de Estadísticos del Estado de 2022. Parte Común.

Ejercicio 1

El marco muestral contiene 70 empresas dedicadas al comercio, agrupadas en 2 estratos de tamaño, medido por el número de asalariados.

(a) Calcule el tamaño muestral en cada estrato, necesario para estimar el total de asalariados, aplicando afijación óptima, con un error de muestreo relativo prefijado del 1,5 %, usando el estimador estratificado de expansión simple sin reemplazamiento. Para ello, se dispone de la siguiente información:

Estrato de tamaño	Tramo de asalariados	Tamaño de la población (N_h)	X= Total de asalariados	Desviación típica de X (S_h)
1	$[1, 9]$	50	250	1,1
2	$[10, 500]$	20	850	8,5
TOTAL		70	1100	

(b) La muestra, estratificada aleatoria simple sin reemplazamiento, obtenida en el apartado anterior se investiga y responden 4 empresas del estrato de tamaño 1 y 7 del tamaño 2; el resto de unidades investigadas se han negado a colaborar. Suponiendo que la falta de respuesta es aleatoria, se repondera. Calcule el estimador del total de la cifra de negocios (Y), usando el estimador estratificado de expansión simple sin reemplazamiento reponderado por falta de respuesta y su error de muestreo relativo o coeficiente de elevación



estimado. Los datos necesarios son los siguientes:

unidades	Estrato de tamaño h h=1	h=2
i	y_{1i}	y_{2i}
1	40	1700
2	70	2000
3	60	1800
4	30	1750
5		2500
6		1700
7		1800

Solución.

(a) Dado que el objetivo es calcular el tamaño muestral en cada estrato necesario para estimar el total de asalariados con un error de muestreo relativo prefijado del 1,5 %, usando el estimador estratificado de expansión simple sin reemplazamiento, se tiene que:

$$e_r = CV(\hat{X}) = \frac{\sqrt{V(\hat{X})}}{E(\hat{X})} \Rightarrow e_r^2 = \frac{V(\hat{X})}{X^2} \Rightarrow V(\hat{X}) = 0{,}015^2 \cdot 1100^2$$

La expresión de la varianza del total según la afijación óptima se obtiene del problema de minimización siguiente:

$$\left.\begin{array}{c} \text{mín}\, V(\hat{X}) \\ \sum_{h=1}^{L} c_h n_h = C \end{array}\right\}$$

Este problema se resuelve aplicando el método de los multiplicadores de



Lagrange, considerando la función lagrangiana siguiente:

$$\Phi = V(\hat{X}) + \lambda\left(\sum_{h=1}^{L} c_h n_h - C\right) = \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) + \lambda\left(\sum_{h=1}^{L} c_h n_h - C\right) =$$
$$= \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{L} N_h S_h^2 + \lambda\left(\sum_{h=1}^{L} c_h n_h - C\right)$$

Para obtener el tamaño muestral en cada estrato, resolviendo el problema de minimización con la función lagrangiana anterior, se tiene que el valor de n_h para afijación óptima en el caso del total es:

$$n_h = n \cdot \frac{N_h S_h/\sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h/\sqrt{c_h}}$$

Por lo tanto, sustituyendo el valor de n_h en la expresión de la varianza, obtenemos la expresión:

$$V(\hat{X}) = \frac{1}{n}\left(\frac{\sum_{h=1}^{L} N_h S_h}{\sqrt{c_h}}\right)\left(\sum_{h=1}^{L} N_h S_h \sqrt{c_h}\right) - \sum_{h=1}^{L} N_h S_h^2$$

En el caso que nos ocupa, hay dos estratos por lo que la expresión anterior se puede escribir como:

$$V(\hat{X}) = \frac{1}{n}\left(\frac{N_1 S_1}{\sqrt{c_1}} + \frac{N_2 S_2}{\sqrt{c_2}}\right)(N_1 S_1 \sqrt{c_1} + N_2 S_2 \sqrt{c_2}) - N_1 S_1^2 - N_2 S_2^2$$

Volviendo a la igualdad en la que se prefija el error de muestreo relativo y sustituyendo con la expresión de la varianza se tiene:

$$\frac{1}{n}\left(\frac{N_1 S_1}{\sqrt{c_1}} + \frac{N_2 S_2}{\sqrt{c_2}}\right)(N_1 S_1 \sqrt{c_1} + N_2 S_2 \sqrt{c_2}) - N_1 S_1^2 - N_2 S_2^2 = 0{,}015^2 \cdot 1100^2$$

Sustituyendo con los valores del enunciado,

$$\frac{1}{n}\left(50 \cdot 1{,}1/\sqrt{c_1} + 20 \cdot 8{,}5/\sqrt{c_2}\right)\left(50 \cdot 1{,}1\sqrt{c_1} + 20 \cdot 8{,}5\sqrt{c_2}\right) - 50 \cdot 1{,}1^2 - 20 \cdot 8{,}5^2 =$$
$$= 0{,}015^2 \cdot 1100^2$$

$$n = \frac{\left(50 \cdot 1{,}1/\sqrt{c_1} + 20 \cdot 8{,}5/\sqrt{c_2}\right)\left(50 \cdot 1{,}1\sqrt{c_1} + 20 \cdot 8{,}5\sqrt{c_2}\right)}{0{,}015^2 \cdot 1100^2 + 50 \cdot 1{,}1^2 + 20 \cdot 8{,}5^2}$$



Dado que no se tiene información sobre los costes en cada estrato, se supone $c_h = k$, para $h = 1, 2$, es decir coste constante en todos los estratos, y se tiene:

$$n = \frac{(50 \cdot 1{,}1 + 20 \cdot 8{,}5)\,(50 \cdot 1{,}1 + 20 \cdot 8{,}5)}{0{,}015^2 \cdot 1100^2 + 50 \cdot 1{,}1^2 + 20 \cdot 8{,}5^2} = 28{,}477$$

De nuevo, como en el caso que nos ocupa no se tiene información del coste por unidad de muestreo en cada estrato, se supone constante en todos los estratos ($c_h = k$ para $h = 1, 2$). Entonces, la afijación óptima coincide con la de mínima varianza y la expresión del tamaño muestral de cada estrato será:

$$n_1 = n \cdot \frac{N_1 S_1}{N_1 S_1 + N_2 S_2} = 28{,}477 \cdot \frac{50 \cdot 1{,}1}{50 \cdot 1{,}1 + 20 \cdot 8{,}5} = 6{,}961$$

$$n_2 = n \cdot \frac{N_2 S_2}{N_1 S_1 + N_2 S_2} = 28{,}477 \cdot \frac{20 \cdot 8{,}5}{50 \cdot 1{,}1 + 20 \cdot 8{,}5} = 21{,}515$$

Por tanto, el tamaño muestral en cada estrato, necesario para estimar el total de asalariados, aplicando afición óptima, con un error de muestreo relativo prefijado del 1,5 % es, respectivamente, $n_1 = 7$ y $n_2 = 22$. Redondeamos al alza para estar seguros de que cumplimos el criterio de afijación. Sin embargo, el tamaño de la población en el estrato 2 es 20 por lo que no se puede tomar un número de empresas mayor a 20. Ante esta circunstancia, seleccionamos para la muestra las 20 unidades del segundo estrato siendo estas autorrepresentadas y así, $n_1 = 9$ y $n_2 = 20$.

(b) Uno de los posibles tratamientos estadísticos de la falta de respuesta es la reponderación, que consiste en ajustar o calibrar los pesos o factores de elevación inicales para compensar la falta de respuesta. Sea el estimador de expansión teórico para el total de Y en el estrato h

$$\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi},$$

supongamos que de las n_h unidades muestrales solo responden n_h^e, entonces



el estimador reponderado será:

$$\hat{Y}_h^r = \frac{N_h}{n_h}\frac{n_h}{n_h^e}\sum_{i=1}^{n_h^e} y_{hi} = \frac{N_h}{n_h^e}\sum_{i=1}^{n_h^e} y_{hi}.$$

Al ser la falta de respuesta aleatoria, el estimador ajustado sigue siendo insesgado. A partir de los datos del problema, la estimación del del total de la cifra de negocios, Y, usando el estimador estratificado de expansión simple sin reemplazamineto reponderado por la falta de respuesta es la siguiente:

$$\hat{Y}^r = \hat{Y}_1^r + \hat{Y}_2^r = \frac{50}{4}(40 + 70 + \cdots + 30) + \frac{20}{7}(1700 + 2000 + \cdots + 18\,000) =$$
$$= \frac{50}{4}200 + \frac{20}{7}13\,250 = 40\,357{,}14.$$

El error de muestreo relativo para el estimador del total de Y^r viene dado por la expresión:

$$CV(\hat{Y}^r) = \frac{\sqrt{V\left(\hat{Y}^r\right)}}{E(\hat{Y}^r)} = \frac{\sqrt{V\left(\hat{Y}^r\right)}}{Y^r}$$

En este caso estamos aplicando que el estimador del total es insesgado, y lo que nos piden es el error relativo de muestreo estimado, es decir:

$$\widehat{CV}(\hat{Y}^r) = \frac{\sqrt{\hat{V}\left(\hat{Y}^r\right)}}{\hat{Y}^r}.$$

La varianza estimada viene dada, una vez calculadas las cuasivarianzas muestrales de cada estrato, por:

$$\hat{V}(\hat{Y}^r) = \sum_{h=1}^{L}\hat{V}(\hat{Y}_h) = \sum_{h=1}^{L} N_h^2\left(1 - \frac{n_h^e}{N_h}\right)\frac{\hat{s}_h^2}{n_h^e} =$$
$$= 50^2\left(1 - \frac{4}{50}\right)\frac{333{,}3333}{4} + 20^2\left(1 - \frac{7}{20}\right)\frac{82\,023{,}8095}{7} =$$
$$= 191\,666{,}667 + 3\,046\,598{,}639 = 3\,238\,265{,}306$$

Así, el error de muestreo relativo es:

$$CV(\hat{Y}^r) = \frac{\sqrt{3\,238\,265{,}306}}{40\,357{,}14} = 0{,}0446$$



El error relativo tiene un valor de $4,46\,\%$, que es superior al prefijado en el cálculo del tamaño muestral, que fuer de un $1,5\,\%$. La falta de respuesta aleatoria aunque no incorpore sesgo incrementa el error de muestreo al reducir los tamaños muestrales.

Ejercicio 2

Proporcione un estimador puntual del parámetro $\theta > 0$ a partir de una observación X con función de densidad

$$f(x|\theta) = \begin{cases} 1/\theta, & \text{si } 0 < x < \theta \\ 0, & \text{en otro caso} \end{cases}$$

siendo la distribución a priori

$$g(\theta) = \begin{cases} \theta \exp(-\theta), & \text{si } \theta > 0 \\ 0, & \text{en otro caso} \end{cases}$$

cuando se emplea como función de pérdida

$$l(\theta, a) = 5(\theta - a)^2, a > 0.$$

Aplíquese al caso en que la observación es 6,1.

Solución.

Es conocido el resultado consistente en que la media de la distribución a posteriori es el estimador Bayes con respecto a la función de pérdida cuadrática.

Recordemos que un estimador de Bayes de θ con respecto a la función de



pérdida $l(\hat{\theta}, \theta)$ minimiza el error esperado con respecto a la distribución a posteriori $f(\theta|x)$, es decir, minimiza:

$$E\left[l(\hat{\theta}, \theta)|x\right] = \int_{-\infty}^{\infty} l(\hat{\theta}, \theta) f(\theta|x) d\theta.$$

Dicho resultado, para el caso absolutamente continuo puede comprobarse fácilmente de la siguiente forma:

$$E\left[l(\hat{\theta}, \theta)|x\right] = \int_{-\infty}^{\infty} l(\hat{\theta}, \theta) f(\theta|x) d\theta = \int_{-\infty}^{\infty} (\hat{\theta} - \theta)^2 f(\theta|x) d\theta$$

Derivando respecto al estimador e igualando la derivada a 0 obtenemos:

$$2\int_{-\infty}^{\infty} (\hat{\theta} - \theta) f(\theta|x) d\theta = 0 \Leftrightarrow \hat{\theta} - \int_{-\infty}^{\infty} \theta f(\theta|x) d\theta = 0$$

y esto nos lleva a que

$$\hat{\theta} = \int_{-\infty}^{\infty} \theta f(\theta|x) d\theta = E(\theta|x).$$

En nuestro caso la función de pérdida es cuadrática y aparece una constante (el número 5) multiplicando, lo cuál no modifica las consideraciones anteriores y por lo tanto se pueden utilizar esos resultados.

Calculemos pues la distribución a posteriori utilizando el dato de que sólo tenemos una observación de la variable aleatoria X y por tanto su función de verosimilitud será $f(x_1, \ldots, x_n|\theta) = f(x|\theta) = \frac{1}{\theta}$ con $0 < x < \theta$:

$$f(\theta|x) = \frac{g(\theta) f(x|\theta)}{\int_{-\infty}^{\infty} g(\theta) f(x|\theta) d\theta} = \frac{\theta e^{-\theta} \frac{1}{\theta}}{\int_{x}^{\infty} g(\theta) f(x|\theta) d\theta} = \frac{e^{-\theta}}{\int_{x}^{\infty} e^{-\theta} d\theta} =$$
$$= \frac{e^{-\theta}}{\left[-e^{-\theta}\right]_{x}^{\infty}} = e^{x-\theta}$$

siempre que $\theta > x$.



Y una vez conocida esta distribución, calculamos su media, que será el estimador Bayes que nos piden, para lo cual se integra por partes.

$$E\left[l(\theta)|x)\right] = \int_x^\infty \theta e^{x-\theta} d\theta = e^x \int_x^\infty \theta e^{-\theta} d\theta = e^x \left[e^{-\theta}(-\theta - 1)\right]_x^\infty =$$
$$= e^x \left[-\frac{\theta+1}{e^\theta}\right]_x^\infty = e^x \left[0 - \left(-\frac{x+1}{e^x}\right)\right] = x + 1$$

Luego tenemos que $\hat{\theta}_{Bayes} = x + 1$, siendo x la única observación de la que disponemos según el enunciado. Si la única observación que tenemos es 6,1 el estimador bayes sería 7,1.

Ejercicio 3

Consideramos un sistema de representación de números enteros en complemento a dos, con 8 bits. Llamemos Entero8bits a este tipo de datos.

(a) ¿Cuántos enteros distintos pueden representarse en este sistema? ¿Cuál es el entero mínimo y el máximo que se pueden representar en dicho sistema?

(b) Represente, en este sistema, las siguientes cantidades enteras:

- 100
- -100

(c) Observe este fragmento de programa en seudocódigo:

```
n: Entero8bits (1)
n = 100        (2)
print(n)       (3)
n = n*2        (4)
print(n)       (5)
```



Es decir, paso a paso, este código expresa lo siguiente:

(1) Se declara la variable n como un entero representado internamente como un Entero8bits.

(2) Se asigna la cantidad 100 a dicha variable n.

(3) Se imprime en efecto el valor 100 asignado.

(4) Se duplica el valor de dicha variable, almacenando dicho valor duplicado en la misma.

(5) Se imprime en efecto el nuevo valor.

Al ejecutarlo, esperamos ver en el output estándar, primero, la cantidad 100, y luego la cantidad 200, pero no es así: vemos la cantidad 100 en primer lugar, pero en segundo, un número negativo: "menos cincuenta y tantos". ¿Qué número es éste? ¿A qué se debe este resultado?

Solución.

a) En complemento a dos, con un número de bits igual a 8, es decir, $n = 8$, se pueden representar $2^n = 2^8$ enteros, desde $-2^{n-1} = -2^7$ hasta $2^{n-1} - 1 = 2^7 - 1$, o sea, el rango $-128, -127, \ldots, +126, +127$.

b) Tenemos que $100 = 64 + 32 + 4 = 2^6 + 2^5 + 2^2 = 01100100$.

Para representar el número -100 seguimos el siguiente algoritmo:
(1) $100 = 01100100$
(2) Cambimos los ceros por unos y viceversa, quedando: 10011011
(3) Sumamos 1 al resultado anterior: $10011100 = -100$

c) Como hemos comentado anteriormente, dado que el mayor número que se puede representar con 8 bits en complemento a dos es el 127, en este



caso el número 200 excede del rango con el que es posible trabajar. De hecho para poder representar el 200 en complemento a dos sería de la siguiente forma:

$200 = 128 + 64 + 8 = 011001000$, siendo el primer cero de la representación imprescindible y necesitando por tanto para ello 9 bits y no 8 como tenemos. Sí únicamente nos quedáramos con los 8 primeros bits de esta representación para cumplir con la restricción que nos impone el problema resulta que sería:

$11001000 = -56$, que sería el "menos cincuenta y tantos"que menciona el enunciado.

Podemos comprobar qué número es 11001000 de la siguiente forma:

- Cambiamos los ceros por unos y viceversa, es decir 00110111
- Sumamos uno al resultado y obtenemos: $00111000 = 32 + 16 + 8 = 56$

El primer uno de la secuencia original nos indica que es un número negativo, y por tanto es el -56.

El programa imprime este número porque tiene en cuenta esos últimos 8 bits para hacerlo, al haber guardado la información como entero en complemento a dos de 8 bits. Lo que se ha producido en este caso es lo que se conoce como desbordamiento u overflow. El desbordamiento aritmético se produce cuando una operación aritmética intenta crear un valor numérico que está fuera del rango que puede representarse con un número dado de dígitos, ya sea mayor que el máximo o menor que el mínimo valor representable. El resultado más común de un desbordamiento es que se almacenan los dígitos representables menos significativos del resultado, que es lo que ha ocurrido en este caso como comentábamos. Esta situación puede dar lugar a resultados que

conducen a un comportamiento no deseado. En particular, si no se ha anticipado esta posibilidad, el desbordamiento puede comprometer la fiabilidad de un programa y su seguridad.

Ejercicio 4

El Producto Interior Bruto a precios de mercado (PIB) de una economía en el año t-1, medido a precios corrientes, se cifró en 525.000 unidades monetarias (u.m)

(a) Se sabe que, en el año t, el excedente bruto de explotación y renta mixta bruta generados en dicha economía se estimaron en 230.000 u.m.

Así mismo, la remuneración a recibir por los asalariados residentes en el conjunto de dicho año t se cifró en 275.000 u.m. (de ellas, 1.500 provenían de empleadores no residentes). Por otro lado, los empleadores residentes debían abonar 1900 u.m a asalariados no residentes en la economía nacional.

Dados además los datos que figuran en la siguiente tabla sobre impuestos devengados y subvenciones a pagar durante el año en dicha economía, determine el crecimiento económico de la misma en el año t respecto al año t-1 en términos nominales.

IVA	4100
Impuestos y derechos sobre las importaciones (excluido el IVA)	200
Otros impuestos sobre los productos	1700
Otros impuestos sobre las producción	1100
Impuesto corriente sobre la renta	5350
Impuestos sobre el capital	410
Subvenciones a las importaciones	50
Otras subvenciones a los productos	600
Otras subvenciones a la producción	1350

(b) Por otro lado, el índice de volumen encadenado del PIB de dicha economía en los años t-1 y t se ha estimado en 110,7 y 108,1, respectivamente. ¿Cuál fue el crecimiento real de dicha economía en el año t respecto a t-1 (es decir, en términos de volumen)? ¿qué podríamos decir sobre la evolución de los precios de los bienes y servicios producidos en dicha economía? ¿a qué tasa evolucionó el deflactor implícito del PIB?



Solución.

(a) Parece conveniente intentar interpretar, antes de proceder a realizar ningún cálculo, los datos de problema con los códigos del SEC2010 y algunas consideraciones sobre la remuneración de asalariados.

Los impuestos y subvenciones del enunciado serían:

- IVA, (4100): D.211
- Impuestos y derechos sobre las importaciones (excluido el IVA), (200): D.212
- Otros impuestos sobre los productos, (1700): D.214
- Otros impuestos sobre la producción, (1100): D.29
- Impuesto corriente sobre la renta, (5350): D.51
- Impuestos sobre el capital (entendemos que son netos), (410): D.91r-D.91p
- Subvenciones a las importaciones, (50): D.311
- Otras subvenciones a los productos, (600): D.319
- Otras subvenciones a la producción, (1350): D.39

En cuanto a la remuneración de asalariados(RA) conocemos lo siguiente:

RA(interior) = RA pagada por residentes a residentes + RA pagada por residentes a no residentes.

RA(nacional) = RA pagada por residentes a residentes + RA pagada por no residentes a residentes.

Como en el enunciado se nos dice que la remuneración a recibir por los asalariados residentes en el año t es 275 000 u.m. entendemos que esta es la nacional, es decir, RA(nacional) = 275 000 u.m. Como además, de

esta remuneración 1500 las habían pagado empleadores no residentes y los empleadores residentes tenían que pagar 1900 u.m. a asalariados no residentes en la economía nacional podemos calcular la interior así:

RA(interior) = (275 000 - 1500) + 1900 = 275 400 u.m., es decir, que la RA pagada por residentes a residentes = 273 500 u.m.

Para determinar el crecimiento económico de la economía en el año t respecto al año t-1 en términos nominales tenemos que calcular el PIB_t, puesto que como nos proporcionan el dato en el año t-1, es decir, como $PIB_{t-1} = 525\,000$ u.m., solo quedaría calcular la tasa de variación para obtener el resultado.

Calculemos pues el PIB_t utilizando la cuenta de explotación:

II. 1.1 Cuenta de explotación

Empleos	**Código**	**Operaciones y otros flujos**	
S.1		**y saldos contables**	**S.1**
Total de la economía			**Total de la economía**
	B.1b/B.1*b	Producto interior bruto	510.500
275.400	D.1	Remuneración de los asalariados	
7.100	D.2	Impuestos sobre la producción y las importaciones	
6.000	D.21	Impuestos sobre los productos	
4.100	D.211	Impuestos del tipo valor añadido (IVA)	
200	D.212	Impuestos y derechos sobre las importaciones, excluido el IVA	
1.700	D.214	Impuestos sobre los productos, excluidos el IVA y los impuestos sobre las importaciones	
1.100	D.29	Otros impuestos sobre la producción	
2.000	D.3	Subvenciones	
650	D.31	Subvenciones a los productos	
50	D.311	Subvenciones a las importaciones	
600	D.319	Otras subvenciones a los productos	
1.350	D.39	Otras subvenciones a la producción	
230.000	**B.2b + B.3b**	**Excedente de explotación bruto + Renta mixta bruta**	

Por lo tanto el $PIB_t = 510\,500$ u.m. puesto que se cumple que:



$PIB_t = D.1 + D.2 - D.3 + (B.2b + B.3b)$ y en la parte derecha de la igualdad todos los datos son conocidos, que son los que figuran en la columna de empleos de la cuenta anterior.

Luego tenemos que $PIB_{t-1} = 525\,000$ u.m. y $PIB_t = 510\,500$ u.m. en términos nominales y a precios de mercado, por lo que la tasa de variación será:

$$TV(PIB_{t-1}^t) = \frac{510500 - 525000}{525000} \cdot 100 = -2{,}7619\,\%$$

es decir, que el procucto interior bruto sufre una disminución en términos nominales de un 2,7619 %.

(b) Los índices de volumen se conocen también con el nombre de índices de cantidad y según el enunciado tenemos que $IQ(PIB_{t-1}) = 110,7$ y que $IQ(PIB_t) = 108,1$ son los índices de volumen encadenados.

También conocemos que se cumple la relación entre índices siguiente en tanto por uno: el índice de valor(IV) es el producto del de precios(IP) por el de cantidad o volumen(IQ), es decir, $IV = IP \cdot IQ$.

El crecimiento real en volumen del PIB entre t-1 y t es la tasa de variación entre los índices encadenados correspondientes, o sea:

$$TV_{volumen}(PIB_{t-1}^t) = \frac{108{,}1 - 110{,}7}{110{,}7} \cdot 100 = -2{,}3486\,\%$$

Puede comprobarse cómo en terminos reales el PIB no disminuye tanto como veíamos en términos nominales en el anterior apartado, y esto seguramente está causado por una disminución de los precios que tendremos que corroborar, es decir, que cuando la tasa del PIB en términos reales es mayor que en términos nominales es porque habrán caído los precios. Sería más correcto hablar no de los precios, sino del deflactor

del PIB, que no es el IPC, porque son conceptos diferentes. De modo resumido podríamos decir que el deflactor del PIB mide los precios de todos los bienes y servicios producidos, mientras que el IPC solo mide los precios de los bienes y servicios comprados por los consumidores, y por tanto vamos a utilizar en las expresiones no el IPC, sino el deflactor.

Nota: En algunos textos la notación PIB_{t-1}^{t} se entiende como el valor del PIB en el año t a precios del año $t-1$, pero la interpretación que nosotros le damos aquí es diferente, que se refiere a la tasa de variación o magnitud entre dos periodos y no deben confundirse. Podría resolverse el ejercicio en estos términos y con esta notación.

Calculemos pues la tasa del deflactor implícito del PIB (DefPIB). Atendiendo a la igualdad sobre índices comentada tendremos equivalentemente que:

$$\frac{TV_{valor}(PIB_{t-1}^{t})}{100}+1=\left(\frac{TV(DefPIB_{t-1}^{t})}{100}+1\right)\left(\frac{TV_{volumen}(PIB_{t-1}^{t})}{100}+1\right)$$

puesto que la tasa de variación y cada índice están íntimamente relacionados.

Así pues:

$$\frac{-2{,}7619}{100}+1=\left(\frac{TV(DefPIB_{t-1}^{t})}{100}+1\right)\left(\frac{-2{,}3486}{100}+1\right)$$

y despejando

$$TV(DefPIB_{t-1}^{t})=\left(\frac{-0{,}027619+1}{-0{,}023486+1}-1\right)\cdot 100=\left(\frac{0{,}972381}{0{,}976514}-1\right)\cdot 100=$$
$$=(0{,}9958-1)\cdot 100=-0{,}423\,\%$$

Luego el deflactor del PIB disminuyó un $0{,}423\,\%$ como era de esperar. No es una caída muy pronunciada porque las tasas de variación en términos nominales y reales no difieren mucho.



Ejercicio 5

De una población se conocen las siguientes tasas de mortalidad referidas al año 2019.

Grupos de edad	${}_nm_x$
Menos de 1 año	0,00431
De 1 a 4 años	0,00015
De 5 a 19 años	0,00013
De 20 a 34 años	0,00024
De 35 a 49 años	0,00087
De 50 a 64 años	0,00408
De 65 a 79 años	0,01557
De 80 a 94 años	0,10733
De 95 años y más	0,29376

Calcule las siguientes series:

(a) Probabilidades de muerte, ${}_nq_x$ (suponga $a_0 = 0,1$ y ${}_na_x = 0,5$ para los demás casos).

(b) Supervivientes, l_x, y defunciones de la tabla, d_x.

(c) Población estacionaria, L_x (suponga $a_0 = 0,1$ y ${}_na_x = 0,5$ para los demás casos).

(d) Tiempo vivido, ${}_nT_x$.

(e) ¿Cuál fue la esperanza de vida al nacer de esta población en 2019?

Solución.

La contestación a los diferentes apartados se basa en la siguiente tabla que vamos a explicar.



Grupo de edad	mx	qx	ax	n	lx	dx	Lx	Tx	ex
Menos de 1	0,00431	0,004293346	0,1	1	100000	429,334611	99613,59885	8237126,141	82,37126141
De 1 a 4	0,00015	0,00059982	0,5	4	99570,66539	59,72448189	398163,2126	8137512,542	81,72600344
De 5 a 19	0,00013	0,001948101	0,5	15	99510,94091	193,8573239	1491210,184	7739349,33	77,7738534
De 20 a 34	0,00024	0,003593532	0,5	15	99317,08358	356,8990826	1487079,511	6248139,146	62,91102115
De 35 a 49	0,00087	0,012965401	0,5	15	98960,1845	1283,058451	1474779,829	4761059,636	48,11086054
De 50 a 64	0,00408	0,059382884	0,5	15	97677,12605	5800,349422	1421654,27	3286279,806	33,64431305
De 65 a 79	0,01557	0,209128965	0,5	15	91876,77663	19214,09521	1234045,935	1864625,536	20,29485148
De 80 a 94	0,10733	0,891951412	0,5	15	72662,68141	64811,58129	603853,3615	630579,601	8,678176868
De 95 en adelante	0,29376	1	3,404139434		7851,100122	7851,100122	26726,23952	26726,23952	3,404139434

Figura 2.1: Tabla de mortalidad abreviada de 2019.

El enunciado proporciona la primera columna, que son las tasas de mortalidad en tanto por uno. Dichas tasas están relacionadas con las probabilidades de morir, es decir, con ${}_nq_x$, mediante la fórmula ${}_nq_x = \frac{n\cdot {}_nm_x}{1+(1-a_x)n\cdot {}_nm_x}$. Con esta fórmula calculamos dichas probabilidades y obtenemos la segunda columna.

Hay que especificar que $q_{95} = 1$, puesto que en el último intervalo es de suponer que todos los supervivientes mueren, y de hecho así suelen publicarse las tablas de mortalidad. Esto hace que con los datos de los que se dispone, para que la tabla sea coherente, podríamos pensar en calcular el último valor de a_x, es decir, a_{95}, despejando de la fórmula anterior, con $a_x = 1 + \frac{{}_nq_x - n\cdot {}_nm_x}{n\cdot {}_nm_x\cdot {}_nq_x}$, que siendo $q_{95} = 1$ queda que $a_{95} = \frac{1}{n\cdot {}_nm_x} = \frac{1}{n\cdot 0{,}2938}$, pero para esto hay que suponer una longitud del último intervalo que no tenemos. Si consideráramos dicha longitud igual a 15 por mantener desde los 5 años las mismas amplitudes, nos estaríamos dejando fuera de la tabla de mortalidad a todas las personas de más de 110 años, que en España es muy pequeño y aportaría un orden de magnitud al numerador añadido a L_{95+} muy bajo en relación a considerar toda la población de ese tramo. Lógicamente es probable que este último intervalo sea más largo y cambiarían un poco los resultados, pero podríamos entender como razonable esta decisión de longitud en el último



tramo. Con este escenario, si lo calculamos obtendríamos que el valor de $a_{95} = 0{,}226942$, al ser el intervalo de longitud 15 significa que de media, el número de años que viven los supervivientes a los 95 años son 3,4041 años aproximadamente.

Lo más usual es que este valor del último a_w se pueda calcular directamente conociendo las personas que mueren de más de 95 años, puesto que se sabe cuánto ha vivido cada una y por lo tanto es factible no tener que utilizar estas fórmulas para ello, y esa parece ser la práctica habitual en diversos institutos de estadística. Puesto que en nuestro ejercicio no tenemos esta información, y para no tener que hacer suposiciones sobre la longitud del último intervalo podemos utilizar la fórmula ${}_nL_x = \frac{{}_nl_x}{{}_nm_x}$ para obtener la población estacionaria, que en breve detallamos.

Las siguientes columnas se calculan de la forma usual. La columna de supervivientes, ${}_nl_x$ comienza en 100 000 y se les van restando las defunciones teóricas ${}_nd_x$ que se calculan como ${}_nd_x = {}_nq_x \cdot {}_nl_x$, de forma que ${}_nl_{x+n} = {}_nl_x - {}_nd_x$. En el último intervalo coinciden el número de supervivientes con las defunciones teóricas como era de esperar.

La población estacionaria ${}_nL_x$ se calcula de abajo hacia arriba, siendo la fórmula general ${}_nL_x = {}_nl_{x+n} \cdot n + {}_nd_x \cdot n \cdot a_x$, siendo para el último intervalo el número de supervivientes a la siguiente edad cero. Para el último intervalo, como hemos dicho, consideramos la fórmula que relaciona los supervivientes con la tasa de mortalidad obtenemos $L_{95} = \frac{l_{95}}{m_{95}} = \frac{7851{,}10012}{0{,}29376} = 26\,726{,}2395$.

La serie ${}_nT_x$ tiene como último valor, el mismo que L_{95}, es decir, 26 726,2395 y en cualquier otra fila se obtiene sumando todas las ${}_nL_x$ desde ese intervalo hasta el final de la tabla, o sea, ${}_nT_x = \sum_x^{95} {}_nL_x$.

Por último la esperanza de vida en cada intervalo se obtiene de hacer

$$ {}_ne_x = \frac{{}_nT_x}{{}_nl_x} $$



Los apartados del problema quedarían así:

(a) Sería la tercera columna de la tabla.

(b) Serían la sexta y séptima columnas de la tabla.

(c) Correspondería a la antepenúltima columna.

(d) El tiempo vivido por los supervivientes a la edad x hasta el final de la generación sería la penúltima columna.

(e) La esperanza de vida al nacer en 2019 sería el primer valor de la última columna, es decir, ${}_1e_0 = 82{,}37126$.

Ejercicio 6

Sea una población de tamaño N, $U = u_1, u_2, ..., u_N$ y sean N valores $\pi_1, \pi_2, ..., \pi_N$, tal que $0 < \pi_k < 1$ para todo $k = 1, 2..., N$. Se generan un conjunto de N realizaciones independientes $\alpha_1, \alpha_2, ..., \alpha_N$ de una distribución uniforme sobre el intervalo $(0, 1)$.

El elemento u_k es seleccionado en la muestra 's' si $\alpha_k < \pi_k$, en caso contrario no se selecciona. Se pide:

(a) Calcular el valor esperado y varianza del tamaño muestral n_s.

(b) Calcular la probabilidad de que la muestra s tenga al menos 1 elemento.

(c) Sea una población formada por 10 elementos, para la cual se dispone de la información de una variable auxiliar x que toma los siguientes valores:



unidad	x
1	10
2	40
3	15
4	20
5	10
6	15
7	50
8	25
9	60
10	40

Se desea seleccionar una muestra de tamaño 5 mediante el método de selcción sistemática con probabilidades proporcionales al tamaño. Suponiendo que el arranque aleatorio seleccionado en el intervalo de muestreo es 35, determine las unidades que entran en la muestra.

Solución.

(a) Dado que el elemento u_k es seleccionado en la muestra 's' si $\alpha_k < \pi_k$, la probabilidad de u_k pertenezca a la muestra viene dada por:

$$P(u_k \in S) = P\left(\alpha_k < \pi_k\right) = P\left(\alpha_k \leq \pi_k\right) = F(\pi_k) = \pi_k$$

donde $F(\cdot)$ es la función de distribución uniforme en el intervalo $[0, 1]$, que al ser una variable aleatoria continua, la probabilidad de cualquier punto en dicho intervalo es cero.

Así, el tamaño muestral n_s es un variable aleatoria discreta que toma valores en $\{0, 1, ..., N\}$. Se define la función indicadora k-ésima, con-

sistente en que toma valor 1 si u_k pertenece a la muestra y 0 en caso contrario:

$$I_k = \begin{cases} 1, & \text{si } u_k \in S \\ 0, & \text{si } u_k \notin S \end{cases}$$

Entonces,

$$n_s = \sum_{k \in U} I_k \Rightarrow E\left[n_s\right] = E\left[\sum_{k \in U} I_k\right] = \sum_{k \in U} E\left(I_k\right) = \sum_{k \in U} \pi_k$$

En cuanto a la varianza

$$V\left(n_s\right) = V\left(\sum_{k \in U} I_k\right) = \sum_{k \in U} V\left(I_k\right)$$

ya que los elementos de la muestra son independientes entre sí, al serlo los u_k. Así, dado que I_k sigue un función de distribución de Bernoulli de parámetros $(1, \pi_k)$, es decir, $I_k \sim \mathcal{B}\left(1, \pi_k\right)$, la varianza de n_s es:

$$V\left(n_s\right) = \sum_{k \in U} \pi_k \left(1 - \pi_k\right)$$

(b) Teniendo en cuenta que $P\left(n_s = 0\right) = \left(1 - \pi_1\right)\left(1 - \pi_2\right)\left(1 - \pi_3\right) \cdots \left(1 - \pi_N\right)$, la probabilidad de que la muestra s tenga al menos 1 elemento, será

$$P\left(n \geq 1\right) = 1 - \prod_{k \in U} \left(1 - \pi_k\right)$$

(c) En el muestreo sistemático con probabilidades proporcionales al tamaño, la probabilidad de que una unidad j pertenezca a la muestra viene dada por

$$\pi_j = n \frac{x_j}{\sum_{j=1}^{10} x_j},$$

que es proporcional a la variable auxiliar x_j. Aquellas unidades cuya probabilidad de inclusión es mayor que 1, serán incluidas en la muestra con certeza y se recalcula la probabilidad de inclusión de las unidades restantes tras excluir las unidades con $\pi_j > 1$.



unidad	x	π_j
1	10	0,1754
2	40	0,7017
3	15	0,2631
4	20	0,3508
5	10	0,1754
6	15	0,2631
7	50	0,8771
8	25	0,4385
9	60	1,0526
10	40	0,7017

En este caso, la unidad 9 entra en la muestra pues $\pi_9 > 1$. Y se recalculan las probabilidades de inclusión mediante la fórmula

$$\pi_j = (n-1)\frac{x_j}{\sum_{j=1,j\neq 9}^{10} x_j},$$

donde ahora hemos quitado un elemento de la población, el 9, para rehacer los cálculos, ya que debe ser $\sum_{j=1}^{10} \pi_j = 5$.

Ahora tenemos que $\sum_{j=1,j\neq 9}^{10} x_j = 285 - 60 = 225$ y por ejemplo, $\pi_2 = 4 \cdot \frac{40}{225} = 0{,}7111$

INSTITUTO NACIONAL DE ESTADISTICA

unidad	x	π_j
1	10	0,1777
2	40	0,7111
3	15	0,2666
4	20	0,3555
5	10	0,1777
6	15	0,2666
7	50	0,8888
8	25	0,4444
10	40	0,7111

Ahora sí, se realiza el muestreo sistemático proporcional al tamaño. En este tipo de muestreo, la selección de las unidades que pertencen a la muestra se realiza usando el intervalo de muestreo, que se calcula dividiendo el total acumulado de la variable auxiliar x y dividiéndolo entre el número de elementos a seleccionar, que en nuestro caso, como la unidad 9 entra en muestra sólo faltan por seleccionar otras 4. En el caso que nos ocupa, el intervalo de muestreo es:

$$h = \frac{\sum_{j=1, j\neq 9}^{10} x_j}{n} = \frac{225}{4} = 56{,}25.$$

A continuación, se selecciona un punto aleatorio de arranque para seleccionar la primera unidad, que en este caso es $r = 35$. El segundo número aleatorio se genera añadiendo el valor del intervalo de muestro al punto aleatorio de arranque. Por tanto, el valor para seleccionar la segunda unidad será $r + k \cdot h = 35 + 1 \cdot 56{,}25 = 91{,}25$. Este proceso de añadir el intervalo de muestro al número anterior se va repitiendo y seleccionado las unidades correspondientes hasta que se alcanza el



tamaño de muestra.

$$\{r, r+h, r+2\cdot h, r+3\cdot h\} = \{35; 91{,}25; 147{,}5; 203{,}75\}$$

unidad	x	suma acumulada	intervalo de inclusión (proporcional al tamaño)	muestra si $r = 35$
1	10	10	0-10	
2	40	50	11-50	x
3	15	65	51-65	
4	20	85	66-85	
5	10	95	86-95	x
6	15	110	96-110	
7	50	160	111-160	x
8	25	185	161-185	
10	40	225	185-225	x

Por tanto, las unidades que pertenecen a la muestra S son las unidades 2, 5, 7, 9 y 10.

Ejercicio 7

En un estudio sobre una enfermedad, se eligen al azar 220 personas adultas que no padecen la misma; se comprueba que 60 de ellas no consumen alcohol. De entre 150 personas adultas con la enfermedad, también elegidas al azar, se observa que 15 de ellas no consumen alcohol.



Discuta a partir de qué valor del nivel de significación se rechazaría la hipótesis nula de que el consumo de alcohol no tiene influencia sobre la presencia de la enfermedad, si se dispone de un software con un comando de sintaxis *qnorm(p)* que proporciona, para una probabilidad p, el correspondiente cuantil de la distribución normal estándar.

Solución.

Sean las hipótesis nula y alternativa:

$$\begin{cases} H_0: & \text{El consumo de alcohol no influye en la presencia de la enfermedad} \\ H_1: & \text{El consumo de alcohol sí influye en la presencia de la enfermedad} \end{cases}$$

Se denota por E el grupo de personas adultas que están enfermas y NE el grupo de personas adultas que no están enfermas. Se denota por A el grupo de personas que consumen alcohol y NA el grupo de personas que no consumen alcohol. Teniendo en cuenta esta nomenclatura, se tiene:

- una muestra de NE formada por $n = 220$ personas, tales que:
  - en A: 160
  - en NA: 60
- una muestra de E formada por $m = 150$ personas, de las cuales:
  - en A: 135
  - en NA: 15

Así, la probabilidad de que una persona que no esté enferma consuma alcohol es $\hat{p}_{NE,A} = \frac{160}{220} = \frac{8}{11}$ y la probabilidad de que una persona que sí está enferma consuma alcohol es $\hat{p}_{E,A} = \frac{135}{150} = \frac{9}{10}$.



Teniendo en cuenta estas probabilidades, las hipótesis nula y alternativa se pueden reescribir como:

$$\begin{cases} H_0: & p_{E,A} = p_{NE,A} \Leftrightarrow p_{E,A} - p_{NE,A} = 0 \\ H_1: & p_{E,A} \neq p_{NE,A} \Leftrightarrow p_{E,A} - p_{NE,A} = 0 \end{cases}$$

El estadísitco de contraste es:

$$T = \frac{\hat{p}_{E,A} - \hat{p}_{NE,A} - (p_{E,A} - p_{NE,A})}{\sqrt{\frac{\hat{p}_o\hat{q}_o}{n} + \frac{\hat{p}_o\hat{q}_o}{m}}} \xrightarrow[n,m\to\infty]{} N(0,1),$$

donde $\hat{p}_o = \frac{n\hat{p}_{NE,A} + m\hat{p}_{E,A}}{n+m}$ y $\hat{q}_o = 1 - \hat{p}_o$.

Por tanto,

$$T = \frac{\frac{9}{10} - \frac{8}{11} - 0}{\sqrt{\frac{0{,}7973 \cdot 0{,}2027}{220} + \frac{0{,}7973 \cdot 0{,}2027}{150}}} = 4{,}0576,$$

donde $\hat{p}_o = \frac{220\frac{8}{11} + 150\frac{9}{10}}{220+150} = 0{,}7973$ y $\hat{q}_o = 1 - 0{,}797 = 0{,}2027$.

La función $qnorm(p)$ proporciona el valor de la variable que deja a su izquierda un área de p, es decir, el cuantil de orden p. Por tanto, tendríamos que tratar de averiguar, mediante algún procedimiento iterativo sencillo para qué valor de p se cumple que $qnorm(p, mean = 0, sd = 1) = 4{,}0576$, siendo el $p - valor = 2 \cdot (1 - p)$, al ser un contraste de dos colas.

Es bien sabido que para un nivel de significación $\alpha = 0{,}05$, el intervalo que se obtiene es $(-1{,}96, 1{,}96)$ y por lo tanto, con los datos que tenemos y el valor de estadístico de contraste, la hipótesis nula se rechazaría, al ser $T = 4{,}0576 > 1{,}96$, siendo el p-valor un valor muy pequeño.

Ejercicio 8

Con la información que se refleja en la tabla siguiente sobre las relaciones exteriores de la economía española en el año 2020, elabore las cuentas del resto del mundo referidas a dicho año.



Desde el año 2012, la economía española ha presentado capacidad de financiación respecto al resto del mundo. ¿Se mantuvo también esa tendencia en el año 2020?

Si dicho saldo se calculara como resultado de la cuenta corriente y de capital de la balanza de pagos, conforme al 6º Manual de Balanza de Pagos y Posición de Inversión Internacional, ¿la conclusión sería la misma?

Operaciones con el resto del mundo	2020
Exportaciones de bienes	265450
Importaciones de bienes (CIF)	281576
Importaciones de bienes (FOB)	274076
Exportaciones de servicios	78890
- De la cuales, gasto en consumo de no residentes en el territorio económico	14055
Importaciones de servicios	53978
- De las cuales, gasto en consumo de residentes fuera del territorio económico	61161
Impuestos menos subvenciones sobre la producción	-3346
Remuneración de asalariados residentes con empleador no residente	2615
Remuneración de asalariados no residentes con empleador residente	311
Rentas de la propiedad a recibir	43210
Rentas de la propiedad a pagar	46124
Impuestos corrientes sobre la renta y el patrimonio a recibir	1379
Impuestos corrientes sobre la renta y el patrimonio a pagar	1544
Cotizaciones sociales netas a recibir	71
Cotizaciones sociales netas a pagar	819
Prestaciones sociales distintas de transferencias sociales en especie a recibir	4055
Prestaciones sociales distintas de transferencias sociales en especie a pagar	563
Otras transferencias corrientes a recibir	11921
Otras transferencias corrientes a pagar	26733
Transferencias de capital a recibir	4550
Transferencias de capital a pagar	636
Adquisiciones menos cesiones de activos no financieros no producidos	1220

Nota: en millones de euros.

Solución.

Las cuentas del resto del mundo registran las operaciones entre unidades residentes y no residentes. La sucesión de cuentas del resto del mundo sigue el mismo esquema general que la de los sectores institucionales: cuentas



corrientes, cuentas de acumulación y balances.

La primera cuenta de la sucesión completa de las cuentas del resto del mundo es la cuenta de intercambios exteriores de bienes y servicios. Las importaciones y exportaciones de bienes y servicios se registrarán por su valor en la frontera. Las importaciones y las exportaciones totales se valoran franco a bordo (fob) en la frontera aduanera del país exportador. Los servicios de transporte exterior y de seguros entre las fronteras del importador y el exportador no se incluyen en el valor de los bienes, sino que se registran como servicios.

Comenzamos calculando la cuenta de intercambios exteriores de bienes y servicios:

I. Cuenta de intercambios exterioriores de bienes y servicios

Empleos TOTAL	Código	Operaciones y otros flujos y saldos contables	Recursos TOTAL
344.340	P.6	Exportaciones de bienes y servicios	
265.450	P.61	Exportaciones de bienes	
78.890	P.62	Exportaciones de servicios	
	P.7	Importaciones de bienes y servicios	328.054
	P.71	Importaciones de bienes	274.076
	P.72	Importaciones de servicios	53.978
-16.286	B.11	**Saldo de intercambios exteriores de bienes y servicios**	

La segunda cuenta de la sucesión es la Cuenta exterior de rentas primarias y transferencias corrientes. Dado que se consideran las operaciones desde la perspectiva del resto del mundo, se recogen en recursos aquellas operaciones que implican crédito del resto del mundo a España (salida de divisas). Por el contrario, se recogen en empleos las operaciones que implican crédito de la economía española al resto del mundo.

Hay que reseñar que el dato del problema de los Impuestos menos subvenciones sobre la producción, que nos lo dan en negativo (-3346), no es correcto, debe ser positivo, como puede comprobarse en las cuentas nacionales españolas, analizando la cuenta de explotación y asignación de la renta



primaria. Puede ser debido a un error de transcripción, y aunque podríamos utilizar el dato en negativo, los resultados variarían y consideramos más oportuno usar el dato correcto.

II. Cuenta exterior de rentas primarias y transferencias corrientes

II. Cuenta de rentas primarias y de transferencias corrientes con el exterior			
Empleos	**Código**	**Operaciones y otros flujos**	**Recursos**
TOTAL		**y saldos contables**	**TOTAL**
	B.11	Saldo de intercambios exteriores de bienes y servicios	-16.286
2.615	D.1	Remuneración de los asalariados	311
	D.2 - D.3	Impuestos netos sobre la producción y las importaciones	-3.346
43.210	D.4	Rentas de la propiedad	46.124
1.379	D.5	Impuestos corrientes sobre la renta, el patrimonio, etc.	1.544
71	D.61	Cotizaciones sociales	819
4.055	D.62	Prestaciones sociales distintas de las transferencias sociales en especie	563
11.921	D.7	Otras transferencias corrientes	26.733
-6.789	**B.12**	**Saldo de las operaciones corrientes con el exterior**	

En tercer lugar, la cuenta de variaciones del patrimonio neto debidas al saldo de las operaciones corrientes con el exterior y a las transferencias de capital

III. 1.1 Cuenta de variaciones del patrimonio neto debidas al saldo de las operaciones corrientes con el exterior y a las transferencias de capital			
Var. de los activos	**Código**	**Operaciones y otros flujos**	**Var. de los pasivos y del patrimonio neto**
TOTAL		**y saldos contables**	**TOTAL**
	B.12	Saldo de las operaciones corrientes con el ext.	-6.789
	D.9	Transferencias de capital, a cobrar	636
	D.9	Transferencias de capital, a pagar	-4.550
-10.703	***B.10.1***	***Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital***	

Y por último, mostramos la cuenta de adquisiciones de activos no financieros. En esta cuenta hay que comentar que el dato de Adquisiciones menos cesiones de activos no producidos, que en el enunciado es positivo, 1220, debería ser negativo como se comprueba también en las cuentas españolas de



2020. Por este motivo lo utilizamos en variación de activos en positivo desde el punto de vista del resto del mundo.

III. 1.2 Cuenta de adquisiciones de activos no financieros

Var. de los activos	Código	Operaciones y otros flujos y saldos contables	Var. de los pasivos y del patrimonio neto
TOTAL			TOTAL
	B.10.1	*Variaciones del patrimonio neto debidas al saldo de las operaciones corrientes con el exterior y a las transferencias de capital*	-10.703
1.220	NP	Adquisiciones menos cesiones de activos no producidos	
-11.923	**B.9**	**Capacidad(+)/Necesidad(-) de financiación**	

Dado que el saldo es negativo, la cuenta refleja la necesidad de financiación del resto del mundo. Es decir, que se mantiene la tendencia de la capacidad de financiación de la economía española en el año 2020.

Las cuentas del sector del resto del mundo recogidas en el SEC son diferentes de las cuentas internacionales que figuran en el 6º Manual de Balanza de Pagos y Posición de Inversión Internacional. Las cuentas internacionales muestran las operaciones entre una economía nacional y las economías extranjeras desde el punto de vista de la economía nacional. De este modo, las importaciones se indican como un empleo (un débito) y las exportaciones como un recurso (un crédito). La segunda gran diferencia entre las cuentas internacionales de la balanza de pagos y las cuentas del sector del resto del mundo en el SEC es el uso de categorías funcionales en las cuentas internacionales, en vez de instrumentos en la nomenclatura de las operaciones financieras del SEC.

Sin embargo, a pesar de estas diferencias la capacidad o la necesidad de financiación del 6º Manual de Balanza de Pagos y Posición de Inversión Internacional tiene el mismo que la partida de las cuentas nacionales para el resto del mundo pero con el signo opuesto, de forma que la conclusión sería la misma pues la capidad o necesidad de financición que proporciona el 6º



Manual de Balanza de Pagos y Posición lo hace desde el punto de vista de la economía nacional.

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA



2.2. Soluciones del cuarto examen del Cuerpo Superior de Estadísticos del Estado de 2022. Especialidad ciencia de datos.

Ejercicio 1

Se quiere estimar el total de una característica Y en una población de tamaño N. Para ello, se tiene información de dos variables auxiliares X_1, total poblacional de la variable X_1 y X_2, total poblacional de la variable X_2. Se pide:

A) Sea el estimador sintético para el total de Y dado por la siguiente expresión:

$$\hat{Y}_R = \alpha \hat{Y}_{R1} + \beta \hat{Y}_{R2}$$

donde $\alpha + \beta = 1$ e $\hat{Y}_{Ri}$ es el estimador del total de Y por el método de la razón usando la variable auxiliar X_i, $i = 1, 2$.

(a) ¿Es el estimador $\hat{Y}_R$ insesgado?

(b) Calcule el valor de α que minimice la varianza de $\hat{Y}_R$

B) Se quiere estimar la superficie media de 100 explotaciones agrarias de una comarca en 2023, media en hectáreas, que se denota por $\overline{Y} = Y/N$. De censos anteriores se conoce la superficie media en 2020 ($\hat{X}_1 = 1100$) y 2010 ($\hat{X}_2 = 900$).

Para ello se extrae una muestra aleatoria simple sin reemplazamiento de tamaño $n = 40$. Los datos disponibles son:

- Estimación de las medias usando un estimador de expansión simple:

$$\hat{\overline{Y}} = 1050; \quad \hat{\overline{X}}_1 = 990; \quad \hat{\overline{X}}_2 = 910;$$

- Cuasidesviaciones muestrales:

$$s_y = 975; \quad s_{x1} = 901; \quad s_{x2} = 820;$$
$$s_{x1y} = 905; \quad s_{x2y} = 850; \quad s_{x1x2} = 800;$$

(c) Calcule la estimación de la media de Y usando el estimador de razón $\hat{Y}_{R1}$ junto a su error de muestreo.

(d) Calcule la estimación de la media de Y usando el estimador de razón $\hat{Y}_{R2}$ junto a su error de muestreo.

(e) Calcule la estimación de la media de Y usando el estimador de razón $\hat{Y}_R$ cuando α toma el valor obtenido a partir de la expresión del apartado A.2), junto a su error de muestreo.

Solución.

A) Dado el estimador sintético para el total de Y, $\hat{Y}_R = \alpha\hat{Y}_{R1} + \beta\hat{Y}_{R2}$, con $\alpha + \beta = 1$:

(a) Se calcula el sesgo del mismo para ver si se trata de un estimador insesgado.

$$sesgo(\hat{Y}_R) = \alpha \cdot sesgo(\hat{Y}_{R1}) + \beta \cdot sesgo(\hat{Y}_{R2}),$$

donde

$$sesgo(\hat{Y}_{Ri}) = sesgo(\hat{R}_i \cdot X_i) = X_i \cdot sesgo(\hat{R}_i)$$

Comenzando por el cálculo de $sesgo(\hat{R}_i)$, se tiene:

$$cov\left(\hat{R}_i, \bar{x}_i\right) = E\left(\hat{R}_i\bar{x}_i\right) - E\left(\hat{R}_i\right)E\left(\bar{x}_i\right)$$



$$
\begin{aligned}
&= E\left(\frac{\bar{y}_i}{\bar{x}_i}\bar{x}_i\right) - E\left(\hat{R}_i\right)E\left(\bar{x}_i\right) \\
&= E\left(\bar{y}_i\right) - \bar{X}_i E\left(\hat{R}_i\right) \\
&= \bar{Y}_i - \bar{X}_i E\left(\hat{R}_i\right) \\
&= R_i\bar{X}_i - \bar{X}_i E\left(\hat{R}_i\right) \\
&= \bar{X}_i\left(R_i - E\left(\hat{R}_i\right)\right) \\
&= -\bar{X}_i B\left(\hat{R}_i\right).
\end{aligned}
$$

De modo que,

$$B\left(\hat{R}_i\right) = -\frac{cov\left(\hat{R}_i, \bar{x}_i\right)}{\bar{X}_i}.$$

Sustituyendo en la expresión obtenida anteriormente,

$$sesgo(\hat{Y}_{Ri}) = X_i \cdot sesgo(\hat{R}_i),$$

se tiene:

$$sesgo(\hat{Y}_{Ri}) = -X_i \cdot \frac{cov\left(\hat{R}_i, \bar{x}_i\right)}{\bar{X}_i} = -X_i \cdot \frac{cov\left(\hat{R}_i, \bar{x}_i\right)}{X_i/N} = -Ncov\left(\hat{R}_i, \bar{x}_i\right)$$

Y por tanto

$$sesgo(\hat{Y}_R) = -N\alpha \cdot cov\left(\hat{R}_1, \bar{x}_1\right) - N\beta \cdot cov\left(\hat{R}_2, \bar{x}_2\right)$$

$$sesgo(\hat{Y}_R) = -N\left(\alpha \cdot cov\left(\hat{R}_1, \bar{x}_1\right) + \beta \cdot cov\left(\hat{R}_2, \bar{x}_2\right)\right)$$

Por tanto, se puede concluir que no se trata de un estimador insesgado.

(b) Ahora se calcula el valor de α que minimiza la varianza del estimador $\hat{Y}_R$.

$$var(\hat{Y}_R) = \alpha^2 \cdot var(\hat{Y}_{R1}) + \beta^2 \cdot var(\hat{Y}_{R2}) + 2\alpha\beta \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right),$$

Y dicha varianza se minimiza mediante la siguiente función Lagrangiana.

$$\begin{aligned} \min_{\alpha} var(\hat{Y}_R) = \quad & \alpha^2 \cdot var(\hat{Y}_{R1}) + \beta^2 \cdot var(\hat{Y}_{R2}) + 2\alpha\beta \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) \\ \text{s.a.} \quad & \alpha + \beta = 1 \end{aligned}$$

Usando el método de los multiplicadores de Lagrangiana:

$$\Phi = \alpha^2 var(\hat{Y}_{R1}) + \beta^2 var(\hat{Y}_{R2}) + 2\alpha\beta \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + \lambda(1 - \alpha - \beta)$$

Resolviendo:

$$\begin{aligned} \frac{\partial \Phi}{\partial \alpha} &= 2\alpha \cdot var(\hat{Y}_{R1}) + 2\beta \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) - \lambda \\ \frac{\partial \Phi}{\partial \beta} &= 2\beta \cdot var(\hat{Y}_{R2}) + 2\alpha \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) - \lambda \end{aligned}$$

Para optimizar, se igualan las derivadas paraciales que se acaban de obterner a 0:

$$\begin{aligned} \frac{\partial \Phi}{\partial \alpha} = 0 \Leftrightarrow \lambda &= 2\alpha \cdot var(\hat{Y}_{R1}) + 2\beta \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) \\ \frac{\partial \Phi}{\partial \beta} = 0 \Leftrightarrow \lambda &= 2\beta \cdot var(\hat{Y}_{R1}) + 2\alpha \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) \end{aligned}$$

De donde se tiene que

$$2\alpha \cdot var(\hat{Y}_{R1}) + 2\beta \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) = 2\beta \cdot var(\hat{Y}_{R2}) + 2\alpha \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right).$$

Simplificando,

$$\alpha \cdot var(\hat{Y}_{R1}) + \beta \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) = \beta \cdot var(\hat{Y}_{R2}) + \alpha \cdot cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right).$$

Sacando factor común,

$$\alpha\left(var(\hat{Y}_{R1}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)\right) = \beta\left(var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)\right)$$



Como se tiene que $\alpha + \beta = 1$, entonces $\beta = 1 - \alpha$ y sustituyendo en la expresión anterior:

$$\alpha \left(var(\hat{Y}_{R1}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)\right) = (1-\alpha)\left(var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)\right)$$

$$\alpha \left[\left(var(\hat{Y}_{R1}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)\right) + \left(var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)\right)\right] = \\ = \left(var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)\right).$$

Despejando α se tiene:

$$\alpha = \frac{var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}{var(\hat{Y}_{R1}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}$$

$$\alpha = \frac{var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}{var(\hat{Y}_{R1}) - 2cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + var(\hat{Y}_{R2})}$$

Para obtener el valor de β se usa la restricción lineal $\beta = 1 - \alpha$, así:

$$\beta = 1 - \frac{var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}{var(\hat{Y}_{R1}) - 2cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + var(\hat{Y}_{R2})}$$

$$\beta = \frac{var(\hat{Y}_{R1}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}{var(\hat{Y}_{R1}) - 2cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + var(\hat{Y}_{R2})}$$

En conclusión, los valores de α y β que minimizan la varianza del estiamdor sintético para el total de Y son:

$$\alpha = \frac{var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}{var(\hat{Y}_{R1}) - 2cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + var(\hat{Y}_{R2})}$$

$$\beta = \frac{var(\hat{Y}_{R1}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}{var(\hat{Y}_{R1}) - 2cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + var(\hat{Y}_{R2})}$$

B) Ahora se estima la superficie media en hectáreas de las 100 explotaciones en el año 2023 según los requerimientos de cada apartado.

NOTA: En el enunciado habla de que se conocen las superficies medias de 2020 y 2010. A pesar de que se utiliza el símbolo del estimador del total, consideramos que se trata de una errata y que, efectivamente, se refiere al valor de la media.

c) La estimación de la media de Y usando el estimador de razón $\hat{Y}_{R1}$ será:

$$\hat{\bar{Y}} = \bar{X}_1 \cdot \hat{R}_1 = \bar{X}_1 \cdot \frac{\hat{Y}_{R1}}{X_1} = \frac{X_1}{N} \cdot \frac{\hat{Y}_{R1}}{X_1} = \frac{\hat{Y}_{R1}}{N} = \frac{1}{100}\hat{Y}_{R1} =$$
$$= \frac{1}{100}\hat{R}_1 X_1 = \frac{1}{100}\frac{\bar{y}}{\bar{x}_1}X_1 = \frac{1}{100}\frac{1050}{990}\bar{X}_1 \cdot 100 = \frac{1050}{990}1100 = 1166{,}66$$

Su error de muestreo será $\sigma\left(\hat{\bar{Y}}\right) = \sqrt{V\left(\hat{\bar{Y}}\right)}$. Se puede estimar mediante:

$$\hat{V}\left(\hat{\bar{Y}}\right) = \bar{X_1}^2 \cdot \hat{V}\left(\hat{R}_1\right) = \bar{X_1}^2 \frac{1}{\bar{X_1}^2}\left(\hat{V}\left(\bar{y}\right) + \hat{R}_1^2\hat{V}\left(\bar{x}_1\right) - 2\hat{R}_1 c\hat{o}v\left(\bar{x}_1, \bar{y}\right)\right)$$

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{(1-f)}{n}\left(s_y^2 + \hat{R}_1^2 s_{x1}^2 - 2\hat{R}_1 s_{x1y}\right)$$

Sustituyendo,

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{\left(1 - \frac{40}{100}\right)}{40}\left(975^2 + \frac{1050^2}{990^2}901^2 - 2\frac{1050}{990}905\right) = 27\,928{,}32$$

Y el error de muestreo estimado es:

$$\hat{\sigma}\left(\hat{\bar{Y}}\right) = \sqrt{27928{,}32} = 167{,}117$$

d) La estimación de la media de Y usando el estimador de razón $\hat{Y}_{R2}$ será:

$$\hat{\bar{Y}} = \bar{X}_2 \cdot \hat{R}_2 = \bar{X}_2 \cdot \frac{\hat{Y}_{R2}}{X_2} = \frac{X_2}{N} \cdot \frac{\hat{Y}_{R2}}{X_2} = \frac{\hat{Y}_{R2}}{N} = \frac{1}{100}\hat{Y}_{R2} =$$
$$= \frac{1}{100}\hat{R}_2 X_2 = \frac{1}{100}\frac{\bar{y}}{\bar{x}_2}X_2 = \frac{1}{100}\frac{1050}{910}\bar{X}_2 \cdot 100 = \frac{1050}{910}900 = 1038{,}46$$

Su error de muestreo será $\sigma\left(\hat{\bar{Y}}\right) = \sqrt{V\left(\hat{\bar{Y}}\right)}$. Con el mismo esquema que en el apartado anterior se estima mediante:

$$\hat{V}\left(\hat{\bar{Y}}\right) = \bar{X_2}^2 \cdot \hat{V}\left(\hat{R}_2\right) = \bar{X_2}^2 \frac{1}{\bar{X_2}^2}\left(\hat{V}\left(\bar{y}\right) + \hat{R}_2^2\hat{V}\left(\bar{x}_2\right) - 2\hat{R}_2 c\hat{o}v\left(\bar{x}_2, \bar{y}\right)\right)$$



$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{(1-f)}{n}\left(s_y^2 + \hat{R}_2^2 s_{x2}^2 - 2\hat{R}_2 s_{x2y}\right)$$

Sustituyendo,

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{\left(1-\frac{40}{100}\right)}{40}\left(975^2 + \frac{1050^2}{910^2}820^2 - 2\frac{1050}{910}850\right) = 27\,658{,}05$$

Y el error de muestreo es:

$$\hat{\sigma}\left(\hat{\bar{Y}}\right) = \sqrt{27658{,}05} = 166{,}307$$

e) Para obtener la estimación de la media de Y usando el estimador de razón $\hat{Y}_R$ se obtienen en primer lugar los valores de α y β.

$$\alpha = \frac{var(\hat{Y}_{R2}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}{var(\hat{Y}_{R1}) - 2cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + var(\hat{Y}_{R2})}$$

$$\beta = \frac{var(\hat{Y}_{R1}) - cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right)}{var(\hat{Y}_{R1}) - 2cov\left(\hat{Y}_{R1}, \hat{Y}_{R2}\right) + var(\hat{Y}_{R2})}$$

Donde:

$$\begin{aligned}
\widehat{var}(\hat{Y}_{R1}) &= \widehat{var}(\hat{R}_1 \cdot X_1) = X_1^2 \widehat{var}(\hat{R}_1) = \\
&= X_1^2 \frac{1}{\bar{X}_1^2}\frac{(1-f)}{n}\left(s_y^2 + \hat{R}_1^2 s_{x1}^2 - 2\hat{R}_1 s_{x1y}\right) = \\
&= N^2 \frac{(1-f)}{n}\left(s_y^2 + \hat{R}_1^2 s_{x1}^2 - 2\hat{R}_1 s_{x1y}\right) = \\
&= 100^2 \frac{\left(1-\frac{40}{100}\right)}{40}\left(975^2 + \frac{1050^2}{990^2}\cdot 901^2 - 2\cdot\frac{1050}{990}\cdot 905\right) = \\
&= 100^2 \cdot 27928{,}32 = 279\,283\,136{,}9
\end{aligned}$$

$$\begin{aligned}
\widehat{var}(\hat{Y}_{R2}) &= \widehat{var}(\hat{R}_2 \cdot X_2) = X_2^2 \widehat{var}(\hat{R}_2) = \\
&= X_2^2 \frac{1}{\bar{X}_2^2}\frac{(1-f)}{n}\left(s_y^2 + \hat{R}_2^2 s_{x2}^2 - 2\hat{R}_2 s_{x2y}\right) = \\
&= N^2 \frac{(1-f)}{n}\left(s_y^2 + \hat{R}_2^2 s_{x2}^2 - 2\hat{R}_2 s_{x2y}\right) =
\end{aligned}$$

$$= 100^2 \frac{\left(1 - \frac{40}{100}\right)}{40} \left(975^2 + \frac{1050^2}{910^2} \cdot 820^2 - 2 \cdot \frac{1050}{910} \cdot 850\right) =$$
$$= 100^2 \cdot 27658{,}05 = 276\,580\,584{,}3$$

$$\begin{aligned}
cov(\hat{Y}_{R1}, \hat{Y}_{R2}) &= E\left[\left(\hat{Y}_{R1} - Y\right)\left(\hat{Y}_{R2} - Y\right)\right] = \\
&= E\left[\left(\hat{R}_1 \cdot X_1 - Y\right)\left(\hat{R}_2 \cdot X_2 - Y\right)\right] = \\
&= E\left[\left(\frac{\bar{y}}{\bar{x}_1} \cdot \bar{X}_1 - \bar{Y}\right)\left(\frac{\bar{y}}{\bar{x}_2} \cdot \bar{X}_2 - \bar{Y}\right)\right] = \\
&= \bar{Y}^2 E\left[\left(\frac{\bar{y}}{\bar{x}_1} \cdot \bar{X}_1 \frac{1}{\bar{Y}} - 1\right)\left(\frac{\bar{y}}{\bar{x}_2} \cdot \bar{X}_2 \frac{1}{\bar{Y}} - 1\right)\right] = \\
&= \bar{Y}^2 E\left[\left(\frac{\bar{y}}{\bar{Y}} \cdot \frac{\bar{X}_1}{\bar{x}_1} - 1\right)\left(\frac{\bar{y}}{\bar{Y}} \cdot \frac{\bar{X}_2}{\bar{x}_2} - 1\right)\right] = \\
&= \bar{Y}^2 E\left[\left(\left(1 + \frac{\bar{y} - \bar{Y}}{\bar{Y}}\right)\left(\frac{1}{1 + \frac{\bar{x}_1 - \bar{X}_1}{\bar{X}_1}}\right) - 1\right)\right. \\
&\quad \left.\left(\left(1 + \frac{\bar{y} - \bar{Y}}{\bar{Y}}\right)\left(\frac{1}{1 + \frac{\bar{x}_2 - \bar{X}_2}{\bar{X}_2}}\right) - 1\right)\right]
\end{aligned}$$

Se denotan

$$e_1 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \qquad e_2 = \frac{\bar{x}_1 - \bar{X}_1}{\bar{X}_1} \qquad e_2' = \frac{\bar{x}_2 - \bar{X}_2}{\bar{X}_2}$$

y se puede reescribir la covarianza entre $\hat{Y}_{R1}$ y $\hat{Y}_{R2}$ como:

$$\begin{aligned}
cov(\hat{Y}_{R1}, \hat{Y}_{R2}) &= \bar{Y}^2 E\left[\left((1 + e_1)\left(\frac{1}{1 + e_2}\right) - 1\right)\left((1 + e_1)\left(\frac{1}{1 + e_2'}\right) - 1\right)\right] = \\
&= \bar{Y}^2 E\left[\left((1 + e_1)(1 + e_2)^{-1} - 1\right)\left((1 + e_1)(1 + e_2')^{-1} - 1\right)\right]
\end{aligned}$$

Si se desarrolla $(1 + e_2)$ y $(1 + e_2')$ en serie de Taylor y se retienen solo los términos de orden inferior o igual a dos en e_1, e_2, e_2' se llega a que

$$cov(\hat{Y}_{R1}, \hat{Y}_{R2}) \simeq \frac{1 - f}{n} Y^2 \left(\frac{s_y}{\bar{Y}} - \rho_2 \frac{s_y}{\bar{Y}} \frac{s_{x2}}{\bar{X}_2} - \rho_1 \frac{s_y}{\bar{Y}} \frac{s_{x1}}{\bar{X}_1} + \rho_{12} \frac{s_{x1}}{\bar{X}_1} \frac{s_{x2}}{\bar{X}_2}\right)$$

Y así se obtiene una aproximación de orden $\mathcal{O}(n^{-2})$ del valor de la covarianza entre $\hat{Y}_{R1}$ y $\hat{Y}_{R2}$, donde



- ρ_i: coeficiente de correlación entre y y x_i con $i = 1, 2$,
- ρ_{12}: coeficiente de correlación entre x_1 y x_2.

Es decir:

$$\rho_1 = \frac{s_{x1y}}{s_{x1}s_y} = \frac{905}{901 \cdot 975} = 0{,}0010301$$
$$\rho_2 = \frac{s_{x2y}}{s_{x2}s_y} = \frac{850}{820 \cdot 975} = 0{,}0010631$$
$$\rho_{12} = \frac{s_{x1x2}}{s_{x1}s_{x2}} = \frac{800}{901 \cdot 820} = 0{,}0010828$$

Sustituyendo en la aproximación obtenida se obtiene un estimador de la covarianza:

$$\begin{aligned}\widehat{cov}(\hat{Y}_{R1}, \hat{Y}_{R2}) =& \frac{1 - \frac{40}{100}}{40}(100 \cdot 1050)^2 \left(\frac{975}{1050} - 0{,}0010631\frac{975}{1050}\frac{820}{910} \right. \\ & \left. -0{,}0010301\frac{975}{1050}\frac{901}{990} + 0{,}0010828\frac{901}{990}\frac{820}{910} \right) = \\ =& \frac{3}{200}(105000)^2\,(0{,}92769) = \\ =& \, 153\,418\,260{,}5\end{aligned}$$

Sustituyendo los valores obtenidos en las ecuaciones de α y β se obtienen los valores estimados:

$$\alpha = \frac{276\,580\,584{,}3 - 153\,418\,214{,}9}{279\,283\,136{,}9 - 2 \cdot 153\,418\,214{,}9 + 276\,580\,584{,}3} = 0{,}494573582$$
$$\beta = \frac{279\,283\,136{,}9 - 153\,418\,214{,}9}{279\,283\,136{,}9 - 2 \cdot 153\,418\,214{,}9 + 276\,580\,584{,}3} = 0{,}505426418$$

Y se tiene que efectivamente α y β cumplen la siguiente restricción lineal: $\alpha + \beta = 1$.

Atendiendo a los resultados obtenidos, la estimación de la media de Y usando el estimador de razón $\hat{Y}_R$ es:

$$\hat{\bar{Y}} = \frac{\hat{Y}_R}{N} = \frac{\alpha\hat{Y}_{R1} + \beta\hat{Y}_{R2}}{N},$$



donde

$$\hat{Y}_{R1} = \hat{R}_1 X_1 = \frac{\bar{y}}{\bar{X}_1} N \bar{X}_1 = \frac{1050}{990} \cdot 100 \cdot 1100 = 116\,666{,}6666$$
$$\hat{Y}_{R2} = \hat{R}_2 X_2 = \frac{\bar{y}}{\bar{X}_2} N \bar{X}_2 = \frac{1050}{910} \cdot 100 \cdot 900 = 103\,846{,}1538$$

y así,

$$\hat{Y}_R = 0{,}494573782 \cdot 116\,666{,}6666 + 0{,}505426218 \cdot 103\,846{,}1538 = 110\,186{,}8408$$

En consecuencia,

$$\hat{\bar{Y}} = \frac{\hat{Y}_R}{N} = \frac{110186{,}8434}{100} = 1101{,}8684$$

En cuanto a su error de muestreo,

$$\begin{aligned}
\widehat{var}(\hat{\bar{Y}}) &= \frac{1}{N^2} \widehat{var}(\hat{Y}_R) = \\
&= \frac{1}{N^2} \left[\alpha^2 \widehat{var}(\hat{Y}_{R1}) + \beta^2 \widehat{var}(\hat{Y}_{R2}) + 2\alpha\beta \widehat{cov} \left(\hat{Y}_{R1}, \hat{Y}_{R2} \right) \right] = \\
&= \frac{1}{100^2} \left[0{,}494573782^2 \cdot 279\,283\,136{,}9 + 0{,}505426218^2 \cdot 276\,580\,584{,}3 + \right. \\
&\left. + 2 \cdot 0{,}494573782 \cdot 0{,}505426218 \cdot 153\,418\,214{,}9 \right] = \\
&= \frac{1}{100^2} \left[0{,}494573782^2 \cdot 279\,283\,136{,}9 + 0{,}505426218^2 \cdot 276\,580\,584{,}3 + \right. \\
&\left. + 2 \cdot 0{,}494573782 \cdot 0{,}505426218 \cdot 153\,418\,214{,}9 \right] = \\
&= \frac{1}{100^2} \cdot 215\,667\,705{,}4 \\
&= 21\,566{,}775\,27
\end{aligned}$$

y por tanto,

$$\hat{\sigma}(\hat{\bar{Y}}) = \sqrt{21\,566{,}770\,54} = 146{,}8563.$$

Ejercicio 2

Suponga que las observaciones (x, y) satisfacen el modelo

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j + \epsilon_i$$



donde $\epsilon_1, \ldots, \epsilon_n$ se distribuyen independientemente según una $N(0, \sigma^2)$.

(a) Escriba la verosimilitud de los datos.

(b) Suponga que a priori $(\beta_1, ..., \beta_p)$ son independientes e idénticamente distribuidas según una distribución de densidad $p(\beta) = \frac{1}{2b} \exp(-|\beta|/b)$, con parámetro b común. Argumente que el estimador lasso surge como moda de la correspondiente distribución a posteriori.

(c) Suponga, ahora, que a priori $(\beta_1, ..., \beta_p)$ son independientes e idénticamente distribuidas según una distribución normal de media 0 y varianza c. Argumente que el estimador de regresión *ridge* surge como moda de la correspondiente distribución a posteriori.

Solución.

1. (a) La función de verosimilitud de los datos puede escribirse de la siguiente forma:

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{x}, \beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right)$$

2. (b) La función de densidad de las β_i es una doble exponencial con media cero y parámetro de escala común b. Vamos a escribir la distribución a posteriori de β salvo una constante de proporcionalidad:

$$f(\beta|\boldsymbol{x}, \boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{y}|\boldsymbol{x}, \beta)p(\beta) = \left(\tfrac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^{n}\epsilon_i^2\right)\left[\tfrac{1}{2b}\exp\left(-\tfrac{|\beta|}{b}\right)\right]$$

Para mostrar que el estimador lasso es la moda para β bajo la distribución a posteriori necesitamos demostrar que el valor más probable de β se da cuando se cumple la solución lasso. Para lograrlo reorganizamos

los términos de la expresión anterior:

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{x},\beta)p(\beta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \epsilon_i^2\right)\left[\frac{1}{2b}\exp\left(-\frac{|\beta|}{b}\right)\right] =$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{2b}\right)\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n \epsilon_i^2 - \frac{|\beta|}{b}\right)$$

Vamos a trabajar con el logaritmo de esta expresión puesto que para maximizarla, al ser el logaritmo una función estrictamente creciente, los máximos coincidirán y es más sencillo el tratamiento utilizando logaritmos:

$$Ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{2b}\right)\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^n \epsilon_i^2 + \frac{|\beta|}{b}\right)$$

Por lo tanto, nuestro objetivo será:

$$\underset{\beta}{maximizar}\left\{Ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{2b}\right)\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^n \epsilon_i^2 + \frac{|\beta|}{b}\right)\right\}$$

lo cual es equivalente a minimizar la expresión:

$$\left(\frac{1}{2\sigma^2}\sum_{i=1}^n \epsilon_i^2 + \frac{|\beta|}{b}\right)$$

De esta forma argumentamos lo siguiente:

$$arg\ \underset{\beta}{min}\left(\frac{1}{2\sigma^2}\sum_{i=1}^n \epsilon_i^2 + \frac{|\beta|}{b}\right) = arg\ \underset{\beta}{min}\left(\frac{1}{2\sigma^2}\sum_{i=1}^n \epsilon_i^2 + \frac{1}{b}\sum_{j=1}^p |\beta_j|\right) =$$
$$= arg\ \underset{\beta}{min}\frac{1}{2\sigma^2}\left(\sum_{i=1}^n \epsilon_i^2 + \frac{2\sigma^2}{b}\sum_{j=1}^p |\beta_j|\right) = arg\ \underset{\beta}{min}\left(\sum_{i=1}^n \epsilon_i^2 + \lambda\sum_{j=1}^p |\beta_j|\right) =$$
$$= arg\ \underset{\beta}{min}\left(RSS + \lambda\sum_{j=1}^p |\beta_j|\right)$$

Pero minimizar esta expresión es precisamente el problema de optimización LASSO, tomando $\lambda = \frac{2\sigma^2}{b}$. Podemos recordar que precisamente en el método de mínimos cuadrados ordinarios elegimos el conjunto de β_j de tal forma que minimice la suma de los errores al cuadrado,



es decir, RSS. Si a este método añadimos el factor de penalización $\sum_{j=1}^{p} |\beta_j|$ lo que se obtiene es lo que se conoce como regresión LASSO. Por este motivo, si resolvieramos el problema de optimización LASSO obtendríamos los valores de β que maximizan la distribución a posteriori, es decir, que se puede entender como moda de la distribución a posteriori.

3. (c) Al igual que en el apartado anterior obtenemos la distribución a posteriori del siguiente producto: $f(\beta|\boldsymbol{x},\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{y}|\boldsymbol{x},\beta)p(\beta)$

   Vamos a calcular la distribución a priori primeramente:

$$p(\beta) = \prod_{i=1}^{p} p(\beta_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi c}} \exp\left(-\frac{\beta_i^2}{2c}\right) = \left(\frac{1}{\sqrt{2\pi c}}\right)^p \exp\left(-\frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right)$$

   De esta forma, la distribución a posteriori queda de esta forma:

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{x},\beta)p(\beta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2\right)\left(\frac{1}{\sqrt{2\pi c}}\right)^p \exp\left(-\frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) =$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2\pi c}}\right)^p \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 - \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right)$$

   Con el mismo razonamiento anterior, tomamos logaritmos:

$$Ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\frac{1}{\sqrt{2\pi c}}\right)^p\right] - \left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 + \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right)$$

   Para comprobar que el estimador ridge es la moda para β en la distribución a posteriori necesitamos demostrar que el valor más probable para β se encuentra como solución del siguiente problema de optimización:

$$arg\ \min_{\beta}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 + \frac{1}{2c}\sum_{i=1}^{p}\beta_i^2\right) = arg\ \min_{\beta}\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\epsilon_i^2 + \frac{2\sigma^2}{2c}\sum_{i=1}^{p}\beta_i^2\right) =$$
$$= arg\ \min_{\beta}\left(\sum_{i=1}^{n}\epsilon_i^2 + \lambda\sum_{i=1}^{p}\beta_i^2\right) = arg\ \min_{\beta}\left(RSS + \lambda\sum_{i=1}^{p}\beta_i^2\right)$$

   Pero resulta que esta formulación del problema es justo la de la regresión ridge con $\lambda = \frac{\sigma^2}{c}$. Al igual que en el apartado (b) el método de



mínimos cuadrados ordinarios elige las β_j de tal forma que se minimice RSS, y cuando se añade el factor de penalización $\sum_{j=1}^{p} |\beta_j|^2$ obtenemos la regresión ridge.

Ejercicio 3

Tenemos una base de datos con datos geográficos del mundo: contiene información sobre países, con su código correspondiente, la población total del país, la superficie y la renta per cápita. También contiene información sobre ciudades, con su nombre, un código, su población y el país al que pertenecen. Es posible que nuestra base de datos no contenga la población de alguna ciudad. En cambio, los demás datos (nombres de las ciudades y países, códigos, etc.) son obligatorios.

(a) Proporcione un ejemplo de un documento XML bien formado para este caso, e indique las reglas que cumple para que se considere un documento bien formado.

(b) Dé instrucciones SQL de creación de dichas tablas, sin olvidar vincularlas con las claves externas necesarias, en su caso.

(c) Elabore una consulta SQL para averiguar cuáles son las ciudades (5 máx.) más pobladas del país del mundo con la renta per cápita más alta registrado en nuestra base de datos.

(d) Elabore una consulta SQL para eliminar de nuestra base de datos el país de menor superficie. Esta operación debe evitar dejar inconsistencias en la base de datos con respecto a las ciudades del país eliminado.



Solución.

La base de datos tendrá un par de tablas:

`Pais`: la clave primaria es `cod_pai`.

`Ciudad`: la clave primaria es `cod_ciu`. Además, en `Ciudad` el campo `cod_pai` es clave foránea apuntando al `cod_pai` de `Pais`.

El esquema de las tablas con información adicional aparece en la Figura 2.2.

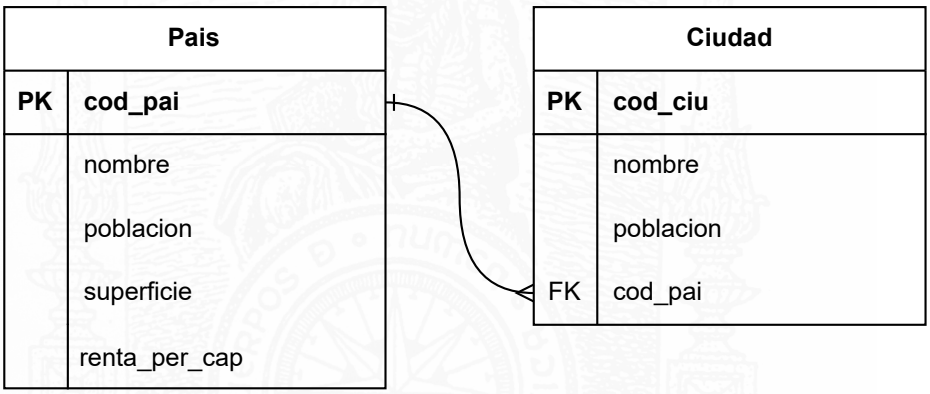


Figura 2.2: Esquema para la bbdd geográfica.

(a) Un posible `XML` de ejemplo con su DTD de reglas incrustado sería

```
<?xml version="1.0"?>
<!DOCTYPE mundo[
    <!ELEMENT mundo (ciudades, paises)>
    <!ELEMENT ciudades (ciudad*)>
    <!ELEMENT paises (pais*)>
    <!ELEMENT ciudad (nombre_c, poblacion_c?)>
    <!ELEMENT pais (nombre_p, poblacion_p, superficie, renta_per_cap)>
    <!ELEMENT cod_ciu (#PCDATA)>
```




```
        <!ELEMENT nombre_c (#PCDATA)>
        <!ELEMENT nombre_p (#PCDATA)>
        <!ELEMENT poblacion_c (#PCDATA)>
        <!ELEMENT poblacion_p (#PCDATA)>
        <!ELEMENT cod_pai (#PCDATA)>
        <!ELEMENT superficie (#PCDATA)>
        <!ELEMENT renta_per_cap (#PCDATA)>
        <!ATTLIST ciudad cod_ciu ID #REQUIRED>
        <!ATTLIST pais cod_pai ID #REQUIRED>
        <!ATTLIST ciudad cod_pai IDREF #REQUIRED>
]>
<mundo>
    <ciudades>
        <ciudad cod_ciu="C546" cod_pai="P13">
            <nombre_c>Talavera de la Reina</nombre_c>
            <poblacion_c>83009</poblacion_c>
        </ciudad>
        <ciudad cod_ciu="C591" cod_pai="P13">
            <nombre_c>Toro</nombre_c>
            <poblacion_c>8448</poblacion_c>
        </ciudad>
        <ciudad cod_ciu="C1257" cod_pai="P16">
            <nombre_c>Verona</nombre_c>
        </ciudad>^^I
    </ciudades>
    <paises>
        <pais cod_pai="P13">
            <nombre_p>Espa^^c3^^b1a</nombre_p>
            <poblacion_p>47423134</poblacion_p>
```




```
            <superficie>505990.1</superficie>
            <renta_per_cap>30103.51</renta_per_cap>
        </pais>
        <pais cod_pai="P16">
            <nombre_p>Italia</nombre_p>
            <poblacion_p>59113567</poblacion_p>
            <superficie>302073.3</superficie>
            <renta_per_cap>35657.50</renta_per_cap>
        </pais>
    </paises>
</mundo>
```

Está bien formado porque:

- Comienza por la cláusula que especifica la versión de `XML`.
- Todos los tags que se abren, se cierran posteriormente.
- El anidamiento de elementos es correcto: los tags se cierran en orden inverso al que se abren.
- Solo existe un elemento en la raíz.

Además, el fichero `XML` es validable puesto que se conforma a la estructura determinada por el DTD incrustado. Sigue la estrategia habitual para codificar bases de datos en cuatro niveles: base de datos, tabla, registro y campo. En particular, en el DTD se especifican los campos obligatorios y optativos de ambas tablas, se definen las claves como atributos (no elementos), y se garantiza la integridad referencial asociada a la clave ajena `cod_pai` de `ciudad` mediante un `IDREF`.

(b) Podemos crear una base de datos `Mundo` y construir sus tablas con el siguiente script `SQL`:




```sql
CREATE TABLE Mundo.Pais (
    cod_pai INT PRIMARY KEY,
    nombre VARCHAR (50) NOT NULL,
    poblacion INT NOT NULL,
    superficie FLOAT NOT NULL,
    renta_per_cap FLOAT NOT NULL
);

CREATE TABLE Mundo.Ciudad (
    cod_ciu INT PRIMARY KEY,
    nombre VARCHAR (50) NOT NULL,
    poblacion INT,
    cod_pai INT NOT NULL,
    FOREIGN KEY (cod_pai) REFERENCES mundo.Pais (cod_pai)
);
```

Estamos utilizando códigos de identificación enteros, como es habitual, porque resultan más eficientes.

(c) El SQL para limitar del número de registros no es estándar, depende del software. Usamos el LIMIT de MySQL.

Una primera versión con varias subconsultas:

```sql
SELECT C.nombre FROM Ciudad C
WHERE C.cod_pai = (
    SELECT cod_pai FROM Pais
    WHERE renta_per_cap = (
        SELECT MAX(renta_per_cap)
        FROM Pais
    )
```




```
)
ORDER BY C.poblacion DESC
LIMIT 5;
```

Otra query equivalente usando un join:

```
SELECT C.nombre
FROM Ciudad C
INNER JOIN Pais P on P.cod_pai = C.cod_pai and P.renta_per_cap = (
    SELECT MAX(renta_per_cap) FROM Pais)
ORDER BY C.poblacion DESC
LIMIT 5;
```

(d) Una query que obtiene el código del país con menor superficie es

```
SELECT cod_pai
FROM Pais
WHERE superficie = (SELECT MIN(superficie) FROM Pais));
```

Lo que queremos es borrar de la tabla `Pais` el registro con ese código. Para ello la envolvemos en un `DELETE`.

```
DELETE
FROM Pais
WHERE cod_pai = (
    SELECT cod_pai
    FROM Pais
    WHERE superficie = (SELECT MIN(superficie) FROM Pais))
);
```

Para que el borrado se produzca en cascada, y que las ciudades de este país también sean eliminadas de la base de datos (conservando así la



integridad referencial) podemos especificarlo en la creación de la tabla. A la sintaxis dada en b) le añadiríamos la coletilla `ON DELETE CASCADE`:

```
CREATE TABLE Mundo.Ciudad (
    cod_ciu INT PRIMARY KEY,
    nombre VARCHAR (50) NOT NULL,
    poblacion INT,
    cod_pai INT NOT NULL,
    FOREIGN KEY (cod_pai) REFERENCES mundo.Pais (cod_pai)
    ON DELETE CASCADE
);
```

Como esta cláusula puede llevar recursivamente a eliminaciones masivas de datos, a veces no se nos permite usar borrados en cascada. En ese caso, antes de eliminar el país tendremos que borrar «a mano» las ciudades asociadas con otra query muy similar:

```
DELETE
FROM Ciudad
WHERE cod_pai = (
    SELECT cod_pai
    FROM Pais
    WHERE superficie = (SELECT MIN(superficie) FROM Pais))
);
```

Ejercicio 4

Diseñe una función que simule una v. a. exponencial, cuya función de densidad es la siguiente:



$$f(x) = \lambda e^{-\lambda x}, \quad \text{para } x \geq 0.$$

Para la simulación, podemos suponer definida una función $U(0,1)$, esto es, una función que simula una v. a. real, uniformemente distribuida en el intervalo $[0,1]$. El diseño de este algoritmo puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R.

Solución.

Lo que se nos pide es algo muy habitual, utilizar el método de inversión para simular una variable aleatoria con cierta función de densidad o distribución a partir de una uniforme.

Teoría en que nos basamos

Si $X \sim U(0,1)$ y $F : \mathbb{R} \to \mathbb{R}$ es una función continua e invertible que reune las características de las funciones de distribución, al definir la variable aleatoria $Y = F^{-1}(X)$ resulta que su función de distribución es justamente F.

Aplicación

Para simular lo primero que necesitamos es conocer la función de distribución F de una $\text{Exp}(\lambda)$ y nos dan la de densidad f. Si no la recordamos, la obtenemos por integración:

$$F(x) = \int_{-\infty}^{x} f(u)du = \int_{0}^{x} \lambda e^{-\lambda u} du = [-e^{-\lambda u}]_0^x = [e^{-\lambda u}]_x^0 = 1 - e^{-\lambda x}.$$

Busquemos la expresión de F^{-1}, que es la que deberemos usar en nuestra implementación.

$$y = F(x) = 1 - e^{-\lambda x} \iff e^{-\lambda x} = 1 - y \iff x = -\frac{\ln(1-y)}{\lambda}.$$



Por tanto, $F^{-1}(x) = -\frac{\ln(1-x)}{\lambda}$.

Llamemos X a la variable uniforme de partida. Si aplicáramos el método de inversión directamente, definiríamos $Y = F^{-1}(X) = -\frac{\ln(1-X)}{\lambda}$ y simularíamos usando esta fórmula, pero es posible ser un poco más eficiente si observamos que también $1 - X \sim U(0, 1)$, es decir, ese reflejo de la uniforme respecto a su centro es totalmente evitable. Por tanto, definimos directamente

$$Y = -\frac{\ln(X)}{\lambda}$$

y esto es lo que vamos a programar. Presentamos un script en `Python`:

```
import numpy as np
import numpy.random as ra

def simula_expon(lamb, nsims):
    return -np.log(ra.uniform(size=nsims)) / lamb

# Ejemplo de uso
exp_muestra = simula_expon(lamb=3.0, nsims=10**6)
```

Ejercicio 5

Una población está formada por $M = 3000$ viviendas (unidades elementales) agrupadas en 30 municipios de tamaños desiguales ($M_i, i = 1, 2, \ldots, 30$). Se quiere estimar la proporción de viviendas que pertenecen a una cierta clase y para ello se usa un muestreo de conglomerados con submuestreo.

En ambas etapas se emplea un procedimiento de selección con probabilidades iguales sin reposición.

En la primera etapa se seleccionan 3 municipios con los siguientes valores de M_i: 10; 50; 100.



En la segunda etapa, realizada con una fracción de muestreo de $f_{2i} = 4/M_i$, se obtienen en los 3 municipios muestrales los valores 2; 2; 3 respectivamente para el número de viviendas que pertenecen a la clase.

Se pide:

(a) Una estimación insesgada de la mencionada proporción y su error de muestreo.

(b) Una estimación, usando el estimador de la razón al tamaño, de la proporción mencionada y su error de muestreo.

(c) Comentar ventajas e inconvenientes del estimador de (b) frente al estimador de (a).

Solución.

a) Del enunciado se tiene:

$$N = 30 \qquad M = 3000 \qquad n = 3 \qquad f_{2i} = \frac{4}{M_i} \Rightarrow m_i = 4 \qquad \forall i = 1, 2, 3$$

Una estimación insesgada de la proporción de viviendas de la clase referida vendría dada por:

$$\widehat{P} = \frac{\widehat{A}}{M} = \frac{1}{M} \sum_{i=1}^{n} \frac{M_i \widehat{P}_i}{\pi_i} = \frac{1}{M} \frac{N}{n} \sum_{i=1}^{n} M_i \widehat{P}_i$$

Puesto que $\widehat{P}_i = \frac{n_i}{m_i}$, donde n_i es el número de viviendas de la clase estudiada en cada conglomerado de la muestra, para cada uno de los conglomerados se tiene

$$\widehat{P}_1 = \frac{2}{4} \qquad \widehat{P}_2 = \frac{2}{4} \qquad \widehat{P}_3 = \frac{3}{4}$$

y así la estimación insesgada de la proporción es:

$$\widehat{P} = \frac{1}{M} \frac{N}{n} \sum_{i=1}^{3} M_i \widehat{P}_i = \frac{1}{3000} \frac{30}{3} \left(10\frac{2}{4} + 50\frac{2}{4} + 100\frac{3}{4} \right) = \frac{105}{300} = 0{,}35.$$



La estimación de la varianza de la proporción ha de hacerse teniendo en cuenta que se trata de muestreo bietápico sin reposición en las dos etapas con probabilidades iguales para conglomerados de distinto tamaño. Así:

$$\widehat{Var}(\widehat{P}) =$$

$$= \frac{1}{M^2}\left[\frac{N^2(1-f_1)}{n}\frac{\sum_i^n\left(M_i\widehat{P}_i - \frac{1}{n}\sum_i^n M_i\widehat{P}_i\right)^2}{n-1} + \frac{N}{n}\sum_i^n M_i^2(1-f_{2i})\frac{\widehat{P}_i\widehat{Q}_i}{m_i-1}\right]$$

$$= \frac{1}{3000^2}\left[\frac{30^2(1-\frac{3}{30})}{3}\frac{(5-35)^2+(25-35)^2+(75-35)^2}{2}\right] +$$

$$+ \frac{1}{3000^2}\cdot\frac{30}{3}\left(100\frac{6}{4}\frac{1}{12} + 2500\frac{46}{50}\frac{1}{12} + 10000\frac{96}{100}\frac{1}{16}\right) =$$

$$= \frac{1}{3000^2}\left[283500 + 5416{,}6666\right] = 0{,}03210$$

Y el error de muestreo será

$$\widehat{\sigma}(\widehat{P}) = \widehat{\sigma}(\widehat{P}) = \sqrt{\widehat{Var}(\widehat{P})} = \sqrt{0{,}03210} = 0{,}17917$$

b) Para el cálculo de la estimación de la proporción de viviendas de la clase en estudio utilizando el método de la razón se utiliza

$$\widehat{P}_R = \frac{\sum_{i=1}^3 A_i}{\sum_{i=1}^3 M_i} = \frac{\sum_{i=1}^3 M_i\widehat{P}_i}{\sum_{i=1}^3 M_i} = \frac{10\frac{1}{2} + 50\frac{1}{2} + 100\frac{3}{4}}{10+50+100} = \frac{105}{160} = 0{,}65625$$

donde se añade el subíndice R para hacer referencia al cálculo de la estimación utilizando el método de la razón. En cuanto a su error de muestreo, se puede aproximar mediante la expresión:

$$\begin{aligned}\widehat{Var}(\widehat{P}_R) \approx & \frac{1}{(\bar{M})^2}\left[\frac{1}{n}(1-f_1)\frac{1}{n-1}\left(\sum_{i=1}^3\left(\frac{M_i\cdot n_i}{m_i} - \widehat{P}_R\cdot M_i\right)^2\right) + \right.\\ & \left. + \frac{1}{nN}\sum_{i=1}^3\frac{M_i(M_i-m_i)\widehat{P}_i\left(1-\widehat{P}_i\right)}{m_i-1}\right] =\\ & = \frac{1}{\frac{M^2}{N^2}}\left[\frac{1}{n}(1-f_1)\frac{1}{n-1}\left(\sum_{i=1}^3\left(\frac{M_i\cdot n_i}{m_i} - \widehat{P}_R\cdot M_i\right)^2\right) + \right.\\ & \left. + \frac{1}{nN}\frac{1}{m-1}\sum_{i=1}^3 M_i(M_i-m_i)\widehat{P}_i\left(1-\widehat{P}_i\right)\right]\end{aligned}$$



donde,

$$\sum_{i=1}^{3}\left(\frac{M_i\cdot n_i}{m_i}-\widehat{P}_R\cdot M_i\right)^2=\left(\frac{10\cdot 2}{4}-\frac{105}{160}\cdot 10\right)^2+\left(\frac{50\cdot 2}{4}-\frac{105}{160}\cdot 50\right)^2+$$
$$+\left(\frac{100\cdot 3}{4}-\frac{105}{160}\cdot 100\right)^2=$$
$$=\left(5-\frac{105}{16}\right)^2+\left(25-\frac{525}{16}\right)^2+\left(75-\frac{1050}{16}\right)^2=$$
$$=\frac{38750}{256}$$

$$\sum_{i=1}^{3}M_i\left(M_i-m_i\right)\widehat{P}_i\left(1-\widehat{P}_i\right)=10\cdot 6\cdot\frac{1}{2}\frac{1}{2}+50\cdot 46\cdot\frac{1}{2}\cdot\frac{1}{2}+$$
$$+100\cdot 96\cdot\frac{3}{4}\cdot\frac{1}{4}=$$
$$=15+575+1800=2390.$$

De esta forma,

$$\widehat{Var}(\widehat{P}_R)=\frac{30^2}{3000^2}\left[\frac{1}{3}\cdot\left(1-\frac{3}{30}\right)\cdot\frac{1}{2}\cdot\left(\frac{38750}{256}\right)+\frac{1}{3\cdot 30}\cdot\frac{1}{3}\cdot 2390\right]=$$
$$=0{,}003155.$$

Y el error de muestreo será

$$\widehat{\sigma}(\widehat{P}_R)=\widehat{em}(\widehat{P}_R)=\sqrt{\widehat{var}(\widehat{P}_R)}=\sqrt{0{,}003155}=0{,}05617.$$

Una estimación de la varianza más sencilla se puede obtener trabajando con el método de conglomerados últimos, que permite considerar el muestreo polietápico como un caso especial del muestreo de conglomerados sin submuestreo, es decir, en una etapa. La aplicación del método es simple y es conveniente cuando no se necesiten estimaciones separadas de las contribuciones a la varianza de cada una de las distintas etapas del muestreo.



Tenemos que:

$$\widehat{P} = \frac{1}{n}\sum_{i=1}^{3}\frac{M_i\widehat{P}_i N}{M} = \frac{1}{n}\sum_{i=1}^{3}\hat{\theta}_i = \frac{1}{3}\left(10\frac{2}{4}\frac{30}{3000} + 50\frac{2}{4}\frac{30}{3000} + 100\frac{3}{4}\frac{30}{3000}\right) =$$
$$= \frac{1}{3}(0{,}05 + 0{,}25 + 0{,}75) = \frac{1{,}05}{3} = 0{,}35.$$

y un estimador insesgado de la varianza es:

$$\widehat{Var}(\widehat{P}) = \frac{\sum_{i=1}^{3}\left(\hat{\theta}_i - \widehat{P}\right)^2}{n(n-1)} = \frac{(0{,}05-0{,}35)^2 + (0{,}25-0{,}35)^2 + (0{,}75-0{,}35)^2}{3\cdot 2} =$$
$$= \frac{0{,}09 + 0{,}01 + 0{,}16}{6} = 0{,}0433$$

que es una peor aproximación del error, pero puede ser útil en muchos casos.

c) El cálculo de este estimador $\widehat{P}_R$ no requiere el conocimiento del número de conglomerados total ni el número total de elementos de la población, sino solo la información obtenida de la muestra, pero suele ser sesgado, aunque dicho sesgo puede ser reducido incrementando el tamaño muestral. El estimador $\widehat{P}$ del primer apartado es insesgado, aunque suele ser menos preciso que el de la razón al tamaño.

El denominador del estimador $\widehat{P}_R$ es $\sum_{i=1}^{3} M_i$, mientras que en el de las dos etapas $\widehat{P}$ aparece en el denominador $M \cdot \frac{n}{N}$ y de ahí su diferencia. En el segundo caso tiene en cuenta tantos elementos como la proporción de conglomerados muestreados por el total de elementos de la población, y en el primero ajusta el denominador por el total de elementos de los conglomerados muestrados concretamente.

En general, también será más eficiente el estimador de razón, puesto que hace posible corregir los desequilibrios de la muestra de unidades primarias. El estimador insesgado suele ser más ineficiente porque es muy sensible a los valores del tamaño de cada conglomerado.



Ejercicio 6

Se modeliza la ocurrencia diaria de lluvia en otoño en cierta estación meteorológica mediante una cadena de Markov con dos estados (1) no llueve y (2) llueve. Se han recogido los siguientes datos de lluvia en 49 días:

```
2222222222 1121111111 2211112221 2111112111 111111111.
```

Se pide:

(a) Defina una matriz de transición para la cadena.

(b) Estime mediante máxima verosimilitud las entradas de la matriz de la cadena.

(c) Suponga distribuciones a priori uniformes para los elementos de la diagonal de la matriz.

- Calcule las distribuciones a posteriori para los elementos de la cadena.
- Calcule las probabilidades de transición esperadas a posteriori.
- Calcule la probabilidad predictiva a posteriori de lluvia para el día 50.

(d) Calcule la probabilidad predictiva a posteriori de que no llueva los días 50 y 51.

Solución.

La serie de códigos de pluviosidad para los 49 días consecutivos mencionados es

222222222211211111112211112222121111121111111111111.

Al ser un proceso de Markov, lo único relevante son esas 48 transiciones de estado, partiendo desde el estado inicial 2. De entre todas las posibles secuencias de 48 transiciones, denotamos $\mathbf{x}$ a la dada en el texto. Denotamos por t_{ij} el número de transiciones entre el estado i y el j, así que el conteo resulta $t_{11} = 25, t_{12} = 5, t_{21} = 6$ y $t_{22} = 12$.

(a) definimos la matriz de transición del proceso de Markov

$$M = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix},$$

donde p_{ij} será la probabilidad de pasar del estado i al j, $1 \leq i, j \leq 2$. Las probabilidades de salida desde cada estado deben sumar 1, de modo que en realidad solo hay dos valores a determinar, p_{11} y p_{22}, que son las probabilidades de permanecer en sendos estados. La matriz de transición queda

$$M = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{pmatrix}.$$

Se trata de la cadena de Markov general con dos estados, representada en la Figura 2.3.

(b) Determinemos las estimaciones de máxima verosimilitud $\widehat{p_{11}}, \widehat{p_{22}}$.

A partir del día inicial, la probabilidad de la secuencia de transiciones del enunciado es

$$h(p_{11}, p_{22}) = p_{11}^{t_{11}} \cdot (1 - p_{11})^{t_{12}} \cdot (1 - p_{22})^{t_{21}} \cdot p_{22}^{t_{22}}.$$

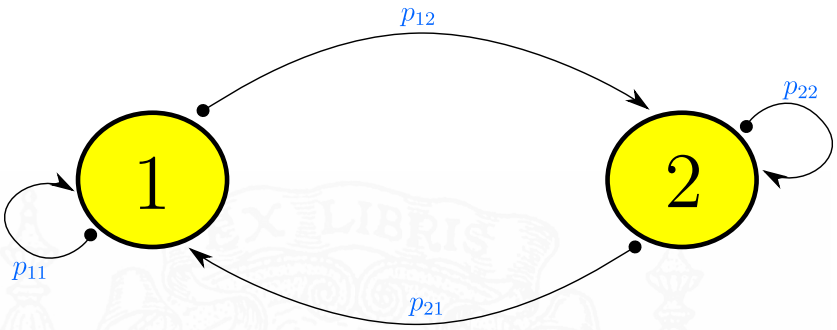


Figura 2.3: Cadena de Markov general con dos estados. Las probabilidades de transición se denotan en azul.

El par $(\widehat{p_{11}}, \widehat{p_{22}})$ es justamente el punto donde h alcanza su máximo. Resulta conveniente tomar logaritmos, que por ser una función estrictamente monónotona no alteran la ubicación de extremos. Por tanto, en vez de h consideramos

$$\begin{aligned}\log h(p_{11}, p_{22}) &= \ln(h(p_{11}, p_{22})) \\ &= t_{11} \ln p_{11} + t_{12} \ln(1 - p_{11}) + t_{21} \ln(1 - p_{22}) + t_{22} \ln p_{22}.\end{aligned}$$

Al ser derivable en ambas variables, su máximo debe alcanzarse en un punto donde ambas derivadas parciales se anulen. En realidad esta función es un producto de una función en p_{11} y otra en p_{22}, con lo que vamos a comprobar que el máximo se da maximizando en cada una de las variables por separado.

$$\frac{\partial \log h(p_{11}, p_{22})}{\partial p_{11}} = \frac{t_{11}}{p_{11}} - \frac{t_{12}}{1 - p_{11}},$$

por lo que

$$\frac{\partial \log h(p_{11}, p_{22})}{\partial p_{11}} = 0 \iff \frac{t_{11}}{p_{11}} - \frac{t_{12}}{1 - p_{11}} = 0 \iff \frac{t_{11}}{p_{11}} = \frac{t_{12}}{1 - p_{11}} \iff$$
$$t_{11}(1 - p_{11}) = p_{11} t_{12} \iff p_{11} = \frac{t_{11}}{t_{11} + t_{12}}.$$

Por tanto, nuestra estimación máximo verosímil para la probabilidad de pasar de 1 a 1 es

$$\widehat{p_{11}} = \frac{t_{11}}{t_{11} + t_{12}} = \frac{25}{25 + 5} = \frac{5}{6}.$$

De forma absolutamente análoga, imponer

$$\frac{\partial \text{logh}(p_{11}, p_{22})}{\partial p_{22}} = 0$$

nos lleva a

$$\widehat{p_{22}} = \frac{t_{22}}{t_{22} + t_{21}} = \frac{12}{12 + 6} = \frac{2}{3}.$$

Para las probabilidades de cambio de estado es trivial $\widehat{p_{12}} = 1 - \widehat{p_{11}} = \frac{1}{6}$ y $\widehat{p_{21}} = 1 - \widehat{p_{22}} = \frac{1}{3}$.

Es decir, para cada transición $i \mapsto j$, la estimación máximo verosímil de la probabilidad de la transición coincide con la proporción veces observadas que estando en i se transicionó a j, lo que parece razonable.

(c) En este apartado la estimación es otra, de carácter bayesiano. Según el enunciado, a priori las probabilidades de no-cambio para ambos estados son P_{11} y P_{22}, variables aleatorias $U(0, 1)$, que además puesto que el problema no menciona ninguna dependencia entre ellas vamos a asumir independientes.

Se nos pide averiguar algunas cuestiones sobre las probabilidades de transición *futuras* o *a posteriori* una vez se ha observado la secuencia de transiciones $\mathbf{x}$. Las denotaremos P_{11}^F y P_{22}^F.

Las probabilidades P_{11} y P_{22} siguen una distribución continua, mientras que el experimento informativo (la secuencia de transiciones de estado)



es de naturaleza discreta. En este caso, la función de densidad conjunta de (P_{11}^F, P_{22}^F) resulta de multiplicar la de (P_{11}, P_{22}) por la probabilidad del suceso observado $\mathbf{x}$, y luego normalizar. Formalmente

$$\begin{aligned} f^F(p_{11}, p_{22}) &\propto f(p_{11}, p_{22}) \cdot \mathbb{P}(\mathbf{x} | P_{11} = p_{11} \wedge P_{22} = p_{22}) \\ &= \mathbb{P}(\mathbf{x} | P_{11} = p_{11} \wedge P_{22} = p_{22}) \quad \text{por ser } f = 1 \text{ en } [0,1]^2 \\ &= p_{11}^{t_{11}} \cdot (1 - p_{11})^{t_{12}} \cdot p_{22}^{t_{22}} \cdot (1 - p_{22})^{t_{21}} \quad \text{si } 0 \leq p_{11}, p_{22} \leq 1. \end{aligned}$$

Esta expresión coincide con la función h del apartado (b), por lo que el EMV se da en el punto en que la densidad es máxima. Es un producto de una función en p_{11} por otra en p_{22}, lo que refleja que las variables a posteriori P_{11}^F y P_{22}^F son independientes entre sí, al igual que sucedía con las probabilidades a priori P_{11} y P_{22}. Natural, tratándose de una cadena de Markov.

Ahora conviene recordar que la función de densidad de una distribución $\text{Beta}(\alpha, \beta)$ es

$$f_{\text{B}}(x; \alpha, \beta) \propto x^{\alpha - 1}(1 - x)^{\beta - 1}$$

en el intervalo $[0, 1]$, y 0 fuera de él. Utilizando esto, resulta

$$f^F(p_{11}, p_{22}) \propto f_{\text{B}}(p_{11}; t_{11} + 1, t_{12} + 1) \cdot f_{\text{B}}(p_{22}; t_{22} + 1, t_{21} + 1).$$

Es más, como tanto el término izquierdo como el derecho son funciones de densidad, al ser proporcionales deben coincidir:

$$f^F(p_{11}, p_{22}) = f_{\text{B}}(p_{11}; t_{11} + 1, t_{12} + 1) \cdot f_{\text{B}}(p_{22}; t_{22} + 1, t_{21} + 1).$$

1) Esto permite concluir que las distribuciones buscadas son

$$P_{11}^F \sim \text{Beta}(\alpha = t_{11} + 1, \beta = t_{12} + 1) = \text{Beta}(\alpha = 26, \beta = 6)$$

$$P_{22}^F \sim \text{Beta}(\alpha = t_{22} + 1, \beta = t_{21} + 1) = \text{Beta}(\alpha = 13, \beta = 7).$$

En cuanto a los otros dos elementos en la matriz de transición, fuera de la diagonal principal, son triviales a partir de estos. Sucede que $P_{12}^F = 1 - P_{11}^F$ y $P_{21}^F = 1 - P_{22}^F$, y es claro que si $X \sim \text{Beta}(\alpha, \beta)$, entonces $1 - X \sim \text{Beta}(\beta, \alpha)$. Por tanto,

$$P_{12}^F \sim \text{Beta}(\alpha = t_{12} + 1, \beta = t_{11} + 1) = \text{Beta}(\alpha = 6, \beta = 26)$$
$$P_{21}^F \sim \text{Beta}(\alpha = t_{21} + 1, \beta = t_{22} + 1) = \text{Beta}(\alpha = 7, \beta = 13).$$

2) Las probabilidades de transición esperadas a posteriori no son más que las esperanzas $\mathbb{E}[P_{ij}^F]$, con $1 \leq i, j \leq 2$. Como se distribuyen según Betas, basta recordar que la esperanza de una $\text{Beta}(\alpha, \beta)$ es $\frac{\alpha}{\alpha+\beta}$. Utilizando esto,

$$\mathbb{E}[P_{11}^F] = \frac{t_{11} + 1}{t_{11} + t_{12} + 2} = \frac{25 + 1}{25 + 5 + 2} = \frac{13}{16}$$
$$\mathbb{E}[P_{22}^F] = \frac{t_{22} + 1}{t_{22} + t_{21} + 2} = \frac{12 + 1}{6 + 12 + 2} = \frac{13}{20}$$
$$\mathbb{E}[P_{12}^F] = \mathbb{E}[1 - P_{11}^F] = 1 - \frac{13}{16} = \frac{3}{16}$$
$$\mathbb{E}[P_{21}^F] = \mathbb{E}[1 - P_{22}^F] = 1 - \frac{13}{20} = \frac{7}{20}.$$

La representación gráfica de los estimadores y funciones de densidad aparece en la Figura 2.4.

3) Puesto que el proceso es Markov, la probabilidad pedida es la de pasar del último estado 1 al estado 2, es decir,

$$P_{12}^F = 1 - P_{11}^F \sim \text{Beta}(\alpha = 6, \beta = 26).$$



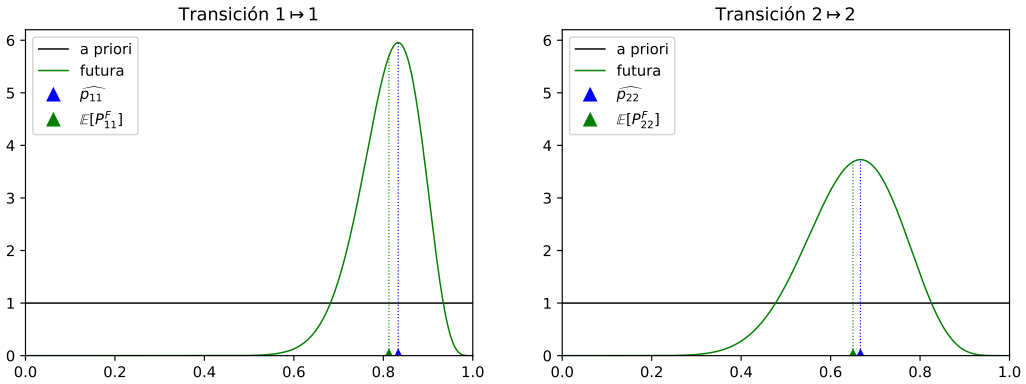


Figura 2.4: Distribuciones a priori/posteriori y EMV de las transiciones de no-cambio. Los máximos de las funciones de densidad a posteriori coinciden con el EMV, mientras que las esperanzas resultan ligeramente más cercanas a $\frac{1}{2}$ (que era la esperanza a priori).

Obsérvese que es una variable aleatoria, no un valor concreto.

(d) Ahora nos piden la probabilidad predictiva de que no llueva ni el día 50 ni el 51. Es decir, puesto que el día 49 tampoco llovió, que la cadena transicione dos veces seguidas desde el estado 1 a sí mismo. Al ser Markov, esa probabilidad es la variable aleatoria $P^F_{11} \cdot P^F_{11} = (P^F_{11})^2$.

Ejercicio 7

Una agencia de viajes necesita desarrollar un sistema para gestionar información sobre pisos turísticos, incluyendo el país, ciudad y dirección en que se ubican, así como otras características tales como número de dormitorios y baños, disponibilidad de cocina, etc.

(a) Defina XML, diga qué reglas debe cumplir un documento bien formado y proporcione un ejemplo de un documento XML bien formado para el ejemplo anterior.



(b) Se desea diseñar una aplicación web para gestionar la información de nuestra agencia. Describa los conceptos de frontend, backend, lado del cliente y lado del servidor, apoyándose en nuestra agencia de viajes. En cada uno de estos conceptos, mencione alguna tecnología que tendría sentido usar.

Solución.

(a) `XML` es el acrónimo para *lenguaje de marcas extendido*. Es un formato de fichero en texto plano que mediante un sistema de tags, elementos, atributos y anidamiento permite codificar información compleja de una forma sistemática. Protocolos secundarios como `XPath` nos ayudan a explotarla de forma eficiente. Está ampliamente difundido y la mayoría de lenguajes de propósito general incorporan librerías paga gestionar estos ficheros con facilidad.

Para el fichero `XML` concreto asociado a nuestra agencia de viajes, la decisión de las tablas a utilizar queda abierta en el enunciado, un diseño muy simple podría consistir en una única tabla `Piso` con toda la información especificada. Por dar algo más de integridad a la información, supondremos que se maneja un catálogo de ciudades y otro de países, de modo que cada piso debe estar en una de las ciudades ya catalogadas. Entonces tenemos tres tablas: `Piso`, `Ciudad` y `Pais`. El esquema de las tablas con información adicional aparece en la Figura 2.5.

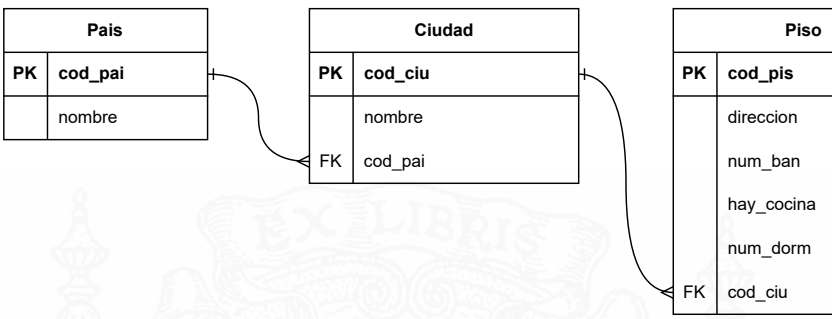


Figura 2.5: Esquema para la bbdd de pisos turísticos.

Un fichero `XML` bien formado muy sencillo, con en elemento de cada, podría ser:

```
<?xml version="1.0"?>
<!DOCTYPE pistur[
        <!ELEMENT pistur (pisos, ciudades, paises)>
        <!ELEMENT pisos (piso*)>
        <!ELEMENT ciudades (ciudad*)>
        <!ELEMENT paises (pais*)>
        <!ELEMENT piso (direccion, num_dorm, num_ban, hay_cocina)>
        <!ELEMENT ciudad (nombre_c)>
        <!ELEMENT pais (nombre_p)>
        <!ELEMENT nombre_c (#PCDATA)>
        <!ELEMENT nombre_p (#PCDATA)>
        <!ELEMENT direccion (#PCDATA)>
        <!ELEMENT num_dorm (#PCDATA)>
        <!ELEMENT num_ban (#PCDATA)>
        <!ELEMENT hay_cocina (#PCDATA)>
        <!ATTLIST piso cod_pis ID #REQUIRED>
        <!ATTLIST ciudad cod_ciu ID #REQUIRED>
        <!ATTLIST pais cod_pai ID #REQUIRED>
```




```
        <!ATTLIST ciudad cod_pai IDREF #REQUIRED>
        <!ATTLIST piso cod_ciu IDREF #REQUIRED>
]>
<pistur>
    <pisos>
        <piso cod_pis="S436" cod_ciu="C546">
            <direccion>Rue del Percebe 13</direccion>
            <num_dorm>9</num_dorm>
            <num_ban>1</num_ban>
            <hay_cocina>True</hay_cocina>
        </piso>
    </pisos>
    <ciudades>
        <ciudad cod_ciu="C546" cod_pai="P13">
            <nombre_c>Talavera de la Reina</nombre_c>
        </ciudad>
    </ciudades>
    <paises>
        <pais cod_pai="P13">
            <nombre_p>Espanna</nombre_p>
        </pais>
    </paises>
</pistur>
```

Está bien formado porque:

- Comienza por la cláusula que especifica la versión de `XML`.
- Todos los tags que se abren, se cierran posteriormente.
- El anidamiento de elementos es correcto: los tags se cierran en



orden inverso al que se abren.

- Solo existe un elemento en la raíz.

Además, el fichero `XML` es validable puesto que se conforma a la estructura determinada por el DTD incrustado. Sigue la estrategia habitual para codificar bases de datos en cuatro niveles: base de datos, tabla, registro y campo. En particular, en el DTD se especifican los campos obligatorios y optativos de ambas tablas, se definen las claves como atributos (no elementos), y se garantiza la integridad referencial asociada a claves ajenas mediante un `IDREF`.

(b) Al montar una aplicación en una arquitectura cliente/servidor,

- el lado del servidor es la máquina (o máquinas) que aloja las bases de datos necesarias, junto con los scripts en lenguajes de programación adecuados para ser capaz de proporcionar al cliente los servicios que se le ofrecen. Por ejemplo, lo que es el sistema de bases de datos podría ser gestionado por un `Oracle`, y los algoritmos con la lógica de gestión en un lenguaje de propósito general como `C++` o `Python`, con soporte de `SQL`.
- el lado cliente es la interfaz con la que el cliente va a poder interactuar con la aplicación, haciendo peticiones de información (como búsqueda de pisos en una localidad o en un radio en torno a un punto, filtrados por características, etc.) o modificaciones (por ejemplo, reservando una habitación). En una aplicación web como la que tenemos en mente, del lado del cliente lo que se necesita es un web-browser que soporte las *cookies* necesarias, como `Chrome` o `Firefox`.



- En una aplicación web, el front-end es la parte del desarrollo relacionada con los aspectos visuales de la web: aquella con la que los usuarios interactúan. Las ventanas donde introducir o recibir información, *combos*, botones, pestañas, desplegables, etc. Existen múltiples herramientas de desarrollo de front-end, como por ejemplo `jQuery`.

- Por otro lado el back-end es donde se detallan la estructura de la web, los datos y la lógica del programa. Por ejemplo, qué pasos se siguen exactamente cuando un cliente da de alta una reserva, o la cancela, cómo filtrar los pisos que se encuentran a menos de cierta distancia de un punto dado, etc. El lenguaje `Python` es muy popular a este respecto, y aporta librerías como `Django` que sistematizan bastante la estructura de las aplicaciones, permitiendo que sean muy homogéneas y modulares.

Ejercicio 8

Deseamos obtener la mediana de una secuencia de números enteros. Los números enteros no están ordenados. La secuencia puede ser un vector o una lista, posiblemente de gran tamaño, por lo que nuestro algoritmo ha de ser lo más eficiente posible. El diseño de este algoritmo puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R. No estaría permitido usar ninguna función predefinida que ordene dicha secuencia. Adicionalmente, se pide discutir la complejidad en tiempo de dicho algoritmo, o al menos un boceto de dicho cálculo.



Solución.

Consideraciones previas

Una primera estrategia podría ser aplicar un algoritmo de ordenación eficiente, como *quicksort*, y una vez ordenada la lista tomar el valor central (o promedio de los dos centrales, si la cantidad es par). Si denotamos por n al tamaño de la lista/vector a procesar, la complejidad algorítmica resultaría $\mathcal{O}(n \cdot \ln(n))$. Sin embargo, eso supone llevar a cabo mucho más trabajo del necesario, y de hecho vamos a presentar un algoritmo $\mathcal{O}(n)$, es decir, complejidad lineal respecto al tamaño de la lista de datos.

Algoritmo propuesto

El algoritmo presentado es capaz de obtener, en una lista posiblemente desordenada y para cualquier k, el elemento que ocuparía la k-ésima posición si la ordenáramos de forma creciente. Consiste en elegir un dato *pivote* al azar de la lista y dividirla en tres sublistas: la de valores menores al pivote, la de valores iguales, y la de mayores. El k-ésimo elemento buscado estará en uno de esos tres grupos, y es en él en el que seguiremos buscando de forma recursiva mientras descartamos completamente los restantes (a diferencia de como funciona *quicksort*, que actúa en cada subgrupo).

Debe tenerse en cuenta un *corner case*: si todos los números son iguales el proceso de trocear no subdivide la lista. Este caso se controla específicamente: el valor a devolver es obviamente ése que se repite en toda la lista.

Complejidad

Sea n el tamaño de la lista. A continuación analizamos la complejidad en función del caso de interés.



Complejidad en el escenario promedio

Al tomar el pivote al azar, sabemos que en término medio el tamaño del trozo en que queda el valor a buscar es $\frac{n}{2}$. De este modo, en cada paso el número de elementos a seguir procesando es la mitad que en la pasada anterior. Por tanto, la suma del número total de operaciones es del orden de $n + \frac{n}{2} + \frac{n}{4} + \frac{n}{8} + \cdots = 2n = \mathcal{O}(n)$. Así que en promedio el tiempo de ejecución es lineal, o dicho de otra manera, proporcional al tamaño de la lista donde se realiza la búsqueda.

Complejidad en el peor escenario

El peor escenario posible se daría cuando los valores son distintos, y el pivote elegido al azar resulta siempre (¡malísima suerte!) un extremo distinto del valor buscado. En ese caso en cada pasada solo eliminamos un valor de la lista, y el número total de operaciones resulta del orden de $n + (n-1) + (n-2) + \cdots + 3 + 2$, cuyo orden es $\mathcal{O}(n^2)$.

Implementación

Presentamos un script `Python`.

```
import numpy.random as ra


def trocear(nums):
    pivote = ra.choice(nums)
    menores, iguales, mayores = [], [], []
    for e in nums:
        if e < pivote:
            menores.append(e)
        elif e == pivote:
```




```
            iguales.append(e)
        else:
            mayores.append(e)
    trozos = [menores, iguales, mayores]
    return [tr for tr in trozos if tr]  # Devuelve los trozos no vacios


def busca_kesimo(nums, k):
    """Devuelve el k-esimo elemento en rango de nums. 0 para el primero,
    1 para el segundo..."""
    # Caso base 1 de la recursion: todos los numeros son iguales
    if all(nums[0] == e for e in nums):
        return nums[0]
    # Caso base 2 de la recursion: el dato buscado es un extremo
    if k == 0:
        return min(nums)
    if k == len(nums) - 1:
        return max(nums)
    # Caso recursivo
    for tr in trocear(nums):
        if len(tr) > k:  # El numero buscado esta en el trozo tr
            return busca_kesimo(tr, k)
        else:  # El numero buscado esta en alguno de los trozos restantes
            k -= len(tr)

def mediana(nums, uso_promedio=True):
    """Devuelve la mediana.
    Cuando el numero de datos es impar, es el central. Si son pares, en sentido
    estricto deberia devolverse el promedio de los dos centrales, pero a veces
```




```
    se admite tomar uno de ellos. Eso lo controla el flag uso_promedio.
    """
    if (len(nums) % 2) or (not uso_promedio):  # numero impar de datos, o no se pide prome
        return busca_kesimo(nums=nums, k=len(nums) // 2)
    else:  # numero par de datos, y se requiere promediar los dos centrales
        v1 = busca_kesimo(nums=nums, k=len(nums) // 2 - 1)
        v2 = busca_kesimo(nums=nums, k=len(nums) // 2)
        return (v1 + v2) / 2


# Ejemplos de uso:
nums = [2, 10, 7, 1, 1, 4, 10, 7, 6, 5]
print(mediana(nums=nums, uso_promedio=False))
print(mediana(nums=nums, uso_promedio=True))
print(mediana(nums=[12, 12, 12, 12, 12]))
```

Versión iterativa

La implementación recursiva anterior es fácil de entender por ser de alto nivel, y de hecho en muchas ocasiones la recursión es muy natural y difícil de evitar. No es éste el caso, resulta trivial transformar `busca_kesimo` a una versión iterativa que lleva a cabo exactamente el mismo algoritmo. Las implementaciones iterativas suelen ejecutarse algo más rápidamente, porque la computadora se ahorra el *overhead* por apilamiento/desapilamiento de los contextos funcionales. Además, ya no corremos el riesgo de *stack overflow* cuando la recursión resulta demasiado profunda, aunque eso hubiera sido sumamente improbable en el programa de este ejercicio.

Versión iterativa para `busca_kesimo`:

```
def busca_kesimo(nums, k):
    while True:
```




```
    # Todos los numeros iguales
    if all(nums[0] == e for e in nums):
        return nums[0]
    # Caso simple al que se llega en algun momento: el k buscado
    # es uno de los extremos de nums.
    if k == 0:
        return min(nums)
    if k == len(nums) - 1:
        return max(nums)
    # Caso no-simple: hay que reemplazar nums por el trozo en el
    # que se debe seguir buscando
    for tr in trocear(nums):
        if len(tr) > k:  # El numero buscado esta en el trozo tr
            nums = tr
            break
        else:  # El numero buscado esta en alguno de los trozos restantes
            k -= len(tr)
```

Sobre la elección de pivotes

La elección de los pivotes al azar conlleva que nuestro programa no sea determinista, y supone un coste añadido respecto a tomar, por ejemplo, siempre el primer elemento como pivote. Pero evita que en ciertos casos habituales se pierda la complejidad lineal. Por ejemplo, si el pivote siempre fuera el primer elemento de la lista, en los casos en que la lista estuviera ordenada (de forma creciente o decreciente) la complejidad sería cuadrática, $\mathcal{O}(n^2)$. Por eso merece la pena elegir el pivote al azar, y asumir ese coste.



2.3. Soluciones del segundo examen del Cuerpo de Diplomados en Estadística del Estado de 2022.

Ejercicio 1

Una variable aleatoria continua tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \frac{1}{2}e^{x} & \text{si } x \leq 0 \\ \frac{1}{2}e^{-x} & \text{si } x > 0 \end{cases}$$

Obtenga:

1. La función de distribución de X
2. La probabilidad de que X sea menor o igual que 1
3. La probabilidad del suceso $\{|X| \leq 2 \cup (X > 0)\}$
4. Si sabemos que X es mayor que 1, la probabilidad de que X sea menor que 2
5. La esperanza de X

Solución.

1) Tenemos una variable aleatoria X con función de densidad:

$$f(x) = \begin{cases} \frac{1}{2}e^{x} & \text{si } x \leq 0 \\ \frac{1}{2}e^{-x} & \text{si } x > 0 \end{cases}$$

cuya gráfica, que ayuda a entender mejor todas las probabilidades que tengamos que calcular a posteriori puede verse en la figura 2.6.

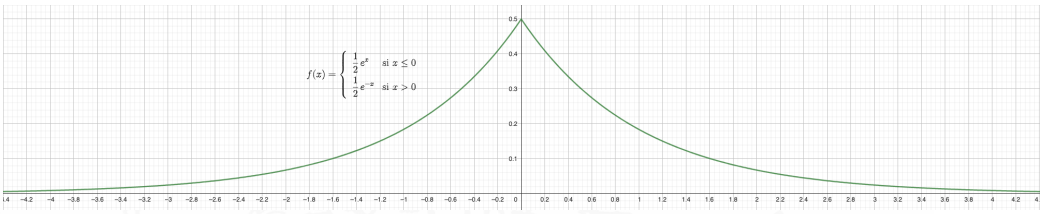


Figura 2.6: Representación gráfica de la función de densidad.

Si $x \leq 0$ tenemos que:

$$F(x) = \int_{-\infty}^{x} \frac{1}{2}e^{x}dx = \frac{1}{2}\left(e^{x}\right)_{-\infty}^{x} = \frac{1}{2}\left(e^{x} - e^{-\infty}\right) = \frac{1}{2}e^{x}$$

Si $x > 0$, como $f(x)$ es simétrica respecto al eje OY la función de distribución valdrá:

$$F(x) = \frac{1}{2} + \int_{0}^{x} \frac{1}{2}e^{-x}dx = \frac{1}{2} + \frac{1}{2}\left(-e^{-x}\right)_{0}^{x} = \frac{1}{2} + \frac{1}{2}\left(-e^{-x} - \left(-e^{0}\right)\right) =$$
$$= \frac{1}{2} + \frac{1}{2}\left(1 - e^{-x}\right) = \frac{1}{2} + \frac{1}{2} - \frac{1}{2}e^{-x} = 1 - \frac{1}{2}e^{-x}$$

2)
$$P(X \leq 1) = F(1) = 1 - \frac{1}{2}e^{-1} = 1 - \frac{1}{2e} = \frac{2e-1}{2e}$$

3) Este apartado nos pide calcular la probalidad del conjunto siguiente:

$$\{|X| \leq 2 \quad \cup \quad (X > 0)\} = \{-2 \leq X \leq 2 \quad \cup \quad (X > 0)\}$$

por lo tanto, el suceso que tendremos que utilizar es $\{-2 \leq X \leq \infty\}$

$$P(-2 \leq X \leq \infty) = P(-2 \leq X \leq 0) + \frac{1}{2} = \int_{-2}^{0} \frac{1}{2}e^{x}dx + \frac{1}{2} =$$
$$= \frac{1}{2}\left(e^{x}\right)_{-2}^{0} + \frac{1}{2} = \frac{1}{2}\left(e^{0} - e^{-2}\right) + \frac{1}{2} =$$
$$= \frac{1}{2}\left(1 - e^{-2}\right) + \frac{1}{2} = \frac{1}{2} - \frac{1}{2e^{2}} + \frac{1}{2} = 1 - \frac{1}{2e^{2}}$$



4) Si sabemos que $X > 1$, la probabilidad de que la variable aleatoria X sea menor que 2 será:

$$P(X < 2/X > 1) = \frac{P(X < 2 \cap X > 1)}{P(X > 1)} = \frac{P(1 < X < 2)}{P(X > 1)} = \frac{\int_1^2 \frac{1}{2}e^{-x}}{\int_1^\infty \frac{1}{2}e^{-x}} =$$
$$= \frac{(-e^{-x})_1^2}{(-e^{-x})_1^\infty} = \frac{-e^{-2} + e^{-1}}{-e^{-\infty} + e^{-1}} = \frac{-e^{-2} + e^{-1}}{e^{-1}} = 1 - \frac{\frac{1}{e^2}}{\frac{1}{e}} =$$
$$= 1 - \frac{1}{e} = \frac{e-1}{e}$$

5) Al ser la variable aleatoria simétrica respecto del origen, si existe, la esperanza vale cero. Se puede calcular de la forma:

$$E(X) = \int_{-\infty}^{0} x\frac{1}{2}e^{x}dx + \int_{0}^{\infty} x\frac{1}{2}e^{-x}dx$$

y calculando las integrales por partes, por ejemplo, y operando se suprimen ambos sumandos y queda cero.

La matización anterior es necesaria porque hay distribuciones, como la de Cauchy, que son simétricas respecto del origen y no tienen momento de orden 1, es decir, no existe su esperanza.

Ejercicio 2

Una fábrica produce determinados artículos que se agrupan en cajas de 15 unidades. En un control de defectuosos de una partida de 140 cajas se obtienen los siguientes datos



N. artículos defectuosos	N de cajas
0	57
1	57
2	18
3	5
4	3
5 o más	0

Calcule:

1. El número medio de defectuosos por caja
2. La varianza y la desviación típica del número de defectuosos por caja
3. La mediana, el primer y tercer cuartil y el percentil del 95 % del número de defectuosos por caja
4. La probabilidad de que un artículo sea defectuoso

Solución.

Utilizaremos la variable aleatoria $X =$ "número de artículos defectuosos por caja". Antes de nada, hagamos la tabla con el acumulado de X.

n. artículos defectuosos	n cajas (n_i)	N_i
0	57	57
1	57	114
2	18	132
3	5	137
4	3	140
≥ 5	0	140



1) Para calcular el número medio de defectuosos por caja promediamos la muestra, obteniendo:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{140}\big(57\cdot 0 + 57\cdot 1 + 18\cdot 2 + 5\cdot 3 + 3\cdot 4\big) = \frac{120}{140} = \frac{6}{7}$$

2) La varianza y la desviación típica del número de defectuosos por caja podría estimarse con la cuasivarianza muestral que es un estimador insesgado de la varianza poblacional, aunque el valor sería muy parecido que calculando la varianza muestral, que por simplicidad vale:

$$\begin{aligned}
&\widehat{\mathrm{V}(X)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \\
&= \frac{1}{140}\left(57\cdot\left(0-\frac{6}{7}\right)^2 + 57\cdot\left(1-\frac{6}{7}\right)^2 + 18\cdot\left(2-\frac{6}{7}\right)^2\right) + \\
&+ \frac{1}{140}\left(5\cdot\left(3-\frac{6}{7}\right)^2 + 3\cdot\left(4-\frac{6}{7}\right)^2\right) \\
&\approx 0{,}851020
\end{aligned}$$

En cuanto a la desviación típica muestral:

$$\widehat{\sigma(X)} = \sqrt{\widehat{\mathrm{V}(X)}} \approx 0{,}922508.$$

3) La mediana es el percentil 50 %, P_{50}, al haber $n = 140$ datos se alcanza en la posición $140\frac{50}{100} = 70$. Ordenados los valores muestrales, en la posición 70 hay un 1.

De forma análoga $Q_1 = P_{25}$ es el valor alcanzado en la posición $140\frac{25}{100} = 35$, que es un 0, con lo que $Q_1 = 0$. Q_3 se alcanza en la posición 105, así que $Q_3 = 1$, y finalmente P_{95} se alcanza en la posición 133, resulta $P_{95} = 3$.

4) En este apartado se ignoran las cajas, se consideran los artículos sueltos.

$$\mathbb{P}(\text{“un articulo es defectuoso”}) = \frac{\text{“num articulos defectuosos”}}{\text{“num articulos”}} =$$



$$= \frac{57 \cdot 0 + 57 \cdot 1 + 18 \cdot 2 + 5 \cdot 3 + 3 \cdot 4}{140 \cdot 15} =$$
$$= \frac{120}{140 \cdot 15} = \frac{2}{35} \approx 0{,}05714.$$

También puede obtenerse dividiendo el número medio de defectos por caja, $\overline{x}$, por el de elementos en cada caja, es decir,

$$\mathbb{P}(\text{“un articulo es defectuoso”}) = \frac{\overline{x}}{15}.$$

Ejercicio 3

Sea $X_1, \ldots, X_n$ una muestra aleatoria simple de una población que sigue una distribución Normal $N(0, \sigma^2)$ siendo σ^2 desconocido. Calcular:

1. El estimador máximo verosímil de σ^2
2. ¿Es un estimador insesgado?
3. Intervalo de confianza de nivel 0,95 para σ^2. Indique el intervalo en función de los valores de la distribución del estimador.

Solución.

1) La función de densidad de la $N(0, \sigma^2)$ es

$$h_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}.$$

La función de verosimilitud de la m.a.s. es por tanto

$$f_\sigma(x_1, \ldots, x_n) = \prod_{i=1}^{n} h_\sigma(x_i) = \frac{1}{\sigma^n \sqrt{2\pi}^n} e^{-\frac{1}{2}\frac{\sum_{i=1}^n x_i^2}{\sigma^2}}.$$



Para obtener el EMV lo que hacemos es buscar el valor de σ que maximiza esta expresión. Es posible ignorar el factor $\sqrt{2\pi}^n$ sin que afecte al máximo buscado, al igual que tomar logaritmos, con lo que buscamos el máximo en σ de

$$L_\sigma(x_1, \ldots, x_n) = -n \ln \sigma - \frac{1}{2}\sigma^{-2} \sum_{i=1}^{n} x_i^2.$$

Al ser la expresión derivable, el máximo se alcanzará en algún punto donde la derivada parcial respecto a σ se anule:

$$\frac{\partial L_\sigma(x_1, \ldots, x_n)}{\partial \sigma} = -n\frac{1}{\sigma} - \frac{1}{2} \cdot (-2) \cdot \sigma^{-3} \sum_{i=1}^{n} x_i^2 = 0.$$

Multiplicando por σ^3:

$$-n\sigma^2 + \sum_{i=1}^{n} x_i^2 = 0,$$

de donde al despejar σ^2 la única solución da lugar al estimador

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

2) Para ver si el estimador es sesgado o no calculemos su esperanza.

$$\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{\sum_{i=1}^n X_i^2}{n}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{E}(X_i^2)\right) = \mathbb{E}(X_1^2) = \sigma^2$$

Así que el sesgo es $\mathbb{E}(\hat{\sigma}^2) - \sigma^2 = 0$, el estimador es en efecto insesgado.

3) Para buscar un intervalo de confianza para σ^2, utilizamos que $X_1, \ldots, X_n$ son independientes, y si se dividen por σ son normales estándar. Por tanto

$$\sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \sim \chi_n^2.$$

Con esto, y seleccionando los puntos de la distribución de la χ_n^2 que acumulan áreas de $\frac{\alpha}{2}$ y $1 - \frac{\alpha}{2}$, es claro que



$$\mathbb{P}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2 \in (\chi^2_{n;1-\frac{\alpha}{2}}, \chi^2_{n;\frac{\alpha}{2}})\right) = 1-\alpha,$$

donde definimos $\chi^2_{n;x}$ como el punto tal que una χ^2_n acumula por encima de él exactamente una probabilidad x.

Ahora bien,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i^2 \in (\chi^2_{n;1-\frac{\alpha}{2}}, \chi^2_{n;\frac{\alpha}{2}})$$

sucede sí y solo si (invirtiendo)

$$\frac{\sigma^2}{\sum_{i=1}^{n} X_i^2} \in \left(\frac{1}{\chi^2_{n;\frac{\alpha}{2}}}, \frac{1}{\chi^2_{n;1-\frac{\alpha}{2}}}\right)$$

lo que equivale a

$$\sigma^2 \in \left(\frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{n;\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{n;1-\frac{\alpha}{2}}}\right)$$

Hemos demostrado que

$$\mathbb{P}\left(\sigma^2 \in \left(\frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{n;\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{n;1-\frac{\alpha}{2}}}\right)\right) = 1-\alpha,$$

con lo que $I^{1-\alpha}_{\sigma^2} = \left(\frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{n;\frac{\alpha}{2}}}, \frac{\sum_{i=1}^{n} X_i^2}{\chi^2_{n;1-\frac{\alpha}{2}}}\right)$ es un intervalo de confianza válido para σ^2 al nivel de confianza $1-\alpha$. Eso sí, no es el de longitud mínima.

Ejercicio 4

En una federación deportiva el número de jugadores y el de entrenadores se distribuye según se indica en esta tabla de frecuencias:

Años	Jugadores (en miles) X_i	Entrenadores (en miles) Y_i
2017	80	3,1
2018	85	3,3
2019	93	5,2
2020	105	6,3
2021	116	7,1

Se pide:

1. Ecuación de la recta de regresión de Y sobre X
2. Interpretación del coeficiente de regresión de Y sobre X
3. Calcular el coeficiente de correlación lineal de Pearson. Interpretación.
4. Calcular el coeficiente de determinación. Interpretación.
5. Predicción del número de entrenadores en 2022 para un número de jugadores estimados en ese año de 125 000.

Solución.

1) La ecuación de la recta de regresión viene dada por

$$y - \overline{y} = \frac{S_{xy}}{S_x^2}(x - \overline{x}),$$

donde

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y}$$

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

Los cálculos parciales son

$$\overline{x} = \frac{80+85+93+105+116}{5} = 95{,}8$$
$$\overline{y} = \frac{3{,}1+3{,}3+5{,}2+6{,}3+7{,}1}{5} = 5$$
$$S_{xy} = \frac{3{,}1\cdot 80 + 3{,}3\cdot 85 + 5{,}2\cdot 93 + 6{,}3\cdot 105 + 7{,}1\cdot 116}{5} - \overline{x}\cdot\overline{y} = 20{,}44$$
$$S_x^2 = \frac{80^2+85^2+93^2+105^2+116^2}{5} - \overline{x}^2 = 173{,}36$$

de donde resulta la ecuación para la recta de regresión

$$f(x) = 0{,}117905x - 6{,}295293$$

La representamos para comprobar que proporciona un buen ajuste a los 5 puntos conocidos, en la Figura 2.7.

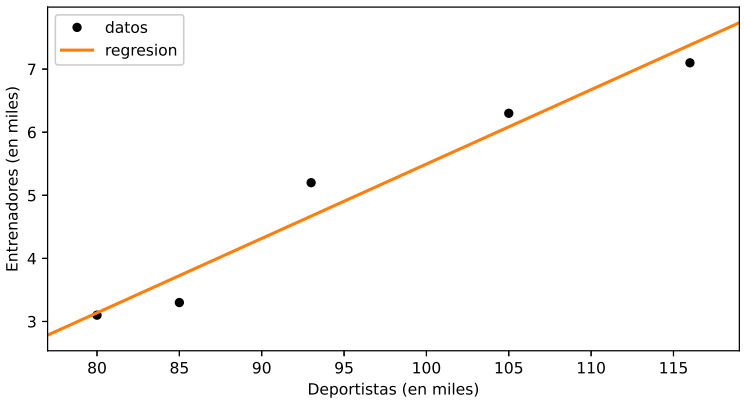


Figura 2.7: Regresión Lineal para los datos conocidos.

2) El coeficiente de la regresión de Y sobre X es la pendiente de la recta de regresión. Asumiendo que ese modelo se ajusta bien a los datos, da la cantidad adicional de miles de entrenadores cuando el número de jugadores aumenta en un millar. Lo que es lo mismo, la cantidad adicional de entrenadores cuando el número de jugadores aumenta en una unidad.

3) El coeficiente de correlación lineal de Pearson viene dado por

$$r = \frac{S_{xy}}{S_x \cdot S_y},$$

donde análogamente a como definimos para la variable X,

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2 = 2{,}528.$$

El resultado que se obtiene es

$$r = \frac{S_{xy}}{S_x \cdot S_y} \approx \frac{20{,}44}{13{,}166624 \cdot 1{,}589968} \approx 0{,}976377.$$

Un valor tan cercano a la unidad indica que la relación entre las variables es fuertemente lineal y positiva, en el sentido de que ambas crecen y decrecen juntas.

4) En general el coeficiente de determinación R^2 viene dado por

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - f(x_i))^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}.$$

Pero también coincide con el cuadrado del coeficiente de Pearson, que ya tenemos calculado. Así, $R^2 = r^2 = 0{,}953313$, lo que indica que más del 95 % de la variación de la variable dependiente Y queda explicada por la variable independiente X.



5) La predicción viene simplemente de aplicar la recta de regresión como función. El dato de que es para el año 2022 es irrelevante, lo esencial es que se espera $x_6 = 125$ y por tanto la predicción según la regresión es

$$f(x_6) = 0{,}117905 \cdot 125 - 6{,}295293 = 8{,}442824.$$

Puesto que estamos hablando de entrenadores, que es un número natural, serán 8443 entrenadores según este modelo de regresión.

Ejercicio 5

El mercado de altramuces en el país B tiene las siguientes curvas de oferta y demanda:

$$X_{Oferta} = 1000 + 400 \cdot P$$
$$X_{Demanda} = 2000–200 \cdot P$$

Nota: Los altramuces son legumbres características de la zona Mediterránea. Estas semillas comestibles se tomaban habitualmente como aperitivo.

1) Razone si el paradigma del mercado de altramuces es un monopolio, oligopolio o un mercado de competencia perfecta.

2) Determine la expresión matemática de la curva de demanda individual de cada consumidor y la oferta individual de cada productor, sabiendo que en el mercado hay 200 consumidores idénticos y 50 productores con las mismas características.

3) Calcule el punto de equilibrio del mercado.

4) Si el precio se fija en 2 unidades monetarias indique que ocurrirá en el mercado.



5) Represente en un solo gráfico las respuestas a los dos apartados anteriores.

Solución.

1) Dadas las caracterísitcas del enunciado podemos decir que el mercado de altramuces es un mercado de competencia perfecta. La competencia perfecta es un tipo de mercado caracterizado por tener un número grande de productores, un producto estandardizado y homogéneo y la libre entrada y salida de vendedores en el mercado. Todas estas caracterísitcas podemos verlas en la descripción del mercado en el enunciado: los altramuces son un bien completamente indistinguible uno de otro, es decir para el cliente no importa qué empresa específica venda el producto ya que todos son idénticos, existen 50 empresas productoras con características similares y no hay grandes barreras a la entrada y salida de productores en el mercado.

2) La curva de demanda de mercado se construye sumando en cantidades las curvas de demanda individuales. Dado que en el enunciado se proporciona la expresión de la curva de demanda de mercado, $X_D = 2000 - 200p$, así como el número de consumidores, 200, para calcular la expresión de la curva de demanda individual se divide la demanda de mercado por el número de consumidores:

$$x_d = \frac{X_D}{200} = \frac{2000 - 200p}{200} = 10 - p.$$

La curva de oferta de mercado se construye sumando en cantidades las curvas de oferta individuales. Dado que en el enunciado se proporciona la expresión de la curva de oferta del mercado de los altramuces, $X_S = 1000 + 400p$, así como el número de productores, 50, para calcular la



expresión de la curva de oferta individual se divide la oferta de mercado por el número de productores:

$$x_s = \frac{X_S}{50} = \frac{1000 + 400p}{50} = 20 + 8p.$$

3) La conidción de equilibrio del mercado supone que para el precio de equilibrio, p_x^{eq}, las cantidades demanda y ofrecida se igualan, vaciando el mercado: $X_D = X_S = X^{eq}$. Igualnado las cantidades demanda y ofertada se obtiene el precio de equilibrio.

$$\begin{aligned} X_D &= X_S \\ 2000 - 200p &= 1000 + 400p \\ 2000 - 1000 &= 400p + 200p \\ 1000 &= 600p \\ p_x^{eq} &= \frac{10}{6} \simeq 1{,}66 u.m. \end{aligned}$$

La cantidad de equilibrio se obtiene sustituyendo el precio de equilibrio en la curva de demanda de mercado o en la curva de oferta de mercado.

$$X_D^{eq} = 2000 - 200p_x^{eq} = 2000 - 200 \cdot \frac{10}{6} = \frac{5000}{3} \simeq 1666{,}66 uds.$$

$$X_S^{eq} = 1000 + 400p_x^{eq} = 1000 + 400 \cdot \frac{10}{6} = \frac{5000}{3} \simeq 1666{,}66 uds.$$

4) Si el precio se fija en 2 unidades monetarias,

$$\begin{aligned} X_D(2) &= 2000 - 200 \cdot 2 = 1600 \\ X_S(2) &= 1000 + 400 \cdot 2 = 1800, \end{aligned}$$

y por tanto, dado que $X_D = 1600 < X_S = 1800 \Rightarrow X_S - X_D = 1800 - 1600 = 200$, en el mercado se produce una situación de exceso de oferta de doscientas unidades.

5) Si se representan gráficamente las respuestas a los dos apartados anteriores resulta la figura 2.8.

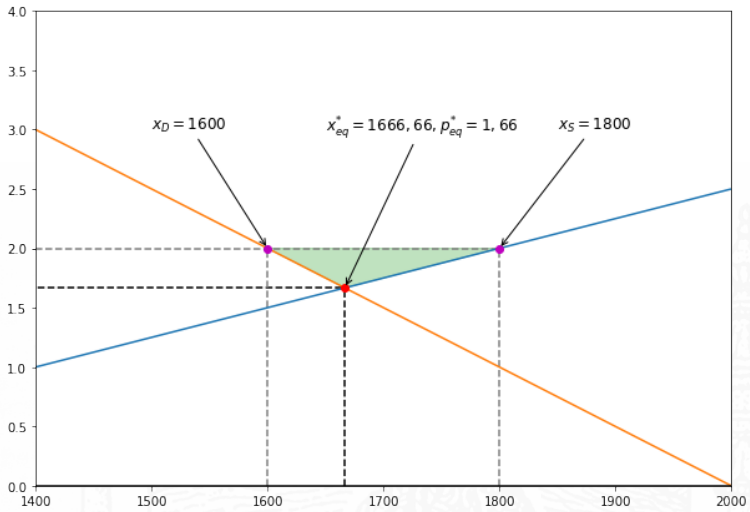


Figura 2.8: Representación gráfica apartados 3 y 4.

El punto en rojo refleja el equilibrio del apartado 3 y el área sombreada de color verde refleja el exceso de oferta de mercado.

Ejercicio 6

A partir de los datos de la siguiente tabla de mortalidad referida a 2020, completar las celdas sombreadas (*). Las tasas de mortalidad y los riesgos de muerte en la tabla están expresados en tantos por mil.

(*) Nota: Expresar los resultados redondeados a dos decimales.



Año 2020	Tasa de mortali-dad	Promedio de años vividos el último año de vida	Riesgo de muerte	Supervi-vientes	Defuncion es teóricas	Población estaciona-ria	Tiempo por vivir	Esperanza de vida
Ambos sexos								
0 años	2,52	0,13	2,51	(1)	251,40	(5)	8.232.951,82	82,33
1 año	0,20	0,45	0,20	(2)	(3)	99.737,61	8.133.171,13	81,54
2 años	0,10	0,49	0,10	99.728,70	10,39	99.723,43	(4)	80,55
3 años	0,11	0,52	0,11	99.718,31	10,73	99.713,18	7.933.710,08	(6)
4 años	(7)	(9)	0,05	99.707,59	5,03	99.705,15	7.833.996,90	78,57
5 años	0,07	0,53	(8)	99.702,55	6,49	99.699,51	7.734.291,75	77,57
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
100 y más años	490,28	1,97	(10)	2.177,59	(11)	4.282,44	(12)	1,97

Solución.

1) Los supervivientes a los cero años son $l_0 = 100\,000$

2) $l_1 = 100000 - 251{,}40 = 99\,748{,}6$

3) $d_1 = l_1 \cdot q_1 = 99748{,}6 \cdot 0{,}0002 = 19{,}949$

4) $T_2 = T_1 - L_1 = 8133171{,}13 - 99737{,}61 = 8\,033\,433{,}52$

5) $L_0 = l_1 + d_0 \cdot a_0 = 99748{,}6 + 251{,}40 \cdot 0{,}13 = 99\,781{,}28$

6) $e_3 = \frac{T_3}{l_3} = \frac{7933710{,}08}{99718{,}31} = 79{,}56$

7) $m_4 = \frac{d_4}{L_4} \cdot 1000 = \frac{5{,}03}{99705{,}15} \cdot 1000 = 0{,}0504‰$

8) $q_5 = \frac{d_5}{l_5} \cdot 1000 = \frac{6{,}49}{99702{,}55} \cdot 1000 = 0{,}065‰$

9) Tenemos que $L_4 = l_5 + d_4 \cdot a_4$, luego despejando tendremos que

$a_4 = \frac{L_4 - l_5}{d_4} = \frac{99705{,}15 - 99702{,}55}{5{,}03} = 0{,}5169$

10) La probabilidad de muerte en el último intervalo de edad, que comprende a todas las personas que tengan 100 o más años es 1, puesto que en este intervalo todos morirán, por lo tanto $q_{100} = 1$

11) Las defunciones teóricas en este último intervalo coinciden con los supervivientes, es decir, $d_{100+} = 2177{,}59$

12) $T_{100+} = L_{100+} = 4282{,}44$, puesto que en el último intervalo de la tabla de mortalidad estas dos series biométricas coinciden.

Ejercicio 7

A partir de la información de las rúbricas SEC2010 de una economía imaginaria que se indican en la siguiente tabla:

Rúbrica SEC2010	Unidades Monetarias. Precios corrientes de 2021
Consumo individual efectivo	1.340
Formación neta de capital	195
Exportaciones de bienes y servicios	500
Excedente Bruto de Explotación / Renta Mixta Bruta	790
VAB a precios básicos	2.150
Consumo colectivo efectivo	660
Consumo de capital fijo	50
Otros impuestos netos sobre la producción	290
Variación de existencias	15
Impuestos sobre productos e importaciones	455
Adquisiciones menos cesiones de objetos valiosos	25
Producción	3.280
Importaciones de bienes y servicios	200
Consumos intermedios	1.130

Calcule:

1) El gasto en consumo final

2) La formación bruta de capital fijo

3) El saldo neto exterior de bienes y servicios

4) Sabiendo que el PIB a precios de mercado de 2021 valorado a precios constantes de 2015 es de 2120 unidades monetarias, cuál sería el deflactor implícito del PIB para 2021 en base 2015=100

5) El valor añadido bruto a precios de mercado (o precios de adquisición)

6) Los impuestos netos sobre productos e importaciones



7) Las subvenciones a los productos

8) La remuneración de los factores de producción

9) La remuneración de los asalariados

10) Compruebe que es posible determinar el PIB a precios de mercado valorado a precios corrientes de 2021 por las tres vías: producción, gasto y renta.

Solución.

1) Sabemos que para el total de la economía, el gasto en consumo final(GCF) coincide con el consumo final efectivo(GCFE), que es la suma del consumo individual efectivo(GCIE) y el consumo colectivo efectivo(GCCE), es decir,
$GCF = GCFE = GCFI + GCFC = GCIE + GCCE = 1340 + 660 = 2000um$

2) $FBC = FBCF + VE + (Adq - Ces\ de\ objetos\ valiosos)$
Primeramente calculamos la formación bruta de capital, que se obtiene como la suma del consumo de capital fijo y la formación neta de capital, que es un dato del problema:
$FBC - CCF = FNC$, es decir que $FBC = 50 + 195 = 245$, por lo tanto $FBCF = 245 - 15 - 25 = 205$

3) El saldo neto exterior de bienes y servicios es la diferencia entre exportaciones e importaciones, es decir:
$SNE = Exportaciones - Importaciones = 500 - 200 = 300$

4) Tenemos que ${}_{base15}PIB_{pm}^{2021} = 2120$ y si calculamos el PIB a precios corrientes en 2021 queda:
$PIB_{pm}^{2021} = GCF + FBC + X - M = 2000 + 245 + 300 = 2545$, por lo

tanto el deflactor pedido será:

$IPC_{2015}^{2021} = \frac{PIB_{pm}^{2021}}{_{base15}PIB_{pm}^{2021}} \cdot 100 = \frac{2545}{2120} \cdot 100 = 120{,}047$

5) Este apartado se contesta directamente, puesto que el $VAB_{pm} = PIB_{pm}$ y por lo tanto $VAB_{pm} = 2545$.

No obstante, vamos a tratar de obtenerlo de otra forma y comprobar que el resultado es el mismo. Primeramente vamos a calcular los impuestos netos sobre los productos, es decir, D.21-D.31, mediante la cuenta de bienes y servicios de la economía:

R	Cuenta de bienes y servicios		E
3280	Producción (P.1)	Consumos intermedios (P.2)	1130
X	Imps. netos sobre los	Gasto en consumo final (P.3)	2000
	productos (D.21 - D.31)	Formación bruta de capital (P.5)	245
200	Importaciones de bienes y	Exportaciones de bienes y	500
	servicios (P.7)	servicios (P.6)	
3875			3875

Despejando de dicha cuenta obtenemos que D.21-D.31 = 395, y por lo tanto:

$VAB_{pm} = VAB_{pb} + (D.21 - D.31) = 2150 + 395 = 2545$

6) Este apartado pide los impuestos netos sobre productos e importaciones, es decir, $D.21 - D.31$, que aunque ya ha sido calculado antes y valía 395, se podría calcular también utilizando la cuenta de producción, según la cual:

$PIB_{pm} = P - CI + (D.21 - D.31)$ y despejando de esta fórmula tenemos que: $D.21 - D.31 = PIB_{pm} - P + CI = 2545 - 3280 + 1130 = 395$



También podríamos calcular, aunque no es objeto de este apartado, los impuestos sobre la producción y las importaciones, es decir, $D.2 - D.3$. Veamos:

$$\begin{aligned}&\text{Impuestos Netos sobre Producción e Importaciones} = D.2 - D.3 = \\ &= D.21 + D.29 - D.31 - D.39 = (D.21 - D.31) + (D.29 - D.39) = \\ &= 395 + 290 = 685\end{aligned}$$

puesto que el valor 290 es un dato del enunciado y se corresponde con otros impuestos netos sobre la producción.

7) Las subvenciones a los productos se calculan de forma sencilla despejando de los impuestos netos sobre los productos:

$$\begin{aligned}\text{Subvenciones a los productos(D.31)} &= -395 + \\ &+ \text{Impuestos sobre los productos(D.21)} = \\ &= -395 + 455 = 60\end{aligned}$$

8) La remuneración de los factores de producción consisten en la remuneración del trabajo y del capital. Los salarios serían la remuneración correspondiente al factor trabajo, y el valor añadido que no se utiliza en pagar salarios quedaría como remuneración del capital, es decir, el EBE/RMB. El PIB por la vía de la renta es la suma de las rentas que perciben el trabajo y el capital por su contribución a los procesos productivos, a lo que hay que añadir los impuestos netos sobre producción y la importación para obtener la magnitud a precios de adqusición. Luego:

$$\begin{aligned}PIB_{pm} &= RA + (EBE/RMB) + \\ &+ \text{Impuestos netos sobre producción e importaciones(D.2-D.3)}\end{aligned}$$



y así:

$$RA + (EBE/RMB) = PIB - (D.2 - D.3) =$$
$$= 2545 - 685 = 1860$$

9) La remuneración de asalariados(RA) de despeja fácilmente de la expresión del apartado anterior:
$RA = 1860 - (EBE/RMB) = 1860 - 790 = 1070$

10) Calculemos el PIB a precios de mercado valorados a precios de 2021 por las tres vías:

$$PIB^{2021}_{producción} = VAB_{pb} + \text{impuestos netos sobre los productos(D.21-D.31)} =$$
$$= 2150 + 395 = 2545$$
$$PIB^{2021}_{renta} = RA + (EBE/RM) +$$
$$+ \text{impuestos netos sobre la producción e importaciones(D.2 - D.3)} =$$
$$= 1070 + 790 + 685 = 2545$$
$$PIB^{2021}_{gasto} = GCF + FBC + (X - M) =$$
$$= 2000 + 245 + 300 = 2545$$

Ejercicio 8

Una población de tamaño $N = 500$ se divide en dos estratos. Obtenemos una muestra estratificada sin reemplazamiento de tamaño $n = 50$ con los siguientes datos:



Estrato	N_h	n_h	$\overline{x}_h$	S_h
1	100	15	30	15
2	400	35	10	8

Obtener:

1) La estimación de la media poblacional

2) La estimación de la varianza de la media anterior

3) ¿Son los tamaños muestrales seleccionados en cada estrato compatibles con la afijación óptima?

4) Si consideramos además los costes de entrevista por unidad en cada estrato $C_1 = 25$ y $C_2 = 4$, ¿cuál sería la afijación optima para el tamaño muestral total de $n = 50$?

Solución.

Se trata de un problema de muestreo estratificado sin reemplazamiento de tamaño $n = 50$, siendo $N = 500$ el tamaño de la población. Por tanto,

1) La estimación de la media poblacional viene dada por

$$\widehat{\overline{x}} = \sum_{h=1}^{2} W_h \overline{x}_h = \frac{1}{5} \cdot 30 + \frac{4}{5} \cdot 10 = 6 + 8 = 14.$$

2) La estimación de la varianza de la media anterior es

$$\widehat{V(\overline{x})} = \sum_{h=1}^{2} W_h^2 \left(1 - f_h\right) \frac{S_h^2}{n_h} = \left(\frac{1}{5}\right)^2 \cdot \left(1 - \frac{15}{100}\right) \cdot \frac{15^2}{15} + \\ + \left(\frac{4}{5}\right)^2 \cdot \left(1 - \frac{35}{400}\right) \cdot \frac{8^2}{35} = 1{,}577.$$



3) La afijación óptima, en el supuesto de que el coste es constante en todos los estratos, viene dada por

$$n_h = n \cdot \frac{N_h S_h}{\sum_{h=1}^{2} N_h S_h}$$

Por tanto, para el primer y el segundo estrato respectivamente, se tiene

$$n_1 = 50 \cdot \frac{100 \cdot 15}{100 \cdot 15 + 400 \cdot 8} = 15{,}95$$
$$n_2 = 50 \cdot \frac{400 \cdot 8}{100 \cdot 15 + 400 \cdot 8} = 34{,}04$$

De este modo, los tamaños muestrales seleccionados en cada estrato no son compatibles con la afijación óptima. La afijación óptima para el tamaño muestral $n = 50$ sería $n_1 = 16$ y $n_2 = 34$.

4) Teniendo en cuenta ahora los costes, la afijación óptima es

$$n_h = n \cdot \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{2} N_h S_h / \sqrt{c_h}}$$

Por tanto, para el primer y el segundo estrato respectivamente, se tiene

$$n_1 = 50 \cdot \frac{(100 \cdot 15)/\sqrt{25}}{(100 \cdot 15)/\sqrt{25} + (400 \cdot 8)/\sqrt{4}} = 7{,}8947$$
$$n_1 = 50 \cdot \frac{(400 \cdot 8)/\sqrt{4}}{(100 \cdot 15)/\sqrt{25} + (400 \cdot 8)/\sqrt{4}} = 42{,}1055$$

Por tanto, los tamaños muestrales compatibles con la afijación óptima son $n_1 = 8$ y $n_2 = 42$.

Ejercicio 9

Se dispone de los siguientes datos de una población.



Grupos de edad	Nacimientos en 2020 por edad de la madre		Población residente a 1 de enero de 2020		Población residente a 1 de enero de 2021	
	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres
<15	0	0	6720	6800	6800	6512
15-19	16	10	3244	3150	3250	3165
20-24	35	30	3192	2914	4000	2956
25-29	32	40	2800	2880	2865	2889
30-34	45	47	3096	2853	3098	2855
35-39	58	55	2950	2900	2876	2942
40-44	38	32	2970	3013	2983	3000
45-49	30	25	3050	3140	3013	3150
50-54	0	0	3200	3220	3390	3267
55-59	0	0	3091	3131	3035	3100
60+	0	0	4400	4600	4400	4600
TOTAL	254	239	38713	38601	39710	38436

Calcular:

1) La tasa bruta de natalidad

2) La tasa global de fecundidad

3) La tasa específica de fecundidad para el grupo 35-39 años

4) La tasa específica de fecundidad de niñas para el grupo 20-24

5) El ratio de masculinidad al nacimiento

Solución.

1) La tasa bruta de natalidad viene dada por

$$TBN^t = \frac{N^t}{P^t_{med}} \cdot 1000$$

donde N^t es el número de nacimientos de madre residente en dicha población en el año t y P^t_{med} es la población residente media en el año t. Así,

$$TBN^{20} = \frac{N^{20}}{P^{20}_{med}} \cdot 1000 = \frac{N^{20}}{\frac{P^{20}+P^{21}}{2}} \cdot 1000 =$$
$$= \frac{254+239}{\frac{(38713+38601)+(39710+38436)}{2}} \cdot 1000 = \frac{493}{77730} \cdot 1000 = 6{,}342‰.$$

2) La tasa global de fecundidad mide la frecuencia de nacimientos entre las mujeres en edad fértil. Se calcula según

$$TGF^t = \frac{N^t}{M^t_{med,15-49}} \cdot 1000,$$

donde N^t es el número de nacimientos de madre residente en la población en el año t y $M^t_{med,15-49}$ es la población media de mujeres de entre 15 y 49 años en el año t. Así,

$$TGF^{20} = \frac{N^{20}}{M^{20}_{med,15-49}} \cdot 1000 \frac{N^{20}}{\frac{M^{20}_{15-49}+M^{21}_{15-49}}{2}} \cdot 1000 =$$
$$= \frac{254+239}{\frac{(3150+\cdots+3140)+(3165+\cdots+3150)}{2}} \cdot 1000 = \frac{493}{\frac{20850+20957}{2}} \cdot 1000 =$$
$$= \frac{493}{20903,5} \cdot 1000 = 23{,}584‰.$$

3) La tasa específica de fecundidad para el grupo 35-39 relaciona los nacimientos correspondientes a madres de entre 35 y 59 años con la población media de mujeres de esa misma edad. Así,

$$TEF^t_{35-39} = \frac{N^t_{35-39}}{M^t_{med,35-39}} \cdot 1000$$

donde N^t_{35-39} es el número de nacimientos correspondientes a madres de entre 35 y 39 años de edad en el año t y $M^t_{med,35-39}$ es la población media de mujeres de entre 35 y 39 años en el año t. Por tanto,

$$TEF^{20}_{35-39} = \frac{N^{20}_{35-39}}{M^{20}_{med,35-39}} \cdot 1000 = \frac{N^{20}_{35-39}}{\frac{M^{20}_{35-39}+M^{21}_{35-39}}{2}} \cdot 1000 =$$



$$= \frac{58+55}{\frac{2900+2942}{2}} \cdot 1000 = \frac{113}{\frac{5842}{2}} \cdot 1000 = 38{,}685‰.$$

4) La tasa específica de fecundidad de niñas para el grupo 20-24 relaciona las niñas nacidas de madres de entre 20 y 24 años con la población media de mujeres de esa misma edad. Así,

$$_{niñas}TEF^{t}_{20-24} = \frac{_{niñas}N^{t}_{20-24}}{M^{t}_{med,20-24}} \cdot 1000$$

donde $_{niñas}N^{t}_{20-24}$ es el número de niñas nacidas de madres de entre 20 y 24 años de edad en el año t y $M^{t}_{med,20-24}$ es la población media de mujeres de entre 20 y 24 años en el año t. Por tanto,

$$_{niñas}TEF^{20}_{20-24} = \frac{_{niñas}N^{20}_{20-24}}{M^{20}_{med,20-24}} \cdot 1000 = \frac{_{niñas}N^{20}_{20-24}}{\frac{M^{20}_{20-24}+M^{21}_{20-24}}{2}} \cdot 1000 =$$
$$= \frac{30}{\frac{2914+2956}{2}} \cdot 1000 = \frac{30}{\frac{5870}{2}} \cdot 1000 = 10{,}221‰.$$

5) El ratio de masculinidad al nacimiento se define como el total de nacimientos de niños por cada 100 nacimientos de niñas, registrados entre los nacimientos de madres pertenecientes a la población en el año t. Se calcula usando

$$RMN^{t} = \frac{N^{t}_{hombres}}{N^{t}_{mujeres}} \cdot 100$$

donde $N^{t}_{hombres}$ es el número de nacimientos de hombres en el año t y $N^{t}_{mujeres}$ es el número de nacimientos de mujeres en el año t. Por tanto, para el caso que nos ocupa,

$$RMN^{20} = \frac{N^{20}_{hombres}}{N^{20}_{mujeres}} \cdot 100 = \frac{254}{239} \cdot 100 = 106{,}276\,\%,$$

es decir, en el año 2020 por cada 100 mujeres nacen 106 hombres en la población.



Ejercicio 10

El índice al Comercio al por Menor (ICM) es un índice tipo Laspeyres encadenado a diciembre. La serie en base 2015 para los años 2018 y 2019 es la siguiente:

Periodo	ICM precios corrientes	Periodo	ICM precios corrientes
2018M01	107,932	2019M01	110,592
2018M02	92,094	2019M02	94,749
2018M03	103,085	2019M03	104,867
2018M04	101,836	2019M04	105,41
2018M05	106,685	2019M05	110,619
2018M06	111,649	2019M06	111,75
2018M07	115,606	2019M07	120,931
2018M08	106,53	2019M08	109,39
2018M09	104,006	2019M09	106,667
2018M10	110,587	2019M10	112,339
2018M11	110,459	2019M11	113,286
2018M12	127,888	2019M12	130,688

La muestra utilizada para el cálculo del índice en el año 2020 fue de 12 006 empresas. Conocemos los siguientes datos adicionales:



	Facturación en miles de euros
Volumen de facturación en comercio al por menor en enero de 2020 con una muestra de 12 006 empresas	185 776,98
Volumen de facturación en comercio al por menor en diciembre de 2019 con una muestra de 12 006 empresas	215 964,12
Volumen de facturación en comercio al por menor en diciembre de 2019 con una muestra de 12 158 empresas	221 476,38

1) Calcular el índice de Comercio al por Menor (ICM) en base 2015 para el periodo de enero de 2020. Justifica su elaboración. Justifica cómo lo has elaborado.

2) ¿Cómo calcularías el coeficiente de enlace para pasar esta serie a base 2019? ¿Qué condición tiene que cumplir el año base?

3) Sabiendo que el IPC en base 2015 asociado a los productos que recoge el ICM tiene como valor índice 103 813 en diciembre de 2019, calcula el índice deflactado. ¿Qué conseguimos con la deflación del índice a precios corrientes?

4) Calcule la tasa de variación mensual del índice de precios corrientes de diciembre de 2019 y explique su significado.

Solución.

1) En primer lugar, se calcula el índice no publicable (INP). El índice no publicable es el cociente entre la facturación del mes en curso y la facturación de diciembre del año anterior (en la muestra seleccionada para el año en curso).

$$_{dic,19}INP^{ene,20} = \frac{F^{ene,20}}{F^{dic,19}_{rotado}} \cdot 100 = \frac{185776{,}98}{215946{,}12} \cdot 100 = 86{,}029$$

donde $F^{ene,20}$ y $F^{dic,19}$ hacen referencia a las facturaciones del mes de enero del año 2020 y el mes de diciembre del año 2019. Por otra parte, "rotado" en el denominador hace referencia a que ambas facturaciones se han calculado con el mismo conjunto de empresas, esto es tras haber realizado la rotación, usando $n = 12006$ empresas. A continuación, se calcula el índice publicable (IP) encadenando el INP calculado a la serie existente. Así,

$$\begin{aligned} _{base15}IP^{ene,20} &= {}_{base15}IP^{dic,19} \cdot \frac{_{dic,19}INP^{ene,20}}{100} = \\ &= 130{,}688 \cdot \frac{86{,}029}{100} = 112{,}42. \end{aligned}$$

2) El año base de la serie en un índice encadenado como este es aquel en el que la media de los 12 meses se hace igual a 100. Esto es, para que el año base fuera 2019 se tiene que cumplir que

$$\frac{\sum_{m=1}^{12} {}_{base15}IP^{m,19} \cdot \text{Coeficiente de enlace}}{12} = 100.$$

De esta forma, el coeficiente de enlace para pasar esta serie a base 2019 será:

$$\begin{aligned} \text{Coeficiente de enlace} &= \frac{1200}{\sum_{m=1}^{12} {}_{base15}IP^{m,19}} = \\ &= \frac{1200}{110{,}592 + 94{,}749 + \cdots + 130{,}688} = \end{aligned}$$



$$= \frac{1200}{1331{,}288} = 0{,}90138$$

siendo ${}_{base15}IP^{m,19}$ el índice del mes m del año 2019 en base 2015. Así, el índice enlazado en base 2019 del mes m del año t sería

$${}_{base19}IE^{m,19} = {}_{base15}IP^{m,t} \cdot \frac{1200}{\sum_{m=1}^{12} {}_{base15}IP^{m,19}} = {}_{base15}IP^{m,t} \cdot 0{,}90138$$

3) El índice deflactado de diciembre de 2019 se obtiene como el cociente entre el ICM a precios corrientes y el deflactor del IPC para productos asociados, ambos en diciembre del año 2019 y en base 2015.

$$IP^{dic19}_{\text{precios constantes}} = \frac{{}_{base15}IP^{dic,19}}{{}_{base15}IPC^{dic,19}} \cdot 100 = \frac{130{,}688}{103{,}813} \cdot 100 = 125{,}888.$$

Con la deflación del índice a precios corrientes conseguimos obtener un índice de cantidades o volumen que mide la evolución de la producción de servicios de comercio al por menor en volumen, sin tener en cuenta el efecto de los precios.

4) La tasa de variación mensual del índice de precios corrientes de diciembre de 2019 es

$$TV(IP)^{dic,19} = \left(\frac{{}_{base15}IP^{dic,19}}{{}_{base15}IP^{nov,19}} - 1\right) \cdot 100 = \left(\frac{130{,}688}{113{,}286} - 1\right) \cdot 100 =$$
$$= 15{,}36\,\%.$$

Esta tasa de variación refleja que la facturación del sector, medida en euros corrientes, se ha incrementado un 15.36 % en diciembre respecto del mes anterior.



EX LIBRIS
INSTITUTO NACIONAL
DE
ESTADISTICA



Capítulo 3

Año 2022. Promoción interna.

"I play each point like my life depends on it."

—Rafael Nadal.



Pablo Picasso, born in a poor family in southern Spain in 1881, started as a child prodigy and ended as the acknowledged greatest painter of his century. After some early training with his father, a provincial drawing teacher, Picasso showed that he had thoroughly grasped naturalistic conventions–the ways that artists make a picture look realistic"–at a very young age. After some incomplete sessions of art school in Barcelona and Madrid, Picasso spent his adolescence associating with the group of Catalan modernists who gathered at Els Quatre Gats in Barcelona. From there he moved to Paris, where he quickly found like-minded poets and painters. His work began to attract serious critical attention and praise by the time he was twenty. Picasso made valuable contributions to art throughout his entire life, but it was the invention of Cubism that secured his immortality.

$2022 = 2 \cdot 3 \cdot 337$

2022 is a modest number, since divided by 22 gives 20 as remainder.

It is a Harshad number since it is a multiple of its sum of digits (6).



3.1. Soluciones del segundo examen del Cuerpo Superior de Estadísticos del Estado de 2022. Parte común. Promoción interna.

Ejercicio 1

Se quiere estimar la superficie total de olivar de una provincia en el 2023. Se dispone como marco de muestreo el listado de explotaciones agrarias del Censo Agrario 2020. Se forman tres estratos de tamaño y en cada uno de ellos se obtiene una muestra aleatoria simple sin reemplazamiento. Se dispone de la siguiente información:

Estratos	Número de explotaciones		Superficie media muestral (Ha.)	Desviación típica muestral (Ha.)
	Población (N_h)	Muestra (n_h)		
1	10200	100	2,5	1
2	5000	1200	10	4,5
3	500	250	50	15

Se pide:

- Para cada uno de los estratos:

  (a) Estime de manera insesgada el total de hectáreas (Ha) de olivar junto a su error de muestreo relativo.

  (b) Calcule los tamaños muestrales necesarios para realizar las estimaciones anteriores con unos errores de muestreo estimados inferiores



a 1000 Ha en el estrato 1, a 500 Ha en el estrato 2, y a 10 Ha en el estrato 3.

- Para el total provincial:

  (c) Estime de manera insesgada el total de hectáreas (Ha) de olivar junto a su error de muestreo relativo.

  (d) Calcule los tamaños muestrales por estrato, necesarios para realizar la estimación anterior con un error de muestreo estimado inferior a 500 Ha para el total provincial.

Solución.

(a) Para estimar de manera insesgada el total de hectáreas de olivar en cada estrato y calcular el error de muestreo relativo, se usa la fórmula del estimador insesgado para el total poblacional (T) y el error de muestreo relativo en cada estrato. El estimador insesgado del total en cada estrato es

$$\hat{T}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} Y_i = N_h \overline{y}_i$$

y su varianza correspondiente será

$$\hat{V}_{\hat{T}_h} = N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{S}_h^2}{n_h}.$$

Sin embargo, el valor del estimador de S_h^2 en la estimación de la varianza $\hat{V}_{\hat{T}_h}$ es desconocido para cada uno de los estratos, pero se puede obtener su valor a partir de la desviación típica muestral proporcionada en el enunciado, puesto que dicho estimador es la cuasivarianza muestral. Por lo que se puede reescribir la expresión anterior como:

$$\hat{V}_{\hat{T}_h} = N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{c,h}^2}{n_h} = N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h - 1},$$



puesto que sabemos que la relación entre la varianza muestral y cuasi-varianza muestral es $n_h \cdot s_h^2 = (n_h - 1) \cdot s_{c,h}^2$.

Finalmente, el error de muestreo relativo vienen dado por:

$$CV_{\hat{T}_h} = \frac{\hat{\sigma}\left(\hat{T}_h\right)}{\hat{T}_h}.$$

Por tanto, para $h = 1$:

$$\hat{T}_1 = 10200 \cdot 2{,}5 = 25500$$

su varianza

$$\hat{V}_{\hat{T}_1} = 10200^2 \left(1 - \frac{100}{10200}\right) \frac{1^2}{100 - 1} = 1040606,061,$$

y el error de muestreo relativo:

$$CV_{\hat{T}_1} = \frac{\sqrt{1040606,061}}{25500} \simeq 0{,}04 = 4\,\%.$$

Para $h = 2$:

$$\hat{T}_2 = 5000 \cdot 10 = 50000,$$

su varianza

$$\hat{V}_{\hat{T}_2} = 5000^2 \left(1 - \frac{1200}{5000}\right) \frac{4{,}5^2}{1200 - 1} = 320892,4103,$$

y el error de muestreo relativo:

$$CV_{\hat{T}_2} = \frac{\sqrt{320892,4103}}{50000} \simeq 0{,}01132 = 1{,}132\,\%.$$

Para $h = 3$:

$$\hat{T}_3 = 500 \cdot 50 = 25000,$$

su varianza

$$\hat{V}_{\hat{T}_3} = 500^2 \left(1 - \frac{250}{500}\right) \frac{15^2}{250 - 1} \cdot = 112951,8072,$$

y el error de muestreo relativo:

$$CV_{\hat{T}_3} = \frac{\sqrt{112951,8072}}{25000} \simeq 0{,}01344 = 1{,}344\,\%.$$



(b) Para calcular los tamaños muestrales necesarios para lograr errores de muestreo inferiores a ciertos valores en cada estrato, podemos utilizar la fórmula del error de muestreo estimado y despejar n_h. La fórmula del error de muestreo para cada estrato es:

$$e_h^2 = \sqrt{N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{S}_h^2}{n_h}}.$$

Despejando n_h:

$$n_h = \frac{N_h^2 \hat{S}_h^2}{e_h^2 + N_h \hat{S}_h^2}$$

Donde, de nuevo, utilizando el razonamiento del apartado anterior

$$n_h \cdot s_h^2 = (n_h - 1) \cdot s_{c,h}^2,$$

se puede reescribir la expresión del valor de n_h:

$$n_h = \frac{N_h^2 \left(\frac{n_h s_h^2}{n_h - 1}\right)}{e_h^2 + N_h \left(\frac{n_h s_h^2}{n_h - 1}\right)}$$

Utilizando esta fórmula para calcular los tamaños muestrales necesarios en cada estrato para cumplir con los errores de muestreo especificados se tiene:

$$n_1 = \frac{10200^2 \cdot \left(\frac{100 \cdot 1^2}{100-1}\right)}{1000^2 + 10200 \left(\frac{100 \cdot 1^2}{100-1}\right)} = 104{,}0191$$

$$n_2 = \frac{5000^2 \cdot \left(\frac{1200 \cdot 4{,}5^2}{1200-1}\right)}{500^2 + 5000 \cdot \left(\frac{1200 \cdot 4{,}5^2}{1200-1}\right)} = 1442{,}1364$$

$$n_3 = \frac{500^2 \cdot \left(\frac{250 \cdot 15^2}{250-1}\right)}{10^2 + 500 \cdot \left(\frac{250 \cdot 15^2}{250-1}\right)} = 499{,}5577$$

De donde se tiene que los tamaños muestrales para lograr los errores de muestreo inferiores a los proporcionados en el enunciado son $n_1 = 105$, $n_2 = 1443$ y $n_3 = 500$.



(c) Para estimar de manera insesgada el total de hectáreas de olivar para el total provincial, se usa la fórmula del estimador insesgado para el total poblacional (T) en muestreo estratificado sin reposición.. El estimador insesgado del total es

$$\hat{T} = \sum_{h=1}^{L} N_h \overline{y}_h$$

y la estimación de su varianza será

$$\hat{V}_{\hat{T}} = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{S}_h^2}{n_h} = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h - 1},$$

Sustituyendo,

$$\hat{T} = 10200 \cdot 2{,}5 + 5000 \cdot 10 + 500 \cdot 50 = 100500$$

y su varianza correspondiente será

$$\hat{V}_{\hat{T}} = 10200^2 \cdot \left(1 - \frac{100}{10200}\right) \frac{1^2}{100 - 1} + \cdots + 500^2 \cdot \left(1 - \frac{250}{500}\right) \frac{15^2}{250 - 1} =$$

$$= 1474450{,}278$$

De forma que el error de muestreo relativo:

$$CV_{\hat{T}} = \frac{\hat{\sigma}\left(\hat{T}\right)}{\hat{T}} = \frac{\sqrt{1474450{,}278}}{100500} \simeq 0{,}01208 = 1{,}208\,\%.$$

(d) Para el cálculo de los tamaños muestrales por estrato, teniendo en cuenta que el error de muestreo estimado sea inferior a 500 Ha y utilizando la afijación de mínima varianza, viende dada por:

$$e^2 = \hat{V}\left(\hat{X}\right) = \frac{1}{n}\left(\sum_{h=1}^{L} N_h \hat{S}_h\right)^2 - \sum_{h=1}^{L} N_h \hat{S}_h^2$$

Despejando n, se tiene:

$$n = \frac{\left(\sum_{h=1}^{L} N_h \hat{S}_h\right)^2}{e^2 + \sum_{h=1}^{L} N_h \hat{S}_h^2} = \frac{\left(\sum_{h=1}^{L} N_h \frac{\sqrt{n_h} s_h}{\sqrt{n_h - 1}}\right)^2}{e^2 + \sum_{h=1}^{L} N_h \left(\frac{n_h s_h^2}{n_h - 1}\right)}$$



y para el caso que nos ocupa:

$$n = \frac{(22802{,}78593)^2}{500^2 + 224589{,}2829} = 1095{,}614808 \simeq 1096.$$

Y en cada estrato se reparte según

$$n_h = n \cdot \frac{N_h \hat{S}_h}{\sum_{h=1}^{L} N_h \hat{S}_h} = \frac{N_h \frac{\sqrt{n_h} s_h}{\sqrt{n_h - 1}}}{\sum_{h=1}^{L} N_h \frac{\sqrt{n_h} s_h}{\sqrt{n_h - 1}}},$$

donde:

$$\begin{aligned}
n_1 =& 1096 \cdot \frac{10200 \cdot \frac{\sqrt{100} \cdot 1}{\sqrt{100-1}}}{10200 \cdot \frac{\sqrt{100} \cdot 1}{\sqrt{100-1}} + \cdots + 500 \cdot \frac{\sqrt{250} \cdot 15}{\sqrt{250-1}}} = \\
&= 1096 \cdot \frac{10303{,}03}{40351{,}91} = 279{,}84 \simeq 280, \\
n_2 =& 1096 \cdot \frac{5000 \cdot \frac{\sqrt{1200} \cdot 4{,}5}{\sqrt{1200-1}}}{10200 \cdot \frac{\sqrt{100} \cdot 1}{\sqrt{100-1}} + \cdots + 500 \cdot \frac{\sqrt{250} \cdot 15}{\sqrt{250-1}}} = \\
&= 1096 \cdot \frac{22518{,}76}{40351{,}91} = 611{,}63 \simeq 612, \\
n_3 =& 1096 \cdot \frac{500 \cdot \frac{\sqrt{250} \cdot 15}{\sqrt{250-1}}}{10200 \cdot \frac{\sqrt{100} \cdot 1}{\sqrt{100-1}} + \cdots + 500 \cdot \frac{\sqrt{250} \cdot 15}{\sqrt{250-1}}} = \\
&= 1096 \cdot \frac{7530{,}12}{40351{,}91} = 204{,}52 \simeq 205.
\end{aligned}$$

Por lo que los tamaños muestrales por estrato necesarios para lograr la estimación insesgada del total con un error de muestreo inferior a 500 Ha son $n_1 = 280$, $n_2 = 612$ y $n_3 = 205$.

Ejercicio 2

Considere tres variables aleatorias A, B y C tales que su distribución conjunta tiene la forma $p(a, b, c) = p(a) \times p(c) \times p(b|a, c)$. A y C tienen distribuciones de Bernoulli de parámetros θ_A y θ_C, respectivamente. Dados a y c, B tiene distribución de Bernoulli.



(a) Dadas n observaciones independientes idénticamente distribuidas según A, derive el estimador de máxima verosimilitud de θ_A.

(b) Dadas las 5 observaciones de (A, C, B) siguientes

$$((0,1,1),(0,0,0),(1,0,1),(0,0,0),(0,1,0)),$$

calcule el estimador de máxima verosimilitud de θ_A, el de θ_C y, finalmente, el del parámetro de Bernoulli de la distribución de B, dados los antecesores a y c, siempre que sea posible. Comente el resultado. Suponga que se dispone de información a priori para cada uno de los seis parámetros antes mencionados, modelizada en los seis casos mediante distribuciones uniformes $U(0,1)$ (e independientes).

(c) Dadas n observaciones independientes idénticamente distribuidas según A, derive la media de la distribución a posteriori de θ_A.

(d) Dadas las mismas 5 observaciones de (A, C, B), calcule las medias de las distribuciones a posteriori de θ_A, de θ_C y, finalmente, del parámetro de Bernoulli de la distribución de B, dados los antecesores a y c. Comente el resultado.

Solución.

(a) Sea $A_1, A_2, \ldots, A_n$ la muestra aleatoria observada de A.

$$\mathbb{P}(A_1 = a_1 \cap A_2 = a_2 \cap \cdots \cap A_n = a_n) = \theta_A^{n_1} \cdot (1-\theta_A)^{n-n_1},$$

donde $n_1 = \sum_{i=1}^{n} a_i$ es el número de unos en $\{a_1, \ldots, a_n\}$.

Considerándola como función de θ_A,

$$f(\theta_A) = \theta_A^{n_1} \cdot (1-\theta_A)^{n-n_1}.$$



El estimador máximo verosímil para el parámetro será el punto donde f alcanza su máximo. Por ser continua y derivable, eso debe suceder en un extremo ($\theta_A \in \{0,1\}$) o bien en un punto donde $f'(\theta_A) = 0$.

Puesto que la función logaritmo es monónota, $lnf(\theta_A) = \ln(f(\theta_A))$ alcanza máximos y mínimos en los mismos puntos que f, y va a resultar más conveniente para analizar.

$$lnf(\theta_A) = n_1 \ln \theta_A + (n - n_1) \ln(1 - \theta_A)$$

Si $n_1 = n$ resulta $lnf(\theta_A) = n \ln \theta_A$, que claramente se hace máxima con $\theta_A = 1$.

Si $n_1 = 0$ resulta $lnf(\theta_A) = n \ln(1 - \theta_A)$, que se hace máxima con $\theta_A = 0$.

Si $0 < n_1 < n$, la función lnf tiene asíntotas verticales en $A_\theta \in \{0,1\}$, donde tiende a $-\infty$. Por tanto, el máximo deberá darse en un punto interior, donde la derivada se anule.

$$lnf'(\theta_A) = n_1 \frac{1}{\theta_A} - (n - n_1) \frac{1}{1 - \theta_A}$$

$$lnf'(\theta_A) = 0 \iff n_1(1 - \theta_A) - (n - n_1)\theta_A = 0 \iff \theta_A = \frac{n_1}{n}.$$

Es decir, el estimador máximo verosímil es $\hat{\Theta}_A = \frac{\sum_i A_i}{n}$ y la estimación $\hat{\theta}_A = \frac{\sum_i a_i}{n}$. Nótese que esto sirve también para los casos $n_1 \in \{0, n\}$ ya analizados. En la Figura 3.1 se representa el grafo de lnf para un ejemplo concreto.

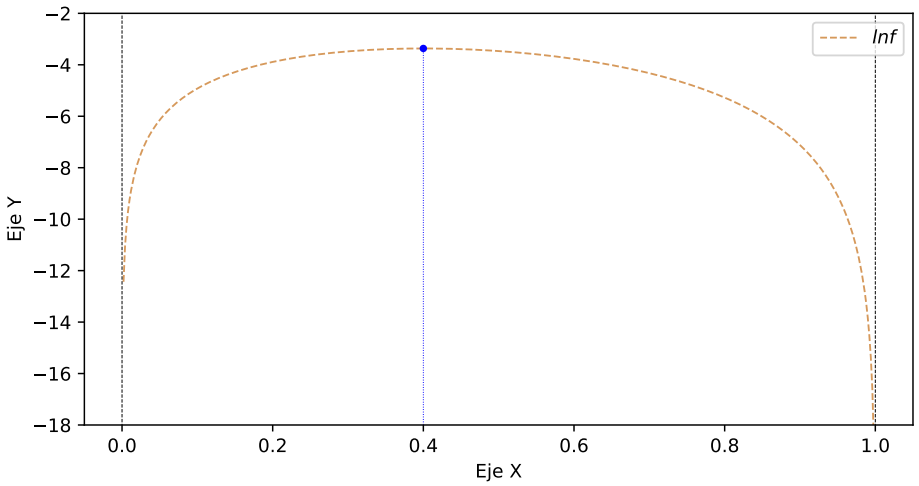


Figura 3.1: Grafo de la función lnf cuando $n = 5$ y $n_1 = 2$. Se aprecian las asíntotas verticales mencionadas, y cómo el máximo se alcanza en $\frac{n_1}{n} = \frac{2}{5}$.

(b) Según indica el problema, A y C son Bernoullis independientes de parámetros θ_A y θ_C, respectivamente. Una vez conocidas ambas, la B también es una Bernoulli. Denotemos θ_B^{ac} al parámetro de la Bernoulli que sigue B una vez se sabe que $A = a$ y $C = c$, para $a, c \in \{0, 1\}$. Esto da lugar a cuatro parámetros para B: $\theta_B^{00}, \theta_B^{01}, \theta_B^{10}, \theta_B^{11}$.

La muestra observada para las variables (A, C, B) es:

$$((0,1,1),(0,0,0),(1,0,1),(0,0,0),(0,1,0)).$$

Según lo visto en el apartado (a), el estimador máximo verosímil para A es

$$\hat{\Theta}_A = \frac{A_1 + A_2 + A_3 + A_4 + A_5}{5},$$

y la estimación resultante

$$\hat{\theta}_A = \frac{a_1 + a_2 + a_3 + a_4 + a_5}{5} = \frac{0+0+1+0+0}{5} = \frac{1}{5}.$$



Análogamente para C,

$$\hat{\Theta}_C = \frac{C_1 + C_2 + C_3 + C_4 + C_5}{5},$$

y la estimación resultante

$$\hat{\theta}_C = \frac{c_1 + c_2 + c_3 + c_4 + c_5}{5} = \frac{1 + 0 + 0 + 0 + 1}{5} = \frac{2}{5}.$$

En cuanto a los parámetros para B, debemos tratar los 4 casos posibles por separado. En todos ellos, por estar fijados $A = a$ y $C = c$, la distribución de B ya es una Bernoulli y por tanto, igual que para θ_A y para θ_B, aplica lo deducido en el apartado (a).

$A = C = 0$) Dos de las muestras caen en este caso, y para ambas se observó $B = 0$. Por tanto,

$$\hat{\theta}_B^{00} = \frac{0 + 0}{2} = 0.$$

$A = 0$ **y** $C = 1$) Otras dos muestras caen en este caso, en una de ellas $B = 0$ y en la otra $B = 1$. Por tanto,

$$\hat{\theta}_B^{01} = \frac{0 + 1}{2} = \frac{1}{2}.$$

$A = 1$ **y** $C = 0$) Una sola muestra cae en este caso, y para ella $B = 1$. Por tanto,

$$\hat{\theta}_B^{10} = \frac{1}{1} = 1.$$

$A = C = 1$) Ninguna muestra cae en este caso. Al no disponer de información, no es posible obtener una estimación máximo verosímil.

(c) En este apartado la estimación es otra, de carácter bayesiano. Según el enunciado, a priori el parámetro de la Bernoulli que sigue A de distribuye según una uniforme, de modo que $\Theta_A \sim U(0, 1)$. Se nos pide



la esperanza de la distribución de Θ_A a posteriori, que denotaremos Θ_A^F, tras disponer de las observaciones $a_1, \ldots, a_n$ de una muestra aleatoria simple $A_1 \ldots, A_n$.

La variable Θ_A es continua, mientras que el experimento informativo (muestreo de A) es de naturaleza discreta. En este caso, la función de densidad de Θ_A^F resulta de multiplicar la de Θ_A por la probabilidad del suceso observado, y luego normalizar. Formalmente,

$$f^F(\theta_A) \propto f(\theta_A)\mathbb{P}(A_1 = a_1 \cap \ldots \cap A_n = a_n | \Theta_A = \theta_A) = \theta_A{}^{\sum_i a_i}(1-\theta_A)^{n-\sum_i a_i}.$$

Conviene recordar que la función de distribución de una Beta(α, β) es $f_B(x; \alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$. Es el mismo patrón que sigue $f^F(\theta_A)$, de lo que se sigue que

$$\Theta_A^F \sim \text{Beta}\Big(\alpha = 1 + \sum_i a_i, \beta = n + 1 - \sum_i a_i\Big).$$

Puesto que la esperanza de una Beta(α, β) es $\frac{\alpha}{\alpha+\beta}$, resulta

$$\mathbb{E}[\Theta_A^F] = \frac{1 + \sum_i a_i}{n+2}.$$

(d) Por lo visto en el apartado anterior, $\mathbb{E}[\Theta_A^F] = \frac{1+1}{5+2} = \frac{2}{7}$.

Este resultado aplica igualmente al resto de estimadores Θ_C, Θ_B^{00},Θ_B^{01},Θ_B^{10} y Θ_B^{11}, ya que todos ellos a priori también se distribuyen según una $U(0,1)$. Los resultados son:

$$\mathbb{E}[\Theta_C^F] = \frac{2+1}{5+2} = \frac{3}{7}$$
$$\mathbb{E}[\Theta_B^{00F}] = \frac{0+1}{2+2} = \frac{1}{4}$$
$$\mathbb{E}[\Theta_B^{01F}] = \frac{1+1}{2+2} = \frac{1}{2}$$
$$\mathbb{E}[\Theta_B^{10F}] = \frac{1+1}{1+2} = \frac{2}{3}$$



$$\mathbb{E}[\Theta_B^{11^F}] = \frac{0+1}{0+2} = \frac{1}{2}.$$

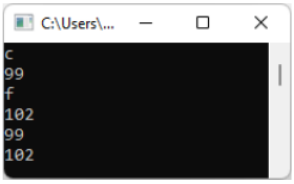
Este último caso es interesante. Puesto que ninguna de las 5 muestras cae en el caso $A = C = 1$, no han aportado información adicional. Aún en ese caso la distribución a posteriori sí existe, y es la misma que la a priori, que puesto que se distribuye según una $U(0,1)$ tiene como esperanza $\frac{1}{2}$.

Ejercicio 3

Observe este fragmento de programa en C/C++. Consiste simplemente en una secuencia de instrucciones. Hemos añadido un breve comentario a su derecha explicando su funcionamiento:

```
char car = 'c';          // Sea la variable car, de tipo carácter, con el valor inicial 'c'
printf(" %c \n", car);   // Escribir el valor de car como un carácter
printf(" %d \n", car);   // Escribir el valor de car como un entero
car = car + 3;           // Aumentar el valor de car en 3 unidades
printf(" %c \n", car);   // Escribir el valor de car como un carácter
printf(" %d \n", car);   // Escribir el valor de car como un entero
int n = int('c');        // Sea la variable entera n, que toma el valor inicial del
                         // código ASCII de 'c'
printf(" %d \n", n);     // Escribir el valor de n como un entero
n = n + 3;               // Aumentar el valor de n en 3 unidades
printf(" %d \n", n);     // Escribir el valor de n como un entero
```

Cada instrucción printf contiene el símbolo '\n', que indica un salto de fin de línea, de manera que la ejecución de este fragmento de programa genera la salida siguiente, una línea por cada instrucción printf:



Ahora, vamos con la representación interna de la información que registra caracteres y enteros. Sabemos que la variable 'car' se está almacenando en una palabra (1 byte) y que el código ASCII del carácter 'c' es 99.

(a) Dé la representación, en binario y en hexadecimal, del entero 99.

(b) Dé la representación, en binario y en hexadecimal, del carácter 'c'.

(c) ¿Por qué la instrucción siguiente no genera un error?

car = car + 3;

(d) En una máquina cuyas palabras de memoria son de 1 byte, tenemos una palabra con el siguiente valor almacenado: "01100001". ¿Qué puede representar?

Solución.

(a) El número 99 en binario se escribe como 1100011, puesto que $99 = 64+32+2+1$, y en hexadecimal se expresa como $63_{(16}$ ya que $99 = 16 \cdot 6+3$.

(b) La representación en binario del carácter 'c' utilizando un byte es



01100011, es decir, que colocamos un cero al principio de la cadena para completar los 8 bits del byte, y en hexadecimal se escribe como un 63.

(c) El tipo carácter en C ocupa un byte de extensión y puede interpretarse con y sin signo. En esencia una variable tipo 'char' es un tipo entero que puede guardar números en un rango desde -128 hasta 127. Al asignar a la variable car la letra c con comillas simples indica que lo que contiene es un único carácter. Las letras tienen un número asociado interno, que es el que conoce el ordenador, y que se ordena en lo que se conoce como tabla ASCII.

INSTITUTO NACIONAL DE ESTADISTICA

Binario	Dec	Hex	Representación
0010 0000	32	20	espacio ( )
0010 0001	33	21	!
0010 0010	34	22	"
0010 0011	35	23	#
0010 0100	36	24	$
0010 0101	37	25	%
0010 0110	38	26	&
0010 0111	39	27	'
0010 1000	40	28	(
0010 1001	41	29	)
0010 1010	42	2A	*
0010 1011	43	2B	+
0010 1100	44	2C	,
0010 1101	45	2D	-
0010 1110	46	2E	.
0010 1111	47	2F	/
0011 0000	48	30	0
0011 0001	49	31	1
0011 0010	50	32	2
0011 0011	51	33	3
0011 0100	52	34	4
0011 0101	53	35	5
0011 0110	54	36	6
0011 0111	55	37	7
0011 1000	56	38	8
0011 1001	57	39	9
0011 1010	58	3A	:
0011 1011	59	3B	;
0011 1100	60	3C	<
0011 1101	61	3D	=
0011 1110	62	3E	>
0011 1111	63	3F	?

Binario	Dec	Hex	Representación
0100 0000	64	40	@
0100 0001	65	41	A
0100 0010	66	42	B
0100 0011	67	43	C
0100 0100	68	44	D
0100 0101	69	45	E
0100 0110	70	46	F
0100 0111	71	47	G
0100 1000	72	48	H
0100 1001	73	49	I
0100 1010	74	4A	J
0100 1011	75	4B	K
0100 1100	76	4C	L
0100 1101	77	4D	M
0100 1110	78	4E	N
0100 1111	79	4F	O
0101 0000	80	50	P
0101 0001	81	51	Q
0101 0010	82	52	R
0101 0011	83	53	S
0101 0100	84	54	T
0101 0101	85	55	U
0101 0110	86	56	V
0101 0111	87	57	W
0101 1000	88	58	X
0101 1001	89	59	Y
0101 1010	90	5A	Z
0101 1011	91	5B	[
0101 1100	92	5C	\
0101 1101	93	5D	]
0101 1110	94	5E	^
0101 1111	95	5F	_

Binario	Dec	Hex	Representación
0110 0000	96	60	`
0110 0001	97	61	a
0110 0010	98	62	b
0110 0011	99	63	c
0110 0100	100	64	d
0110 0101	101	65	e
0110 0110	102	66	f
0110 0111	103	67	g
0110 1000	104	68	h
0110 1001	105	69	i
0110 1010	106	6A	j
0110 1011	107	6B	k
0110 1100	108	6C	l
0110 1101	109	6D	m
0110 1110	110	6E	n
0110 1111	111	6F	o
0111 0000	112	70	p
0111 0001	113	71	q
0111 0010	114	72	r
0111 0011	115	73	s
0111 0100	116	74	t
0111 0101	117	75	u
0111 0110	118	76	v
0111 0111	119	77	w
0111 1000	120	78	x
0111 1001	121	79	y
0111 1010	122	7A	z
0111 1011	123	7B	{
0111 1100	124	7C	|
0111 1101	125	7D	}
0111 1110	126	7E	~

Figura 3.2: Tabla con el código ASCII



Al escribir car + 3, en car está almacenado un número, el 99, ya que es el que corresponde a la letra 'c', y si sumamos un 3 obtenemos el número 102, que es la letra 'f' en el código ASCII (ver figura 3.2). Por este motivo no da error. Realmente el ordenador no trabaja con datos de tipo char. Lo que hace es almacenar internamente, como decíamos, una tabla que asocia a cada caracter un número. Cuando guardas un caracter en una variable, realmente guardas un número. Cuando muestras el valor de una variable tipo char realmente recuperas un número y con la tabla interna de conversión el ordenador sabe qué número mostrar, de manera que cuando operas con caracteres, en realidad estás operando con números.

Sumar caracteres o números a caracteres es equivalente a sumar números. Es lo que se llama casting esta conversión de char a int y como lo realiza el ordenador internamente se llama casting implícito.

La expresión car + 3 no se interpreta como sumar 3 unidades al caracter c, sino como pasar al caracter que ocupa 3 posiciones más en la tabla interna y como un caracter es un número, la expresión se interpreta como pasar al número 3 unidades más grande del número asociado a la 'c' dentro de la tabla interna.

(d) Si las palabras son de 1 byte, entonces la que contiene "01100001"se corresponde con la letra 'a' si el tipo de variable es char, puesto que 01100001 será el número 97 según la tabla ASCII, que está asociado a la letra 'a'. Si el tipo de variable es de tipo entero sin signo (unsigned int) sería el número 97, al igual que si la variable es de tipo entero con signo (signed int) que utilizaría la codificación complemento a 2 y también representaría el número 97.



Ejercicio 4

Dada la población:

u_i	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
x_i	2	7	2	3	5	6	4	1

se desea obtener una muestra sistemática de tamaño 3 para lo cual se determina un periodo $k = N/n$, redondeado al entero más próximo y un arranque aleatorio entre 1 y K.

Se pide:

(a) Determine el espacio muestral y la función de probabilidad.

(b) Calcule la probabilidad π_i que tiene la unidad u_i de pertenecer a la muestra.

(c) Calcule la $V(\hat{X})$ siendo $\hat{X}$ el estimador insesgado del total.

(d) A partir de la distribución en el muestreo de $\hat{X}$, calcule $E(\hat{X})$ y $V(\hat{X})$.

Solución.

(a) El conjunto de muestras que se obtienen de la población, teniendo en cuenta el tipo de muestreo, y siendo $k = \frac{N}{n} = \frac{8}{3} = 2{,}66666$, que redondeando al entero más próximo es 3, tenemos que son:

$$(\tilde{u}_1) = u_1, u_4, u_7$$
$$(\tilde{u}_2) = u_2, u_5, u_8$$
$$(\tilde{u}_3) = u_3, u_6$$

El muestreo sistemático es un muestreo sin reposición, y se cumple que $P(\tilde{u}_i) = \frac{1}{k} = \frac{1}{3}$

(b) Tenemos que calcular $\pi_i = P(u_i \in \text{muestra})$ y es claro que para cualquier $i \in U$, es decir, para cualquier elemento de la población se cumple



que $\pi_i = \frac{1}{k}$. Además se cumple que:

$$\pi_{ij} = \begin{cases} \frac{1}{k} & \text{si i,j pertenecen a la misma muestra, o sea, } |i-j| = k \\ 0 & \text{en el resto de los casos} \end{cases}$$

(c) Cuando el tamaño de la población N es múltiplo del tamaño de muestra n es conocido que el estimador insesgado del total $X = \sum_{i=1}^{N} X_i$ tiene la expresión siguiente: $\hat{X} = \sum_{i=1}^{n} \frac{X_i}{\pi_i}$. En el caso del muestreo sistemático, $\hat{X} = \sum_{i=1}^{n}\sum_{j=1}^{1} \frac{X_{ij}}{\frac{1}{k}}$ es insesgado de $X = \sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}$, y tenemos que:

$$\hat{X} = \sum_{i=1}^{n}\sum_{j=1}^{1} \frac{X_{ij}}{\frac{1}{k}} = \sum_{i=1}^{n} k \cdot X_{ij} = \frac{N}{n}\sum_{i=1}^{n} X_{ij} = N \cdot \overline{x}_j$$

En este problema no se cumple que el tamaño poblacional sea múltiplo de n, puesto que 8 no es múltiplo de 3, por lo que puede comprobarse fácilmente que ese estimador es sesgado con la siguiente tabla:

S(X)	P(S(X))	$\hat{X} = N \cdot \overline{x}_j$	$\hat{\overline{X}} = \overline{x}_j$
(2,3,4)	$\frac{1}{3}$	24	3
(7,5,1)	$\frac{1}{3}$	$\frac{104}{3}$	$\frac{13}{3}$
(2,6)	$\frac{1}{3}$	32	4

utilizando las distribuciones en el muestreo de los estimadores de la media y del total poblacional, obteniendo:

$$E(\hat{X}) = 24\tfrac{1}{3} + \tfrac{104}{3}\tfrac{1}{3} + 32\tfrac{1}{3} = 8 + \tfrac{104}{9} + \tfrac{32}{3} = 30{,}2222.$$

que es diferente de $X = 30$. Por este motivo, debemos utilizar otro estimador que sí sea insesgado en cualquier situación, y para ello tenemos el siguiente:

$\hat{X}_{HT} = k \cdot X_s = k \cdot \sum_{h \in s} X_h$, siendo s uno de los elementos del conjunto de posibles muestras $\{s_1, \ldots, s_r, \ldots, s_k\}$, siendo

$$s_r = \{m : m = r + (j-1)k \leq N; j = 1, 2, \ldots, n_s\}$$

La varianza de este estimador viene dada por

$V_{sys}(\hat{X}_{HT}) = k \cdot \sum_{r=1}^{k}(X_{s_r} - \overline{X}_{tm})^2$ donde $\overline{X}_{tm} = \frac{\sum_{r=1}^{k} X_{s_r}}{k} = \frac{X}{k}$ es una media de los totales muestrales $X_{s_r} = \sum_{k \in s_r} X_k$.

Vamos a calcular esta varianza, que es lo que pide este apartado haciendo uso de la tabla correspondiente:

s_r	X_{s_r}	$P(s_r)$	$\hat{X}_{HT} = k \cdot X_{s_r}$
(2,3,4)	9	$\frac{1}{3}$	$3 \cdot 9 = 27$
(7,5,1)	13	$\frac{1}{3}$	$13 \cdot 3 = 39$
(2,6)	8	$\frac{1}{3}$	$3 \cdot 8 = 24$

Como $\overline{X}_{tm} = \frac{\sum_{r=1}^{k} X_{s_r}}{k} = \frac{9+13+8}{3} = \frac{30}{3} = 10$ entonces podemos escribir

$$V_{sys}(\hat{X}_{HT}) = k \cdot \sum_{r=1}^{k}(X_{s_r} - \overline{X}_{tm})^2 = 3 \cdot \left[(9-10)^2 + (13-10)^2 + (8-10)^2\right] =$$
$$= 3 \cdot (1 + 9 + 4) = 42$$

Si hubiéramos utilizado el estimador que es sesgado, se podría haber calculado la varianza teniendo en cuenta que $S_{bs}^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\overline{x}_j - \overline{X}\right)^2}{k-1}$ es la fuente de variación entre las muestras y que $\overline{X} = \frac{30}{8}$ de esta forma:



$$
\begin{aligned}
V(\hat{X}) =& V(N \cdot \overline{x}_j) = N^2 \cdot V(\overline{x}_j) = N^2 \cdot \left(1 - \frac{n}{N}\right) \frac{S_{bs}^2}{n} = \\
=& N^2 \left(\frac{N-n}{N}\right) \frac{1}{n} \frac{n}{k-1} \sum_{j=1}^{k} \left(\overline{x}_j - \overline{X}\right)^2 = \\
=& N^2 \left(\frac{N-n}{N}\right) \frac{1}{n} \frac{n}{\frac{N-n}{n}} \sum_{j=1}^{k} \left(\overline{x}_j - \overline{X}\right)^2 = \\
=& N^2 \cdot \frac{n}{N} \sum_{j=1}^{k} \left(\overline{x}_j - \overline{X}\right)^2 = N \cdot n \sum_{j=1}^{k} \left(\overline{x}_j - \overline{X}\right)^2
\end{aligned}
$$

(d) Con la tabla construida en el apartado (c) para el estimador $\hat{X}_{HT}$ tenemos la distribución en el muestreo de los estimadores del total poblacional, por lo que:

$E(\hat{X}) = 27 \cdot \frac{1}{3} + 39 \cdot \frac{1}{3} + 24 \cdot \frac{1}{3} = \frac{90}{3} = 30$, por lo tanto es un estimador insesgado. La varianza, utilizando de nuevo la distribución en el muestreo de $\hat{X}_{HT}$ vale:

$V_{sys}(\hat{X}_{HT}) = \frac{27^2+39^2+24^2}{3} - 30^2 = 42$

que lógicamente coincide con la ya obtenida de forma teórica anteriormente.

Ejercicio 5

Se tienen datos $x_1, x_2, ..., x_n$ independientes idénticamente distribuidos de una normal de media μ y varianza conocida e igual a 9. Se determina que, a priori, μ se distribuye según una distribución normal de media 10 y varianza 4.

(a) Calcule la verosimilitud del modelo.

(b) Calcule el estimador de máxima verosimilitud de μ.

(c) Calcule la distribución a posteriori de μ.

(d) Calcule la moda a posteriori de μ.

(e) Proporcione la expresión de un intervalo de probabilidad 0.8 a posteriori para μ suponiendo que se tiene acceso a una rutina $qnorm(p)$ que, dada una probabilidad p, nos da el correspondiente cuantil de la distribución normal estándar.

Solución.

(a) La función de densidad de la variable aleatoria $X \sim N(\mu, \sigma^2 = 9)$ es

$$f(x,\mu) = \frac{1}{3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\cdot 9}} = \frac{1}{3\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{18}}.$$

Para obtener la verosimilitud del modelo, se considera el caso general de una muestra de tamaño n y vendrá definida por

$$\begin{aligned} L\left(x_1, x_2, ..., x_n; \mu\right) &= \prod_{i=1}^{n} f\left(x_i; \mu\right) = \\ &= \prod_{i=1}^{n} \frac{1}{3\sqrt{2\pi}} \exp\left\{-\frac{\left(x_i-\mu\right)^2}{18}\right\} = \\ &= \left(\frac{1}{3\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{18}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right\} = \\ &= \left(\frac{1}{3\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{18}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right\} \exp\left\{-\frac{n}{18}(\mu-\overline{x})^2\right\} \end{aligned}$$

(b) El estimador de máxima verosimilitud de μ es el valor $\hat{\mu}_{MV}$ que maximiza la verosimilitud del modelo. Es decir:

$$\hat{\mu}_{MV} \text{ tq } \max_{\mu} L\left(x_1, x_2, ..., x_n; \mu\right) = \max_{\mu} \left(\frac{1}{3\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{18}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right\}$$

Dado que ln es una transformación creciente, encontrar el máximo de la expresión anterior es equivalente a:

$$\max_{\mu} \ln\left(L\left(x_1, x_2, ..., x_n; \mu\right)\right) = \max_{\mu} \ln\left(\left(\frac{1}{3\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{18}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right\}\right) =$$



$$= \max_{\mu} \left\{ -n \ln \left(3\sqrt{2\pi} \right) - \frac{1}{18} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

Y el valor de μ que maximiza esta función es aquel tal que

$$\frac{\partial}{\partial \mu} \ln(L(\mu)) = 0$$

:

$$\begin{aligned}
\frac{\partial}{\partial \mu} \ln(L(\mu)) = 0 &\Leftrightarrow -\frac{1}{18} \sum_{i=1}^{n} -2 (x_i - \mu) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} -2 (x_i - \mu) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} (x_i - \mu) = 0 \\
&\Leftrightarrow -n\mu + \sum_{i=1}^{n} x_i = 0 \\
&\Leftrightarrow n\mu = \sum_{i=1}^{n} x_i \\
&\Leftrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \\
&\Leftrightarrow \hat{\mu} = \overline{x}
\end{aligned}$$

(c) Para encontrar la distribución a posteriori de μ, primero necesitamos especificar la forma de la distribución a priori y luego actualizarla utilizando los datos observados. Dado que se especifica que la distribución a priori de μ es una distribución normal con media $\mu_0 = 10$ y $\sigma_0^2 = 4$, la función de densidad de probabilidad a priori de μ es

$$f\left(\mu | \mu_0, \sigma_0^2\right) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\mu - \mu_0)^2}{2 \cdot \sigma_0^2}} = \frac{1}{2\sqrt{2\pi}} \exp\left\{ -\frac{1}{8} (\mu - 10)^2 \right\}.$$

Sabiendo que los datos observados siguen una distribución normal con media μ y varianza conocida $\sigma^2 = 9$, podemos actualizar la distribución a priori utilizando el Teorema de Bayes, obteniendo la distribución a

posteriori. La distribución a posteriori de μ, denotada como $f\left(\mu|x\right)$, es aquella que resulta de aplicar conjuntamente la probabilidad a priori y la verosimilitud de los datos, entre la probabilidad de los propios datos.

$$\pi\left(\mu|x\right) = \frac{f\left(x|\mu\right)\pi(\mu)}{\int f\left(x|\mu\right)\pi(\mu)d\mu}.$$

Como el denominador se vuelve una constante, podemos decir que

$$\pi\left(\mu|x\right) \propto f\left(x|\mu\right)\pi(\mu).$$

Por tanto, la distribución a posteriori será de la forma

$$\pi\left(\mu|x\right) = k\exp\left\{-\frac{1}{8}\left(\mu-10\right)^2\right\} exp\left\{-\frac{n}{18}(\mu-\overline{x})^2\right\}$$

donde todos los factores independientes de μ han sido englobados en la constante k puesto que se trata de obtener una función de μ. Entonces,

$$\begin{aligned}
\pi\left(\mu|x\right) &= k_1\exp\left\{-\frac{1}{8}\left(\mu^2-2\cdot 10\cdot\mu\right)-\frac{n}{18}\left(\mu^2-2\overline{x}\mu\right)\right\} \\
&= k_1\exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{1}{4}+\frac{n}{9}\right)-2\mu\left(\frac{10}{4}+\frac{n\overline{x}}{9}\right)\right]\right\} \\
&= k_1\exp\left\{-\frac{1}{2}\left(\frac{1}{4}+\frac{n}{9}\right)\left[\mu^2-2\mu\frac{\left(\frac{10}{4}+\frac{n\overline{x}}{9}\right)}{\left(\frac{1}{4}+\frac{n}{9}\right)}\right]\right\} \\
&= k_2\exp\left\{-\frac{1}{2}\left(\frac{1}{4}+\frac{n}{9}\right)\left[\mu^2-\theta^*\right]^2\right\}
\end{aligned}$$

siendo

$$\theta^* = \frac{\left(\frac{10}{4}+\frac{n\overline{x}}{9}\right)}{\left(\frac{1}{4}+\frac{n}{9}\right)}$$

que se puede reescribir como

$$\theta^* = \frac{\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{x}}{9}\right)}{\left(\frac{1}{\sigma_0^2}+\frac{n}{9}\right)}$$

donde

- $\overline{x}$: es la media de los datos observados, la media muestral



- μ_0: es la media de la distribución a priori, que en este caso $\mu_0 = 10$
- σ^2: es la varianza de los datos observados
- σ_0^2: es la varianza de la distribución a priori, que en este caso $\sigma_0^2 = 4$
- n: es el número de datos observados

Lo cual muestra que la distribución a posteriori de μ es una distribución

$$N\left(\frac{\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{x}}{9}\right)}{\left(\frac{1}{\sigma_0^2}+\frac{n}{9}\right)}, \frac{1}{\sqrt{\frac{1}{\sigma_0^2}+\frac{n}{9}}}\right) = N\left(\frac{\left(\frac{10}{4}+\frac{n\overline{x}}{9}\right)}{\left(\frac{1}{4}+\frac{n}{9}\right)}, \frac{1}{\sqrt{\frac{1}{4}+\frac{n}{9}}}\right)$$

y se observa que dado que tanto la verosimilitud como la distribución a priori son normales, la distribución a posteriori también sigue una distribución normal.

Para apartados posteriores, operando en los parámetros de la distribución a posteriori, denotamos $\mu_n = \frac{90+4n\overline{x}}{9+4n}$ y $\sigma_n^2 = \frac{36}{9+4n}$.

(d) La moda de la distribución a posteriori de μ corresponde al valor de μ que maximiza la función de densidad de probabilidad a posteriori. Dado que estamos trabajando con una distribución normal, la moda coincide con la media de la distribución. Entonces, la moda a posteriori de μ será la misma que su media, que se calcula utilizando la fórmula utilizada en el apartado anterior, cuya expresión general es:

$$\text{Media distribución a posteriori} = \frac{\mu_0\sigma^2 + n\sigma_0^2\overline{x}}{\sigma^2 + n\sigma_0^2}$$

y por tanto la moda de la distribución a posteriori es

$$\mu_n = \frac{90+4n\overline{x}}{9+4n}$$

(e) Para obtener un intervalo de probabilidad del 80 % a posteriori para μ, se pueden utilizar los cuantiles de la distribución normal. Dado que la

distribución a posteriori de μ es normal, podemos expresar este intervalo de probabilidad utilizando la media y la desviación estándar de esta distribución. La media ya la hemos detallado en el apartado anterior y la varianza de la distribución a posteriori vale $\sigma_n^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}$. El intervalo de probabilidad del 80 % a posteriori para μ estará centrado en la media a posteriori μ_n y abarcará el rango que corresponde al cuantil del 10 % y el 90 % de la distribución normal. Para calcular los límites de este intervalo, necesitamos encontrar los cuantiles de la distribución normal estándar correspondientes al 10 % y al 90 %. Dado que tenemos acceso a una rutina $qnorm(p)$ que nos da el cuantil de la distribución normal estándar para una probabilidad p, podemos usarla para encontrar estos valores. Entonces, los límites del intervalo de probabilidad del 80 % a posteriori para μ serán:

$$
\begin{aligned}
&[\mu_n + qnorm(0{,}1)\cdot\sigma_n\ ,\ \mu_n + qnorm(0{,}9)\cdot\sigma_n] = \\
&\qquad = \left[\frac{90+4n\overline{x}}{9+4n} + qnorm(0{,}1)\cdot\sqrt{\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}}\ ,\right. \\
&\qquad\quad \left.\frac{90+4n\overline{x}}{9+4n} + qnorm(0{,}9)\cdot\sqrt{\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}}\right] = \\
&\qquad = \left[\frac{90+4n\overline{x}}{9+4n} + qnorm(0{,}1)\cdot\sqrt{\frac{36}{9+4n}}\ ,\right. \\
&\qquad\quad \left.\frac{90+4n\overline{x}}{9+4n} + qnorm(0{,}9)\cdot\sqrt{\frac{36}{9+4n}}\right]
\end{aligned}
$$

Ejercicio 6

En una localidad turística hay 12 empresas de alojamiento dedicadas a la hostelería, de las que se conoce su cifra de negocio del último mes, expresada en u.m.



Empresa	Cifra de negocios	Grupos de actividad CNAE
A	25.000	551 Hoteles y alojamientos similares
B	30.000	552 Alojamientos turísticos y otros alojamientos de corta estancia
C	15.000	553 Campings y aparcamientos para caravanas
D	12.000	559 Otros alojamientos
E	18.000	551 Hoteles y alojamientos similares
F	13.000	551 Hoteles y alojamientos similares
G	3.000	552 Alojamientos turísticos y otros alojamientos de corta estancia
H	2.000	552 Alojamientos turísticos y otros alojamientos de corta estancia
I	11.000	559 Otros alojamientos
J	2.000	552 Alojamientos turísticos y otros alojamientos de corta estancia
K	2.000	552 Alojamientos turísticos y otros alojamientos de corta estancia

Se pretende publicar una tabla que refleje la cifra de negocio de dichas empresas atendiendo al grupo CNAE en el que se engloba su actividad y respetando la confidencialidad de los informantes. La tabla propuesta es la siguiente:

Grupos de actividad CNAE	Cifra de negocio
551 Hoteles y alojamientos similares	56.000
552 Alojamientos turísticos y otros alojamientos de corta estancia	39.000
553 Campings y aparcamientos para caravanas	15.000
559 Otros alojamientos	23.000

Justifique razonadamente si la tabla propuesta resulta segura en un contexto de control de secreto estadístico atendiendo a los siguientes criterios:

(a) Según la regla de la frecuencia (o valor límite), para $k = 2$.



(b) Según la regla de dominancia, para $n = 1$, $k = 75\,\%$.

(c) Según la regla del porcentaje, para $p = 10\,\%$.

(d) Proponga alguna alternativa concreta para solucionar los problemas de identificación encontrados.

Solución.

(a) La regla de la frecuencia para un valor k considera que la tabla no es segura en contexto de control de secreto estadístico si cada fila de tabla implica a k o menos registros. Considerando el valor $k = 2$, se observa que en la tabla original no cumplen la condición de la regla de la frecuencia las filas que hacen referencia al grupo *553 Campings y aparcamientos para caravanas* y al grupo *559 Otros alojamientos*, pues ambos grupos están formados por dos y una empresa respectivamente. En la tabla propuesta se han unido los datos de los registros del grupo *559 Otros alojamientos*, pero del *553 Campings y aparcamientos para caravanas* sólo había una empresa y sigue habiendo un registro solo de este grupo, por lo que no se ha resuelto el problema de la sensitividad de esas celdas.

(b) La regla de la dominancia establece que un registro es sensitivo en un contexto de control estadístico si n o menos registros del conjunto constituyen más del $k\,\%$ de la suma total. En la tabla siguiente se calcula el porcentaje sobre el total de cada grupo de actividad que supone cada una de las empresas:



Empresa	Porcentaje de la cifra de negocios sobre el total del grupo	Grupos de actividad CNAE
A	44,64 %	551 Hoteles y alojamientos similares
B	76,92 %	552 Alojamientos turísticos y otros alojamientos de corta estancia
C	100 %	553 Campings y aparcamientos para caravanas
D	52,17 %	559 Otros alojamientos
E	32,14 %	551 Hoteles y alojamientos similares
F	23,21 %	551 Hoteles y alojamientos similares
G	7,69 %	552 Alojamientos turísticos y otros alojamientos de corta estancia
H	5,12 %	552 Alojamientos turísticos y otros alojamientos de corta estancia
I	47,82 %	559 Otros alojamientos
J	5,12 %	552 Alojamientos turísticos y otros alojamientos de corta estancia
K	5,12 %	552 Alojamientos turísticos y otros alojamientos de corta estancia

Se observa que tanto la empresa B dentro del grupo *552 Alojamientos turísticos y otros alojamientos de corta estancia*, como la empresa C dentro del grupo *553 Campings y aparcamientos para caravanas*, suponen más del 75 % del grupo. Por tanto, atendiendo a la regla de dominancia, para n=1 y k=75 %, la tabla inicial del problema no resulta segura en un contexto de control de secreto estadístico, pues en los grupos *552 Alojamientos turísticos y otros alojamientos de corta estancia* y *553 Campings y aparcamientos para caravanas* existe una empresa que contribuye con más del 75 % del grupo. En la tabla propuesta se han fusionado las empresas del grupo *552 Alojamientos turísticos y otros alojamientos de corta estancia* para paliar esta problemática, pero sigue habiendo un registro con la única empresa del *553 Campings y*



aparcamientos para caravanas, por lo que los problemas persisten con esta alternativa.

(c) La regla del porcentaje para $p = 10\,\%$ se basa en la siguiente inecuación.

$$X - x_{(1)} - x_{(2)} < \frac{p}{100} \cdot x_{(1)},$$

donde X es el total publicado en la tabla para un grupo en concreto, $x_{(1)}$ es la empresa que más contribuye a ese grupo y $x_{(2)}$ es la segunda empresa que más contribuye a ese grupo. De esta forma, una celda se considera sensitiva en un contexto de secreto estadístico si el total de la celda menos los dos principales contribuyentes es menor que el p por ciento del principal contribuyente. Si para una celda se satisface la inecuación anterior, la celda es sensitiva porque la empresa con el segundo valor más alto podría calcular el valor de la mayor con un error pequeño y comprometería la integridad de la tabla.

En la tabla siguiente se recoge el valor de los componentes de la inecuación:

Grupo actividad CNAE	X	$x_{(1)}$	$x_{(2)}$
551 Hoteles y alojamientos similares	56000	25000	18000
552 Alojamientos turísticos y otros alojamientos de corta estancia	39000	30000	3000
553 Campings y aparcamientos para caravanas	15000	15000	
559 Otros alojamientos	23000	12000	11000

Operando, se obtienen los valores de los lados izquierdo y derecho de la inecuación:



Grupo actividad CNAE	$X - x_{(1)} - x_{(2)}$	$\frac{p}{100} \cdot x_{(1)}$	
551 Hoteles y alojamientos similares	13000	2500	$X - x_{(1)} - x_{(2)} > \frac{p}{100} \cdot x_{(1)}$
552 Alojamientos turísticos y otros alojamientos de corta estancia	6000	3000	$X - x_{(1)} - x_{(2)} > \frac{p}{100} \cdot x_{(1)}$
553 Campings y aparcamientos para caravanas	0	1500	$X - x_{(1)} - x_{(2)} < \frac{p}{100} \cdot x_{(1)}$
559 Otros alojamientos	0	1200	$X - x_{(1)} - x_{(2)} < \frac{p}{100} \cdot x_{(1)}$

A la vista de los resultados obtenidos, la tabla es sensitiva en términos del secreto estadístico para los grupos *553 Campings y aparcamientos para caravanas* y *559 Otros alojamientos*.

(d) Proponga alguna alternativa concreta para solucionar los problemas de identificación encontrados.

La regla de la frecuencia en el primer apartado y la regla del porcentaje del tercero, ponen de manifiesto problemas con los grupos *553 Campings y aparcamientos para caravanas* y *559 Otros alojamientos*. Una posible solución es agregar los grupos problemáticos de la tabla. De esta forma, al tener una información más agregada de la población, se limita la información detallada sobre características de empresas concretas. Dado que la tabla contiene celdas tanto unitarias (solo una empresa contribuye a la celda) como binarias (dos empresas contribuyen a la celda), el foco de atención se desvía hacia respuestas individuales con el consiguiente riesgo de identificación. Agrupar los grupos de actividad CNAE *553 Campings y aparcamientos para caravanas* y *559 Otros alojamientos* en una única celda podría ser una solución. De esta forma, la variable Grupos de actividad CNAE se recodificaría en tres categorías y la tabla quedaría así:



Grupos de actividad CNAE	Cifra de negocio
551 Hoteles y alojamientos similares	56.000
552 Alojamientos turísticos y otros alojamientos de corta estancia	39.000
553 Campings y aparcamientos para caravanas y 559 Otros alojamientos	38.000

No obstante, el grupo *551 Hoteles y alojamientos similares* no presenta en la tabla original problema alguno y no parece haber motivo para tener que agregar todas sus celdas, puesto que no entra en conflicto con ninguno de los criterios, por lo que si el objetivo es preparar un fichero de microdatos "tapando" lo mínimo posible de manera que no se comprometa el secreto estadístico los registros de ese grupo podrían permanecer. Ahora bien, puesto que sólo hay 3 empresas en el mismo, lo más prudente probablemente sería agruparlas.

En muchas publicaciones se postula que parece preferible la regla del p % a la regla de dominancia porque la regla de dominancia es más conservadora que la regla del p %, es decir, proporciona más supresiones de celdas que la regla del p %. Por ejemplo, en "Handbook on statistical disclosure control" [Hundepool et al, 2010] y en el documento ""Statistical disclosure control in tabular data" [Castrol, 2009], la regla de dominancia con n=2 se compara con la regla del p %, ya que cuando n=2 la regla de la dominancia se parece mucho a la formulación de la regla del p %. Existe todo un abanico de teorías y bibliografía al respecto de esta problemática, pero entendemos que con estas reflexiones es suficiente para proporcionar una respuesta razonable al problema



3.2. Soluciones del cuarto examen del Cuerpo Superior de Estadísticos del Estado de 2022. Ciencia de datos. Promoción interna.

Ejercicio 1

Para estimar el número medio de portátiles por vivienda en una determinada zona rural se realiza un diseño en dos etapas:

Primera etapa: Se obtiene una muestra aleatoria simple con reemplazamiento de m=3 pueblos entre un total de M=300. El tamaño de cada pueblo (N_i) se mide por el número de viviendas.

Segunda etapa: Se selecciona una muestra aleatoria simple de 5 viviendas en cada uno de los pueblos seleccionados en primera etapa.

El número total de viviendas es de N=9000 .

Los datos disponibles son:

Unidades primarias (pueblos)	Tamaños (N_i)	Número de portátiles observados en las viviendas de la muestra m_i=5
1	50	1;0;3;2;2
2	60	2;2;1;2;3
3	65	2;4;3;2;1

Se pide:

(a) Calcular un estimador insesgado para el número medio de portátiles por vivienda en los siguientes casos:

- Muestreo con probabilidades iguales en las dos etapas.
- Muestreo con probabilidades proporcionales al tamaño en la primera etapa.

(b) Estimar el error de muestreo en ambos casos. ¿Cuál de los dos estimadores es preferible?

Solución.

Denotamos por $N = 300$ el número de pueblos o conglomerados, $\overline{M} = 30$ el tamaño medio de los pueblos, $M = 9000$ el número total de viviendas, y por M_i el número de viviendas de cada pueblo, con $n = 3$ el número de

pueblos de la muestra y $m_i = 5$ el número de viviendas muestreadas en cada uno de los pueblos de la muestra.

(a) En el caso de que el muestreo sea con probabilidades iguales en las dos etapas tenemos que $\pi_i = n \cdot P_i = n \cdot \frac{1}{N} = \frac{n}{N}$, de manera que:

$$\widehat{\widehat{\mathrm{X}}} = \sum_{i=1}^{n} \frac{\hat{X}_i}{\pi_i} = \sum_{i=1}^{n} \frac{M_i \cdot \overline{x}_i}{\pi_i} = \frac{300}{3}\left(50 \cdot \frac{8}{5} + 60 \cdot \frac{10}{5} + 65 \cdot \frac{12}{5}\right) =$$
$$= 100(80 + 120 + 156) = 35\,600$$

De esta forma, el estimador insesgado para el número medio de portátiles por vivienda es:

$$\widehat{\overline{\mathrm{X}}} = \frac{\widehat{\widehat{\mathrm{X}}}}{M} = \frac{\widehat{\widehat{\mathrm{X}}}}{N \cdot \overline{M}} = \frac{35\,600}{300 \cdot 30} = 3{,}995$$

Hay que observar que el número medio de viviendas en los pueblos de la muestra es bastante mayor que 30, que es el que corresponde a todos los de la población. En efecto, $\overline{m} = \frac{50+60+65}{3} = 58{,}33$, lo cual influye en ese valor del estimador, que parece estar alejado de un valor digamos razonable en torno a 2, que es lo que la muestra parece indicar.

Si el muestreo es con probabilidades proporcionales al tamaño en la primera etapa tendríamos que $\pi_i = n \cdot P_i = n \cdot \frac{M_i}{M}$, por lo que:

$$\widehat{\widehat{\mathrm{X}}} = \sum_{i=1}^{n} \frac{\hat{X}_i}{\pi_i} = \sum_{i=1}^{n} \frac{M_i \cdot \overline{x}_i}{\pi_i} = \frac{M}{n} \sum_{i=1}^{n} \overline{x}_i =$$
$$= \frac{50 \cdot \frac{8}{5}}{3 \cdot \frac{50}{9000}} + \frac{60 \cdot \frac{10}{5}}{3 \cdot \frac{60}{9000}} + \frac{65 \cdot \frac{12}{5}}{3 \cdot \frac{65}{9000}} =$$
$$= \frac{9000}{3}\left(\frac{8}{5} + \frac{10}{5} + \frac{12}{5}\right) = \frac{54\,000}{3} = 18\,000$$

Por tanto, el estimador insesgado para el número medio de portátiles por vivienda es:

$$\widehat{\overline{\mathrm{X}}} = \frac{\widehat{\widehat{\mathrm{X}}}}{M} = \frac{\widehat{\widehat{\mathrm{X}}}}{N \cdot \overline{M}} = \frac{18\,000}{9000} = 2$$



(b) Para estimar el error de muestreo en ambos casos vamos a utilizar el método de los conglomerados últimos, según el cual, siendo $\hat{\theta}_i$ un estimador insesgado de θ a partir de los datos de la unidad primaria i-ésima, y $\hat{\hat{\theta}} = \sum_{i=1}^{n} \frac{\hat{\theta}_i}{n}$ es

$$\hat{V}(\hat{\theta}) = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\hat{\theta}})^2}{n(n-1)}$$

En el caso de probabilidades iguales en las dos etapas tendremos que $\hat{\theta}_i = \frac{\hat{X}_i}{P_i}$, $\hat{\hat{\theta}} = \hat{\hat{X}} = \sum_{i=1}^{n} \frac{\hat{\theta}_i}{n}$, y al ser $P_i = \frac{1}{N}$:

$$\hat{V}(\hat{\hat{\mathrm{X}}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{X}_i}{P_i} - \hat{\hat{X}} \right)^2 = \frac{\sum_{i=1}^{n} \left(\frac{\hat{X}_i}{\frac{1}{N}} - N \sum_{i=1}^{n} \frac{M_i \cdot \overline{x}_i}{n} \right)^2}{n(n-1)} =$$

$$= \frac{N^2}{n(n-1)} \sum_{i=1}^{n} \left(\hat{X}_i - \sum_{i=1}^{n} \frac{\hat{X}_i}{n} \right)^2 =$$

$$= \frac{300^2}{3 \cdot 2} \left[\left(50 \cdot \frac{8}{5} - 118,666 \right)^2 + \left(60 \cdot \frac{10}{5} - 118,666 \right)^2 + \left(65 \cdot \frac{12}{5} - 118,666 \right)^2 \right] =$$

$$= \frac{9000}{6} \left[1485{,}111 + 1{,}7778 + 1393{,}777 \right] = 43\,360\,001{,}063$$

puesto que

$$\sum_{i=1}^{n} \frac{\hat{X}_i}{n} = \frac{50 \cdot \frac{8}{5} + 60 \cdot \frac{10}{5} + 65 \cdot \frac{12}{5}}{3} = \frac{80 + 120 + 156}{3} = 118{,}6666$$

En realidad el valor de la estimación de la varianza toma el valor 43 360 000, pero el uso de decimales en el valor de $\hat{\hat{\mathrm{X}}}$ hace que el cálculo sea un poquito mayor, lo cual no influye en el resultado final del ejercicio en absoluto.

Cuando las probabilidades son proporcionales al tamaño en primera etapa las expresiones con el método de conglomerados últimos quedan

de la siguiente forma, utilizando ahora que $P_i = \frac{M_i}{M}$:

$$\hat{V}(\hat{\overline{\mathrm{X}}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{X}_i}{P_i} - \hat{\hat{X}} \right)^2 = \frac{\sum_{i=1}^{n} \left(\frac{\hat{X}_i}{\frac{M_i}{M}} - \frac{M}{n} \sum_{i=1}^{n} \overline{x}_i \right)^2}{n(n-1)} =$$

$$= \frac{\sum_{i=1}^{n} \left(\frac{M_i \cdot \overline{x}_i}{\frac{M_i}{M}} - \frac{M}{n} \sum_{i=1}^{n} \overline{x}_i \right)^2}{n(n-1)} = \frac{M^2}{n(n-1)} \sum_{i=1}^{n} \left(\overline{x}_i - \frac{\sum_{i=1}^{n} \overline{x}_i}{n} \right)^2 =$$

$$= \frac{9000^2}{3 \cdot 2} \left[\left(\frac{8}{5} - 2 \right)^2 + \left(\frac{10}{5} - 2 \right)^2 + \left(\frac{12}{5} - 2 \right)^2 \right] =$$

$$= \frac{81\,000\,000}{3 \cdot 2} (0{,}16 + 0 + 0{,}16) = 4\,320\,000$$

habiendo utilizado que $\frac{\sum_{i=1}^{n} \overline{x}_i}{n} = \frac{\frac{8}{5} + \frac{10}{5} + \frac{12}{5}}{3} = \frac{\frac{30}{5}}{3} = 2$

Puede apreciarse como el error de muestreo en el caso de utilizar probabilidades proporcionales a los tamaños es mucho menor que en el caso de probabilidades iguales, por lo que es preferible el segundo de los estimadores.

Ejercicio 2

Se formula un modelo (de regresión no lineal) de red neuronal de dos capas ocultas completamente conectada con las siguientes características:

i) entradas $\boldsymbol{x} \in \mathbb{R}^6$;

ii) salida $y \in \mathbb{R}$;

iii) primera capa oculta con cuatro nodos y función de activación logística;

iv) segunda capa oculta con dos nodos y función de activación ReLU.



Se pide:

(a) Representar el modelo gráficamente.

(b) Formular la transformación asociada a un nodo de la primera capa oculta. Formule la transformación asociada a un nodo de la segunda capa oculta.

(c) Representar la relación entre la respuesta y la salida de la segunda capa oculta.

(d) Se dispone de datos $(\boldsymbol{x}_i, y_i)_{i=1}^n$ para entrenamiento de la red. Se desea ajustar la red mediante mínimos cuadrados con regularizador de caída de pesos (weight decay). Formule la función objetivo del problema que debería resolverse para ajustar el modelo.

(e) Describir analíticamente una iteración genérica del algoritmo del descenso del gradiente para ese problema. Describa qué rol tendría el método de retropropagación en tal algoritmo.

(f) Describir qué problemas se dan cuando el tamaño n de los datos es grande y el modelo de red neuronal es profundo. Describa cómo pueden mitigarse esos problemas con un método de descenso del gradiente estocástico.

(g) Se dispone de datos $(\boldsymbol{x}_i, y_i)_{i=1}^n$ para ajustar la red. Explique cómo procedería, y con qué propósito, a dividir el conjunto en uno de entrenamiento y otro de contraste para un mejor ajuste del modelo.

Solución.

(a) Presentamos en la Figura 3.3 el esquema para la red neuronal. En amarillo las capas de entrada y salida, con los valores asociados. En verde las dos capas ocultas.

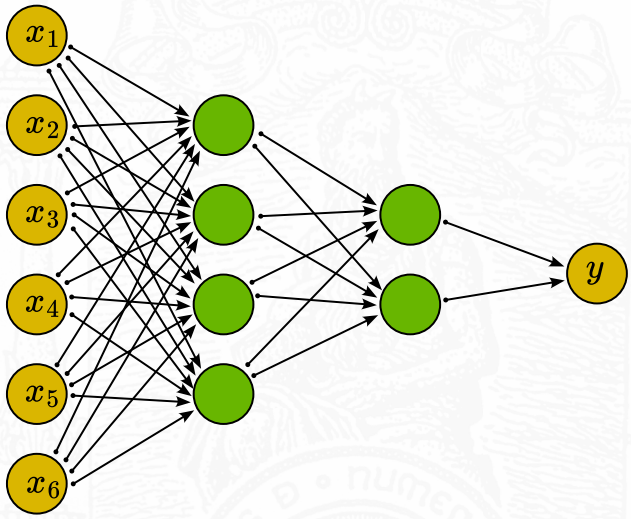


Figura 3.3: Esquema red neuronal.

(b) Denotemos con x_j^i al valor asociado al nodo n_j^i, el j-ésimo de la i-ésima capa, siendo la capa 1 la de inputs, las 2 y 3 las ocultas, y la 4 la de salida. Así $x_1^4 = y$ y $x_3^1 = x_3$.

Entonces, la transformación pedida para el j-ésimo nodo de la primera capa oculta es

$$x_j^2 = \sigma\left(b_j^2 + \sum_{k=1}^{6} \omega_{j,k}^1 x_k^1\right),$$

donde:

- b_j^l es el *bias* asociado al nodo n_j^l.
- $\omega_{j,k}^l$ es el peso que tiene el nodo n_k^l en el j-ésimo nodo de la siguiente capa.



- σ es la sigmoide, dada por $\sigma(x) = \frac{1}{1+e^{-x}}$.

Análogamente, para el j-ésimo nodo de la segunda capa oculta:

$$x_j^3 = \text{relu}\left(b_j^3 + \sum_{k=1}^{4} \omega_{j,k}^2 x_k^2\right),$$

donde

$$\text{relu}(x) = \begin{cases} x & \text{si } x \geq 0 \\ 0 & \text{en otro caso} \end{cases}$$

(c) El enunciado no especifica la función de activación de la última capa (la de salida), estamos asumiendo se trata de la sigmoide, como en la primera capa oculta. De ahí que

$$x_1^4 = \sigma\left(b_1^4 + \sum_{k=1}^{2} \omega_{1,k}^3 x_k^3\right).$$

Por otro lado, tampoco especifica cuál es la función de transformación de los valores de la última capa en una respuesta. Cuando la última capa tiene un solo nodo, como es el caso, es habitual que esa función sea simplemente la identidad. Entonces la respuesta de la red coincide con el valor almacenado en el último nodo:

$$y = x_1^4 = \sigma\left(b_1^4 + \sum_{k=1}^{2} \omega_{1,k}^3 x_k^3\right).$$

(d) Denotemos por μ a la función de regresión asociada a la red, es decir, $\mu(\boldsymbol{\omega}, \boldsymbol{x})$ es el resultado de alimentar la primera capa de la red con los inputs de $\boldsymbol{x}$, propagar hacia adelante utilizando los pesos de $\boldsymbol{\omega}$, y tomar el resultado en la neurona de la última capa.

Considerando fija la muestra de entrenamiento $(\boldsymbol{x}_i, y_i)_{i=1}^n$, para minimizar errores cuadrados, la función de error sería

$$E(\boldsymbol{\omega}) = \sum_{i=1}^{n} (\mu(\boldsymbol{\omega}, \boldsymbol{x}_i) - y_i)^2.$$



Aplicar además un *weight decay* λ supone utilizar, en su lugar,

$$E_\lambda(\boldsymbol{\omega}) = \sum_{i=1}^{n} (\mu(\boldsymbol{\omega}, \boldsymbol{x}_i) - y_i)^2 + \lambda \sum_{\omega \in \boldsymbol{\omega}} \omega^2,$$

es decir, sumar los cuadrados de los pesos multiplicados por ese λ.

(e) En general, si no se aplicara weight decay, el método de gradiente daría lugar a actualizar cada peso ω de cada capa con

$$\omega \leftarrow \omega - \eta \frac{\partial E}{\partial \omega},$$

donde η es el *learning rate* del paso.

Si además hay un weight decay λ, el gradiente da lugar a

$$\omega \leftarrow \omega - \eta \frac{\partial E}{\partial \omega} - \eta \lambda w.$$

El rol del *backpropagation* es calcular eficientemente todas esas derivadas parciales. Como hay dependencia entre ellas, proporciona un orden efectivo para obtenerlas, y los productos y sumas matriciales que permiten calcular de forma sencilla unas a partir de otras.

(f) En principio, el error E depende de la muestra completa, y toda ella debe ser utilizada en el cálculo de cada derivada parcial en el apartado anterior. Eso implica que el algoritmo pueda llegar a ser bastante lento, máxime cuando la red es profunda y consta de numerosas neuronas. El método de gradiente estocástico consiste en seleccionar una submuestra del total en cada paso de recálculo de pesos. La muestra se particiona aleatoriamente en submuestras, y cada una de ellas se utiliza en un paso de aprendizaje por descenso de gradiente. Se denomina *epoch* al conjunto de pasos tras el que todos los elementos de la muestra han sido utilizados.



Esta técnica resulta útil porque las submuestras representan razonablemente bien a la muestra total, con lo que proporcionan direcciones de descenso útiles, y al ser mucho menores se computan considerablemente más rápido.

(g) Se procede eligiendo la proporción de datos que va a formar parte del batch de entremamiento (el resto irá al de validación). Una vez decidida, se construyen ambos aleatoriamente a partir de los datos originales. Durante el entrenamiento el aprendizaje de la red se lleva a cabo utilizando el batch de entrenamiento, pero el grado de precisión se mide sobre el batch de validación. Esto es necesario para evitar *overfitting*, que es el fenómeno que se produce cuando el entrenamiento da lugar a precisiones muy altas sobre los datos utilizados, pero no sobre otros datos con el mismo origen. Sucede cuando el modelo aprende las idiosincrasias de esos datos en particular, en vez de captar las características generales de la población de la que provienen.

Ejercicio 3

Tenemos el siguiente esquema relacional, con información de una pequeña biblioteca:

```
Socio(cód socio, nombre, dirección. email)
Libro(núm registro, título, autor, isbn, ubicación)
EnPréstamo(núm registro, cód socio, fecha préstamo)
```

Se pide diseñar consultas de álgebra relacional para averiguar la siguiente información:

(a) Qué títulos están actualmente en préstamo.



(b) Qué socios tienen más de tres ejemplares en préstamo.

(c) Nombre y direcciones de email de los socios que tienen libros desde hace más de 30 días. (Admitimos la operación "fecha + 30" o "fecha – 30" para indicar fechas posteriores o anteriores en 30 días a una fecha dada y las operaciones "fecha1 < fecha2" para expresar que una fecha es anterior a otra.)

(d) Un socio devuelve un libro. ¿Cuál será la operación de álgebra relacional que actualiza nuestro esquema?

Solución.

Por comodidad, renombremos las relaciones Socio como S, Libro como L y EnPréstamo como P.

a) $\Pi_{\text{num\_registro}} P$

Esta expresión se puede interpretar como tomar toda la relación de préstamos y quedarse con el conjunto de valores distintos del campo `num_registro`.

b) $\Pi_{\text{cod\_socio}}\ \sigma_{\text{nlibros}>3}\ (\rho(\text{cod\_socio}, \text{nlibros})\ \text{cod\_socio}\ \text{COUNT num\_registro(P)})$

Esta expresión se puede interpretar como agrupar en la relación de préstamos por el campo `cod_socio`, contando en cada grupo el número de registros y almacenándolo en un nuevo campo `nlibros`. A continuación, filtrar solo los casos para los que ese `nlibros` sea mayor que tres, y finalmente quedarse solo con el conjunto de valores del campo `cod_socio`.

c) $\Pi_{\text{nombre,email}}(S \bowtie \Pi_{\text{cod\_socio}}\ \sigma_{\text{fecha\_prestamo}+30<\text{hoy}} P)$



Esta expresión se puede interpretar como filtrar en la relación de préstamos por aquellos en los que el campo `fecha_prestima` es más de 30 días más antiguo que el día hoy, luego quedarse solo con el conjunto de socios. Posteriormente cruzar con la relación de socios para tener todos los datos de cada uno de esos socios, y finalmente quedarse solo con las columnas `nombre` e `email`.

d) El álgebra relacional no tiene operaciones que modifiquen el esquema. No obstante, si interpretamos la cuestión como qué operación podría fabricar, a partir de la relación de préstamos P original, una nueva versión P' que ya refleje la devolución del libro, la operación podría ser una diferencia. En concreto,

$$P' \leftarrow P - \sigma_{\text{num\_registro=num\_reg\_devuelto} \wedge \text{cod\_socio=cod\_soc\_devuelve}} P.$$

Lo que hace es filtrar en la relación de préstamos aquellos en los que tanto el socio como el libro coinciden con el de la devolución, y posteriormente filtrar una nueva copia de la relación de préstamos eliminando los primeros.

Ejercicio 4

Tenemos un archivo de texto plano, llamado “población.csv”, con datos de poblaciones del mundo. He aquí el inicio de su contenido:




```
7.975.105.155
China,1.425.887.337
India,1.417.173.173
Estados Unidos,338.289.857
Indonesia,275.501.339
```

La primera línea es la población mundial total; las siguientes, las de cada país. El nombre del país puede contener más de una palabra (ejemplos: Corea del Norte, Islas Vírgenes de los Estados Unidos). Las cifras de poblaciones vienen en el formado mostrado, con puntos separando los miles, millones, etc. Los nombres de los países y las cifras vienen separados con una coma.

Deseamos elegir el nombre de un país aleatoriamente, con probabilidades proporcionales a sus poblaciones. Se pide diseñar un programa o algoritmo para ello.

El diseño pedido puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R. En la simulación de la variable aleatoria que se necesite, se podrá asumir que se tiene definida y disponible una función básica que simule una variable aleatoria uniforme continua U(0, 1). No se podrá usar ninguna otra función aleatoria.

Solución.

A continuación, se escribe el diseño pedido en pseudocódigo. Sería similar en el resto de lenguajes.



1: Leer el archivo “poblacion.csv”
2: Inicializar una lista vacía llamada “paises”
3: Inicializar una lista vacía llamada “poblaciones”
4: **for** cada línea en el archivo “poblacion.csv”, excepto la primera línea **do**
5: Separar la línea en dos partes usando la coma como delimitador
6: Añadir el nombre del país a la lista “paises”
7: Añadir la población del país a la lista “poblaciones”
8: **end for**
9: Sumar todas las poblaciones en la lista “poblaciones” para obtener la población total
10: Generar un número aleatorio uniforme continuo entre 0 y 1 con la función proporcionada
11: Multiplicar este número por la población total para obtener un valor dentro del rango de las poblaciones
12: Inicializar una variable llamada “acumulador” a 0
13: Inicializar una variable llamada “indice” a 0
14: **while** “acumulador” sea menor que el valor generado **do**
15: Sumar la población del país en el índice actual de la lista “poblaciones” al “acumulador”
16: Incrementar el “indice” en 1
17: **end while**
18: Devolver el país en el índice correspondiente en la lista “paises”



En el código, en primer lugar, se lee el archivo CSV de población. A continuación, se inicializan dos listas: una para almacenar los nombres de los países y otra para almacenar sus poblaciones. Para cada línea en el archivo CSV, exceptuando la primera línea que contiene la población mundial total, se separan el país y su población, haciendo uso del delimitador y se van añadiendo, respectivamente, el nombre del país a la lista de nombres y la población a la lista de poblaciones. De esta forma, para una posición dada, en la lista de países se encuentra el nombre de un país y en la lista de poblaciones su correspondiente población. Sumando la lista de poblaciones, se calcula la población total. Después se genera un número aleatorio uniforme continuo entre 0 y 1 haciendo uso de la función proporcionada y se multiplica este número por la población total recien calculada. De esta forma, se obtiene un número aleatorio dentro del rango de las poblaciones. Después de generar el número aleatorio, hay que seleccionar un país utilizando dicho número aleatorio. Para ello, se utiliza un bucle que itera sobre la lista de poblaciones de los países. El bucle recorre cada país en orden y va sumando las poblaciones al acumulador. Dicho acumulador se compara con el número aleatorio generado y cuando alcanza o supera el valor del número aleatorio, se detiene el bucle y se devuelve el país correspondiente a ese índice. Esto asegura que los países con mayores poblaciones contribuyan más al acumulador, aumentando así las probabilidades de ser seleccionados.

La implementación en Python de este caso concreto podría ser la siguiente, que utiliza una variable menos en el planteamiento respecto del desarrollo del pseudocódigo anterior, puesto que al decrementar el umbral en cada iteración no se necesita la variable acumulativa.

```
from scipy.stats import uniform
```




```python
def lee_fichero(nombre):
    """Funcion de lectura del fichero a bajo nivel. Evitamos usar
      pandas, que lo haria muy sencillo."""
    pobs = []
    paises = []
    with open(nombre) as f:
        next(f)  # Ignoramos la primera linea con la poblacion total
        for line in f:
            line = line.strip()
            if line:
                line = line.replace('.', '')  # Quitamos los separadores de miles
                pais, pob = line.split(',')
                paises.append(pais)
                pobs.append(int(pob))
    return paises, pobs


def sorteo_ponderado(paises, pobs):
    umbral = uniform.rvs() * sum(pobs)
    for pais, pob in zip(paises, pobs):
        umbral -= pob
        if umbral <= 0:
            return pais


if __name__ == "__main__":
    paises, pobs = lee_fichero('datos.csv')
    pais_elegido = sorteo_ponderado(paises, pobs)
    print(f"El pais elegido fue '{pais_elegido}'.")
```



Ejercicio 5

Sea una población formada por N=40 empresas. Se quiere estimar el total de su cifra de negocios (CN) y para ello se obtiene una muestra aleatoria simple (sin reemplazamiento) de n=4 empresas.

Se conoce el número de ocupados (x) de cada una de las empresas de la población. El total de ocupados en la población es X=160.

Los datos disponibles son los siguientes:

Unidades de la Muestra (i)	Y (CN, miles de euros)	X (ocupados)
1	100	3
2	200	5
3	500	7
4	300	5

Se pide:

(a) Estimar el total de la Y (CN) por el método de la razón.

(b) Estimar el sesgo del estimador del apartado anterior.

(c) Estimar su varianza.

(d) Explicar si es o no relevante el efecto del sesgo en el intervalo de confianza a partir de los resultados de los apartados anteriores.

Solución.

(a) El estimador del total poblacional de la cifra de negocios obtenido por le método de la razón se escribe como sigue:

$$\widehat{Y}_R = \frac{y}{x}X = \frac{\overline{y}}{\overline{x}}X = \hat{R}X = \frac{\frac{1100}{4}}{\frac{20}{4}} \cdot 160 = 55 \cdot 160 = 8800$$

(b) Calculemos la esperanza del estimador anterior para analizar el sesgo y calcularlo. Sabemos que el estimador de la razón es sesgado y llamaremos sesgo a $B(\bullet)$.

$$E\left(\widehat{Y}_R\right) = E(\hat{R}X) = XE(\hat{R}) = X(R + B(\hat{R})) = XR + XB(\hat{R})$$

Como $R = \frac{Y}{X}$ el sesgo del estimador es $XB(\hat{R})$ y como conocemos el valor de X, para estimar el sesgo en realidad lo que tenemos que



estimar es el sesgo del estimador de razón, es decir,

$$\hat{B}(\hat{R}) = \frac{(1-f)}{n\overline{X}^2}\left(\hat{R}\hat{S}_x^2 - \hat{S}_{xy}\right) = \frac{\left(1-\frac{4}{40}\right)}{4\cdot\left(\frac{160}{40}\right)^2}\left(55\cdot\frac{8}{3} - \frac{800}{3}\right) =$$
$$= \frac{0{,}9}{64}\left(\frac{-360}{3}\right) = -1{,}6875$$

donde hemos utilizado los resultados siguientes, considerando que $n = 4$:

$$\hat{S}_x^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1} = \frac{1}{3}\left[(3-5)^2+(5-5)^2+(7-5)^2+(5-5)^2\right] = \frac{8}{3}$$

$$\hat{S}_{xy} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1} =$$
$$= \frac{1}{3}\left[(3-5)(100-275)+(5-5)(200-275)+(7-5)(500-275)+(5-5)(300-275\right.$$
$$= \frac{1}{3}\left[350+450\right] = \frac{800}{3}$$

Por tanto el sesgo de $\widehat{Y}_R$ será $B(\widehat{Y}_R) = X\cdot B(\hat{R})$ y su estimación

$$\hat{B}(\widehat{Y}_R) = X\cdot\hat{B}(\hat{R}) = 160\cdot(-1{,}6875) = -270$$

(c) La expresión de la varianza de $\widehat{Y}_R$ es $V(\widehat{Y}_R) = V(\hat{R}X) = X^2\cdot V(\hat{R})$ y su estimación vale

$$\widehat{V}(\widehat{Y}_R) = X^2\cdot\widehat{V}(\hat{R}) = X^2\frac{(1-f)}{n\cdot\overline{X}^2}\left(\hat{S}_y^2+\hat{R}^2\hat{S}_x^2-2\hat{R}\hat{S}_{xy}\right) =$$
$$= N^2\frac{(1-f)}{n}\left(\hat{S}_y^2+\hat{R}^2\hat{S}_x^2-2\hat{R}\hat{S}_{xy}\right) =$$
$$= 40^2\frac{\left(1-\frac{4}{40}\right)}{4}\left[29\,166{,}6666+55^2\frac{8}{3}-2\cdot55\cdot\frac{800}{3}\right] =$$
$$= 1600\frac{0{,}9}{4}\left[29\,166{,}6666+8066{,}6666-29\,333{,}3333\right] =$$
$$= 1600\frac{0{,}9}{4}\left[7899{,}999999\right] = 2\,843\,999{,}9999 \approx 2\,844\,000$$

donde hemos utilizado el dato de $\hat{S}_y^2$, que se calcula como

$$\hat{S}_y^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1} = \frac{1}{3}\left[(100-275)^2 + (200-275)^2 + (500-275)^2 + (300-275)^2\right] =$$
$$= \frac{1}{3}\left[30\,625 + 5625 + 50\,625 + 625\right] = \frac{87\,500}{3} = 29\,166{,}6666$$

y la desviación típica estimada será $\widehat{\sigma}(\widehat{Y}_R) = \sqrt{2\,844\,000} = 1683,449$

En este cálculo de la estimación de la varianza utilizamos las varianzas estimadas de cada una de las variables, además de la covarianza estimada entre ellas, pero hay una forma, teniendo en cuenta los datos que tenemos del problema, que no requeriría del conocimiento de las mismas, aliviando en cierto sentido las operaciones, que sería utilizando la fórmula:

$$\widehat{V}(\widehat{Y}_R) = N^2\frac{(1-f)}{n}\left(\hat{S}_y^2 + \hat{R}^2\hat{S}_x^2 - 2\hat{R}\hat{S}_{xy}\right) =$$
$$= N^2\frac{1-f}{n(n-1)}\left[\sum_{i=1}^{n} y_i^2 + \hat{R}^2\sum_{i=1}^{n} x_i^2 - 2\hat{R}\sum_{i=1}^{n} x_i \cdot y_i\right]$$

en la cual, las sumas que aparecen son muy sencillas y rápidas de calcular.

(d) Para determinar si el sesgo del estimador del total es influyente en el intervalo de confianza lo que suele hacerse es calcular el sesgo relativo,

$$\left|\frac{\hat{B}(\widehat{Y}_R)}{\hat{\sigma}(\hat{Y}_R)}\right| = \left|\frac{-270}{1683,449}\right| = 0{,}1604$$

Se ha podido comprobar experimentalmente que la región del error aumenta conforme este cociente aumenta. Cochran obtiene como regla de trabajo que el efecto del sesgo en la exactitud de un estimador es despreciable si este cociente es menor que 0,1, aunque indica que es



posible llegar al valor 0,2. En el caso que nos ocupa precisamente está entre estos dos valores, por lo que parece que sí puede llegar a ser influyente en el cálculo de un intervalo de confianza.

Ejercicio 6

Suponga que las observaciones (x, y) de un problema pueden modelizarse mediante regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, donde $\epsilon_1, ..., \epsilon_n$ se distribuyen independientemente según una $N(0, \sigma^2)$. Para simplificar, suponga que σ es conocida e igual a 1. Se pide:

(a) Escribir la log-verosimilitud de los datos.

(b) Suponga que, a priori, (β_0, β_1) son independientes e idénticamente distribuidas según nuna distribución normal de media 0 y varianza c^2. Calcule la distribución a posteriori de los parámetros.

(c) Basándose en tal distribución a posteriori, explique cómo contrastaría la hipótesis nula $H_0 : \beta_1 = 0$.

(d) Calcular la distribución predictiva de y para una nueva observación x_{n+1}.

(e) Dar un intervalo predictivo de probabilidad 0.88 para y dado x_{n+1}, basado en la salida de una rutina $rnorm(s)$ que permite generar s observaciones de una normal estándar.

Solución.

(a) Sea $\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}$ la función de densidad de las variables de error ϵ_i. Puesto que $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, la función de densidad de y_i condicionada

a x_i y los parámetros será

$$f\left(y_i|\beta_0,\beta_1,x_i\right)=\phi(y_i-(\beta_0+\beta_1x_i))=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y_i-(\beta_0+\beta_1x_i))^2}{2\sigma^2}}.$$

Dado que los errores son independientes e idénticamente distribuidos (iid), la verosimilitud de los datos es el producto de las densidades de probabilidad de cada observación. Por lo tanto, la función de verosimilitud $L(\beta_0,\beta_1)$ es:

$$\begin{aligned}L(\beta_0,\beta_1|y)&=\prod_{i=1}^{n}f\left(y_i|\beta_0,\beta_1,x_i\right)=\\&=\frac{1}{\sigma^n(\sqrt{2\pi})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-(\beta_0+\beta_1x_i)\right)^2\right)\end{aligned}$$

y la log-verosimilitud $l(\beta_0,\beta_1|y)$ es el logaritmo natural de la función de verosimilitud:

$$l(\beta_0,\beta_1|y)=\log(L(\beta_0,\beta_1|y))=\sum_{i=1}^{n}\log\left(f\left(y_i|\beta_0,\beta_1,x_i\right)\right)$$

Sustituyendo:

$$l(\beta_0,\beta_1|y)=-n\log\sigma-\frac{n}{2}\log\left(2\pi\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-(\beta_0+\beta_1x_i)\right)^2.$$

(b) Para encontrar la distribución a posteriori de los parámetros (β_0,β_1), partamos de la distribución a priori conjunta de dichos parámetros, cuya función de densidad viene dada por

$$p(\beta_0,\beta_1)=p(\beta_0)\cdot p(\beta_1)=\frac{1}{c\sqrt{2\pi}}e^{-\frac{\beta_0^2}{2c^2}}\cdot\frac{1}{c\sqrt{2\pi}}e^{-\frac{\beta_1^2}{2c^2}},$$

cuya expresión es proporcional a

$$p(\beta_0,\beta_1)\propto\exp\left(-\frac{\beta_0^2+\beta_1^2}{2c^2}\right).$$

De hecho la distribución conjunta de (β_0,β_1) es una normal $N(\mathbf{0},c^2\boldsymbol{I})$.



Ahora, utilizando el Teorema de Bayes, la función de densidad de la distribución a posteriori es proporcional al producto de la verosimilitud y la función de densidad de la distribución a priori:

$$\pi(\beta_0, \beta_1|x, y) \propto p(\beta_0, \beta_1) \cdot L(\beta_0, \beta_1|y).$$

Sustituyendo sus expresiones obtenemos

$$\pi(\beta_0, \beta_1|x, y) \propto \exp\left(-\frac{\beta_0^2 + \beta_1^2}{2c^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2\right). \quad (3.1)$$

Podríamos detenernos aquí, pero si queremos entender realmente lo que está pasando, puesto que se trata de la exponencial de un polinomio de segundo grado sobre β_0 y β_1, cabe preguntarse si se trata de una distribución binormal (una distribución normal multivariante con dos componentes). De ser así, su función de densidad debería ser proporcional a una expresión del tipo

$$\exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}) \cdot \boldsymbol{Q} \cdot (\boldsymbol{\beta} - \boldsymbol{\mu})^\top\right), \quad (3.2)$$

donde

- $\boldsymbol{\beta} = (\beta_0, \beta_1)$
- $\boldsymbol{\mu} = (\mu_0, \mu_1)$ es el posible vector de esperanzas
- $\boldsymbol{Q} = \begin{pmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{pmatrix}$ es la inversa de la posible matriz de covarianzas Σ.

Para ver si en efecto la expresión (3.1) puede encajar con la expresión (3.2) necesitamos poner ambas como polinomio en $\{\beta_0, \beta_1\}$ e igualar coeficientes.

La expresión (3.1), tras operar y agrupar, tomando logaritmos y eliminando ese factor $-\frac{1}{2}$ se reduce a

$$\beta_0^2\left(\frac{1}{c^2}+\frac{n}{\sigma^2}\right)+\beta_1^2\left(\frac{1}{c^2}+\frac{\sum_i x_i^2}{\sigma^2}\right)+\beta_0\beta_1\left(2\frac{\sum_i x_i}{\sigma^2}\right)+$$
$$\beta_0\left(-2\frac{\sum_i y_i}{\sigma^2}\right)+\beta_1\left(-2\frac{\sum_i x_i y_i}{\sigma^2}\right)+\left(\frac{\sum_i y_i^2}{\sigma^2}\right).$$

La expresión (3.2), tras operar y agrupar, tomando logaritmos y eliminando ese factor $-\frac{1}{2}$ se reduce a

$$\beta_0^2\left(q_{11}\right)+\beta_1^2\left(q_{22}\right)+\beta_0\beta_1\left(2q_{12}\right)+\beta_0\left(-2q_{11}\mu_0-2q_{12}\mu_1\right)+$$
$$\beta_1\left(-2q_{12}\mu_0-2q_{22}\mu_1\right)+\left(q_{11}\mu_0^2+2q_{12}\mu_0\mu_1+q_{22}\mu_1^2\right).$$

Igualando coecifientes resulta que es posible satisfacer todas las ecuaciones tomando:

- $\boldsymbol{Q}=\begin{pmatrix}\frac{1}{c^2}+\frac{n}{\sigma^2} & \frac{\sum_i x_i}{\sigma^2}\\ \frac{\sum_i x_i}{\sigma^2} & \frac{1}{c^2}+\frac{\sum_i x_i^2}{\sigma^2}\end{pmatrix}$
- $\boldsymbol{\mu}=\boldsymbol{Q}\cdot\begin{pmatrix}\frac{\sum_i y_i}{\sigma^2}\\ \frac{\sum_i x_i y_i}{\sigma^2}\end{pmatrix}$

Es decir, la distribución conjunta de (β_0,β_1) a posteriori es una binormal $\mathcal{N}_2(\boldsymbol{\mu},\boldsymbol{\Sigma}=\boldsymbol{Q}^{-1})$. Debemos invertir $\boldsymbol{Q}$ para obtener la matriz de covarianzas Σ.

Para facilitar la comprensión trabajaremos con unos datos concretos que utilizaremos para realizar algunos cálculos y plots. Hemos elegido $c=0{,}25$. Los puntos (x,y) los obtenemos sumando una muestra de errores normales estándar independientes a las ordenadas de una recta cuyas β_0 y β_1 sorteamos según la distribución a priori, y resultaron ser $\beta_0=-0{,}3562, \beta_1=0{,}22356$. La secuencia de puntos resultante fue la de la Figura 3.4.



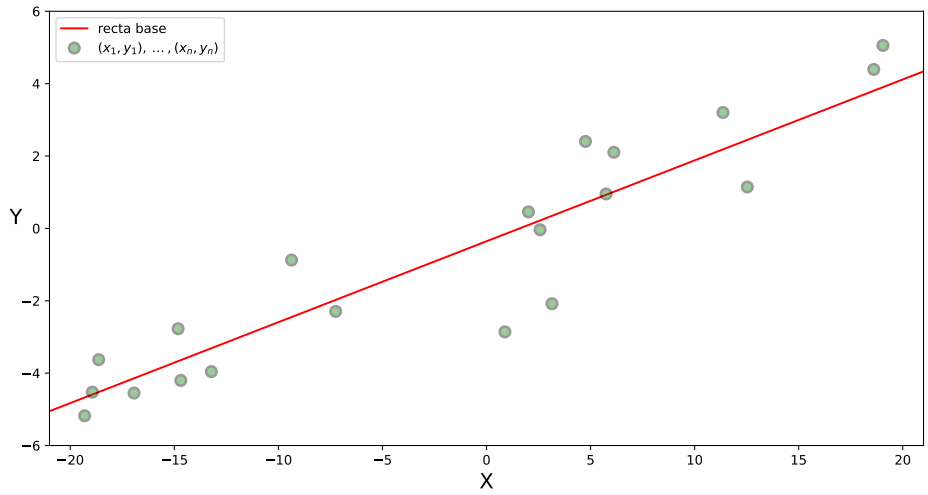


Figura 3.4: Muestra $(x_1, y_1), \ldots, (x_n, y_n)$ que utilizaremos para los cálculos y plots ilustrativos.

Para los datos mencionados, las funciones de densidad a priori y a posteriori son las mostradas en la Figura 3.5. Se obtienen valores para el vector de esperanzas $\boldsymbol{\mu} = (-0{,}180, 0{,}232)$, Y las desviaciones típicas de β_0, β_1 a posteriori se han reducido desde el $c = 0{,}25$ inicial a 0,168 y 0,017, respectivamente.

Otro aspecto interesante puede ser graficar muestras de rectas provenientes de las distribuciones a priori y a posteriori de (β_0, β_1), cosa que hacemos en la Figura 3.6.

(c) Va a resultar sencillo contrastar la hipótesis nula $H_0 : \beta_1 = 0$ basándose en los cálculos de la distribución a posteriori de β_1 obtenida en el apartado anterior, ya que llegamos a la conclusión de que es una $\mathcal{N}(\mu_1, \sigma^2 = \Sigma_{22})$.

Dado un nivel de significación α, rechazaremos la hipótesis nula cuando el punto 0 esté en una de las dos colas de la distribución de β_1 con

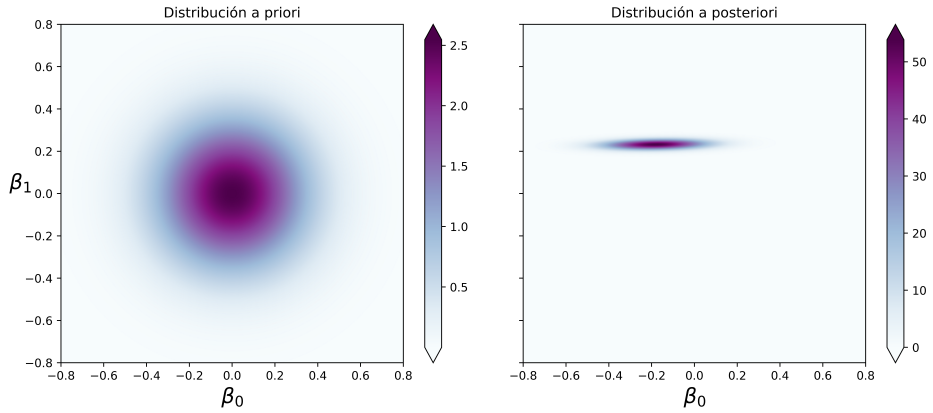


Figura 3.5: Distribuciones del vector (β_0, β_1). El plot de la izquierda muestra la distribución a priori, en la que ambas variables son independientes, tienen esperanza 0 y la misma varianza. El plot derecho muestra la distribución condicionada al conocimiento de los puntos de la Figura 3.4. Se aprecia cómo tanto β_0 como β_1 se mueven ya en una zona mucho más reducida, y también cómo la pendiente β_1 se va conociendo mucho mejor que la ordenada en el origen β_0.

área $\frac{\alpha}{2}$. En caso contrario, la aceptaríamos, lo que puede expresarse formalmente como

$$\Phi^{-1}\left(\frac{\alpha}{2}\right) \leq \frac{\mu_1}{\sigma_1} \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right),$$

donde Φ es la función de distribución de una normal estándar. Puesto que $\Phi^{-1}\left(1-\frac{\alpha}{2}\right) = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$ la condición anterior se reduce a

$$\left|\frac{\mu_1}{\sigma_1}\right| \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right).$$

(d) Dada x_{n+1}, buscamos la distribución predictiva para la Y asociada. En este apartado asumiremos siempre que nos estamos refiriendo a las distribuciones a posteriori. Tenemos

$$Y = \beta_0 + \beta_1 x_{n+1} + \epsilon_{n+1},$$

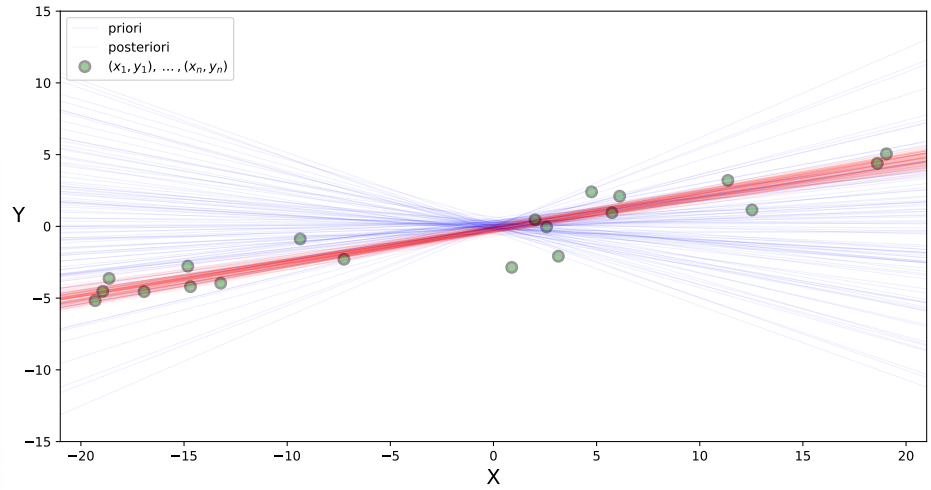


Figura 3.6: Distribuciones de un muestreo de 100 rectas asociadas al vector (β_0, β_1). En azul aparecen ejemplos provenientes de la distribución a priori (que ignora los puntos de la muestra), mientras que en rojo aparece la muestra de rectas a posteriori que sí los tiene en cuenta.

por lo visto en los apartados anteriores sabemos que las distribuciones a posteriori de β_0 y β_1 son normales, con lo que al tratarse de una suma de normales de nuevo la distribución predictiva de la Y buscada será una normal. Determinemos su esperanza y su varianza.

En cuanto a la esperanza, es inmediato:

$$\mathbb{E}[Y] = \mathbb{E}[\beta_0] + x_{n+1}\mathbb{E}[\beta_1] + \mathbb{E}[\epsilon_{n+1}] = \mu_0 + x_{n+1}\mu_1.$$

Respecto a la varianza, la matriz 3×3 de covarianzas para el vector $(\beta_0, \beta_1, \epsilon_{n+1})$ es

$$\begin{pmatrix} \Sigma & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

donde Σ es la matriz 2×2 de covarianzas para (β_0, β_1) tal como se calculó en el apartado b).



Ahora utilizamos un resultado conocido: si $\boldsymbol{X}$ es una normal multivariante, para cualquier matriz M se tiene que

$$\mathrm{Cov}(M \cdot \boldsymbol{X}) = M \cdot \mathrm{Cov}(\boldsymbol{X}) \cdot M^{\top}.$$

Puesto que

$$Y = \begin{pmatrix} 1 & x_{n+1} & 1 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \epsilon_{n+1} \end{pmatrix}$$

es sencillo calcular su varianza con la fórmula anterior, resultando

$$\mathrm{Var}(Y) = \begin{pmatrix} 1 & x_{n+1} & 1 \end{pmatrix} \cdot \begin{pmatrix} \Sigma & 0 \\ 0 & \sigma^2 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ x_{n+1} \\ 1 \end{pmatrix}$$

(e) En principio, para calcular el intervalo predictivo podríamos hacerlo de forma exacta a partir de los cálculos del apartado anterior, ya que la distribución de Y es normal y conocemos su esperanza y varianza. Bastaría tomar como extremos del intervalo de confianza los puntos donde la función de distribución de esa normal acumula áreas de $\frac{\alpha}{2} = 0{,}06$ y $1 - \frac{\alpha}{2} = 0{,}94$.

Ahora bien, puesto que el enunciado indica que se utilize la rutina `rnorm`, se asume que espera una estimación del intervalo por Monte Carlo con s simulaciones. Para ello necesitaremos generar una muestra con s valores de la variable Y, y los extremos del intervalo predictivo serán los percentiles 6 y 94 de esa muestra. Tenemos dos formas de obtener la muestra:

- Si se permite utilizar lo calculado en el apartado d), bastará tomar los normales estándar fabricados por `rnorm`, multiplicarlos



por la desviación típica calculada para Y y sumar finalmente su esperanza.

- En caso contrario, debemos hacer la simulación primeramente de los (β_0, β_1) a posteriori. Puesto que conocemos su vector de esperanzas $\boldsymbol{\mu}$ y matriz de covarianzas Σ, puede hacerse mediante:

$$\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \mu^\top + \text{cholesky}(\Sigma) \cdot \begin{pmatrix} N_0 \\ N_1 \end{pmatrix}$$

donde N_0 y N_1 son normales estándar independientes fabricadas por `rnorm`.

Una vez sorteadas β_0 y β_1, debe sortearse otra normal independiente y multiplicarla por σ para obtener el valor de ϵ_{n+1}, y tomar para la distribución empírica que estamos construyendo el valor $y = \beta_0 + \beta_1 x_{n+1} + \epsilon_{n+1}$. La operación completa (el sorteo de las betas, el error epsilon y la suma final) debe hacerse s veces, para tener la muestra mencionada.

Ejercicio 7

Un supermercado necesita diseñar una base de datos para gestionar su inventario de productos y controlar las ventas. Cada producto se registra con su código, una descripción sucinta, su precio y las cantidades de dicho producto mínima (para reponer existencias cuando baje de dicha cantidad), real (la que hay realmente en los estantes) y máxima (la que cabe en los estantes). Cada cliente se registra con su email, su nombre y su domicilio). Cada vez que un cliente acude al supermercado, al pagar se genera un único tique, con la fecha, la lista de productos y la cantidad de cada uno, y el



importe total de la factura. El supermercado registra, además, lógicamente, el cliente que ha realizado dicha compra. Se pide:

(a) Realizar un diseño E/R de dicha base de datos.

(b) Definir, con este diseño, los siguientes conceptos, si están presentes, o inventa un ejemplo sencillo pero apropiado para cada uno de ellos:

- Entidades fuertes, entidades débiles.
- Atributos, claves.
- Relación uno a uno, uno a muchos, muchos a muchos.

(c) Participación total o parcial en una relación.

(d) Atributos de una relación.

(e) Atributo multivalorado, atributo calculado.

Solución.

(a) Un posible diagrama E/R para el enunciado aparece en la Figura 3.7.

(b) Una entidad es fuerte si las tuplas quedan totalmente identificadas a través del valor de un campo clave, sin que sea necesario relacionarlas con elementos de otras entidad para conseguir esa identificación total. En caso contrario se dice que la entidad es débil. Por ejemplo, la entidad `Producto` o `Cliente` de nuestro esquema son fuertes. La entidad `Ticket` hubiera sido débil si no hubiéramos introducido el atributo clave `código` como identificador único. Si no estuviera, puesto que dos tickets podrían corresponder a la misma fecha e importe, la única forma de saber exactamente de qué ticket estamos hablando sería cruzar con `Cliente` para ver quién lo pagó, y con los productos que contiene.

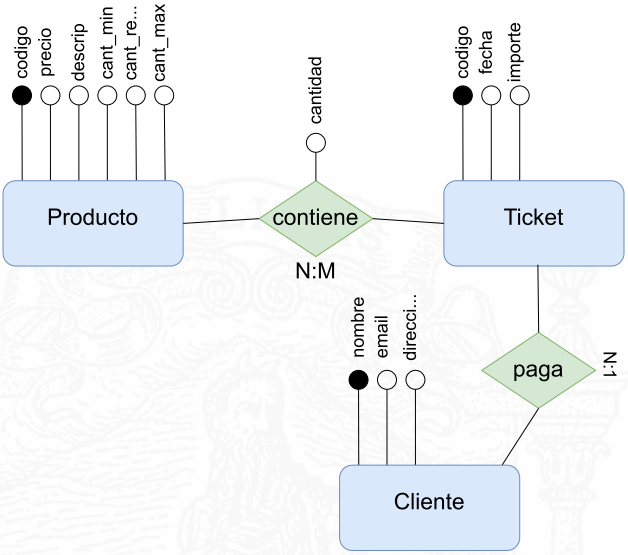


Figura 3.7: Esquema E/R para la bbdd del supermercado.

Atributos son los distintos campos que se introducen en cada tupla de cada entidad, y son los mismos para todas ellas.

Un atributo es clave candidata si no puede haber dos tuplas en la entidad que coincidan en los valores de ese atributo. Es clave ajena si hace referencia a otro atributo de otra entidad.

Una relación entre dos entidades es "uno a uno" si cada elemento de una de las entidades se relaciona como mucho con uno de la otra. Es "uno a muchos" si los elementos de la primera pueden relacionarse con varios de la segunda simultáneamente, pero no al contrario. Es "muchos a muchos" cuando cabe que tanto los de la primera como los de la segunda entidad se relacionen con varios de la otra entidad.

(c) Una entidad participa en una relación de forma total si es obligatorio que cada tupla en la entidad participe en la relación. En caso contrario, se dice que la participación es parcial.



(d) Una relación puede tener atributos, que aportan un valor o valores adicionales a la forma en que los elementos de las entidades se relacionan. En el E/R del primer apartado, la relación `contiene` tiene un atributo `cantidad`, que indica cuántas veces se compra cada producto en el ticket.

(e) Un atributo multivalorado es aquel que para una misma tupla podría tener varios valores. Por ejemplo, el campo `direccion` del cliente podría ser multivaluado, indicando que es posible que un cliente introduzca varias direcciones si dispone de varios domicilios.

Un atributo es calculado si es función del resto de valores y podría ser deducido.

Ejercicio 8

Tenemos un archivo de texto plano, llamado "pps.csv"(abreviando país, población, superficie, separados por comas), con datos de poblaciones y superficies de países del mundo. He aquí el inicio de su contenido:

```
China,1.425.887.337,9.562.900
India,1.417.173.173,3.287.300
Estados Unidos,338.289.857,9.831.500
Indonesia,275.501.339,1.913.600
```

Cada línea consigna el nombre de cada país, seguido de su población y su superficie en kilómetros cuadrados. El nombre del país puede contener más de una palabra (ejemplos: Corea del Norte, Islas Vírgenes de los Estados Unidos). Las cifras de poblaciones vienen en el formado mostrado, con puntos separando los miles, millones, etc. Los nombres de los países y las cifras vienen



separados con una coma. En las cantidades, se usa el punto para separar miles, millones, etc. Se pide:

(a) Diseñar un programa o algoritmo para obtener el coeficiente de correlación entre ambas variables, superficie y población.

(b) Como sabes, es posible realizar esto recorriendo el archivo dos veces, una para calcular las medias y otra para recopilar el resto de la información. Pero la lectura de archivos es bastante costosa computacionalmente. Por eso, en este segundo apartado se pide diseñar un programa o algoritmo para obtener obligatoriamente este coeficiente leyendo el archivo en una sola pasada. Si ya lo has hecho así en tu primer apartado, basta con que en éste lo menciones.

El diseño puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R. No está permitido usar ninguna función predefinida que calcule la media ni otras medidas estadísticas.

Solución.

Veamos una implementación del algoritmo pedido, en Python. Se lee el fichero una sola vez, creando una lista de superficies y otra de poblaciones. Lo natural en Python habría sido usar `pandas.read_csv`, por si se prefiere una versión más a bajo nivel se ha programado la lectura en la función `lee_fichero`.

Recordemos que la fórmula del coeficiente de Correlación de Pearson muestral entre dos vectores $\boldsymbol{x}$ e $\boldsymbol{y}$ es

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum_i (x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i (x_i-\overline{x})^2}\sqrt{\sum_i (y_i-\overline{y})^2}},$$

donde $\overline{x} = \frac{\sum_{i=1}^n x_i}{n}$, $\overline{y} = \frac{\sum_{i=1}^n y_i}{n}$.




```python
import math


def lee_fichero(nombre):
    """Funcion de lectura del fichero a bajo nivel. Evitamos usar pandas.read_csv,
    que lo haria muy sencillo, por si no esta permitido."""
    pobs = []
    supefs = []
    with open(nombre) as f:
        for line in f:
            line = line.strip()
            if line:
                line = line.replace('.', '')  # Quitamos los separadores de miles
                _, pob, sup = line.split(',')
                pobs.append(int(pob))
                supefs.append(int(sup))
return pobs, supefs



def suma(datos):
    """Suma los elementos en datos.
    Es como el 'sum' de Python, la programamos por si no esta permitido usarla."""
    total = 0
    for d in datos:
        total += d
    return total

def media(datos):
    """Calcula la media de los elementos en datos."""
    return suma(datos) / len(datos)
```




```python
def correlacion(xs, ys):
    """Calcula el coeficiente de correlacion de las series xs e ys."""
    xmed = media(xs)
    ymed = media(ys)
    totxy = suma((x - xmed) * (y - ymed) for (x, y) in zip(xs, ys))
    totxx = suma((x - xmed)**2 for x in xs)
    totyy = suma((y - ymed)**2 for y in ys)
    return totxy / math.sqrt(totxx * totyy)


# Codigo con una invocacion de prueba:
if __name__ == "__main__":
    pobs, superfs = lee_fichero('datos.csv')
    corr = correlacion(pobs, superfs)
    print(f"La correlacion calculada es {corr}.")
```

Aclaraciones a la solución

Aclaramos algunas construcciones de Python utilizadas.

1. El `zip` es la función *cremallera* que empareja elementos de distintos iterables para formar tuplas. En el ejemplo, `zip(xs, ys)` es un nuevo iterable que empareja el primer elemento en `xs` con el primero en `ys`, luego los segundo elementos, etc.

2. La asignación

   ```
   _, pob, sup = line.split(',')
   ```

   trocea la cadena `line` según el separador coma, se asegura de que salen exactamente tres subcadenas, asocia la variabla `pob` a la segunda de ellas y `sup` a la tercera.



3. La construcción

   ```
   [(x - xmed)**2 for x in xs]
   ```

   dentro de

   ```
   suma((x - xmed)**2 for x in xs)
   ```

   es una *lista intensional* o *comprehension list*. Consiste en la lista que resulta al tomar cada elemento `x` de `xs`, restarle `xmed` y elevar el resultado al cuadrado.

   Por ejemplo,

   ```
   [x * x + 1 for x in [2, 3, 5, 7]]
   ```

   produciría la lista `[5, 10, 26, 50]`.



3.3. Soluciones del segundo examen del Cuerpo de Diplomados en Estadística del Estado de 2022. Promoción interna del grupo C1.

La promoción desde el grupo C1 al cuerpo de diplomados en estadística del estado puede hacerse desde dos vías. O bien desde el grupo C1 general o bien desde el grupo C1 especialidad estadística. Desde el primero de ellos el segundo examen de problemas consta de 10 problemas, que resolvemos a continuación, y para la especialidad estadística el examen solo tiene 8 problemas, pero todos ellos son comunes a los incluídos en el grupo general. Especificaremos para cada ejercicio si es común a los dos grupos o es específico del grupo general y está excluido para los C1 de la especialidad estadística.

Ejercicio 1.

Este ejercicio el mismo que el ejercicio 1 del examen de promoción interna desde C1 especialidad estadística.

Un colectivo de personas ha recibido las rentas anuales que se presentan en la siguiente distribución de frecuencias:



Rentas (en 10^4€)	Nº de personas
4–6	12
6–8	13
8–10	35
10–12	25
12–14	15
	100

Calcular:

1) La renta media

2) La moda de la distribución de renta

3) La mediana de la distribución de renta

4) La concentración de renta en el colectivo (índice de Gini).

Solución.

Antes de nada, completemos la tabla con algunas columnas adicionales que serán de mucha ayuda.

Rentas (en 10^4€)	i	n_i	x_i	N_i	p_i	$x_i n_i$	u_i	q_i
4–6	1	12	5	12	0,12	60	60	0,0641
6–8	2	13	7	25	0,25	91	151	0,16132
8–10	3	35	9	60	0,6	315	466	0,49786
10–12	4	25	11	85	0,85	275	741	0,79167
12–14	5	15	13	100	1	195	936	1

donde:

- $k = 5$ es el número de intervalos



- n_i es el número de personas
- x_i es la marca de clase, el centro de los intervalos
- $N_i = \sum_{j=1}^{i} n_j$ es el acumulado de n_i, y $N = N_k$ el total
- $p_i = \frac{N_i}{N}$
- $u_i = \sum_{j=1}^{i} x_j n_j$ es el acumulado de $x_i n_i$
- $q_i = \frac{u_i}{u_k}$.

1) Calculamos la renta media usando las marcas de clase:

$$N = \sum_i n_i = 12 + 13 + 35 + 25 + 15 = 100.$$
$$\overline{x} = \frac{\sum_i n_i x_i}{N} = \frac{12 \cdot 5 + 13 \cdot 7 + 35 \cdot 9 + 25 \cdot 11 + 15 \cdot 13}{100} = \frac{936}{100} = 9{,}36.$$

2) El intervalo modal es aquel donde n_i se hace máximo, es decir, el $(8, 10)$, donde $n_i = 35$.

   En cuanto al valor concreto de la moda, depende del criterio, vamos a calcularlo para dos diferentes.

   a) Bajo la hipótesis de uniformidad para los valores dentro del intervalo, la moda es el punto dentro del intervalo modal cuya distancia a los extremos es inversamente proporcional a la frecuencia de los intervalos vecinos. Sea (L_{i-1}, L_i) el intervalo modal $(8, 10)$ y $n_{i-1} = 13$ es la frecuencia del intervalo anterior y $n_{i+1} = 25$ la del siguiente. Si $\text{Mo} = L_{i-1} + m$, entonces la regla de proporcionalidad anterior se traduce en que

$$m \cdot n_{i-1} = (L_i - L_{i-1} - m) \cdot n_{i+1},$$



de donde

$$m = \frac{(L_i - L_{i-1}) \cdot n_{i+1}}{n_{i-1} + n_{i+1}} = \frac{2 \cdot 25}{13 + 25} \approx 1{,}31579$$

y

$$\text{Mo} = L_{i-1} + m \approx 8 + 1{,}31579 = 9{,}31579.$$

b) En otro criterio utilizado habitualmente es que la moda es el punto del intervalo modal cuya distancia a los extremos es directamente proporcional a la diferencia de frecuencia entre la del intervalo modal y la del de ese extremo. Para esa otra moda $\text{Mo}' = L_{i-1} + m'$, la regla de proporcionalidad anterior se traduce en que

$$\frac{m'}{n_i - n_{i-1}} = \frac{L_i - L_{i-1} - m'}{n_i - n_{i+1}},$$

de donde

$$m' \cdot (n_i - n_{i+1}) = (L_i - L_{i-1} - m') \cdot (n_i - n_{i-1}),$$

y despejando m',

$$m' = (L_i - L_{i-1}) \cdot \frac{n_i - n_{i-1}}{2n_i - n_{i+1} - n_{i-1}} = 2 \cdot \frac{35 - 13}{2 \cdot 35 - 13 - 25} = 1{,}375$$

y

$$\text{Mo}' = L_{i-1} + m' = 8 + 1{,}375 = 9{,}375.$$

3) El intervalo donde se ubica la mediana es el primero en que N_i alcanza $\frac{N}{2} = 50$, como podemos ver en la tabla también sucede en 8–10. Para hallar el valor exacto, de nuevo bajo la hipótesis de valores uniformemente distribuidos en este intervalo, resulta:

$$\text{Me} = L_{i-1} + (L_i - L_{i-1}) \cdot \frac{\frac{N}{2} - N_{i-1}}{n_i} = 8 + 2 \cdot \frac{50 - 25}{35} \approx 9{,}42857.$$



4) Para calcular el índice Gini aplicamos directamente la fórmula:

$$IG = \frac{\sum_{i=1}^{N-1} p_i - q_i}{\sum_{i=1}^{N-1} p_i} = \frac{\sum_{i=1}^{k-1} p_i - \sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i}.$$

Calculamos ambos sumatorios:

$$\sum_{i=1}^{k-1} p_i = 0{,}12 + 0{,}25 + 0{,}6 + 0{,}85 = 1{,}82$$
$$\sum_{i=1}^{k-1} q_i = 0{,}0641 + 0{,}16132 + 0{,}49786 + 0{,}79167 = 1{,}51496$$

Sustituyendo:

$$IG = \frac{\sum_{i=1}^{k-1} p_i - \sum_{i=1}^{k-1} q_i}{\sum_{i=1}^{k-1} p_i} = \frac{1{,}82 - 1{,}51496}{1{,}82} = 0{,}16761.$$

Ejercicio 2

Este ejercicio el mismo que el ejercicio 2 del examen de promoción interna desde C1 especialidad estadística.

Dado el vector aleatorio (X, Y) con función de densidad conjunta:

$$f(x,y) = \begin{cases} \frac{3}{8}(x^2 + y^2) \to -1 \leq x \leq 1; -1 \leq y \leq 1 \\ 0 \to \text{resto} \end{cases}$$

Se pide:

1) Funciones de densidad marginales de X y de Y

2) Coeficiente de correlación lineal entre X e Y

3) Indique si X e Y son variables aleatorias independientes

4) Varianza de la variable $Z = 2X - 3Y$

Solución.

Denotemos por $f_{X,Y} = f$ la función de densidad conjunta de X e Y.

a) La función de densidad marginal de X se define como

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy = \int_{-1}^{1} \frac{3}{8}(x^2+y^2)dy = \frac{3}{8}\Big[yx^2 + \frac{1}{3}y^3\Big]_{-1}^{1}$$
$$= \frac{3}{8}\left(x^2 + \frac{1}{3} - (-x^2 - \frac{1}{3})\right) = \frac{3}{4}\left(x^2 + \frac{1}{3}\right) = \frac{3x^2+1}{4},$$

para valores de $-1 \leq x \leq 1$.

Dado que en $f_{X,Y}$ las variables son intercambiables sin que cambie la fórmula, por simetría

$$f_Y(y) = \frac{3y^2+1}{4}, \text{ con } -1 \leq y \leq 1.$$

b) El coeficiente de correlación lineal entre X e Y se define como

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

Calculemos primeramente el numerador $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

En cuanto a las esperanzas de las variables sueltas,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \cdot f_X(x)dx = \int_{-1}^{1} \frac{3x^3+x}{4}dx = \Big[\frac{\frac{3}{4}x^4 + \frac{1}{2}x^2}{4}\Big]_{-1}^{1} = 0,$$

lo que no debe sorprender, ya que la función de densidad f_X es simétrica respecto al cero, se integra en una región finita y no hay asíntotas verticales que pudieran hacer que la esperanza no existiera. Por simetría $\mathbb{E}[Y] = 0$.

En cuanto a la esperanza del producto de variables,

$$\mathbb{E}[XY] = \int_{\mathbb{R}}\int_{\mathbb{R}} x \cdot y \cdot f_{X,Y}dxdy = \int_{-1}^{1} x\Big(\int_{-1}^{1} \frac{3}{8}(yx^2+y^3)dy\Big)dx$$



$$= \int_{-1}^{1} x\left[\frac{3}{8}\left(\frac{1}{2}y^2x^2 + \frac{1}{4}y^4\right)\right]_{y\to -1}^{y\to 1} dx = \int_{-1}^{1} x \cdot 0 \cdot dx = 0.$$

Por tanto, $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 0 \cdot 0 = 0$, y

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{0}{\sigma_X \cdot \sigma_Y} = 0.$$

Es decir, las variables X e Y no están correladas. Hay que notar que las variables X e Y tienen varianza finita y por eso el cociente anterior es nulo, lo cuál puede comprobarse de la siguiente forma:

$$V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mathbb{E}(X^2) = \int_{\mathbb{R}} x^2 \cdot f_X(x)dx =$$
$$= \int_{-1}^{1} \frac{3x^4 + x2}{4} dx = \left[\frac{\frac{3}{5}x^5 + \frac{1}{3}x^3}{4}\right]_{-1}^{1} = 2 \cdot \frac{\frac{3}{5} + \frac{1}{3}}{4} = \frac{7}{15} \neq 0$$

c) Las variables X e Y serían independientes si y solo si su función de densidad conjunta $f_{X,Y}(x,y)$ se pudiera poner como producto de una función de x y otra de y, que de hecho serían las marginales. Pero

$$f_{X,Y}(x,y) = \frac{3}{8}(x^2 + y^2) \neq f_X(x) \cdot f_Y(y) = \frac{3x^2+1}{4} \cdot \frac{3y^2+1}{4}.$$

Esto nos permite concluir que X e Y no son independientes.

d) Se define $Z = 2X - 3Y$. Calculemos su varianza.

$$\mathrm{Var}(Z) = \mathrm{Var}(2X - 3Y) = 4\mathrm{Var}(X) + 9\mathrm{Var}(Y) - 12\mathrm{Cov}(X,Y).$$

En cuanto a la covarianza, ya vimos que es cero. Por simetría sabemos que $\mathrm{Var}(X) = \mathrm{Var}(Y)$, que la habíamos calculado anteriormente. Por tanto

$$\mathrm{Var}(Z) = 4\mathrm{Var}(X) + 9\mathrm{Var}(Y) = 13\frac{7}{15} = \frac{91}{15} \approx 6{,}0666.$$



Ejercicio 3

En una distribución bidimensional de frecuencias se han calculado la recta de regresión de Y sobre X:

$$Y = 1{,}8X - 5$$

Y la de X sobre Y:

$$X = 0{,}5Y + 3$$

Determinar los valores medios de X y de Y. Deducir si la relación de las variables X e Y es alta o baja.

Solución.

Sabemos que la recta de regresión de Y respecto a X pasa por el punto dado por los valores medios $(\overline{x}, \overline{y})$, y de igual manera la recta de regresión de X respecto a Y pasa por el punto $(\overline{y}, \overline{x})$. Por tanto, el sistema de ecuaciones formado por las dos rectas de regresión tiene como solución el punto dado por esos valores medios.

$$\begin{cases} y = 1{,}8x - 5 \\ x = 0{,}5y + 3 \end{cases} \tag{3.3}$$

Para resolverlo, por ejemplo por sustitución

$$x = 0{,}5y + 3 = 0{,}5(1{,}8x - 5) + 3 = 0{,}9x + 0{,}5,$$

de donde $0{,}1x = 0{,}5$ y por tanto $x = 5$. Aplicando la ecuación para y, $y = 1{,}8x - 5 = 1{,}8 \cdot 5 - 5 = 9 - 5 = 4$. Es decir, $\overline{x} = 5$ y $\overline{y} = 4$.

En cuanto a si la relación es alta o baja, resulta que el r^2 entre ellas viene dado por el producto de las pendientes de las rectas de regresión,

$$r^2 = b \cdot b' = 1{,}8 \cdot 0{,}5 = 0{,}9,$$



siendo $r = \frac{S_{X,Y}}{S_X \cdot S_Y}$. Por tanto, puede decirse que sí, que la relación entre las variables es bastante alta.

Ejercicio 4

Este ejercicio el mismo que el ejercicio 3 del examen de promoción interna desde C1 especialidad estadística.

Se tiene una variable aleatoria X que sigue una distribución Binomial(m, p) con parámetro p desconocido. Se extrae una muestra de tamaño n. Calcular:

1) un estimador de p por el método de los momentos

2) calcular el error cuadrático medio del estimador

3) Hallar la región crítica proporcionada por el test de razón de verosimilitud basada en este estimador para el contraste

$$H_0 : p \leq p_0$$
$$H_1 : p > p_0.$$

Solución.

Tenemos la variable aleatoria $X \sim$ Binomial(m, p) con p desconocido. Sea $X_1, X_2, \ldots, X_n$ la muestra aleatoria simple extraída.

a) Queremos construir un estimador para p con el método de los momentos. Puesto que es un único parámetro, bastará utilizar el primer momento muestral

$$M_1 = \frac{\sum_{i=1}^{n} X_i}{n}.$$



El primer momento teórico de X es

$$\mu_1 = \mathbb{E}[X] = mp.$$

El estimador $\hat{P}$ para p por el método de los momentos resulta de igualar momentos teóricos y muestrales, con lo que $\frac{\sum_{i=1}^n X_i}{n} = m\hat{P}$, de lo que resulta

$$\hat{P} = \frac{\sum_{i=1}^n X_i}{nm}.$$

b) Buscamos el error cuadrático medio del estimador $\hat{P}$, que viene dado por $\text{ECM}(\hat{P}) = \mathbb{E}\left[(\hat{P}-p)^2\right]$. Comprobamos que $\hat{P}$ es insesgado:

$$\mathbb{E}[\hat{P}] = \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{nm}\right] = \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{nm} = \frac{nmp}{nm} = p.$$

Por tanto, el ECM coincide con la varianza, y es sencillo de calcular:

$$\begin{aligned}\text{ECM}(\hat{P}) = \mathbb{E}\left[(\hat{P}-p)^2\right] = \text{Var}[\hat{P}] = \text{Var}\left[\frac{\sum_{i=1}^n X_i}{nm}\right] \\ = \frac{\sum_{i=1}^n \text{Var}[X_i]}{n^2m^2} = \frac{\sum_{i=1}^n mp(1-p)}{n^2m^2} = \frac{p(1-p)}{nm},\end{aligned}$$

donde hemos usado que la varianza de una Binomial(m,p) es $mp(1-p)$.

c) Se nos pide la región crítica del test de razón de verosimilitud para el contraste de hipótesis

$$\begin{cases} H_0 : p \leq p_0 \\ H_1 : p > p_0 \end{cases}$$

expresada en términos de $\hat{P}$, para un $0 \leq p_0 \leq 1$ dado.

La razón de verosimilitud se define como

$$\Lambda(x_1,\ldots,x_n) = \frac{\sup_{p\leq p_0} f_p(x_1,\ldots,x_n)}{\sup_{0\leq p\leq 1} f_p(x_1,\ldots,x_n)},$$



donde, por estar tratando con una distribución discreta,

$$f_p(x_1,\ldots,x_n) = \mathbb{P}[X_1 = x_1 \cap \ldots \cap X_n = x_n | p] = \prod_{i=1}^{n} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_1}.$$

Si definimos $T = \sum_{i=1}^{n} x_i$, $C = \prod_{i=1}^{n} \binom{m}{x_i}$ y $N = nm$, resulta

$$f_p(x_1,\ldots,x_n) = Cp^T(1-p)^{N-T}.$$

Observación

Un hecho relevante: la Λ está siendo escrita en función de T, que a su vez coincide con la estimación dada por el estimador $\hat{P}$ de los apartados anteriores, salvo que multiplicada por N. Esto nos dice que será posible describir la región crítica del contraste de razón de verosimilitud a partir de $\hat{P}$, como nos pide el enunciado.

Busquemos el EMV, que es el valor $\hat{p}$ donde se alcanza el supremo $\sup_{0\leq p\leq 1} f_p(x_1,\ldots,x_n) = Cp^T(1-p)^{N-T}$, considerándola una función de p y buscando ceros de sus derivadas, ya que es continua. Por comodidad, antes tomamos logaritmos que no alteran la ubicación de máximos/mínimos por ser ésta una función estrictamente creciente.

$$\begin{aligned} lnf(p) &= \ln f_p(x_1,\ldots,x_n) = \ln(Cp^T(1-p)^{N-T}) \\ &= \ln C + T\ln p + (N-T)\ln 1 - p, \end{aligned}$$

y tomando derivadas

$$\frac{\partial lnf(p)}{\partial p} = \frac{T}{p} - \frac{N-T}{1-p}.$$

Buscamos los ceros igualando a cero:

$$\frac{\partial lnf(p)}{\partial p} = 0 \iff T(1-p) = p(N-T) \iff p = \frac{T}{N},$$

con lo que $\hat{p} = \frac{T}{N}$ resulta ser el EMV para p.

Comprobamos que ese punto es un máximo, y no un mínimo, calculando la segunda derivada de lnf:

$$\frac{\partial^2 lnf(p)}{\partial p^2} = -\frac{T}{p^2} - \frac{N-T}{(1-p)^2}.$$

Es claro que ambos sumandos son negativos. Eso implica que $\hat{p}$ es realmente el máximo de la función, y puesto que la derivada no se anula en ningún otro punto, no hay máximos locales: $f(p)$ es creciente en $[0, \hat{p}]$ y decreciente en $[\hat{p}, 1]$. Esto nos ayuda a resolver dónde alcanza su supremo el término $\sup_{p \leq p_0} f_p(x_1, \ldots, x_n)$ que aparece en Λ. Claramente, si $\hat{p} \leq p_0$ lo hará en el propio $\hat{p}$, y si no en el punto más cercano a él del intervalo $[0, p_0]$, que es obviamente p_0. Por tanto

$$\Lambda(x_1, \ldots, x_n) = \frac{\sup_{p \leq p_0} f_p(x_1, \ldots, x_n)}{\sup_{0 \leq p \leq 1} f_p(x_1, \ldots, x_n)} = \begin{cases} 1 & \text{si } \hat{p} \leq p_0 \\ \frac{f_{p_0}(x_1, \ldots, x_n)}{f_{\hat{p}}(x_1, \ldots, x_n)} & \text{si } \hat{p} > p_0. \end{cases}$$

Sustituyendo,

$$\frac{f_{p_0}(x_1, \ldots, x_n)}{f_{\hat{p}}(x_1, \ldots, x_n)} = \frac{C{p_0}^T (1-p_0)^{N-T}}{C(\frac{T}{N})^T (1-\frac{T}{N})^{N-T}} = \left(\frac{p_0 N}{T}\right)^T \cdot \left(\frac{(1-p_0)N}{N-T}\right)^{N-T}.$$

Llamemos $g(T)$ a esa función, y comprobemos que es decreciente para $\frac{T}{N} > p_0$. Para ello, podemos simplificar la función utilizando propiedades de logaritmos y derivadas. La función original puede expresarse de la siguiente manera:

$$g(T) = p_0^T \cdot (1-p_0)^{N-T} \cdot \left(\frac{N}{T}\right)^T \cdot \left(\frac{N}{N-T}\right)^{N-T}.$$

Ahora, tomemos el logaritmo natural de $g(T)$ para facilitar la derivación:

$$\ln(g(T)) = T\ln(p_0) + (N-T)\ln(1-p_0) + T\ln\left(\frac{N}{T}\right) + (N-T)\ln\left(\frac{N}{N-T}\right).$$



Ahora, calculemos la derivada con respecto a T:

$$\frac{\partial}{\partial T}\ln(g(T)) = \ln(p_0) - \ln(1-p_0) - \ln\left(\frac{T}{N}\right) + \ln\left(\frac{N-T}{N}\right).$$

Entonces $g(T)$ será decreciente donde esa expresión sea negativa, es decir, tomando exponenciales, para los T tales que

$$\frac{p_0 \frac{N-T}{N}}{(1-p_0)\frac{T}{N}} < 1,$$

o equivalentemente

$$\frac{p_0(N-T)}{(1-p_0)T} < 1.$$

Eso sucede si $p_0(N-T) < (1-p_0)T$, lo que equivale a $p_0N - p_0T < T - p_0T$ y finalmente a $T > p_0N$. Es decir, sí es cierto que $g(T)$ es decreciente para $\frac{T}{N} > p_0$.

La región crítica del test de verosimilitud que buscamos en principio vendrá dada por $\Lambda(x_1, \ldots, x_n) < k$, para alguna k que deberemos calcular de modo que dé lugar al nivel de significación α deseado. Pero por ser Λ decreciente en T, la región crítica de hecho será un intervalo $(c, N]$ para T, con c entera. Al ser la distribución T discreta, las probabilidades alcanzables son un conjunto finito, con lo que se hace necesario transformar esa región crítica en una función crítica:

$$\varphi = \begin{cases} 1 \text{ si } T > c \\ \gamma \text{ si } T = c \\ 0 \text{ si } T < c, \end{cases}$$

con una $0 < \gamma \leq 1$. La condición que deben garantizar c y γ es que

$$\sup_{p \leq p_0} \mathbb{E}_p[\varphi] = \alpha,$$



y claramente ese supremo se alcanza en p_0. El valor de c será aquel tal que $\sum_{i=c}^{N} \mathbb{P}_{p_0}[T=i] > \alpha$ pero $\sum_{i=c+1}^{N} \mathbb{P}_{p_0}[T=i] \leq \alpha$. Entonces, se toma

$$\gamma = \frac{\alpha - \sum_{i=c+1}^{N} \mathbb{P}_{p_0}[T=i]}{\mathbb{P}_{p_0}[T=c]}.$$

Para obtener la región basada en el estimador $\hat{P}$ del apartado a), en vez de en términos de T, basta dividir por N:

$$\varphi = \begin{cases} 1 \text{ si } \hat{P} > \frac{c}{N} \\ \gamma \text{ si } \hat{P} = \frac{c}{N} \\ 0 \text{ si } \hat{P} < \frac{c}{N}. \end{cases}$$



Ejercicio 5

El índice de Precios de un determinado sector presenta las siguientes características técnicas:

- **Tipo Laspeyres**
- **Año Base 2018 con media anual =100**

El índice General de Precios de ese sector consta de 7 agrupaciones diferenciadas, y conocemos los siguientes datos sobre una de esas agrupaciones, la agrupación A1, que consta a su vez de cuatro subgrupos.

Datos sobre la Agrupación A1:

Agrupación A1	Ponderación (tanto por 1)	Índice enero 2022	Ìndice febrero 2022
Subgrupo A11	0,015	121,3	122,5
Subgrupo A21	0,024	117,4	116,3
Subgrupo A31	0,007	106,0	106,0
Subgrupo A31	0,036	115,8	118,3

Conocemos que el Índice General de Precios del sector en febrero 2022 es 111,5. Se pide calcular:

1) La variación mensual (%) del Índice de Precios de la Agrupación A1 en el mes de febrero de 2022.

2) Repercusión que se produciría en el Índice de la Agrupación A1 en el mes de marzo 2022 si el subgrupo A31 sufriera un incremento del 3,5% en sus precios dicho mes. ¿Cuál sería, además, la repercusión en el Índice General de Precios del Sector?

Solución.

1) Para calcular la variación mensual del Índice de Precios de la Agrupación A1 en febrero de 2022 calculamos los índices de A1 en enero y en febrero, teniendo en cuenta que son índices de Laspeyres:

$$IP(A1)_{2018}^{feb2022} = \sum_i I_i^{feb2022} \cdot W_i = \sum_i I_i^{feb2022} \cdot \frac{w_i}{\sum_i w_i} =$$

$$= \frac{122{,}5 \cdot 0{,}015 + 116{,}3 \cdot 0{,}024 + 106 \cdot 0{,}007 + 118{,}3 \cdot 0{,}036}{0{,}015 + 0{,}024 + 0{,}007 + 0{,}036} =$$

$$= \frac{9{,}6295}{0{,}082} = 117{,}43$$

$$IP(A1)_{2018}^{ene2022} = \sum_i I_i^{ene2022} \cdot W_i = \sum_i I_i^{ene2022} \cdot \frac{w_i}{\sum_i w_i} =$$
$$= \frac{121{,}3 \cdot 0{,}015 + 117{,}4 \cdot 0{,}024 + 106 \cdot 0{,}007 + 115{,}8 \cdot 0{,}036}{0{,}015 + 0{,}024 + 0{,}007 + 0{,}036} =$$
$$= \frac{9{,}5479}{0{,}082} = 116{,}44$$

De esta forma, la tasa de variación mensual será:

$$TVM^{feb2022} = \left(\frac{IP(A1)_{2018}^{feb2022}}{IP(A1)_{2018}^{ene2022}} - 1\right) \cdot 100 = \left(\frac{117{,}43}{116{,}44} - 1\right) \cdot 100 =$$
$$= (1{,}0085 - 1) \cdot 100 = 0{,}85\,\%$$

2) Si el subgrupo A31 sufre un incremento de un 3,5 % en el mes de marzo, como el valor del índice en febrero es de 106 entonces dicho índice en marzo valdrá:

$$IP(A31)_{2018}^{mar2022} = 106 \cdot 1{,}035 = 109{,}71$$

De esta forma, el incremento del índice en el subrupo A31 entre febrero y marzo es:

$$\triangle I(A31)_{feb-mar} = 109{,}71 - 106 = 3{,}71$$

Por tanto, la repercusión que se produciría en el índice A1 en marzo con una variación del 3,5 % en el subgrupo A31 sería:

$$R^{A1}(\uparrow A31) = \frac{\triangle I(A31)_{feb-mar} \cdot w_{A31}}{\sum_{i \in Ai1} w_i} = \frac{3{,}71 \cdot 0{,}007}{0{,}082} = 0{,}3167$$

La repercusión en el Índice General de Precios del Sector (IPS) que se produciría con esa subida en el subgrupo A31 sería:

$$R^{IPS}(\uparrow A31) = \frac{\triangle I(A31)_{feb-mar} \cdot w_{A31}}{\sum_i w_i} = \frac{3{,}71 \cdot 0{,}007}{1} = 0{,}026$$

En este caso la suma de las ponderaciones en el denominador es 1 porque estamos trabajando con el índice general, no con un grupo, y las ponderaciones en el ejercicio nos las proporcionan en tanto por uno. Era esperable que la repercusión sobre el índice general sea menor que sobre el grupo al que pertenece.



Ejercicio 6

Este ejercicio el mismo que el ejercicio 5 del examen de promoción interna desde C1 especialidad estadística.

Utilizando los datos siguientes relativos a una economía imaginaria,

Gasto en consumo final	2300
Formación bruta de capital fijo	215
Deflactor implícito del PIB para 2021 en base 2015=100	115,1
Impuestos netos sobre productos	415
Subvenciones a los productos	70
Remuneración de los asalariados	1135
Exportaciones de bienes y servicios	545
Consumo colectivo efectivo	692
Consumo de capital fijo	70
Otros impuestos netos sobre la producción	335
Variación de existencias	16
Adquisiciones menos cesiones de objetos valiosos	27
Producción	3380
Importación de bienes y servicios	210
Consumos intermedios	902

calcule en términos SEC2010:

1) **El consumo individual efectivo**
2) **La formación neta de capital**
3) **El saldo neto exterior de bienes servicios**
4) **El PIBpm a precios constantes de 2015**
5) **El valor añadido bruto a precios de mercado (o precios de adquisición)**
6) **El VAB a precios básicos**
7) **Los impuestos sobre los productos**
8) **La remuneración de los factores de producción**
9) **El Excedente Bruto de Explotación**
10) **Comprueba que es posible determinar el PIBpm a precios corrientes por las tres vías: producción, gasto y renta**

Solución.

1) Sabemos que el gasto en consumo final(GCF) es la suma del gasto en consumo final individual efectivo(CFIE) y del gasto en consumo final



colectivo efectivo(CFCE) por lo tanto:

$$GCF = CFIE + CFCE$$

$$2300 = CFIE + 692$$

Con lo que despejando obtenemos que $CFIE = 2300 - 692 = 1608$

2) Calculamos en primer lugar la formación bruta de capital(FBC):

$$FBC = FBCF + VE + (\text{Adquis - Ces de objetos valiosos}) =$$

$$= 215 + 16 + 27 = 258$$

Y ahora podemos obtener la formación neta de capital(FNC) restando el consumo de capital fijo(CCF) a la cantidad anterior:

$$FNC = FBC - CCF = 258 - 70 = 188$$

3) El saldo exterior de bienes y servicios(SE) es la diferencia entre las exportaciones(X) y las importanciones(M):

$$SE = X - M = 545 - 210 = 335$$

4) Vamos a calcular el producto interior bruto a precios de mercado:

$$PIB_{pm} = GCF + FBC + (X - M) = 2300 + 258 + 335 = 2893$$

Y ahora deflactamos este valor por el índice correspondiente(115,1), que se nos da como dato en el problema, para obtener el PIB a precios constantes de 2015:

$$PIB^{2021}_{pm,2015} = \frac{2893}{1{,}151} = 2513{,}466$$

5) El VAB_{pm} coincide para el total de la economía con el PIB_{pm} que se calculó en el apartado anterior.



6) Sabemos que:

$$PIB_{pm} = VAB_{pb} + \text{Impuestos netos sobre productos}$$

por lo que despejando obtenemos el valor pedido:

$$VAB_{pb} = PIB_{pm} - \text{Impuestos netos sobre productos} = 2893 - 415 = 2478$$

7) Los impuestos netos sobre los productos es la diferencia entre los impuestos y las subvenciones sobre los productos, por lo tanto:

$$\begin{aligned}&\text{Impuestos sobre productos(D21)} = \\ &= \text{Impuestos netos sobre productos(D21-D31)} + \text{Subvenciones sobre productos(D31)} = \\ &= 415 + 70 = 485\end{aligned}$$

8) La remuneración de los factores de producción(RFP) es la suma de la remuneración de asalariados(RA) y el excedente bruto de explotación(EBE) más la renta mixta bruta(RMB), o lo que es lo mismo, utilizando la cuenta de explotación, la diferencia entre el producto interior bruto y los impuestos netos sobre producción e importaciones:

$$\begin{aligned}RFP &= PIB_{pm} - (D2 - D3) = PIB_{pm} - (D21 + D29 - (D31 + D39)) = \\ &= PIB_{pm} - ((D21 - D31) + (D29 - D39)) = 2893 - (415 + 335) = \\ &= 2893 - 750 = 2143\end{aligned}$$

9) El excedente bruto de explotación más la renta mixta bruta se puede calcular con la cuenta de explotación, por ejemplo:

II. 1.1 Cuenta de explotación			
Empleos	**Código**	**Operaciones y otros flujos**	**Recursos**
S.1		**y saldos contables**	**S.1**
	B.1*b	Producto interior bruto	2.893
1.135	D.1	Remuneración de asalariados	
485	D.21	Impuestos sobre los productos	
-70	D.31	Subvenciones a los productos	
335	D.29 - D39	Otros impuestos netos sobre la	
		producción	
1.008	**B.2b + B.3b**	**EBE + RMB**	

o más fácilmente, restando a la remuneración de los factores de producción, calculada en el apartado anterior, la remuneración de asalariados.

10) Calculemos el PIB a precios corrientes por las tres vías:

$$\begin{aligned} PIB_{producción} &= VAB_{pb} + \text{impuestos netos sobre los productos(D21-D31)} = \\ &= 2478 + 415 = 2893 \\ PIB_{renta} &= RA + (EBE/RM) + \\ &+ \text{impuestos netos sobre la producción e importaciones(D2 - D3)} = \\ &= 1135 + 1008 + (415(D21 - D31) + 335(D29 - D39)) = 2893 \\ PIB_{gasto} &= GCF + FBC + (X - M) = \\ &= 2300 + 258 + 335 = 2893 \end{aligned}$$

Ejercicio 7

Este ejercicio el mismo que el ejercicio 4 del examen de promoción interna desde C1 especialidad estadística.



Se quiere comparar la estructura de mortalidad de dos poblaciones distintas A y B. Se dispone de una tercera población C que servirá como referencia. A partir de los datos de la tabla:

Grupos de edad	Población de C en cada grupo de edad	Tasas específicas de mortalidad de la población A	Tasas específicas de mortalidad de la población B
De 0 a 4 años	2152	0,34	0,15
De 5 a 24 años	9943	0,19	0,14
De 25 a 44 años	12840	3,53	1,15
De 45 a 64 años	16942	9,81	12,78
De 65 a 84 años	23354	40,73	33,76
De 85 y más	15004	93,65	98,83
TOTAL	80235		

Calcule las tasas estandarizadas de mortalidad de las poblaciones A y B mediante el método directo de estandarización o 'población tipo' utilizando la población C e indique qué población, A o B, es la que tiene una menor tasa estandarizada.

Solución.

En el método directo de estandarización, se calcula la tasa que se esperaría encontrar en las poblaciones bajo estudio si todas tuvieran la misma composición según la variable cuyo efecto se espera ajustar o controlar. Se utiliza la estructura de una población llamada "estándar", o población tipo, en nuestro caso la población C, cuyos estratos corresponden al factor que se quiere controlar y a la cual se aplica las tasas específicas por esos mismos estratos de las poblaciones estudiadas. De esta forma se obtiene el número de casos "esperado" en cada estrato si la composición fuera la misma en cada población. La tasa ajustada o "estandarizada" se obtiene dividiendo el total de casos esperados por el total de la población estándar. En nuestro ejemplo los casos son defunciones y hablamos de defunciones estimadas o esperadas que se producirían en cada población estudiada.

Los cálculos de las defunciones estimadas en la población A en cada grupo

de edad, por ejemplo serían:

$$DE_A^{0-4} = 2152 \cdot \frac{0{,}34}{1000} = 0{,}731\,68$$
$$DE_A^{5-24} = 9943 \cdot \frac{0{,}19}{1000} = 1{,}889\,17$$

y así sucesivamente, como puede verse en la tabla 3.8.

Grupos de edad	Población de C en cada grupo de edad	Tasas específicas de mortalidad de la población A	Tasas específicas de mortalidad de la población B	Defunciones estimadas que se producirían en la población A	Defunciones estimadas que se producirían en la población B
De 0 a 4 años	2152	0,34	0,15	0,73168	0,3228
De 5 a 24 años	9943	0,19	0,14	1,88917	1,39202
De 25 a 44 años	12840	3,53	1,15	45,3252	14,766
De 45 a 64 años	16942	9,81	12,78	166,20102	216,51876
De 65 a 84 años	23354	40,73	33,76	951,20842	788,43104
De 85 y más	15004	93,65	98,83	1405,1246	1482,84532
TOTAL	80235			2570,48009	2504,27594
Las tasas específicas están en tantos por mil			Tasa Estandarizada	32,03689275	31,21176469
			(En tanto por mil)		

Figura 3.8: Cálculo de las tasas estandarizadas de las poblaciones A y B.

Para calcular las tasas estandarizadas(TE), una vez que tenemos las defunciones esperadas o estimadas en cada grupo de edad, lo que hacemos es sumarlas todas y dividirlas por la población total de C, es decir:

$$TE_A = \frac{2570{,}480}{80\,235} \cdot 1000 = 32{,}037‰$$
$$TE_B = \frac{2504{,}276}{80\,235} \cdot 1000 = 31{,}212‰$$

Se observa que la tasa estandarizada es algo mayor en A que en B. El hecho de que en B haya tasas específicas en algunos grupos de edad mayores que en A no llega a compensar lo que aporta el grupo de edad de 65 a 84 años, que es el más numeroso y ahí la tasa específica de mortalidad en A es mayor que en B.



Ejercicio 8

Este ejercicio el mismo que el ejercicio 6 del examen de promoción interna desde C1 especialidad estadística.

Se considera una población de $N = 100$ hogares todos ellos constituidos por 4 personas. Se quiere estimar la proporción de personas con cierto atributo. Para ello se obtiene una muestra aleatoria sin reemplazamiento de 5 hogares obteniéndose los siguientes valores:

Hogar de la muestra	Nº de personas con atributo
1	2
2	3
3	1
4	2
5	1

(1) Obtenga una estimación de la proporción del atributo en la población

(2) Una estimación de la varianza de la estimación anterior

Solución.

(1) El estimador de la proporción del atributo en la población es:

$$\hat{P} = \frac{\hat{A}}{M} = \frac{\frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{\overline{M}} A_{ij}}{M} = \frac{\frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{\overline{M}} A_{ij}}{N \cdot \overline{M}} =$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{\overline{M}} \frac{A_{ij}}{\overline{M}}\right) = \frac{1}{n}\sum_{i=1}^{n}\hat{P}_i = \frac{\frac{2}{4}+\frac{3}{4}+\frac{1}{4}+\frac{2}{4}+\frac{1}{4}}{5} = \frac{9}{20}$$



(2) La estimación de la varianza del estimador recien calculado, teniendo en cuenta que se trata de un muestro sin reposición es:

$$\hat{V}\left(\hat{P}\right) = (1-f)\,\frac{\hat{P}\hat{Q}}{n-1} = \left(1-\frac{n}{N}\right)\frac{\hat{P}\hat{Q}}{n-1} = \left(1-\frac{5}{100}\right)\frac{\frac{9}{20}\cdot\frac{11}{20}}{5-1} = 0{,}05878$$

Ejercicio 9

Este ejercicio el mismo que el ejercicio 8 del examen de promoción interna desde C1 especialidad estadística.

Una aldea cuenta con 50 personas que todos los días se dedican a la avicultura o a la agricultura. Si utilizan todos sus recursos y teniendo en cuenta la tecnología que poseen, son capaces de llegar a las siguientes combinaciones de patatas y aves.

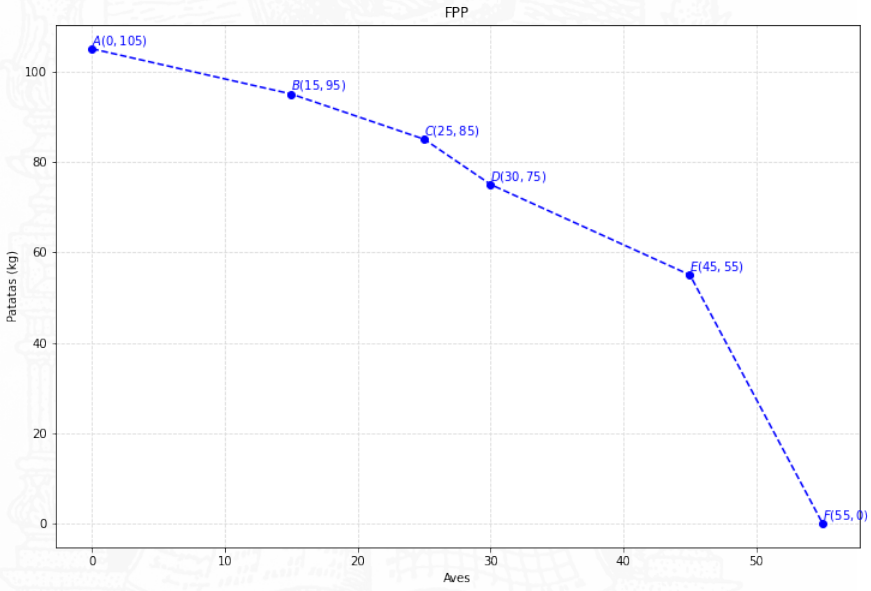
	Aves	kg de patatas
A	0	105
B	15	95
C	25	85
D	30	75
E	45	55
F	55	0

1. Dibuja la frontera de posibilidades de producción (FPP).
2. ¿Cuál es el significado de la FPP de esta aldea?
3. ¿Cuál es el coste de oportunidad para la aldea de pasar de la combinación E a la C?

4. Clasifica las siguientes seis combinaciones de bienes: (25 aves, 100 Kg patatas) ; (30 aves, 75 Kg patatas) ; (45 aves, 55 Kg patatas) ; (15 aves, 95 kg patatas) ; (45 aves, 10 Kg patatas) ; (35 aves, 75 Kg patatas) como ineficientes, inalcanzables o eficientes ¿Qué quieren decir esos términos?

Solución.

1. La frontera de posibilidades de producción de esta aldea es la siguiente:



2. ¿Cuál es el significado de la FPP de esta aldea?

   La frontera de posibilidades de producción (FPP) muestra las distintas combinaciones de producción de aves y kilogramos de patatas que pueden producirse con los recursos y la tecnología disponibles en la al-

dea. Dado que los recursos son limitados, también existe un límite para las cantidades de bienes y servicios que puede producir la aldea. Los puntos que se encuentran sobre la FPP ilustran las combinaciones de producción que son eficientes.

3. ¿Cuál es el coste de oportunidad para la aldea de pasar de la combinación E a la C?

   El coste de oportunidad es el valor de los bienes y servicios a los que se renuncian para adquirir otros. En este caso, si se pasa de la combinación E a la combinación C, se está renunciado a la producción de 20 aves. Esto es, el coste de oportunidad de pasar de E a C es de 20 unidades.

4. Clasifica las siguientes seis combinaciones de bienes: (25 aves, 100 Kg patatas) ; (30 aves, 75 Kg patatas) ; (45 aves, 55 Kg patatas) ; (15 aves, 95 kg patatas) ; (45 aves, 10 Kg patatas) ; (35 aves, 75 Kg patatas) como ineficientes, inalcanzables o eficientes ¿Qué quieren decir esos términos?

   Un punto ineficiente será aquel en el cual la capacidad productiva de la aldea no esté siendo utilizada al máximo, se están despilfarrando recursos , y gráficamaente la combinación producida se encuentra por debajo de la FPP. Un punto es inalcanzable cuando la capacidad productiva de la aldea no permite esa combinación de producciones tan elevada. Gráficamente, un punto inalcanzable se encuentra por encima de la FPP. Finalmente, un punto es eficiente cuando se da una situación en la que el país utiliza todos los recursos disponibles de la mejor forma posible, y son aquellos puntos que se encuentran sobre la FPP.

   En la figura 3.9 se encuentran representadas las diferentes combinaciones propuestas: de color rojo la combinación ineficiente, de color verde

las combinaciones eficientes y de color amarillo las combinaciones inalcanzables.

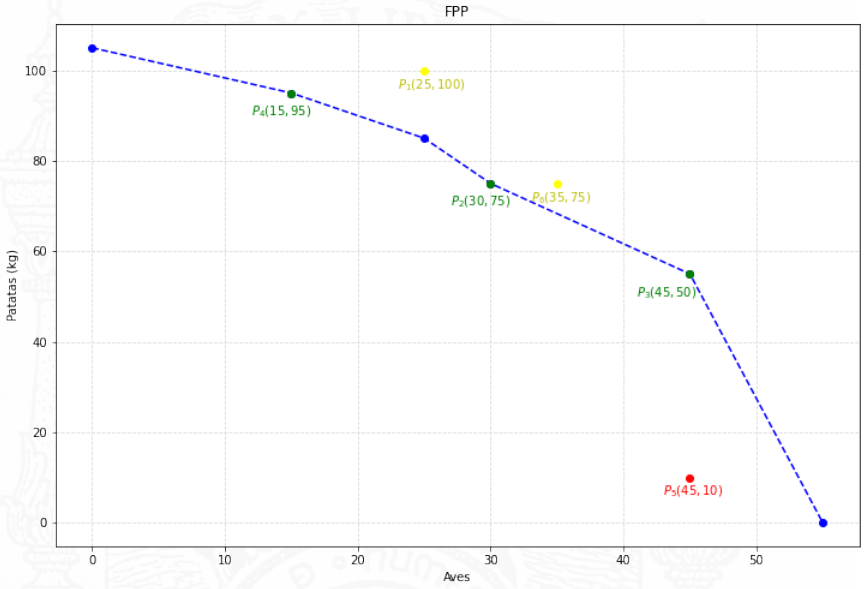


Figura 3.9: Puntos eficientes, ineficientes e inalcanzables en el gráfico de FPP.

Ejercicio 10

Este ejercicio el mismo que el ejercicio 7 del examen de promoción interna desde C1 especialidad estadística.

A partir de la información de la tabla

Grupos de edad	Mujeres residentes (a fecha 1 de julio)	Nacimientos por edad de la madre
De 15 a 19 años	1.188.022	5.470
De 20 a 24 años	1.165.886	23.781
De 25 a 29 años	1.237.778	56.916
De 30 a 34 años	1.357.305	112.874
De 35 a 39 años	1.574.520	102.129
De 40 a 44 años	1.905.822	32.832
De 45 a 49 años	1.958.973	3.094
TOTAL	10.388.306	337.096

Calcule:

1) Las tasas específicas de fecundidad por grupo de edad de la madre.

2) El índice sintético de fecundidad

3) La edad media a la maternidad

Solución.

1) Las tasas específicas de fecundidad por grupo de edad de la madre se calculan como el cociente entre el número de nacimientos de cada grupo de edad y la población media de mujeres de ese grupo determinado:

$$TEF_x = \frac{N_x}{M_x^{01/07}} \cdot 1000.$$

De esta forma,

$$TEF_{15-19}^{01/07} = \frac{N_{15-19}}{M_{15-19}^{01/07}} \cdot 1000 = \frac{5470}{1\,188\,022} = 4{,}6042‰$$



$$TEF_{20-24}^{01/07} = \frac{N_{20-24}}{M_{20-24}^{01/07}} \cdot 1000 = \frac{23\,781}{1\,165\,886} \cdot 1000 = 20{,}3973‰$$

$$TEF_{25-29}^{01/07} = \frac{N_{25-29}}{M_{25-29}^{01/07}} \cdot 1000 = \frac{56\,916}{1\,237\,778} \cdot 1000 = 45{,}9823‰$$

$$TEF_{30-34}^{01/07} = \frac{N_{30-34}}{M_{30-34}^{01/07}} \cdot 1000 = \frac{112\,874}{1\,357\,305} \cdot 1000 = 83{,}1603‰$$

$$TEF_{35-39}^{01/07} = \frac{N_{35-39}}{M_{35-39}^{01/07}} \cdot 1000 = \frac{102\,129}{1\,574\,520} \cdot 1000 = 64{,}8635‰$$

$$TEF_{40-44}^{01/07} = \frac{N_{40-44}}{M_{40-44}^{01/07}} \cdot 1000 = \frac{32\,832}{1\,905\,822} \cdot 1000 = 17{,}2272‰$$

$$TEF_{45-49}^{01/07} = \frac{N_{45-49}}{M_{45-49}^{01/07}} \cdot 1000 = \frac{3094}{1\,958\,973} \cdot 1000 = 1{,}5793‰$$

2) El Índice Sintético de Fecundidad es la suma de las tasas específicas de fecundidad en tanto por uno. Al estar manejando grupos de edad quinquenales, cada mujer "vive" cinco años dentro de cada uno de los intervalos de edad, y por tanto hay que multiplicar por 5 para obtener el número medio de hijos que tendría una mujer a lo largo de los 35 años de vida fértil:

$$\begin{aligned} ICF =& 5 \sum_{x=15}^{49} f_x = \\ =& 5 \cdot (0{,}0046042 + 0{,}0203973 + 0{,}0459823 + 0{,}0831603 + 0{,}0648635 + \\ &+ 0{,}0172272 + 0{,}0015793) = \\ =& 5 \cdot 0{,}2378146 = 1{,}1890731. \end{aligned}$$

En término medio, cada mujer tendrá 1,189 hijos a lo largo de su vida fértil.

3) Para el cálculo de la edad media a la maternidad se utiliza

$$EMM = \frac{\sum_{x=15}^{49} \left(x + \frac{n}{2}\right) f_x}{\sum_{x=15}^{49} f_x}$$



donde x hace referencia al extremo izquierdo del intervalo de cada uno de los grupos de edad y n hace referencia a la amplitud de dichos intervalos. Por tanto, la edad media a la maternidad será

$$EMM = \frac{\left(15+\frac{5}{2}\right)0{,}0046042 + \left(20+\frac{5}{2}\right)0{,}0203973 + \cdots + \left(45+\frac{5}{2}\right)0{,}0015793}{0{,}2378146} =$$

$$= \frac{7{,}7463061}{0{,}2378146} = 32{,}572876$$

Por tanto la edad media a la maternidad es 32,57 años.



Capítulo 4

Año 2023. Turno libre.

"Use the worst colour you can find in each place - it usually is the best."

—Roy Lichtenstein.



Roy Lichtenstein is the most famous artists in the history of Pop Art, and was noted for his contributions to this art genre, particularly with his graphic works known for their bright and catchy colors and images. Roy Lichtenstein was born and raised in New York City on October 27, 1923. As a young boy, he developed an interest in two things - comic books and science. After he graduated from Manhattan's Franklin School for Boys, he enrolled at the Ohio State University to pursue his college education. In the 1960s, Lichtenstein began exploring a wide range of subjects and painting techniques. However, Lichtenstein had a different style from these other artists during his time. Instead of focusing on the more emotional aspect of life in his works, he was more interested in making use of stencils and rather impersonal means of presenting his ideas . This style was often used in commercial art.

$2023 = 17^2 \cdot 7$

2023 is a Harshad number since it is a multiple of its sum of digits(7).

$2023 = (2+0+2+3)(2^2+0^2+2^2+3^2)^2$.



4.1. Soluciones del segundo examen del Cuerpo de Diplomados en Estadística del Estado de 2023.

Ejercicio 1

Sea X la variable aleatoria que sigue una distribución exponencial de parámetro $\lambda > 0$. Sea g función de X tal que $g(X) = X^2$.

1. Halle $E(g(X))$.
2. Calcule $P(g(X) > 8|X > 2)$.
3. Sea Y otra variable aleatoria con la misma distribución que X y sea:

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & si\ x \geq 0, y \geq 0 \\ \\ 0 & en\ otro\ caso \end{cases}$$

   la función de densidad conjunta de (X, Y). ¿Son X e Y independientes? ¿Por qué?
4. Calcule $E(g(X)Y)$.
5. Sean $U = \frac{X}{X+Y}$ y $V = X + Y$. Calcule la función de densidad conjunta de (U, V).

Solución.

1.

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx = \int_{0}^{\infty} x^2 \cdot \lambda \cdot e^{-\lambda x}dx$$



Teniendo en cuenta la función gamma de parámetro p, $\Gamma(p) = \int_0^\infty x^{p-1}e^x dx$, y haciendo el cambio de variable $\begin{cases} \lambda x = y \\ \lambda dx = dy \end{cases}$, se tiene:

$$E(g(X)) = \frac{1}{\lambda^2}\int_0^\infty y^2 \cdot \lambda \cdot e^{-y}\frac{dy}{\lambda} = \frac{1}{\lambda^2}\Gamma(3) = \frac{2}{\lambda^2}.$$

A continuación, en la figura 4.1 se representa gráficamente la función $f(x) = x^2 \cdot \lambda \cdot e^{-\lambda x}$ con $\lambda = 0{,}3, 0{,}7$ y 1.

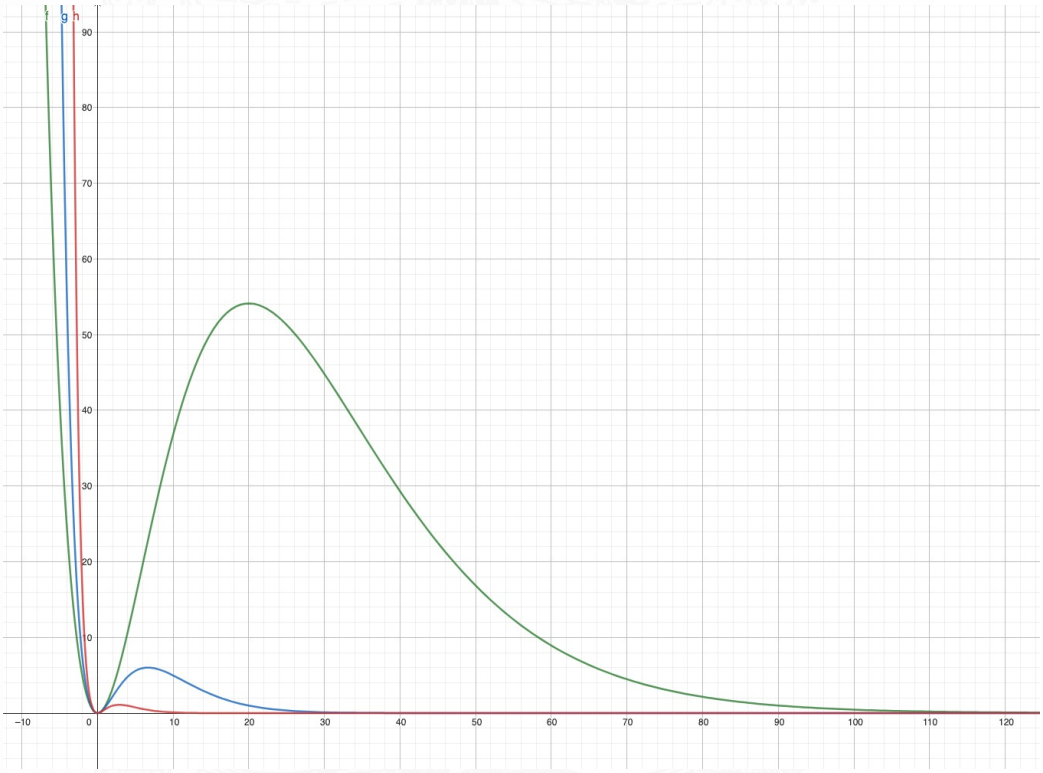


Figura 4.1: Representación gráfica de la función de la cuál hay que calcular la esperanza en el punto 1 del ejercicio.

2.

$$P(g(X) > 8|X > 2) = \frac{P(g(X) > 8 \text{ y } X > 2)}{P(X > 2)}$$

Para que $g(X) > 8$ y $X > 2$, ha de ser: $X^2 > 8 \Leftrightarrow X > \sqrt{8} = 2\sqrt{2} > 2$, luego la condición $g(X) > 8$ y $X > 2$ es equivalente a $X > 2\sqrt{2}$. Por tanto,

$$P(g(X) > 8|X > 2) = \frac{P(X > 2\sqrt{2})}{P(X > 2)} = \frac{\int_{2\sqrt{2}}^{\infty} \lambda e^{-\lambda x} dx}{\int_2^{\infty} \lambda e^{-\lambda x} dx} = \frac{\left[-e^{-\lambda x}\right]_{2\sqrt{2}}^{\infty}}{\left[-e^{-\lambda x}\right]_2^{\infty}} = \\ = \frac{e^{-\lambda \cdot 2\sqrt{2}}}{e^{-\lambda \cdot 2}} = e^{-2\lambda\sqrt{2}+2\lambda} = e^{2\lambda(1-\sqrt{2})}.$$

3. Siendo $f(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & si\ x \geq 0, y \geq 0 \\ 0 & en\ otro\ caso \end{cases}$, se puede reescribir como producto de las funciones de densidad de dos variables aleatorias X e Y con distribuciones exponenciales ambas:

$$\begin{cases} \lambda e^{-\lambda x} \cdot \lambda e^{-\lambda y} & si\ x \geq 0, y \geq 0 \\ 0 & en\ otro\ caso \end{cases},$$

de forma que $f(x,y) = f_1(x) \cdot f_2(y)$ y por tanto X e Y son independientes.

4.

$$E(g(X)Y) = E(X^2Y) = \int_0^{\infty} \int_0^{\infty} x^2 \cdot y \cdot \lambda^2 \cdot e^{-\lambda(x+y)} dxdy = \\ = \lambda \int_0^{\infty} x^2 \cdot e^{-\lambda x} \left(\int_0^{\infty} \lambda \cdot y \cdot e^{-\lambda y} dy \right) dx.$$

Usando el cambio de variable $\begin{cases} \lambda y = z \\ \lambda dy = dz \end{cases}$

$$E(g(X)Y) = \lambda \int_0^{\infty} x^2 \cdot e^{-\lambda x} \left(\frac{1}{\lambda} \int_0^{\infty} z \cdot e^{-z} dz \right) dx \\ = \lambda \int_0^{\infty} x^2 \cdot e^{-\lambda x} \left(\frac{1}{\lambda} \Gamma(2) \right) dx \\ = \lambda \left(\frac{1}{\lambda} \Gamma(2) \right) \int_0^{\infty} x^2 \cdot e^{-\lambda x} dx$$



$$= \int_0^\infty x^2 \cdot e^{-\lambda x} dx$$

Haciendo de nuevo un cambio de variable, $\begin{cases} \lambda x = z \\ \lambda dx = dz \end{cases}$

$$E(g(X)Y) = \int_0^\infty x^2 \cdot e^{-\lambda x} dx = \frac{1}{\lambda^2} \int_0^\infty z^2 \cdot e^{-z} \frac{dz}{\lambda} = \frac{1}{\lambda^3} \Gamma(3) = \frac{2}{\lambda^3}.$$

Este apartado se podía haber resuelto utilizando el resultado del anterior, puesto que al ser las variables aleatorias X e Y independientes, entonces se cumple que :

$$E(g(X)Y) = E(g(X))E(Y) = E(X^2)E(Y) = \frac{2}{\lambda^2} \cdot \frac{1}{\lambda} = \frac{2}{\lambda^3}.$$

5. Siendo $U = \frac{X}{X+Y}$ y $V = X + Y$, se obtiene la siguiente defínción de X e Y en función de U y V: $\begin{cases} X = U \cdot V \\ Y = V - U \cdot V \end{cases}$. La función de densidad conjunta de (U,V) se define como:

$$g(u, v) = f(x(u, v), y(u, v)) \cdot |J|,$$

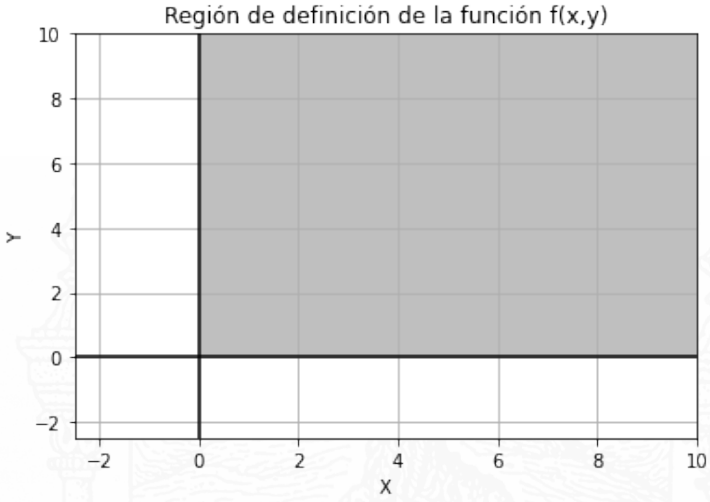
donde

$$|J| = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v(1-u) + uv = v - vu + uv = v,$$

y así,

$$g(u, v) = \lambda^2 \cdot e^{-\lambda(uv+v-uv)} \cdot v = \lambda^2 \cdot e^{-\lambda v} \cdot v$$

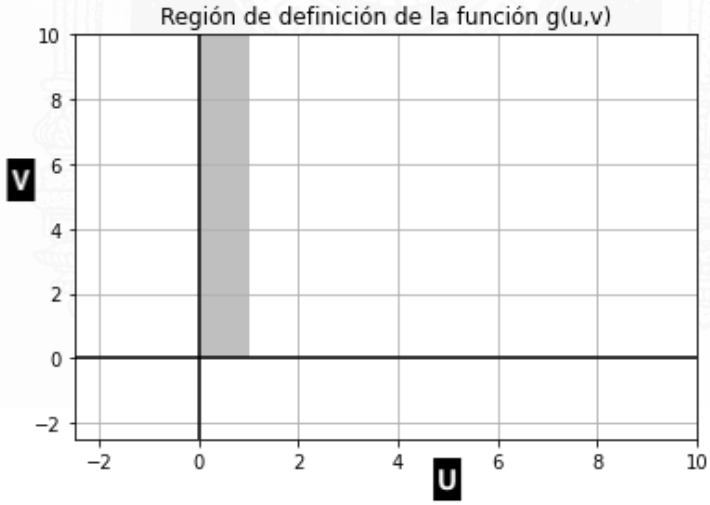
La función de densidad $f(x, y)$ toma valores si $\begin{cases} x \geq 0 \\ y \geq 0 \end{cases}$.



Reescribiendo las condiciones en función de U y V se obtiene la región de definición de la función de densidad conjunta de (U, V): $\begin{cases} uv \geq 0 \\ v - uv \geq 0 \end{cases} \Leftrightarrow$

$$\begin{cases} uv \geq 0, v \geq uv \\ v(1-u) \geq 0 \end{cases} \Leftrightarrow \begin{cases} v \geq 0 \\ (1-u) \geq 0 \end{cases} \Leftrightarrow \begin{cases} v \geq 0 \\ 0 \leq u \leq 1 \end{cases}$$

Es decir, la región de defínción de la función $g(u, v)$ es:





Ejercicio 2

De una población con $N = 6$ unidades, se obtiene una muestra de $n = 2$ unidades sin reposición y con probabilidades iguales. Posteriormente se agregan las unidades en dos estratos:

Estratos (h)	X_{hi}
I	0, 1, 3
II	5, 6, 9

1. Varianza del estimador de la media, $V(\overline{x})$, sin estratificar.
2. Varianza de la media con estratificación, $V(\overline{x}_{st})$, y fracción de muestreo $f_h = \frac{1}{3}$, en cada estrato.
3. Si comparamos las varianzas obtenidas en los apartados a) y b).¿Qué conclusión se podría sacar?

Solución.

1. En un muestreo aleatorio simple, la varianza del estimador de la media se obtiene miedante la expresión

$$V(\overline{x}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}.$$

siendo S^2 la cuasivarianza poblacional, cuyo valor se obtiene de la siguiente forma:

$$S^2 = \frac{\sum_i X_i^2 - N\overline{X}^2}{N-1} = \frac{(0^2+1^2+3^2+5^2+6^2+9^2) - 6\left(\frac{0+1+3+5+6+9}{6}\right)^2}{6-1} =$$
$$= \frac{152 - 6\left(\frac{24}{6}\right)^2}{5} = \frac{56}{5} = 11,2.$$



Por tanto,

$$V(\overline{x}) = \left(1 - \frac{2}{6}\right) \frac{11,2}{2} = 3,73.$$

2. En el muestreo estratificado,

$$V(\overline{x}_{st}) = \sum_h W_h^2 \left(1 - f_h\right) \frac{S_h^2}{n_h}.$$

siendo S_h^2 la cuasivarianza del estrato h, cuyo valor para cada uno de los estratos se obtiene de la siguiente forma:

$$S_1^2 = \frac{\sum_i X_{1i}^2 - N_1 \overline{X}_1^2}{N_1 - 1} = \frac{(0^2 + 1^2 + 3^2) - 3\left(\frac{0+1+3}{3}\right)^2}{3 - 1} =$$
$$= \frac{10 - 3\left(\frac{4}{3}\right)^2}{2} = 2,33.$$
$$S_2^2 = \frac{\sum_i X_{2i}^2 - N_2 \overline{X}_2^2}{N_2 - 1} = \frac{(5^2 + 6^2 + 9^2) - 3\left(\frac{5+6+9}{3}\right)^2}{3 - 1} =$$
$$= \frac{142 - 3\left(\frac{20}{3}\right)^2}{2} = 4,33.$$

Por tanto,

$$V(\overline{x}_{st}) = \frac{1}{4} \cdot \frac{2}{3} \left(2,33 + 4,33\right) = 1,1.$$

3. Si comparamos las varianzas obtenidas en los apartados a) y b), se observa que la varianza del muestreo estratificado es menor que la varianza del muestreo aleatorio simple, ya que los estratos se han agrupado de forma que los valores son parecidos en cada uno de ellos. Cuanto más homogéneos sean los estratos, más precisión se gana respecto al muestreo aleatorio simple.

Ejercicio 3

Sea X_1, X_2, ..., X_n una muestra aleatoria simple de una distribución de Bernoulli de parámetro p con $0 < p < 1$.



1. Utilice el método de los momentos para hallar un estimador de p. También calcule el estimador de p máximo verosímil. ¿Son consistentes?

2. Calcule la Cota de Cramér-Rao y el Estimador Insesgado de Mínima Varianza para estimar el parámetro p.

3. Calcule el error cuadrático medio del estimador obtenido en el apartado 2).

4. Se dispone de una muestra aleatoria simple de viviendas de una urbanización en el 73 A de una calle con datos sobre el número de animales de compañía que conviven en la vivienda almacenada en la siguiente tabla:

NUM.	CALIF.	BLOQ.	PORT.	ESCAL.	PLANTA	PTA.	NUM_ANI
73	A		1		3	D	2
73	A		1		5	B	0
73	A		2		2	C	0
73	A		1		2	A	0
73	A		2		1	C	2
73	A		2		3	A	1
73	A		2		1	D	1
73	A		2		1	B	0
73	A		1		2	C	0
73	A		2		4	B	3
73	A		1		4	D	0
73	A		1		4	E	1
73	A		2		3	C	0
73	A		1		1	E	0
73	A		2		4	C	1
73	A		1		4	B	1
73	A		2		5	E	0
73	A		1		5	A	1
73	A		1		5	D	2

Denotando p como la probabilidad de que en una vivienda habite algún

animal de compañía, calcule la estimación de p para el 73 A, portal 1 usando el estimador obtenido en el apartado 2).

Solución.

1. Por el método de los momentos, se iguala la media muestral con la esperanza de la distribución.

$$\overline{x} = E(X) = p \Rightarrow \widehat{p}_{MM} = \overline{x}$$

Para obtener el estimador según el método de la máxima verosimiliutd, escribimos en primer lugar la función de verosimilitud.

$$f(x_1, \ldots, x_n, p) = p^{x_1}(1-p)^{x_1} \cdot \cdots \cdot p^{x_n}(1-p)^{x_n} = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} =$$
$$= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}$$

A continuación, tomando logaritmos e igualando la derivada a 0:

$$\ln f(x_1, \ldots, x_n, p) = \sum_{i=1}^{n} x_i \cdot \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$
$$\frac{\partial}{\partial p} \ln f(x_1, \ldots, x_n, p) = \frac{\sum_{i=1}^n x_i}{p} + \frac{n - \sum_{i=1}^n x_i}{1-p}(-1) = \frac{\sum_{i=1}^n x_i - np}{p(1-p)} = 0$$
$$\Rightarrow \sum_{i=1}^{n} x_i - np = 0 \Rightarrow \widehat{p}_{MV} = \frac{\sum_{i=1}^n x_i}{n} = \overline{x}$$

Para ver si son consistentes, puesto que ambos estimadores son iguales y coinciden, hay que comprobar que son asintóticamente insesgados y que su varianza tiende a 0. Es decir,

$$E(\widehat{p}) = E(\overline{x}) = \frac{1}{n}\sum_i E(X_i) = \frac{1}{n}\sum_i p = \frac{np}{n} = p$$
$$V(\widehat{p}) = V(\overline{x}) = \frac{1}{n}V(X) = \frac{p(1-p)}{n} \xrightarrow{n\to\infty} 0$$

De modo que el estimador obtenido por ambos métodos es consistente.



2. La Cota de Cramér-Rao viene dada por:

$$f(x,p) = p^x(1-p)^x, \text{ si } x = 0,1$$

$$\ln f(x,p) = x\ln p + (1-x)\ln(1-p)$$

$$\frac{\partial}{\partial p}\ln f(x,p) = \frac{x}{p} + \frac{1-x}{1-p}(-1) = \frac{x-p}{p\,(1-p)}$$

$$E\left[\frac{\partial}{\partial p}\ln f(X,p)\right]^2 = E\left[\frac{X-p}{p\,(1-p)}\right]^2 = \frac{E(X-p)^2}{p^2\,(1-p)^2} = \frac{Var(X)}{p^2\,(1-p)^2} = \frac{1}{p(1-p)}$$

$$C.C.R = \frac{1}{n\frac{1}{p(1-p)}} = \frac{p(1-p)}{n} = \frac{Var(X)}{n} = Var(\overline{x})$$

En el apartado anterior ya vimos que $\widehat{p} = \overline{x}$ es insesgado y como la $C.C.R = Var(\overline{x})$, se tiene que $\widehat{p} = \overline{x}$ es el estimador insesgado de mínima varianza para estimar el parámetro p.

3. El error cuadrático medio del estimador obtenido en el apartado anterior es:

$$ECM(\overline{x}) = E((\overline{x}-p)^2) = B(\overline{x})^2 + V(\overline{x}) = V(\overline{x}) = \frac{p(1-p)}{n}$$

4. Definiendo

$$p = P(\text{“}\textit{en una vivienda habite algún animal de compañía}\text{”}),$$

$$\widehat{p}_{73A,portal1} = \overline{x} = \frac{\sum_i x_i}{n} = \frac{1+0+0+0+0+1+0+1+1+1}{10} = \frac{5}{10} = \frac{1}{2}$$

donde x_i toma valor 1 o 0 si hay o no algún animal de compañía en la vivienda y estamos cogiendo las filas de la tabla del enunciado en las que el portal es el 1, que es la muestra que tenemos disponible. De modo que $\widehat{p} = 0,5$ es la probabilidad estimada de que en una vivienda del 73 A, portal 1, habite algún animal de compañía.



Ejercicio 4

Se conocen los siguientes datos de las cuentas nacionales de una economía para un año t:

Rúbrica SEC2010	Unidades Monetarias. Precios corrientes de t
Formación bruta de capital (P.5g)	300
Otros impuestos (-) subvenciones s/ la producción (D.29 -D.39)	10
Producción (P.1)	2.600
Adquisiciones (-) cesiones de activos no producidos (NP)	5
Consumos intermedios (P.2)	1.400
Exportaciones de bienes y servicios (P.6)	510
Remuneración de los asalariados (D.1)	650
Variación de existencias (P.52)	8
Ahorro nacional neto (B.8n)	110
Transferencias de capital netas del exterior (D.9r -D.9p)	15
Renta disponible ajustada bruta (B.7g)	1.316
Saldo de intercambios exteriores de bienes y servicios (B.11)	26
Consumo de capital fijo (P.51c)	206
Excedente de explotación bruto (B.2g)	410

Para esa economía en el año t se pide:

1) Calcular el PIB a precios de mercado por la perspectiva de demanda.

2) Calcular la cuenta de bienes y servicios y verificar que se cumplen las propiedades de esa cuenta con los datos del problema.

3) Calcular la Renta Mixta Bruta (B.3g). ¿En qué cuentas y para qué sectores institucionales del sistema SEC2010 se calcula la Renta Mixta Bruta?

4) Con los datos que ha obtenido para esta economía ¿se puede calcular el PIB por tres ópticas alternativas? Justificar la respuesta conceptual y numéricamente.

5) Si el PIB pm de esa economía en el año precedente (t-1) era 1250 (a precios corrientes de ese año) y se sabe que la variación del deflactor implícito del PIB entre los dos periodos fue del 2%, calcular cuál fue la tasa de crecimiento real del PIB de esa economía entre t-1 y t.

INSTITUTO NACIONAL DE ESTADISTICA

Solución.

(1) El PIB por el método de la demanda se calcula como:

$$PIB_{pm}(demanda) = GCF + FBC + X - M$$



El saldo de intercambios exteriores de bienes y servicios(B.11) es la diferencia entre importaciones(P.7) y exportaciones(P.6). Como P.6=510 y B.11=26 podemos despejar las importaciones, resultando que

$$P.7 = P.6 + B.11 = 536$$

Además, sabemos que FBC(P.5g) = 300, por lo que el dato que nos falta para poder calcular el producto interior bruto es el gasto en consumo final(P.3). Dicho agregado lo podríamos obtener de la cuenta de utilización de la renta disponible ajustada(II.4.2), pero dado que a nivel de la economía la renta disponible ajustada bruta es igual a la renta disponible bruta(B.6b), trabajaremos con la cuenta de utilización de la renta disponible de la siguiente forma:

II. 4.1 Cuenta de utilización de la renta disponible

Empleos S.1 Total de la economía	Código	Operaciones y otros flujos y saldos contables	Recursos S.1 Total de la economía
	B.6b	Renta disponible bruta	1.316
1.000	P.3	Gasto en consumo final	
	D.8	Ajuste por la variación de los derechos por pensiones	
316	B.8b	**Ahorro bruto**	
110	*B.8n*	***Ahorro neto***	

Figura 4.2: Cálculo del Gasto en consumo final con la cuenta II.4.1.

Como podemos ver en la figura 4.2 los datos en gris son datos que el problema nos proporciona en la tabla inicial y el Ahorro bruto se puede calcular sumando al ahorro neto el consumo de capital fijo(P.51c), que vale 206 según los datos de dicha tabla. El agregado D.8 no afecta a los cálculos, puesto que a nivel de la economía total se compensa en empleos y recursos. Así pues, obtenemos por diferencia el valor de P.3, Gasto en consumo final, cuyo valor es 1.000.

Por tanto, ya podemos calcular el producto interior bruto:

$$PIB_{pm}(demanda) = GCF + FBC + X - M = 1000 + 300 + 510 - 536 =$$
$$= 1300 - 26 = 1274$$

(2) Antes de mostrar la cuenta de bienes y servicios vamos a calcular los impuestos netos sobre los productos utilizando la cuenta de producción:

I. Cuenta de producción

Empleos S.1 Total de la economía	Código	Operaciones y otros flujos y saldos contables	Recursos S.1 Total de la economía
	P.1	Producción	2.600
1.400	P.2	Consumos intermedios	
	D.21- D.31	Impuestos menos subvenciones sobre los productos	74
1.274	**B.1*b**	**Producto interior bruto**	
206	P.51c	Consumo de capital fijo	
1.168	***B.1*n***	***Producto interior neto***	

Figura 4.3: Cálculo de los impuestos menos subvenciones sobre los productos(D.21-D.31) con la cuenta I.

Los datos de Producción(P.1) y de Consumos intermedios(P.2) nos los proporciona el problema y el PIB lo hemos calculado en el apartado anterior, así que despejando obtenemos los impuestos netos sobre los productos.

Ahora ya sí que estamos en disposición de trabajar con la cuenta de bienes y servicios.



R	Cuenta de bienes y servicios		E
2600	Producción (P.1)	Consumos intermedios (P.2)	1400
74	Imps. netos sobre los productos (D.21 - D.31)	Gasto en consumo final (P.3)	1000
		Formación bruta de capital (P.5)	300
536	Importaciones de bienes y servicios (P.7)	Exportaciones de bienes y servicios (P.6)	510
3210			3210

3) En este apartado tenemos que calcular la Renta Mixta Bruta(B.3g), RMB, para lo cual utilizamos la cuenta de explotación de la figura 4.4:

II. 1.1 Cuenta de explotación

Empleos S.1 Total de la economía	Código	Operaciones y otros flujos y saldos contables	Recursos S.1 Total de la economía
	B.1b/B.1*b	Producto Interior Bruto	1.274
650	D.1	Remuneración de los asalariados	
74	D.21 - D.31	Impuestos menos subvenciones sobre los productos	
10	D.29 -D.39	Otros impuestos - subvenciones sobre la producción	
540	**B.2b + B.3b**	**Excedente de explotación bruto + Renta mixta bruta**	

Figura 4.4: Cálculo del EBE+RMB con la cuenta II.1.1.

Podemos comprobar que el saldo de esta cuenta, que es la suma del excedente bruto de explotación y la renta mixta bruta, es decir, EBE+RMB vale 540. Pero como sabemos que EBE=B.2b = 410 podemos despejar y calcular el agregado que se nos pide, resultando que:

$$RMB(B.3b) = 540 - 410 = 130$$

La renta mixta bruta se utiliza en la cuenta de explotación y en la de asignación de la renta primaria, y sólo el sector hogares tiene renta



mixta bruta. Para el total de la economía, este agragado tiene el valor del sector hogares.

4) En los apartados anteriores ya calculamos el PIB por la vía de la demanda. Veamos las otras 2 vías:

$$PIB_{oferta} = \sum VAB + \text{Imp's netos sobre productos} =$$
$$= \text{P.1 - P.2 + D.21 - D.31} =$$
$$= 2600 - 1400 + 74 = 1274$$

$$PIB_{renta} = R.A + \text{Imp's - Sub's sobre producción e importaciones + EBE/RMB} =$$
$$= 650 + 74(\text{D.21 - D.31}) + 10(\text{D.29 - D.39}) + 540 = 1274$$

5) Tenemos que a precios corrientes, los valores del producto interior bruto en los años $t-1$, y t son respectivamente $PIB(t-1) = 1250$ y $PIB(t) = 1274$, y que la variación del deflactor implícito del PIB entre esos dos años es de un 2 %.

Por lo tanto, el valor del PIB en el año t a precios del año t-1, que lo denotamos como PIB^t_{t-1} valdrá:

$$PIB^t_{t-1} = \frac{PIB(t)}{Deflactor^t_{t-1}} = \frac{1274}{1{,}02} = 1249{,}01$$

De esta forma, la tasa de variación a precios constantes del PIB será:

$$TV^t_{t-1}(constantes) = \frac{1249{,}01 - 1250}{1250} \cdot 100 = -0{,}0784\,\%$$

Otra forma de resolver esta cuestión es que podíamos haber calculado la tasa de variación(TV) a precios corrientes, que vale:

$$TV^t_{t-1}(corrientes) = \frac{1274 - 1250}{1250} \cdot 100 = 1{,}92\,\%$$



Como la tasa de variación del deflactor es un 2 % y se cumple la relación:

$$\text{Índice de Valor} = \text{Índice de Precios} \cdot \text{Índice de Volumen}$$

entonces tenemos que $101{,}92 = 102 \cdot$ Índice de Volumen, y por tanto

$$\text{Índice de Volumen} = \frac{101{,}92}{102} \cdot 100 = 0,999215 \cdot 100 = 99{,}9215$$

lo que quiere decir que entre t-1 y t ha disminuido en un 0,0785 %, que es la tasa de variación entre los valores 100 y 99,9215.

Ejercicio 5

Sea una determinada población que a 1 de enero del año 2023 cuenta con 4.111.555 individuos y supongamos que durante el año 2023 se producen 41.198 nacimientos, 37.760 defunciones, 20.084 inmigraciones y 908 emigraciones. Considere los siguientes datos:

$e^{0,0135} = 1,0135$	$\sqrt{1,0055} = 1,0027$	$(1,0027)^5 = 1,0135$

1. Calcule el crecimiento natural, el saldo migratorio y el crecimiento total que dicha población registra durante el año 2023.
2. Utilizando el crecimiento experimentado por esta población en 2023 y suponiendo un crecimiento aritmético de la misma, ¿qué tamaño alcanzaría a 1 de enero de 2029?
3. Utilizando el crecimiento experimentado por esta población en 2023, ¿qué tamaño alcanzaría a 1 de enero de 2029 suponiendo un crecimiento geométrico?



4. Por último, utilizando el crecimiento experimentado por esta población en 2023, realizar el mismo cálculo suponiendo un crecimiento exponencial.

Solución

1. El crecimiento natural de la población en el año 2023 es:

$$CN_{2023} = 41\,198 - 37\,760 = 3438$$

En cuanto al saldo migratorio:

$$SM_{2023} = 20\,084 - 908 = 19\,176$$

Por último, el crecimiento total es

$$CT_{2023} = CN_{2023} + SM_{2023} = 3438 + 19176 = 22\,614$$

2. El tamaño de dicha población a 1 de enero de 2024 es de $4\,111\,555 + 22\,614 = 4\,134\,169$ habitantes. La tasa de crecimiento para el período anual es:

$$r = \left(\frac{4\,134\,169}{4\,111\,555} - 1\right) \cdot 100 = (1{,}0055 - 1) \cdot 100 = (0{,}0055) \cdot 100 = 0{,}55\,\%$$

Por tanto, la población a 1 de enero de 2029 suponiendo un crecimiento aritmético es:

$$P_{01-01-2029} = P_{01-01-2023} \cdot (1 + n \cdot r) = 4\,111\,555 \cdot (1 + 6 \cdot 0{,}0055) =$$
$$= 4\,111\,555 \cdot (1 + 0{,}033) = 4\,111\,555 \cdot 1{,}033 = 4\,247\,236{,}32.$$

3. Suponiendo un crecimiento geométrico,

$$P_{01-01-2029} = P_{01-01-2023} \cdot (1 + r)^n = 4\,111\,555 \cdot (1 + 0{,}0055)^6 =$$



$$= 4\,134\,169 \cdot (1 + 0,0055)^5 = 4\,134\,169 \cdot (1,0055)^5 =$$
$$= 4\,134\,169 \cdot (\sqrt{1,0055})^{10} = 4\,134\,169 \cdot 1,0027^{10} =$$
$$= 4\,134\,169 \cdot (1,0027^5)^2 = 4\,134\,169 \cdot 1,0135^2 =$$
$$= 4\,134\,169 \cdot 1,027\,18 \approxeq 4\,246\,545$$

4. Por último, suponiendo un crecimiento exponencial,

$$P_{01-01-2029} = P_{01-01-2024} \cdot e^{r \cdot n} = 4\,134\,169 \cdot e^{0,0055 \cdot 5} =$$
$$= 4\,134\,169 \cdot e^{0,0275} \approx 4\,134\,169 \cdot (e^{0,0135})^2 =$$
$$= 4\,134\,169 \cdot 1,0135^2 = 4\,134\,169 \cdot 1,027\,18 =$$
$$= 4\,246\,545$$

El motivo de que en este último apartado obtengamos lo mismo que en el anterior es que en el enunciado nos proporcionan el valor de $e^{0,0135}$, que al cuadrado sería $e^{0,027}$, que no es lo que teóricamente deberíamos usar como puede verse, que es $e^{0,0275}$, aunque el resultado sería muy similar, y por eso hacemos uso de los datos del problema disponibles para facilitar los cálculos.

Ejercicio 6

En el mercado de un determinado bien, la función de oferta responde a la forma: $Q_o = 40p$ mientras que la de demanda se estima que es: $Q_d = 4000 - 10p$. Donde Q_d es la cantidad demandada, Q_o es la cantidad ofertada y p es el precio mostrado en unidades monetarias genéricas.

1. Calcule el precio y la cantidad de equilibrio.

2. Represéntelo gráficamente.

3. Halle el excedente del consumidor y del productor.

4. Suponga ahora que el Gobierno fija un impuesto a las empresas por el que la función de oferta varía pasando a ser: $Q'_o = 20p - 200$. Calcule el nuevo precio y la nueva cantidad de equilibrio y represéntelo gráficamente.

5. Calcule la pérdida irrecuperable de eficiencia que se ha producido por la imposición del impuesto.

Solución

1. El precio de equilibrio es aquel en el que la cantidad demandada es igual a la cantidad ofrecida.

$$Q_d = Q_o$$
$$4000 - 10p = 40p$$
$$4000 = 50p$$
$$p^* = 80$$

Consecuentemente, la cantidad de equilibrio es

$$Q(p^*) = 4000 - 10p^* = 4000 - 10 \cdot 80 = 4000 - 800 = 3200 \Rightarrow Q^* = 3200.$$

2. La representación gráfica es la que puede verse en la figura 4.5

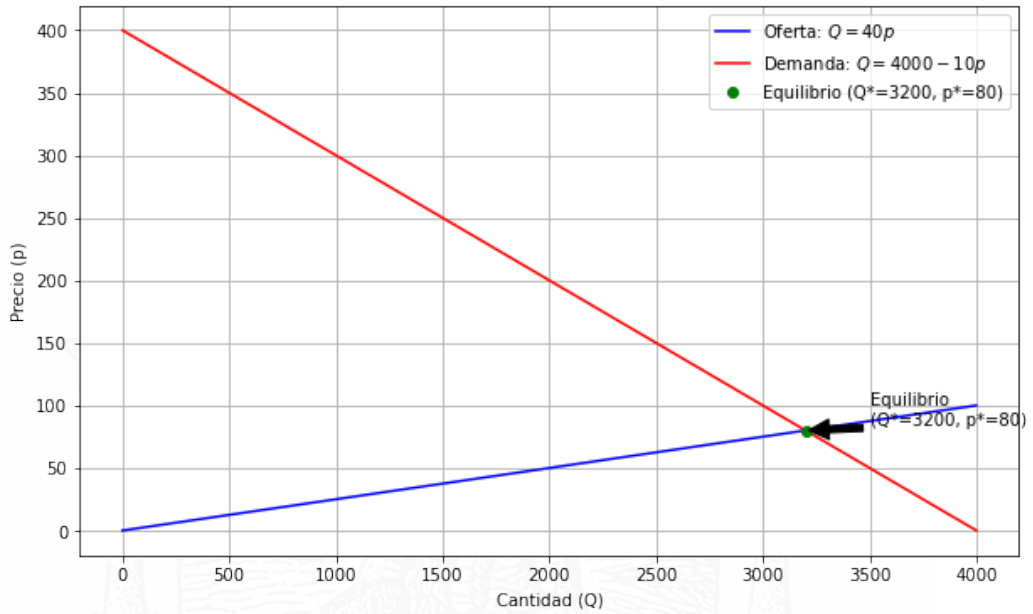


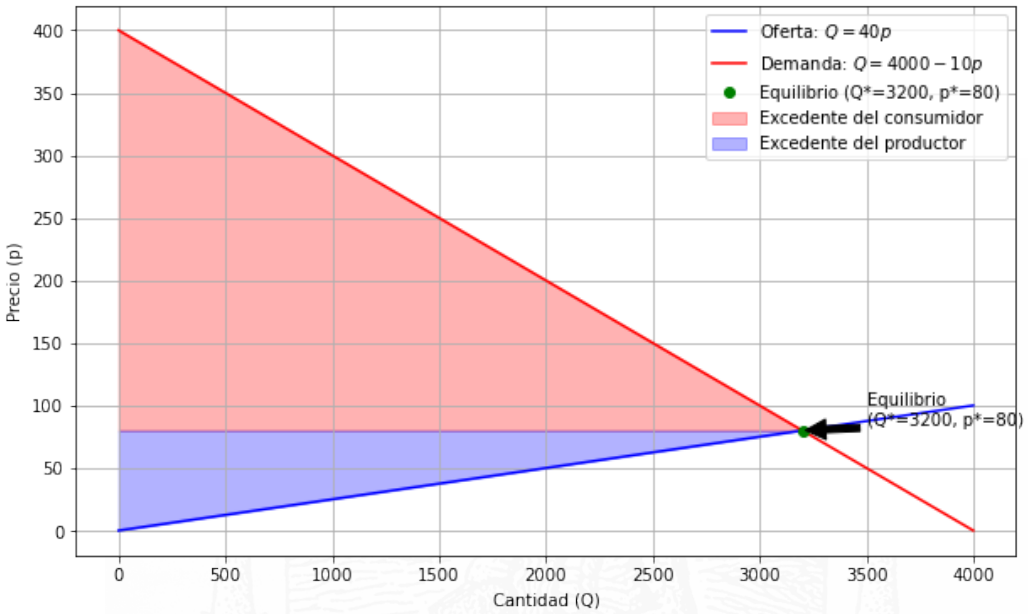
Figura 4.5: Representación gráfica del punto de equilibrio entre la oferta y la demanda.

3. El excedente del consumidor es el área entre la curva de demanda y el precio de equilibrio, es decir, por encima del precio de equilibrio y debajo de la curva de demanda.

$$A_{EC} = \frac{3200 \cdot 320}{2} = 512\,000.$$

El excedente del productor es el área entre la curva de oferta y el precio de equilibrio, esto es, por debajo del precio de equilibrio y por encima de la curva de oferta.

$$A_{EP} = \frac{3200 \cdot 80}{2} = 128\,000.$$



4. El nuevo precio de equilibrio será:

$$Q_d = Q'_o$$

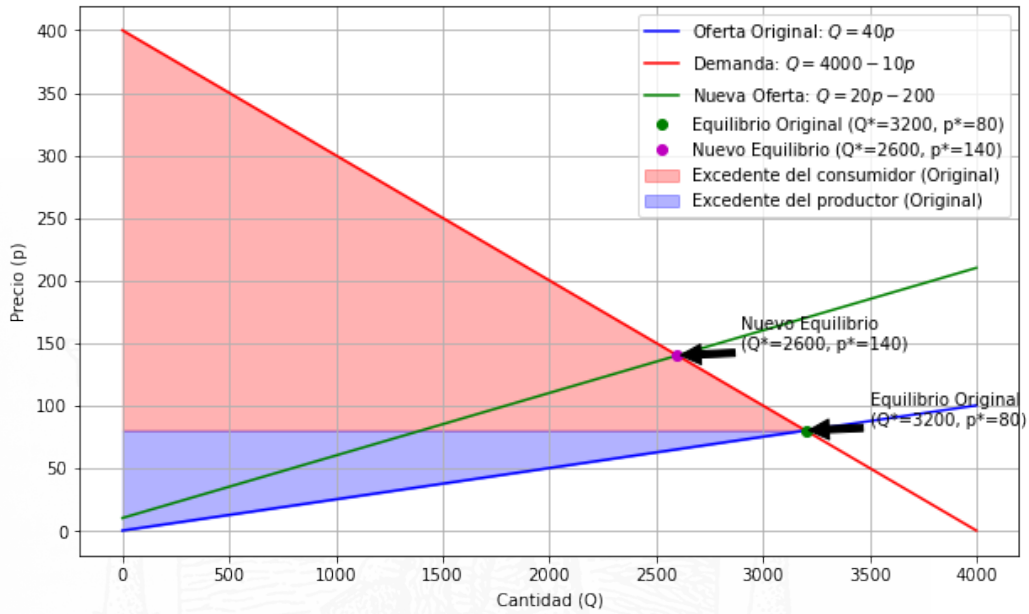
$$4000 - 10p = 20p - 200$$

$$4200 = 30p$$

$$p^* = 140$$

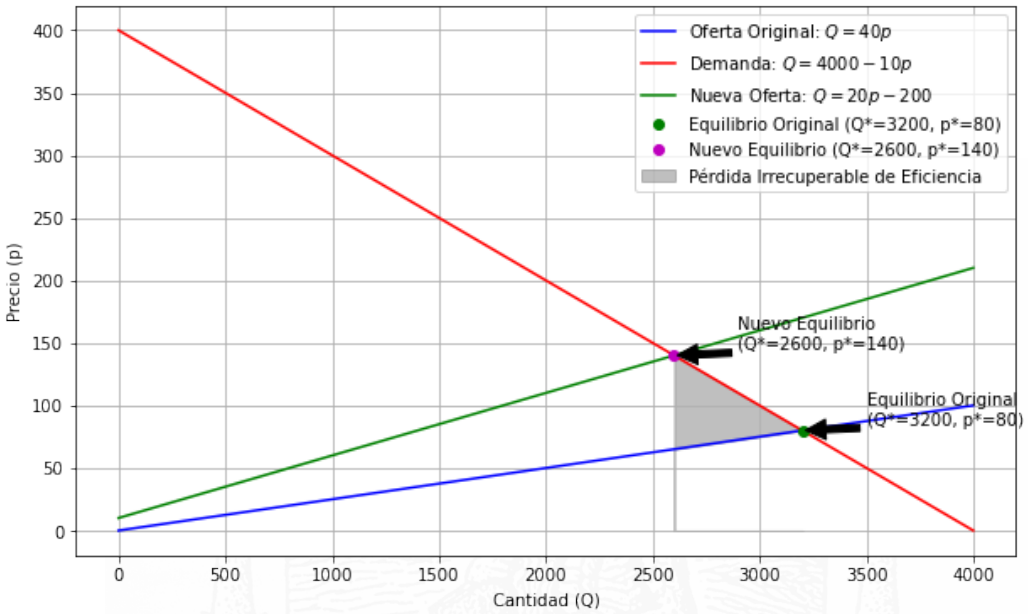
Consecuentemente, la nueva cantidad de equilibrio es

$$Q(p^*) = 4000 - 10p^* = 4000 - 10 \cdot 140 = 4000 - 1400 = 2600 \Rightarrow Q^* = 2600.$$



5. Cuando se fija un impuesto, tanto el excedente del consumidor como el excedente del productor disminuyen, y la recaudación que se obtiene por el impuesto es insuficiente para compensar dicha pérdida. Esta es la pérdida irrecuperable de eficiencia. Se corresponde con el triángulo entre la demanda, la nueva oferta y las cantidades que no se transaccionan debido al impuesto.

$$PIE = \frac{(140 - 65)(3200 - 2600)}{2} = \frac{75 \cdot 600}{2} = 22500$$



Ejercicio 7

Un establecimiento registra el número de unidades vendidas de un producto (X) y el precio unitario en euros (Y), para cada uno de los trimestres de tres años consecutivos:

Año	Trimestre	X	Y
2015	Trimestre 1	80	20
	Trimestre 2	100	30
	Trimestre 3	100	20
	Trimestre 4	110	30
2016	Trimestre 1	90	40
	Trimestre 2	120	20
	Trimestre 3	110	30
	Trimestre 4	90	50
2017	Trimestre 1	100	20
	Trimestre 2	100	40
	Trimestre 3	100	20
	Trimestre 4	100	40

donde:

$$\sum_i \sum_j x_i \cdot n_{ij} = 1200; \quad \sum_i \sum_j x_i^2 \cdot n_{ij} = 121\,200$$

$$\sum_i \sum_j y_i \cdot n_{ij} = 360; \quad \sum_i \sum_j y_i^2 \cdot n_{ij} = 12\,000$$



$$\sum_i \sum_j x_i \cdot y_j \cdot n_{ij} = 35\,700$$

Calcule:

1) La media aritmética y mediana de las ventas trimestrales.

2) En los trimestres en los que se producen 100 unidades, calcule el precio medio pagado por unidad.

3) El coeficiente de variación de las ventas trimestrales.

4) Calcule la moda y la varianza de la variable $Z = 3X + 2$.

5) La covarianza de las variables X e Y.

Solución

Antes de nada, conviene aclarar a qué se refiere el enunciado con la variable *ventas trimestrales*, que tal como está expresado resulta ambiguo. Podría tratarse de la cantidad de ventas (simplemente, la variable X), o de los resultados de ventas entendidos de forma pecuniaria, es decir, la cantidad económica que se deriva de las ventas (variable $X \cdot Y$). Entendemos que se refiere al primero de estos casos, y resolvemos el problema conforme a ese criterio.

La notación con los n_{ij} del enunciado es innecesariamente compleja y no se comprende demasiado bien, no se ha explicado en qué rangos se mueven los índices ni en particular qué significan los dos índices que acompañan a la n. Sorprende especialmente la última fórmula $\sum_i \sum_j x_i \cdot y_j \cdot n_{ij} = 35\,700$, donde se están cruzando valores de X e Y que corresponden a distintos puntos. Por ello, en esta solución vamos a prescindir del conteo de frecuencias en variables



n, denotando simplemente con $\{(x_i, y_i)\}_{i=1}^{12}$ los $N = 12$ puntos conocidos, correspondientes a los 12 cuatrimestres informados.

1) El promedio de ventas es

$$\overline{X} = \frac{1}{N}\sum_i x_i = \frac{80 + 2 \cdot 90 + 6 \cdot 100 + 2 \cdot 110 + 120}{12} = 100.$$

En cuanto a la mediana de X, ordenando los datos resulta la serie ordenada $X_{(i)}$ de $N = 12$ datos:

$$\{80, 90, 90, 100, 100, 100, 100, 100, 100, 110, 110, 120\}$$

La mediana es el valor en el dato central, puede tomarse el valor

$$\text{Me} = x_{(\lceil \frac{N}{2} \rceil)} = x_{(6)} = 100.$$

Por tratarse de un número $N = 12$ par de datos no hay uno que sea el central, a veces en estos casos se prefiere tomar como mediana el promedio de los dos datos centrales, que resulta

$$\text{Me} = \frac{x_{(6)} + x_{(7)}}{2} = \frac{100 + 100}{2} = 100.$$

2) Hay 6 trimestres en los que se producen 100 unidades. La variable Y de precios en ellos es, respectivamente, $\{30, 20, 20, 40, 20, 40\}$. El promedio de esos valores es $\frac{170}{6} \approx 28{,}33$.

3) El coeficiente de variación de X se define como

$$\text{CV}(X) = \frac{\sigma_X}{\overline{X}}.$$

El denominador lo tenemos por el apartado primero. Calculemos el numerador.

$$\text{Var}(X) = \frac{1}{N}\sum_i x_i^2 - \overline{X}^2 = \frac{121\,200}{12} - 100^2 = 10\,100 - 10\,000 = 100.$$



$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{100} = 10.$$

Por tanto resulta $\mathrm{CV}(X) = \frac{\sigma_X}{\overline{X}} = \frac{10}{100} = 0{,}1$.

4) $Z = 3X + 2$. Por ser una transformación inyectiva, la moda se alcanza en la transformada del valor que es la moda de X, el 100 (con 6 apariciones). Por tanto, $\mathrm{Mo}(Z) = 3 \cdot 100 + 2 = 302$. Si queremos hacerlo explícito, los 12 valores para Z son

$$\{242, 302, 302, 332, 272, 362, 332, 272, 302, 302, 302, 302\},$$

cuyo conteo de frecuencias corresponde a la tabla

Z	frecuencia
242	1
272	2
302	6
332	2
362	1

La mayor frecuencia se alcanza para $Z = 302$, confirmando que esa es la moda.

En cuanto a la varianza,

$$\mathrm{Var}(Z) = \mathrm{Var}(3X + 2) = 3^2\mathrm{Var}(X) = 900.$$

5) Aplicamos directamente la fórmula para la covarianza muestral:

$$\begin{aligned}\mathrm{Cov}(X, Y) = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y} &= \frac{\sum_i x_i \cdot y_i}{N} - \overline{X} \cdot \frac{\sum_i y_i}{N} \\ &= \frac{35700}{12} - 100\frac{360}{12} = 2975 - 100 \cdot 30 = -25.\end{aligned}$$

Ejercicio 8

Un determinado IPC sectorial con Base 2019 (media anual=100), tipo Laspeyres, consta de 21 artículos clasificados en siete grupos.

1) Conocidos los siguientes datos en marzo 2024:

Agrupación	Índice (período *t*)	Ponderación (tanto por uno)
IPC General	114,9	1,000
Grupo 6 (telecom.)	119,3	0,180
Artículo 6.1 (telefonía móvil)	132,9	0,045

Se pide calcular la repercusión que supondría en el grupo 6 (telecomunicaciones) un incremento de precios del 7% en el artículo 6.1 (telefonía móvil) en el mes de abril-24.

También se pide calcular la repercusión en el IPC General de esa misma variación del 7% en el artículo 6.1 en abril-24.

2) Conocido ese IPC General = **114,9** en el mes de marzo-24, se pide obtener el nuevo IPC General resultante para marzo-24 tras excluir del cálculo del Índice tres artículos de los veintiunos iniciales. En la tabla de más abajo se indican los datos de los tres artículos a excluir.

Artículo	Índice (período *t*)	Ponderación (tanto por uno)
Art. 3.1	120	0,1
Art. 4.2	107	0,2
Art. 7.1	115	0,1

Calcular el nuevo Índice general en el período *t (marzo-24)*

Solución.

1) Un incremento de precios del 7 % en el artículo 6.1 supone que su índice pasa de 132,9 a $142,2 = 132,9 \cdot 1,07$.

   La repercusión respecto al grupo 6 de este incremento se calcula como:

$$R^6_{6.1} = \frac{I^{abril24}_{6.1} - I^{marzo24}_{6.1}}{I^{marzo24}_{6}} \cdot W_{6,1} \cdot 100$$

   donde: $I^{abril24}_{6.1}$ es el índice del artículo 6.1 en abril de 2024,

   $I^{marzo24}_{6.1}$ es el índice del artículo 6.1 en marzo de 2024,



$I_6^{marzo24}$ es el índice del grupo 6 en marzo de 2024,

$W_{6.1}$ es la ponderación en tanto por uno del artículo 6.1 en el grupo 6.

Por tanto:

$$R_{6.1}^{6} = \frac{I_{6.1}^{abril24} - I_{6.1}^{marzo24}}{I_6^{marzo24}} \cdot W_{6.1} \cdot 100 = \frac{142,2 - 132,9}{119,3} \cdot \frac{0,045}{0,180} \cdot 100 =$$

$$= 0,07795 \cdot 25 = 1,94875$$

Respecto al general tendríamos de una forma análoga que:

$$R_{6.1}^{G} = \frac{I_{6.1}^{abril24} - I_{6.1}^{marzo24}}{I_G^{marzo24}} \cdot W_{6,1} \cdot 100 = \frac{142,2 - 132,9}{114,9} \cdot 0,045 \cdot 100 =$$

$$= 0,3642$$

2) Teniendo en cuenta que:

$$IPC^{marzo24} = \sum_{i=1}^{21} I_i^{marzo24} \cdot W_i$$

podemos escribir

$$IPC^{marzo24} \overset{not}{=} I_G^{marzo24} = 120 \cdot 0,1 + 107 \cdot 0,2 + 115 \cdot 0,1 + IPC_{Restringido}^{marzo24} \cdot 0,6$$

y despejando obtenemos:

$$IPC_{Restringido}^{marzo24} \overset{not}{=} I_{G,R}^{marzo24} = \frac{114,9 - 120 \cdot 0,1 + 107 \cdot 0,2 + 115 \cdot 0,1}{0,6} = 116,66$$

Ejercicio 9

Los datos de producción de energía eléctrica y renta per cápita de un determinado país en el período 2015 a 2019 fueron:



Años	Energía Eléctrica (en miles de mill. de kW/h)	Renta per Cápita (en miles de €)
2015	32	35
2016	38	40
2017	41	43
2018	46	47
2019	52	51

Se pide:

1) Hallar las ecuaciones de las rectas de regresión y dibujarlas

2) Significado de los coeficientes de regresión

3) Calcular el valor del coeficiente de correlación

4) Previsión de la renta per cápita cuando la producción de energía eléctrica sea de 60,000 millones de kW/h

5) ¿Qué confianza merecen las predicciones?

Solución

1) Sea X la energía, Y la renta per cápita y $N = 5$ el número de puntos. La recta de regresión r de la variable Y respecto de X tiene por ecuación

$$r \equiv y - \overline{Y} = \frac{S_{XY}}{S_X^2}(x - \overline{X}).$$

Calculemos los términos:

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{32 + 38 + 41 + 46 + 52}{5} = 41{,}8$$
$$\overline{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{35 + 40 + 43 + 47 + 51}{5} = 43{,}2$$



$$S_X^2 = \mathrm{Var}(X) = \overline{X^2} - \overline{X}^2 = \frac{32^2 + 38^2 + 41^2 + 46^2 + 52^2}{5} - 41{,}8^2 = 46{,}56$$
$$S_Y^2 = \mathrm{Var}(Y) = \overline{Y^2} - \overline{Y}^2 = \frac{35^2 + 40^2 + 43^2 + 47^2 + 51^2}{5} - 43{,}2^2 = 30{,}56$$
$$S_{XY} = \mathrm{Cov}(X,Y) = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y}$$
$$= \frac{32 \cdot 35 + 38 \cdot 40 + 41 \cdot 43 + 46 \cdot 47 + 52 \cdot 51}{5} - 41{,}8 \cdot 43{,}2 = 37{,}64.$$

Ya podemos sustituir en la ecuación de la recta de regresión:

$$r \equiv y - \overline{Y} = \frac{S_{XY}}{S_X^2}(x - \overline{X}) \equiv y - 43{,}2 = \frac{37{,}64}{46{,}56}(x - 41{,}8)$$
$$\equiv y = 0{,}8084x + 9{,}4080.$$

En cuanto a la recta de regresión s de la variable X respecto de la Y, análogamente sigue la ecuación

$$s \equiv x - \overline{X} = \frac{S_{XY}}{S_Y^2}(y - \overline{Y}) \equiv x - 41{,}8 = \frac{37{,}64}{30{,}56}(y - 43{,}2)$$
$$\equiv x = 1{,}2316y - 11{,}4083.$$

En la Figura 4.6 aparecen graficadas ambas regresiones. Los puntos originales están cerca de una alineación perfecta, lo que hace que ambas rectas de regresión se ajusten muy bien a ellos y parezcan casi la misma, pero no es el caso. Para verificarlo, calculamos el producto de las pendientes en su expresión original, que resulta

$$Q = \frac{37{,}64}{46{,}56} \cdot \frac{37{,}64}{30{,}56} \approx 0{,}995710 \neq 1.$$

2 La regresión de Y sobre X es una recta, por tanto se caracteriza por dos parámetros: su pendiente y su ordenada en el orígen. En cuanto a

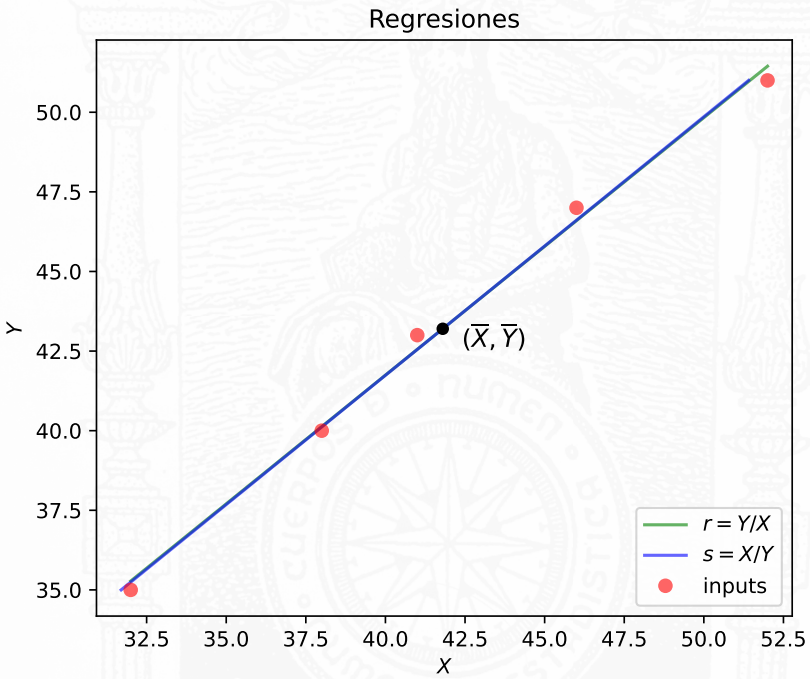


Figura 4.6: Regresiones lineales para los datos conocidos. Se trata de dos rectas distintas, aunque parezcan solaparse.



la pendiente, asumiendo que ese modelo se ajusta bien a los datos, da la cantidad adicional de euros cuando el número de kW/h de energía producida aumenta en un millón. La ordenada en el orígen da la predicción de la renta per cápita según nuestro modelo para el caso de países cuya producción energética es nula.

3) El coeficiente de correlación es $r = \sqrt{Q} \approx \sqrt{0{,}995710} \approx 0{,}997852$.

   De forma alternativa, puede calcularse con la fórmula $r = \frac{S_{XY}}{\sqrt{S_X^2}\sqrt{S_Y^2}}$, dando el mismo resultado.

(4) La previsión pedida resulta de estimar el valor a partir de la recta de regresión r, es decir,

$$r(60) \approx 0{,}8084 \cdot 60 + 9{,}4080 \approx 57{,}91323 \text{ miles de euros} = 57\,913{,}23€.$$

5) La confianza en las estimaciones debe depender del indicador de la bondad del ajuste $R^2 = r^2 = Q = 0{,}995710$. Esto significa que el $99{,}57\,\%$ de la varianza de Y viene explicada por la varianza de X, por lo que la confianza en la predicción debe ser bastante alta.

Ejercicio 10

Dado el proceso:

$Y_t = 1{,}4Y_{t-1} - 0{,}5Y_{t-2} + a_t$

Donde a_t es un proceso de ruido blanco con varianza δ_a^2 se pide:

1) ¿De qué tipo de proceso estamos hablando? Escribir el proceso en notación de retardos.
2) ¿Qué significa que un proceso es estacionario?
3) ¿Cuáles son las condiciones que se deben cumplir para que este proceso sea estacionario?
4) ¿Este proceso es estacionario? Razone la respuesta.

Solución.

1) Se conoce como proceso AR(2) a un proceso cuya expresión responde a:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + a_t + \delta$$

cuya media se puede calcular de forma sencilla y vale $\mu = \frac{\delta}{1-\phi_1-\phi_2}$, por lo que como el proceso que figura en el ejercicio no tiene constante entendemos que el proceso representa un modelo que está expresado en desviaciones respecto a la media, que de forma general se escribe como:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + a_t$$

Denotando por $Ly_t = y_{t-1}$, en forma de retardos, ese proceso se escribiría como:

$$(1 - \phi_1 L - \phi_2 L^2) y_t = a_t$$

siendo $\Phi(L) = 1-\phi_1 L-\phi_2 L^2$ y por tanto, el proceso que se nos presenta un proceso autorregresivo de orden 2.

2) Se dice que un proceso de series de tiempo Y_t es estacionario en covarianzas o débilmente estacionario cuando la media y las autocovarianzas existen, es decir, son finitas y no dependen de la fecha t:

$$E(Y_t) = \mu < \infty, \forall t$$

$$E[(Y_t - \mu)(Y_{t-j} - \mu)] = \gamma_j < \infty, \forall t, j$$

Un proceso Y_t es estrictamente estacionario o fuertemente estacionario cuando para cualquier valor de $j_1, j_2, \cdots, j_n$ la distribución conjunta de $Y_t, Y_{t+j_1}, Y_{t+j_2}, \cdots, Y_{t+j_n}$ depende sólo de los intervalos que separan las fechas $j_1, j_2, \cdots, j_n$ y no de la fecha t. La estacionariedad estricta o fuerte requiere que la distribución conjunta de probabilidades sea la



misma sin importar el punto inicial o final de la serie de tiempo; es decir, el proceso generador de datos es constante en el tiempo.

Generalmente la palabra estacionariedad hará referencia a estacionariedad débil. La estacionariedad es importante para la estimación de los parámetros, porque sugiere que algo es fijo a través de la secuencia de variables aleatorias y que si pudiéramos observar el proceso por un período de tiempo suficientemente largo aprenderíamos sobre el proceso generador de datos subyacente.

3) Para que este proceso sea estacionario hay que comprobar que las raíces del polinomio de retardos caen fuera del círculo unitario.

En general tenemos que:

$$L_1 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}$$

$$L_2 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}$$

En el caso que nos ocupa, $\phi_1 = 1,4$ y $\phi_2 = -0,5$, luego:

$$L_1 = \frac{1,4 + \sqrt{1,96 - 2}}{+1} = 1,4 + \sqrt{-0,04} = 1,4 + i\sqrt{0,04} = 1,4 + 0,2 \cdot i$$

$$L_2 = \frac{1,4 - \sqrt{1,96 - 2}}{+1} = 1,4 - \sqrt{-0,04} = 1,4 - i\sqrt{0,04} = 1,4 - 0,2 \cdot i$$

Al ser raices complejas, el módulo debe ser mayor que 1. Veamos:

$$|L_1| = \sqrt{(1,4)^2 + (0,2)^2} = \sqrt{1,96 + 0,04} = \sqrt{2} > 1$$

De la misma forma que comprueba que $|L_2|$ es también mayor que 1, luego el proceso es estacionario. La condición de estacionariedad a veces se encuentra de la siguiente forma escrita:

$$\begin{cases} \phi_1 + \phi_2 < 1 \\ \phi_2 - \phi_1 < 1 \\ |\phi_2| < 1 \end{cases}$$



que sustituyendo los valores de ϕ_1 y de ϕ_2 se comprueba fácilmente que se cumplen.



4.2. Soluciones del segundo examen del Cuerpo Superior de Estadísticos del Estado de 2023. Parte Común.

Cuestión 1. En una población formada por $N = 8$ unidades $\{1, 2, 3, 4, 5, 6, 7, 8\}$, se considera un diseño muestral que genera las siguientes muestras posibles:

$$p(\{1,3,5,6\}) = \frac{1}{8}; \quad p(\{1,4,6,8\}) = \frac{1}{8}; \quad p(\{4,5,7,8\}) = \frac{1}{8}$$
$$p(\{2,3,7,8\}) = \frac{1}{4}; \quad p(\{2,4,6,8\}) = \frac{3}{8}$$

Se pide:

a) Calcular las probabilidades de inclusión de primer orden, π_i, de todos los elementos de la población.

b) El diseño de la muestra, ¿es aleatorio simple sin reemplazamiento, aleatorio simple estratificado o ninguno de ellos? Razonar la respuesta.

c) Si los tamaños de estas 8 unidades son $\{100, 40, 40, 20, 20, 10, 5, 5\}$ respectivamente, ¿podría aplicarse un diseño muestral sin reposición, con tamaño $n = 4$, de manera que las π_i sean proporcionales a sus tamaños? Justifique su respuesta.

Solución.

a) La probabilidad de inclusión de primer orden, denotada como π_i para $i = 1, \ldots, 8$, representa la probabilidad de que cada elemento de la población pertenezca a la muestra S. En particular, para la primera unidad de la población, esta probabilidad se calcula como:

$$\pi_1 = P(1 \in S) = P(\{1,3,5,6\}) + P(\{1,4,6,8\}) = \frac{1}{8} + \frac{1}{8} = \frac{2}{8}$$

De igual forma se calculan las demás, obteniéndose:

$$\pi_2 = \frac{5}{8}, \pi_3 = \frac{3}{8}, \pi_4 = \frac{5}{8}, \pi_5 = \frac{2}{8}, \pi_6 = \frac{5}{8}, \pi_7 = \frac{3}{8}, \pi_8 = \frac{7}{8}$$

Puesto que, en un diseño muestral con tamaño de muestra fijo, se cumple que $\sum_{i \in U} \pi_i = n$, esta expresión puede utilizarse para verificar la corrección de los cálculos realizados.

En este caso, puede comprobarse que:

$$\sum_{i=1}^{8} \pi_i = \frac{32}{8} = 4,$$

lo cual coincide con el tamaño de las muestras seleccionadas.

b) En un muestreo aleatorio simple sin reposición, las probabilidades de inclusión de primer orden de todos los elementos de la población son iguales. En concreto, se cumple que $\pi_i = \frac{n}{N}$. Dado que $n = 4$ y $N = 8$, se tiene que $\pi_i = \frac{4}{8} = \frac{1}{2}, i = 1, \ldots, 8$. Sin embargo, al analizar las probabilidades en el caso que nos ocupa, se observa que estas no cumplen con dicha relación. Por lo tanto, no puede tratarse de un muestreo aleatorio simple sin reposición.

En el muestreo estratificado, en cada estrato se cumple que $\pi_i = \frac{n_i}{N_i}$, siendo n_i y N_i el número de elementos de la muestra y de la población en ese estrato respectivamente.

En nuestro caso, podríamos agrupar los elementos de la población de forma que tuviesen probabilidades iguales de la siguiente manera:

$$\begin{cases} \{1,5\} \to \frac{2}{8} \\ \{2,4,6\} \to \frac{5}{8} \\ \{3,7\} \to \frac{3}{8} \\ \{8\} \to \frac{7}{8} \end{cases}$$



Como en cada estrato las probabilidades de extraer elementos de él deberían ser iguales, esta sería la única configuración posible de estratos, pero es obvio que no se cumple la condición citada de las π_i, obsérvese el caso concreto de $\frac{7}{8}$ que es fracción irreducible, ni se cumple que el diseño muestral que obtenemos pudiera ser coherente con tener esos 4 posibles estratos, luego tampoco puede ser el muestreo estratificado.

c) Que las probabilidades de inclusión de cada elemento sean proporcionales a los tamaños de los mismos es equivalente a que:

$$\pi_i = n \cdot \frac{x_i}{\sum_{i=1}^{N} x_i}$$

ya que $\sum\limits_{i=1}^{N} \pi_i = n$, donde x_i es el tamaño de la uidad i-ésima, .

En concreto, dado que:

$$\sum_{i=1}^{8} x_i = 100 + 40 \cdot 2 + 20 \cdot 2 + 10 + 5 \cdot 2 = 240,$$

y considerando que $x_1 = 100$, se tendría:

$$\pi_1 = 4 \cdot \frac{100}{240} = \frac{10}{6} > 1.$$

Esto no es posible, ya que las probabilidades no pueden ser mayores que 1. Por lo tanto, no es factible aplicar aplicar un diseño muestral sin reposición de tamaño 4 cuyas probabilidades de inclusión sean proporcionales a estos tamaños de manera estricta.



Cuestión 2. La función de densidad conjunta de las variables (X, Y) es la siguiente:

$$f_a(x, y) = a^2 e^{a(x+y)}, x \geq 0, y \geq 0.$$

Se toma una muestra aleatoria simple de tamaño n.

a) Calcule el estimador máximo verosímil del parámetro a.

b) Utilice dicho estimador para obtener $\hat{a}$ a partir de la información muestral recogida en la siguiente tabla:

x	1	2	3	4	5
y	2	0	1	0	1

Solución.

a) Dada una muestra aleatoria simple de tamaño n escribimos la función de verosimilitud:

$$f_a\left((x_1, y_1), \ldots, (x_n, y_n)\right) = \left(a^2\right)^n \cdot e^{a\sum_{i=1}^{n}(x_i+y_i)} \quad , (x_1, y_1), \ldots, (x_n, y_n) \geq 0$$

Tomando logaritmos tenemos:

$$Lnf_a\left((x_1, y_1), \ldots, (x_n, y_n)\right) = 2nLa + a\sum_{i=1}^{n}(x_i + y_i)$$

Y derivando respecto del parámetro a e igualando a cero queda:

$$\frac{\partial Lnf_a\left((x_1, y_1), \ldots, (x_n, y_n)\right)}{\partial a} = \frac{2n}{a} + \sum_{i=1}^{n}(x_i + y_i) = 0$$

Despejando el parámetro obtenemos el estimador de máxima verosimilitud:

$$\hat{a} = \frac{-2n}{\sum_{i=1}^{n}(x_i + y_i)} = \frac{-2}{\bar{x} + \bar{y}}$$

Calculando la segunda derivada es sencillo ver que se trata de un máximo, al ser negativa, por lo que este es el estimador máximo verosímil.



b) Según los datos de la tabla tenemos que $\sum_{i=1}^{5} x_i = 1+2+3+4+5 = 15$ y sabemos también que $\sum_{i=1}^{5} y_i = 2+0+1+0+1 = 4$, por lo tanto:

$$\hat{a} = \frac{-2n}{\sum_{i=1}^{n}(x_i+y_i)} = \frac{-2}{\bar{x}+\bar{y}} = \frac{-2}{\frac{15}{5}+\frac{4}{5}} = \frac{-2}{3+0{,}8} = -0{,}5263$$

nota: el ejercicio no especifica nada acerca del signo de a, pero está claro que no puede ser positivo, porque entonces la función no sería función de densidad. Hubiera sido más familiar si la expresión de la conjunta hubiera tenido un signo menos en el exponente y se hubiera tratado del producto de dos exponenciales de parámetro a, siendo a positivo. En la estimación lógicamente obtenemos un valor negativo y en el caso que hubieramos esperado dicha estimación sería positiva.

Cuestión 3

a) Sea un sistema de representación numérica de enteros con 8 bits.

- Expresar el número decimal -39, en representación binaria utilizando formatos signo–magnitud (con primer dígito para signo), complemento a 1 y complemento a 2. El dígito de signo se representa como 1 negativo y 0 positivo.
- Si un número binario en complemento a 2 se representa como 10101010, ¿cúal es su valor decimal?
- ¿Cúal es el entero mínimo y máximo representables en complemento a 2 en este sistema (que incluye bit de signo)?

b) En un sistema binario (sin bit para signo), ¿a qué valor decimal exacto equivale el número 110101,111?



c) ¿Cuál es el valor decimal exacto del número binario $0{,}10101101101\cdots = 0{,}1\widehat{0101}$?

Solución

Por claridad, denotaremos $(x)_2$ cuando x deba interpretarse como número en base 2, $(x)_2$ SM como binario en signo-magnitud, $(x)_2$ C1 como binario en complemento a 1 y $(x)_2$ C2 como binario en complemento a 2.

a) En cuanto a la representación en signo-magnitud de -39, comenzamos considerando el valor absoluto del número -39 a expresar. La forma de expresar 39 como potencias de 2 es $39 = 32 + 4 + 2 + 1$, por tanto su representación en binario es $(100111)_2$. Debemos utilizar 7 bits, por lo que pasa a ser $(0100111)_2$. Finalmente para expresar -39 en signo-magnitud debemos anteponer a esto el bit de signo, que por ser negativo es 1, resultando $(10100111)_2$ SM.

   Para obtener la representación binaria de -39 en complemento a uno, se parte de la representación en binario de su valor absoluto 39 con 8 bits, que es $(00100111)_2$, y puesto que deseamos cambiar el signo, se invierten todos los bits, dando lugar a $(11011000)_2$ C1.

   Finalmente, para obtener la representación binaria de -39 en complemento a dos, seguimos los siguientes pasos:

   (1) Partimos de la representación del valor absoluto en binario, $(00100111)_2$

   (2) Puesto que queremos el número en negativo, invertimos todos los bits (cambiando 0s por 1s y 1s por 0s), recuperando la representación en complemento a uno, $(11011000)_2$ C1

   (3) Sumamos 1 al resultado anterior, interpretado como binario sin signo: $(11011001)_2$ C2



Para obtener el valor decimal correspondiente $(10101010)_2$ C2, deshacemos los pasos anteriores, en orden inverso:

(1) Restamos 1 en binario: al restar de $(10101010)_2$ resulta $(10101001)_2$.

(2) Puesto que empieza por 1, sabemos que el número buscado es negativo. Su valor absoluto resulta de invertir todos los bits, dando lugar a $(01010110)_2$

(3) Convertimos ese número binario a decimal: $(01010110)_2 = 2+4+16+64 = 86$.

Por tanto, el número representado es el -86.

Respecto al rango de representación en complemento a dos, con un número de bits igual a 8 (es decir, $n = 8$) se pueden representar $2^n = 2^8$ enteros, desde $-2^{n-1} = -2^7$ hasta $2^{n-1} - 1 = 2^7 - 1$, o sea, el rango $-128, -127, \ldots, +126, +127$.

b) El valor viene de sumar las potencias de 2 en cuya posición hay un uno. Es decir,

$$\begin{aligned}(110101{,}111)_2 &= 2^5 + 2^4 + 2^2 + 2^0 + 2^{-1} + 2^{-2} + 2^{-3} \\ &= 32 + 16 + 4 + 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = 53 + \frac{7}{8} = 53{,}875.\end{aligned}$$

c) Sea $x = (0{,}1\overline{0101})_2$, por ser un número periódico sucede que $x \in \mathbb{Q}$, desearíamos obtener su representación como fracción irreducible. La dificultad estriba en el manejo del periodo, porque contiene una suma infinita de potencias de dos. Para librarnos de ella, consideremos dos múltiplos de x, provenientes de multiplicarlo por 2^2 y 2^5. El primero supone pasar a la izquierda del punto decimal el prefijo de decimales

antes del periodo, ya que resulta $4x = (10.\widehat{101})_2$. El segundo supone, además de eso, pasar a la izquierda del punto decimal un periodo entero completo, puesto que $32x = (10101.\widehat{101})_2$. Ambos cantidades tienen exactamente los mismos dígitos decimales, así que al restarlas obtenemos un entero:

$$32x - 4x = (10101.\widehat{101})_2 - (10.\widehat{101})_2 = (10101)_2 - (10)_2 = 16 + 4 + 1 - 2 = 19,$$

de donde $28x = 19$ y por tanto $x = \frac{19}{28}$.



Cuestión 4. Complete las diez casillas marcadas en la siguiente tabla de mortalidad, indicando la fórmula utilizada en cada caso o razonando la respuesta si no es necesario aplicar ninguna fórmula.

	$_nq_x$	$_nl_x$	$_nd_x$	$_nL_x$	$_nT_x$	$_ne_x$	$_na_x$
0 años	1,8246	100.000,00	182,46	99.859,10	g	83,23	i
De 1 a 4 años	0,4937	99.817,54	49,28	e	8.223.335,74	82,38	0,2768
De 5 a 9 años	a	99.768,26	33,45	498.759,90	7.824.208,14	78,42	0,5133
De 10 a 14 años	0,3820	99.734,81	38,10	498.585,83	7.325.448,24	73,45	0,5369
De 15 a 19 años	1,0772	99.696,71	107,39	498.197,46	6.826.862,41	68,48	0,4672
De 20 a 24 años	1,6008	99.589,32	159,42	497.563,59	6.328.664,95	63,55	0,5195
De 25 a 29 años	1,3157	99.429,90	130,82	496.837,63	5.831.101,36	58,65	0,5232
De 30 a 34 años	2,3522	99.299,08	233,57	495.948,73	5.334.263,73	53,72	0,5319
De 35 a 39 años	3,4211	99.065,51	338,91	494.551,62	4.838.315,00	48,84	0,5421
De 40 a 44 años	4,5478	98.726,60	448,99	492.563,95	4.343.763,38	44,00	0,5238
De 45 a 49 años	8,0801	98.277,61	794,09	f	3.851.199,43	39,19	0,5401
De 50 a 54 años	13,7566	97.483,52	1.341,04	484.447,87	3.361.637,39	34,48	0,5571
De 55 a 59 años	23,8322	96.142,48	2.291,29	475.372,55	2.877.189,52	29,93	0,5339
De 60 a 64 años	34,7539	93.851,19	3.261,69	461.721,45	2.401.816,97	25,59	0,5380
De 65 a 69 años	50,7729	90.589,50	4.599,49	442.359,47	1.940.095,52	h	0,5396
De 70 a 74 años	73,6006	85.990,01	d	415.013,80	1.497.736,05	17,42	0,5280
De 75 a 79 años	121,5686	c	9.684,29	376.070,32	1.082.722,25	13,59	0,5408
De 80 a 84 años	204,4377	69.976,80	14.305,90	317.073,42	706.651,93	10,10	0,5413
De 85 a 89 años	374,6874	55.670,90	20.859,18	228.407,19	389.578,51	7,00	0,5211
De 90 a 94 años	592,7031	34.811,72	20.633,01	119.639,04	161.171,32	4,63	0,4725
95 y más años	b	14.178,71	14.178,71	41.532,28	41.532,28	2,93	j

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA

Solución.

En la figura 4.7 podemos ver el valor calculado para cada una de las celdas propuestas, cuyo desarrollo y explicación detallamos a continuación.

	$_nq_x$	$_nl_x$	$_nd_x$	$_nL_x$	$_nT_x$	$_ne_x$	$_na_x$
0 años	1,8246	100000	182,46	99859,1	8323194,84	83,23	0,22777595
De 1 a 4 años	0,4937	99817,54	49,28	399127,603	8223335,74	82,38	0,2768
De 5 a 9 años	0,33527697	99768,26	33,45	498759,9	7824208,14	78,42	0,5133
De 10 a 14 años	0,382	99734,81	38,1	498585,83	7325448,24	73,45	0,5369
De 15 a 19 años	1,0772	99696,71	107,39	498197,46	6826862,41	68,48	0,4672
De 20 a 24 años	1,6008	99589,32	159,42	497563,59	6328664,95	63,55	0,5195
De 25 a 29 años	1,3157	99429,9	130,82	496837,63	5831101,36	58,65	0,5232
De 30 a 34 años	2,3522	99299,08	233,57	495948,73	5334263,73	53,72	0,5319
De 35 a 39 años	3,4211	99065,51	338,91	494551,62	4838315	48,84	0,5421
De 40 a 44 años	4,5478	98726,6	448,99	492563,95	4343763,38	44	0,5238
De 45 a 49 años	8,0801	98277,61	794,09	489562,04	3851199,43	39,19	0,5401
De 50 a 54 años	13,7566	97483,52	1341,04	484447,87	3361637,39	34,48	0,5571
De 55 a 59 años	23,8322	96142,48	2291,29	475372,55	2877189,52	29,93	0,5339
De 60 a 64 años	34,7539	93851,19	3261,69	461721,45	2401816,97	25,59	0,538
De 65 a 69 años	50,7729	90589,5	4599,49	442359,47	1940095,52	21,4163399	0,5396
De 70 a 74 años	73,6006	85990,01	6328,89685	415013,8	1497736,05	17,42	0,528
De 75 a 79 años	121,5686	79661,1131	9684,29	376070,32	1082722,25	13,59	0,5408
De 80 a 84 años	204,4377	69976,8	14305,9	317073,42	706651,93	10,1	0,5413
De 85 a 89 años	374,6874	55670,9	20859,18	228407,19	389578,51	7	0,5211
De 90 a 94 años	592,7031	34811,72	20633,01	119639,04	161171,32	4,63	0,4725
95 y más años	1000	14178,71	14178,71	41532,28	41532,28	2,93	2,93

Figura 4.7: Tabla de mortalidad con las celdas objeto de cálculo coloreadas.

a) $_5q_5 = \frac{_5d_5}{_5l_5} \cdot 1000 = \frac{33{,}45}{99\,768{,}26} \cdot 1000 = 0,33527697‰$

b) $_+q_{95} = 1000‰$, puesto que la columna de las probabilidades de muerte está en tanto por mil y en el último intervalo la probabilidad de morir es 1.

c) $_5l_{75} = {}_5d_{75} + {}_5l_{80} = 9684{,}29 + 69\,976{,}80 = 79\,661{,}113$

d) $_5d_{70} = {}_5l_{70} \cdot \frac{_5q_{70}}{1000} = 85\,990{,}01 \cdot \frac{73{,}6006}{1000} = 6328{,}896$

e) $_4L_1 = {}_4T_1 - {}_5T_5 = 8\,223\,335{,}74 - 7\,824\,208{,}14 = 399\,127{,}603$



f) ${}_5L_{45} = {}_5T_{45} - {}_5T_{50} = 3\,851\,199{,}43 - 3\,361\,637{,}39 = 489\,562{,}04$

g) ${}_1T_0 = {}_4T_1 + {}_1L_0 = 8\,223\,335{,}74 + 99\,859{,}1 = 8\,323\,194{,}84$

h) ${}_5e_{65} = \frac{{}_5T_{65}}{{}_5l_{65}} = \frac{1\,940\,095{,}52}{90\,589{,}5} = 21{,}416$

i) Para el cálculo del número medio de años que viven los que mueren entre 0 y 1 años utilizamos la fórmula:

$$ {}_1L_0 = {}_4l_1 \cdot 1 + {}_1a_0 \cdot {}_1d_0 $$

y despejando obtenemos:

$$ {}_1a_0 = \frac{{}_1L_0 - {}_4l_1}{{}_1d_0} = \frac{99\,859{,}1 - 99\,817{,}54}{49{,}28} = 0,22777 $$

j) El valor del número medio de años que viven las personas que mueren con más de 95 años es justamente la esperanza de vida media en ese último intervalo, por lo tanto:

$$ {}_+a_{95} = {}_+e_{95} = 2{,}93 $$



Cuestión 5. Teniendo en cuenta la siguiente información contable correspondiente a una economía para un determinado año T (datos en miles de millones EUR):

Remuneración de los asalarados residentes	660
siendo de empleadores no residentes	18
Remuneración de los asalariados no residentes por empleadores residentes	15
Otros impuestos menos subvenciones sobre la producción	13
Saldo de intercambios exteriores de bienes y servicios	-20
Consumo colectivo efectivo	112
Transferencias sociales en especie	159
Renta disponible ajustada bruta del Sector Hogares	1.032
Ahorro bruto del Sector Hogares	75

a) Elabore el siguiente cuadro macroeconómico (en hoja aparte):

(miles de millones EUR)	AÑO T
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO	
Demanda	
Gasto en consumo final	
Gasto en consumo final de los hogares	
Gasto en consumo final de las ISFLSH	15
Gasto en consumo final de las AAPP	
Formación bruta de capital	294
Exportaciones de bienes y servicios	
Importaciones de bienes y servicios	540
Oferta	
Agricultura, ganadería, silvicultura y pesca	33
Industria	158
Construcción	
Servicios	922
Impuestos menos subvenciones sobre los productos	
Rentas	
Remuneración de los asalariados	
Excedente de explotación bruto / Renta mixta bruta	
Impuestos menos subvenciones sobre la producción y las importaciones	133

Suponiendo que, en ese año, el ajuste por la variación de los derechos por pensiones no es significativo.

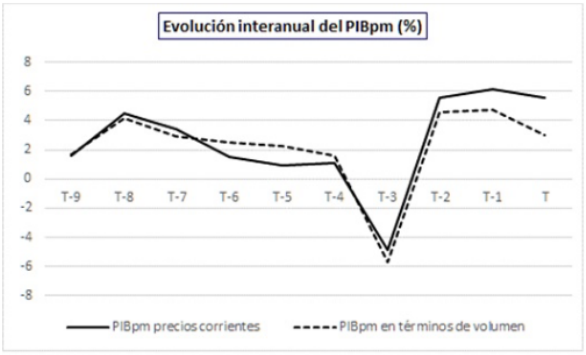
b) Indique cómo se vería afectado el cuadro anterior si, además, se producen conjuntamente las siguientes transacciones económicas:

1) Reducción del tipo del IVA en productos energéticos, que provoca un descenso de 2 miles de millones EUR en los ingresos de las AAPP.

2) Aumento del servicio producido por personal doméstico remunerado en 3 miles de millones EUR.

3) Disminución de la aportación del hogar al pago de medicamentos con receta en 5 miles de millones EUR.

c) Atendiendo al siguiente gráfico, ¿qué se podría decir de la variación del deflactor implícito del PIB en los últimos 10 años en esta economía?



Solución.

a) En la figura 4.9 se puede ver el desglose de los elementos que componen el Producto Interior Bruto según las tres vías, que detallamos a continuación.

Por el método de las rentas, la remuneración de asalariados que hay que considerar en el cálculo es la interior, que se corresponde con la pagada por residentes a residentes más la pagada por residentes a no residentes, que con los datos del problema es 660 + 15 - 18 , es decir, 657.

El saldo de intercambios exteriores de bienes y servicios se refiere a la diferencia entre las exportaciones y las importaciones de bienes y servicios de un país con el resto del mundo. Este concepto forma parte de la balanza de bienes y servicios, que a su vez es un componente

clave de la balanza por cuenta corriente en la balanza de pagos, y como conocemos dicho saldo, que vale -20 y las importaciones de bienes y servicios, que vale 540, despejando obtenemos las exportaciones, que son 520.

Por otra parte:

$$D.2 - D.3 = D.21 - D.31 + D.29 - D.39$$

luego según los datos del problema queda

$$133 = D.21 - D.31 + 13$$

y despejando obtenemos que $D.21 - D.31 = 120$.

El consumo colectivo efectivo es el gasto en consumo final colectivo de las administraciones públicas, que según la primera tabla de datos del problema tiene un valor de 112. Para calcular el gasto en consumo final de las administraciones públicas, el cuál se compone del gasto individual y colectivo necesitamos pues el gasto en consumo final individual. Pero como según el problema las transferencias sociales en especie son 159, que son la suma del gasto en consumo individual de las AAPP y de las ISFLSH, y el gasto en consumo final de las ISFLSH (que es individual íntegramente) vale 15 según el problema, obtenemos por diferencia que el gasto en consumo final individual de las AAPP es 159-15 , es decir, 144. Por tanto, el gasto en consumo final de las AAPP es 112 + 144, es decir, el colectivo más el individual, en total 256.

Si somos capaces de calcular el gasto en consumo individual de los hogares, podremos calcular el PIB por el método de la demanda y como el valor coincide por los otros dos métodos podremos despejar el dato



de la construcción y el dato del EBE/RMB por diferencia, resolviendo el apartado a) completamente.

Para calcular el gasto en consumo individual de los hogares nos ayudamos de la cuenta de redistribución de la renta en especie y de la cuenta de utilización de la renta disponible.

II. 3 Cuenta de redistribución de la renta en especie

Empleos	**Código**	**Operaciones y otros flujos**	**Recursos**
S.1		**y saldos contables**	**S.1**
Total de la economía			**Total de la economía**
	B.6b	Renta disponible bruta	873
	D.63	Transferencias sociales en especie:	159
1.032	**B.7b**	**Renta disponible ajustada bruta**	

II. 4.1 Cuenta de utilización de la renta disponible

Empleos	**Código**	**Operaciones y otros flujos**	**Recursos**
S.1		**y saldos contables**	**S.1**
Total de la economía			**Total de la economía**
	B.6b	Renta disponible bruta	873
798	P.3	Gasto en consumo final	
	D.8	Ajuste por la variación de los derechos	
		por pensiones	
75	**B.8b**	**Ahorro bruto**	

Figura 4.8: Cálculo del Gasto en consumo final de los hogares. Las celdas en gris son datos del problema.

En la figura 4.8 hemos utilizado que según nos dice el enunciado el ajuste por la variación de los derechos por pensiones no es significativo y por este motivo no se refleja ningún valor en la cuenta II.4.1. De la cuenta II.3 obtenemos la renta disponible bruta, que se utiliza en la

siguiente cuenta II.4.1 para obtener el dato del gasto en consumo final de los hogares, que es individual.

Ahora ya con este dato del gasto en consumo final de los hogares (798), podemos completar los datos de la tabla de la figura 4.9 obteniendo el valor de la construcción y del EBE/RMB.

Producto interior bruto a precios de mercado y sus componentes

Precios corrientes

Unidad: miles de millones de euros

Demanda	
Gasto en consumo final	1.069
Gasto en consumo final de los hogares	798
Gasto en consumo final de las ISFLSH	15
Gasto en consumo final de las AAPP	256
Formación bruta de capital	294
Exportaciones de bienes y servicios	520
Importaciones de bienes y servicios	540
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO	1.343

Oferta	
Agricultura, ganadería, silvicultura y pesca	33
Industria	158
Construcción	110
Servicios	922
Impuestos menos subvenciones sobre los productos (D.21-D.31)	120
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO	1.343

Rentas	
Remuneración de los asalariados (660-18+15)	657
Excedente de explotación bruto / Renta mixta bruta	553
Impuestos menos subvenciones sobre la producción y las importaciones (D.2-D.3)	133
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO	1.343

Dato dado por el problema en esta tabla	
Dato obtenido de la tabla de la economía y del cuadro directamente	
Dato que se calcula operando con diversos datos del problema	

Figura 4.9: Cálculo del Producto Interior Bruto por el método de la demanda, de la oferta y de las rentas.



b) 1) Una reducción del tipo del IVA en productos energéticos, que provoca una caída de 2 mil millones de euros en los ingresos de las AAPP hace que disminuya el gasto en consumo final de las AAPP en esa cuantía. Además, esa disminución hace que sea menor el valor de D.21 y por tanto de D.2, con lo cuál , si no hubiera ninguna transacción adicional, haría que disminuyera el valor del PIB en 2000 millones de euros. Por el método de la oferta y las rentas se pueden ver los nuevos valores de cada agregado.

2) Un aumento del servicio producido por personal doméstico remunerado de 3000 millones de euros hace que aumente el gasto en consumo final de los hogares, así como el agregado de servicios en el cuadro de la oferta y la remuneración de asalariados en la parte de las rentas.

3) Una disminución de la aportación del hogar al pago de medicamentos con receta en 5000 millones de euros no parece que tenga efecto en el PIB porque esta partida que no recae sobre los hogares serán las AAPP y las ISFLSH las que tengan que hacer frente a ellas, por lo que cambian los valores del gasto en consumo final de cada sector.

Producto interior bruto a precios de mercado y sus componentes

Precios corrientes

Unidad: miles de millones de euros	**Situación inicial**	**Apartado b)**	
Demanda			
Gasto en consumo final	1.069	1.070	
Gasto en consumo final de los hogares	798	796	2), 3) 798 + 3 - 5 = 796
Gasto en consumo final de las ISFLSH	15	15	
Gasto en consumo final de las AAPP	256	259	1), 3) 256 - 2 + 5 = 259
Formación bruta de capital	294	294	
Exportaciones de bienes y servicios	520	520	
Importaciones de bienes y servicios	540	540	
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO	1.343	1.344	
Oferta			
Agricultura, ganadería, silvicultura y pesca	33	33	
Industria	158	158	
Construcción	110	110	
Servicios	922	925	2)
Impuestos menos subvenciones sobre los productos (D.21-D.31)	120	118	1)
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO	1.343	1.344	
Rentas			
Remuneración de los asalariados	657	660	2)
Excedente de explotación bruto / Renta mixta bruta	553	553	
Impuestos menos subvenciones sobre la producción y las importaciones (D.2-D.3)	133	131	1)
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO	1.343	1.344	

Figura 4.10: Efecto en el Producto Interior Bruto de diversas incidencias en la economía.

En la figura 4.10 se puede ver la situación inicial y la que resultaría de reflejar esos tres condicionantes que nos muestra el problema, comprobando que globalmente el producto interior bruto ha aumentado en 1000 millones de euros.

c) Para dar respuesta a este ejercicio vamos a analizar un caso real con datos de la economía española y la traducción a este caso sería inmediata con las consideraciones que vamos a trabajar en el ejemplo. En el gráfico que nos presenta el apartado c) aparecen las tasas de variación del PIB en términos corrientes y en términos de volumen, puesto que el título del gráfico es la evolución interanual en tanto por ciento del PIB.



Sabemos que se cumple la relación siguiente:

$$\text{Índice}_{valor} = \text{Índice}_{precios} \cdot \text{Índice}_{volumen}$$

es decir, que un índice de valor es el producto del índice de volumen por el índice de precios, y también sabemos cuál es la relación que hay entre un índice y una tasa de variación, puesto que por ejemplo $\left(\frac{\text{Índice}_{valor}}{100} - 1\right) \cdot 100 = \text{TasaVariación}_{valor}$, expresada en tanto por ciento.

Tasas de variación interanuales

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021 (P)	2022 (A)
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO (corrientes)	0,32	-0,83	-3,07	-1,01	1,17	4,40	3,37	4,31	3,56	3,46	-10,16	9,23	10,15
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO(volumen)	0,16	-0,81	-2,96	-1,40	1,40	3,84	3,04	2,98	2,28	1,98	-11,17	6,40	5,77
DEFLACTOR DEL PIB	0,15	-0,02	-0,11	0,40	-0,22	0,55	0,32	1,30	1,25	1,45	1,14	2,66	4,14

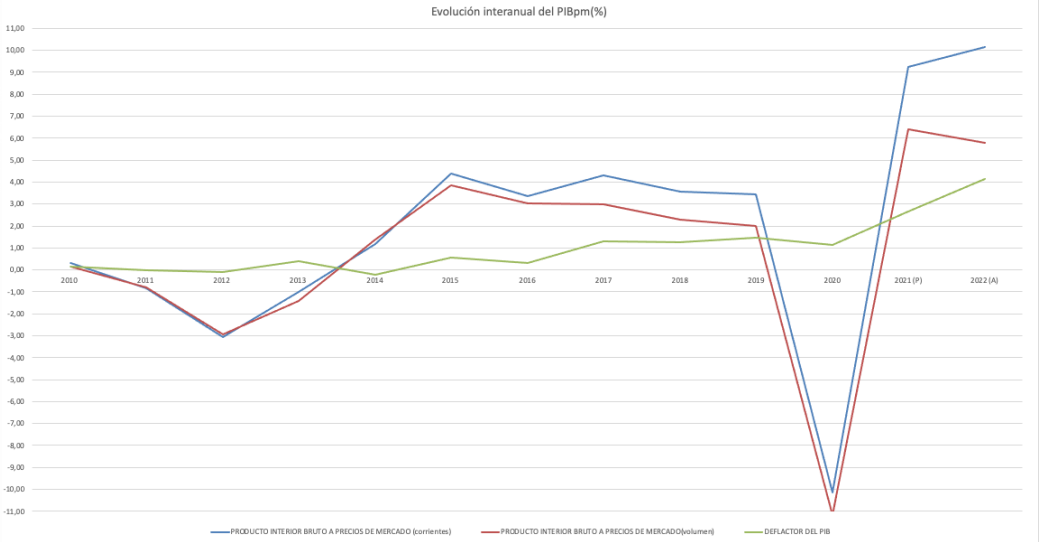


Figura 4.11: Cálculo de los valores del deflactor del PIB dadas las tasas de variación del PIB a precios corrientes y en términos de volumen desde 2010 a 2022 en España.

Con estas premisas podemos calcular el deflactor del PIB en la figura 4.11 con los datos del PIB a precios corrientes y en volumen. Por ejemplo:



En el año 2010 los valores de las tasas de variación del PIB a precios corrientes y en términos de volumen fueron de 0.32 % y de 0.16 % respectivamente y el cálculo que tenemos que hacer es:

$$\left(\frac{0{,}32}{100}+1\right)\cdot 100 = \frac{\text{Deflactor}}{100}\cdot\left(\frac{0{,}16}{100}+1\right)\cdot 100$$

Despejando obtenemos que el valor del deflactor en ese año es aproximadamente 100,15, por lo que la tasa de variación del deflactor, según la fórmula que citábamos antes será 0,15 %. Análogamente se podrían calcular el resto de valores para los años de los que disponemos de datos, resultando el gráfico anterior. Puede apreciarse que cuando la tasa de variación a precios corrientes del PIB es mayor que en términos de volumen esto se traduce en que el deflactor es positivo y si ocurre lo contrario, es decir, que la tasa en términos de volumen es mayor que a precios corrientes entonces la tasa de variación del deflactor es negativa, como ocurre por ejemplo en los años 2011, 2012 y 2014. Cuanta mayor sea la diferencia entre las tasas de variación mayor será el valor de la tasa de variación del deflactor, como ocurre en 2021 y 2022.

Volviendo al gráfico del apartado c) por el que se nos pregunta, y con la experiencia de haber analizado el caso anterior, podemos decir que la tasa de variación del deflactor implícito del PIB es positiva entre T-9 y T-7, cercana a cero y pasa a ser negativa hasta T-4 puesto que la tasa de variación en volumen supera a la tasa a precios corrientes del PIB entre T-7 y T-4. A partir de ese momento vuelve a ser positiva e ir creciendo en los tres últimos años puesto que la tasa en términos de volumen ya no está tan cercana a la tasa a precios corrientes que parece desmarcarse de esos valores.



Cuestión 6

Para el estudio del comportamiento medio de una determinada característica cuantitativa Y en una población, se propone aplicar un muestreo aleatorio estratificado. La población ha sido agrupada en tres estratos de acuerdo a su edad: menos de 20 años, de 20 a 49 años y de 50 o más años. La información de la que se dispone de los estratos es lasiguiente:

Estrato	Tramos de edad	N_h	$\overline{Y}_h$	S_h^2
1	De 0 a 20 años	150	20	36
2	De 20 a 49 años	250	35	144
3	De 50 o más años	200	60	256

Donde:

- N_h = Número de personas en el estrato h.
- $\overline{Y}_h$ = Valor medio de la característica en el estrato h.
- S_h^2 = Cuasivarianza poblacional de Y en el estrato h.

Se realiza un muestreo sin reemplazamiento y con probabilidades iguales en cada estrato. Se pide:

1. Obtener la cuasivarianza poblacional de Y. ¿Qué proporción supone la variación interestrato de la variación total?

2. Comentar a la vista de los resultados obtenidos en el apartado anterior, si la variable de estratificación es apropiada en términos de ganancia en precisión.

3. Calcular el tamaño muestral necesario, para que siendo una muestra autoponderada, se obtenga una varianza para el estimador de la media de 1,45. Asimismo, calcular los tamaños muestrales en los estratos.

Solución

Se cumple que

$$\sigma^2 = \sum_{h=1}^{L} W_h \left(\overline{Y}_h - \overline{Y}\right)^2 + \sum_{h=1}^{L} W_h \sigma_h^2$$

siendo:

$$\sigma_h^2 = \frac{1}{N_h} \sum_{k=1}^{N_h} \left(Y_{h_k} - \overline{Y}_h\right) = S_h^2 \frac{N_h - 1}{N_h}$$

$$S_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} \left(Y_{h_k} - \overline{Y}_h\right)$$

Donde,

$$\overline{Y} = \sum_h W_h \overline{Y}_h \qquad \overline{Y}_h = \frac{Y_h}{N_h} \qquad Y_h = \sum_{k=1}^{N_h} Y_{h_k} \qquad W_h = \frac{N_h}{N}$$

Por una parte,

$$N\sigma^2 = (N-1)\, S^2$$

$$\left(\frac{N-1}{N}\right) S^2 = \sum_{h=1}^{L} W_h \left(\overline{Y}_h - \overline{Y}\right)^2 + \sum_{h=1}^{L} W_h S_h^2 \left(\frac{N_h - 1}{N_h}\right)$$

$$S^2 = \sum_{h=1}^{L} \left(\frac{N_h}{N-1}\right) \left(\overline{Y}_h - \overline{Y}\right)^2 + \sum_{h=1}^{L} \frac{N_h}{N} \frac{N}{N-1} \frac{N_h - 1}{N_h} S_h^2$$

$$S^2 = \underbrace{\sum_{h=1}^{L} \left(\frac{N_h}{N-1}\right) \left(\overline{Y}_h - \overline{Y}\right)^2}_{\text{Variabilidad entre estratos}} + \underbrace{\sum_{h=1}^{L} \left(\frac{N_h - 1}{N-1}\right) S_h^2}_{\text{Variabilidad dentro de los estratos}}$$

Por otra parte,

$$\sigma^2 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} \left(Y_{h_i} - \overline{Y}\right)^2$$



$$N\sigma^2 = \sum_{h=1}^{L}\sum_{i=1}^{N_h}\left(Y_{h_i} - \overline{Y}\right)^2 = \sum_{h=1}^{L}\sum_{i=1}^{N_h}\left(\left(Y_{h_i} - \overline{Y}_h\right)\left(\overline{Y}_h - \overline{Y}\right)\right)^2 =$$
$$= \sum_{h=1}^{L}\sum_{i=1}^{N_h}\left(Y_{h_i} - \overline{Y}_h\right)^2 + \sum_{h=1}^{L}\sum_{i=1}^{N_h}\left(\overline{Y}_h - \overline{Y}\right)^2 + \underbrace{2\sum_{h=1}^{L}\sum_{i=1}^{N_h}\left(Y_{h_i} - \overline{Y}_h\right)\left(\overline{X}_h - \overline{Y}\right)}_{=0}$$
$$= \sum_{h=1}^{L} N_h\sigma_h^2 + \sum_{h=1}^{L} N_h\left(\overline{Y}_h - \overline{Y}\right)^2.$$

1. Para el cálculo de la cuasivarianza poblacional de Y, se necesita:

$$\overline{Y} = \sum_h W_h\overline{Y}_h = \frac{150\cdot 20 + 250\ \cdot 35 + 200\cdot 60}{150+250+600} = \frac{23750}{600} = 39,5833$$

Así,

$$S^2 = \sum_{h=1}^{L}\left(\frac{N_h}{N-1}\right)\left(\overline{Y}_h - \overline{Y}\right)^2 + \sum_{h=1}^{L}\left(\frac{N_h-1}{N-1}\right)S_h^2 =$$
$$= \frac{150}{599}\left(20-39,5833\right)^2 + \frac{250}{599}\left(35-39,5833\right)^2 + \frac{200}{599}\left(60-39,5833\right)^2 +$$
$$+ \left(\frac{149}{599}\right)36 + \left(\frac{249}{599}\right)144 + \left(\frac{199}{599}\right)256 =$$
$$= \underbrace{243,9816}_{\text{Variabilidad entre estratos}} + \underbrace{153,8631}_{\text{Variabilidad dentro de los estratos}} = 397,8447$$

Consecuentemente, la proporción que supone la variación interestrato de la variación total es el 61,32 %:

$$\frac{\text{Variabilidad entre los estratos}}{\text{Variabilidad total}} = \frac{243,9816}{397,8447} = 0,6132$$

2. La estratificación permite reducir la varianza (en comparación con el muestreo aleatorio simple) si la variable de estratificación está correlacionada con la variable de interés. El principio fundamental de la estratificación establece que la ganancia en precisión será mayor cuando las medias de los estratos difieran lo máximo posible y exista una fuerte homogeneidad dentro de cada estrato.

Así, para determinar si la estratificación es apropiada en términos de ganancia en precisión, evaluaremos la proporción de las variaciones entre los estratos y dentro de los estratos con respecto a la variación total. Cuanto más pequeño sea el segundo sumando,

$$S^2 = \sum_{h=1}^{L} \left(\frac{N_h}{N-1} \right) \left(\overline{Y}_h - \overline{Y} \right)^2 + \overbrace{\sum_{h=1}^{L} \left(\frac{N_h - 1}{N-1} \right) S_h^2}^{},$$

que mide la variablidad dentro de los estratos, es decir cuanta más homogeneidad haya dentro de los estratos, mayor será el primer sumando, $\sum_{h=1}^{L} \left(\frac{N_h}{N-1} \right) \left(\overline{Y}_h - \overline{Y} \right)^2$, que mide la heterogeneidad entre los estratos. Cuanto mayor sea la heterogeneidad entre estratos más aconsejable será el uso del muestreo estratificado, puesto que su precisión dependará de la varianza dentro de los estratos.

$$S^2 = \underbrace{\sum_{h=1}^{L} \left(\frac{N_h}{N-1} \right) \left(\overline{Y}_h - \overline{Y} \right)^2}_{\text{Variabilidad entre estratos (61,33 \%)}} + \underbrace{\sum_{h=1}^{L} \left(\frac{N_h - 1}{N-1} \right) S_h^2}_{\text{Variabilidad dentro de los estratos (38,67 \%)}} .$$

La heterogeneidad entre estratos está medida por el primer término y sí parece adecuada para estratificar, aunque parte de la variabilidad total está dentro de los estratos, como es lógico e inevitable. No obstante, sí se ganará en precisión.

3. Como la muestra es autoponderada, el número de elementos asignados a los diversos estratos es proporcional a la representación de los estratos de la población, es decir, el tamaño de la muestra extraída de cada estrato es proporcional al tamaño relativo de ese estrato de la población. La fracción de muestreo es aplicada a cada estrato, dando a cada elemento de la población la misma oportunidad de ser seleccionado. En definitiva, se va a utilizar la afijación porporcional, es decir, a cada estrato se asigna un número de unidades muestrales proporcionales al



tamaño: $n_h = N_h \cdot k$).

$$n = \sum_h n_h = \sum_h N_h \cdot k = N \cdot k \Rightarrow k = \frac{n}{N} = f$$

$$f_h = \frac{n_h}{N_h} = \frac{N_h \cdot k}{N_h} = f$$

$$W_h = \frac{N_h}{N} = \frac{n_h/k}{n/k} = \frac{n_h}{n}$$

luego,

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - f_h\right) \frac{S_h^2}{n_h} = \frac{1-k}{n} \sum_h W_h \cdot S_h^2$$

$$1,45 = \frac{1-k}{N \cdot k} \sum_h W_h \cdot S_h^2$$

$$600 \cdot 1,45 \cdot k = \sum_h W_h \cdot S_h^2 - k \sum_h W_h \cdot S_h^2$$

$$k \left(600 \cdot 1,45 + \sum_h W_h \cdot S_h^2 \right) = \sum_h W_h \cdot S_h^2$$

$$k = \frac{\sum_h W_h \cdot S_h^2}{\sum_h W_h \cdot S_h^2 + 870} = \frac{\left(\frac{150}{600}\right) 36 + \left(\frac{250}{600}\right) 144 + \left(\frac{200}{600}\right) 256}{\left(\frac{150}{600}\right) 36 + \left(\frac{250}{600}\right) 144 + \left(\frac{200}{600}\right) 256 + 870} =$$

$$= \frac{154,3}{154,3 + 870} = 0,150667$$

Por tanto,

$$n = k \cdot N = 0,150667 \cdot 600 = 90,40 \simeq 91.$$

Y siendo $n = 91$,

$$n_1 = 0,150667 \cdot 150 = 22,60 \simeq 23$$

$$n_2 = 0,150667 \cdot 250 = 37,6667 \simeq 38$$

$$n_3 = 0,150667 \cdot 200 = 30,133 \simeq 30$$



Cuestión 7

Considere que la esperanza condicional de una variable dependiente y es la siguiente:

$$E\left(y|x_1, x_2\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

donde las variables son continuas.

1. Calcule qué efecto esperado (efecto parcial) sobre y tiene un cambio de una unidad en la variable x_1, considerando que x_2 permanece constante.

2. Si dispusiera de una muestra de tamaño n, y quisiera estimar el siguiente modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + u_i, i = 1, \cdots, n,$$

donde los errores u_i no son observables, ¿cómo se deduciría y cuál sería la expresión del estimador de mínimos cuadrados ordinarios de $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$?

Solución

1. El efecto esperado sobre y que tiene un cambio de una unidad en la variable x_1, considerando que x_2 permanece constante, es

$$\frac{\partial E\left(y|x_1 x_2\right)}{\partial x_1} = \beta_1 + \beta_3 x_2.$$

Al variar x_1 en una unidad, y varía en media $\beta_1 + \beta_3 x_2$ unidades. Es decir, cuando hay un término de interacción, el efecto de la variable que forma la interacción depende del nivel de la otra variable que la compone. El término de la interacción en el modelo significa que la relación entre x_1 e y cambia dependiendo del valor de x_2 y viceversa. El modelo es más flexible, pero añade complejidad interpretativa.



2. El método de mínimos cuadrados ordinarios consiste en elegir los β de la ecuación anterior de tal modo que la suma de los cuadrados de lo errores, u_i, se minimice. La función de mínimos cuadrados es:

$$L = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \beta_3 x_{1i} x_{2i})^2$$

Esta función L debe minimizarse con respecto a $\beta_0, \beta_1, \beta_2, \beta_3$ de forma que los estimadores de mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ deben satisfacer:

$$\begin{cases} \left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\ldots,\hat{\beta}_3} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_3 x_{1i} x_{2i}\right) = 0 \\ \left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\ldots,\hat{\beta}_3} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_3 x_{1i} x_{2i}\right) x_{21} = 0 \\ \left.\frac{\partial L}{\partial \beta_2}\right|_{\hat{\beta}_0,\ldots,\hat{\beta}_3} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_3 x_{1i} x_{2i}\right) x_{12} = 0 \\ \left.\frac{\partial L}{\partial \beta_3}\right|_{\hat{\beta}_0,\ldots,\hat{\beta}_3} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_3 x_{1i} x_{2i}\right) x_{1i} x_{2i} \end{cases}$$

y simplificando quedaría:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_{1i} + \hat{\beta}_2 \sum_i x_{2i} + \hat{\beta}_3 \sum_i x_{1i} x_{2i} = \sum_i y_i \\ \hat{\beta}_0 \sum_i x_{1i} + \hat{\beta}_1 \sum_i x_{1i}^2 + \hat{\beta}_2 \sum_i x_{1i} x_{2i} + \hat{\beta}_3 \sum_i x_{1i}^2 x_{2i} = \sum_i x_{1i} y_i \\ \hat{\beta}_0 \sum_i x_{2i} + \hat{\beta}_1 \sum_i x_{1i} x_{2i} + \hat{\beta}_2 \sum_i x_{2i}^2 + \hat{\beta}_3 \sum_i x_{1i} x_{2i}^2 = \sum_i x_{2i} y_i \\ \hat{\beta}_0 \sum_i x_{1i} x_{2i} + \hat{\beta}_1 \sum_i x_{1i}^2 x_{2i} + \hat{\beta}_2 \sum_i x_{1i} x_{2i}^2 + \hat{\beta}_3 \sum_i x_{1i}^2 x_{2i}^2 = \sum_i x_{1i} x_{2i} y_i \end{cases}$$

Estas cuatro son las ecuaciones normales de mínimos cuadrados. La solución de este sistema son los estimadores de mínimos cuadrados de los coeficientes de regresión.

Expresando las ecuaciones iniciales en forma matricial:

$$y = X\beta + u$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & x_{11}x_{21} \\ 1 & x_{12} & x_{22} & x_{12}x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{1n}x_{2n} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_3 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$L = \sum_{i=1}^{n} u_i^2 = u^T u = (y - X\beta)^T (y - X\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta =$$
$$= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta,$$

donde $\beta^T X^T y$ es una matriz de dimensión 1×1, es decir, es un escalar, y su traspuesta $\left(\beta^T X^T y\right)^T = y^T X\beta$ es el mismo escalar. Los estimadores de MCO sastisfacen que:

$$\left.\frac{\partial L}{\partial \beta}\right|_{\hat{\beta}} = -2X^T y + 2X^T X\hat{\beta} = 0 \Leftrightarrow X^T X\hat{\beta} = X^T y,$$

luego

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y.$$



Cuestión 8

a) Elabore la sucesión de cuentas corrientes y de capital del sector Hogares de la página siguiente, que corresponde a una economía para un año T, teniendo en cuenta la información (en miles de millones EUR) para el total de la economía que aparece en esta tabla:

Formación bruta de capital	294
Consumo colectivo efectivo	110
Transferencias sociales en especie (TSE)	182
Remuneración de los asalariados por empleadores residentes	655
de la cual, a asalariados no residentes	12
Remuneración de los asalariados residentes por empeadores no residentes	15
Excedente de Explotación Bruto	450
Saldo de intercambios exteriores de bienes y servicios	-18
Impuestos menos subvenciones sobre la producción y las importaciones	133
Cotizaciones Sociales netas (total empleos)	198
Prestaciones Sociales distintas de las TSE (total recursos)	250
Saldo de la Cuenta de Explotación de la rama *servicios de alquiler de las viviendas ocupadas por sus propietarios*	80

Se considera que, para esta economía, las adquisiciones menos cesiones de activos no producidos son despreciables en el año T.

b) Responda a las siguientes cuestiones:

1) ¿Cuál sería la tasa de ahorro de los hogares en el año T? Interprete el resultado.
2) ¿Qué se podría concluir de su Capacidad / necesidad de financiación?
3) Si para dicha economía se realiza una nueva estimación del importe de los servicios de alquileres imputados de las viviendas ocupadas por sus propietarios para ese año, y el resultado supera en 5 miles de millones EUR a la actual, ¿qué saldos y operaciones se verían modificados a la luz de esta nueva información?

INSTITUTO NACIONAL DE ESTADISTICA



Cuentas corrientes

I. Cuenta de producción

Empleos	Código	Operaciones y otros flujos y saldos contables	Recursos
S.14 Hogares			S.14 Hogares
	P.1	Producción	365
	P.11	Producción de mercado	260
	P.12	Producción para uso final propio	105
98	P.2	Consumos intermedios	
	B.1b/	**Valor añadido bruto/**	
	P.51c	Consumo de capital fijo	
	B.1n/	***Valor añadido neto/***	

II. 1.1 Cuenta de explotación

	B.1b/B.1*b	Valor añadido bruto / Producto interior bruto	
	*B.1n/B.1*n*	*Valor añadido neto / Producto interior neto*	
40	D.1		
10	D.2		
-3	D.3		
	B.2b	**Excedente de explotación bruto**	
	B.3b	**Renta mixta bruta**	
	B.2n/B.3n	***Excedente de explotación neto / Renta mixta neta***	

II. 1.2 Cuenta de asignación de la renta primaria

	B.2b	Excedente de explotación bruto	
	B.3b	Renta mixta bruta	
	B.2n/B.3n	*Excedente de explotación neto / Renta mixta neta*	
	D.1	Remuneración de los asalariados	
10	D.4	Rentas de la propiedad	52
	B.5b/B.5*b	**Saldo de rentas primarias bruto**	
	B.5n/B.5*n	***Saldo de rentas primarias neto***	

II. 2 Cuenta de distribución secundaria de la renta

	B.5b/B.5*b	Saldo de rentas primarias bruto / Renta nacional bruta	
	*B.5n/B.5*n*	*Saldo de rentas primarias neto / Renta nacional neta*	
135	D.5	Impuestos corrientes sobre la renta, el patrimonio, etc.	
	D.61	Cotizaciones sociales netas	1
1	D.62	Prestaciones sociales distintas de las TSE	
96	D.7	Otras transferencias corrientes	84
	B.6b	**Renta disponible bruta**	
	B.6n	***Renta disponible neta***	

II. 4.1 Cuenta de utilización de la renta disponible

	B.6b	Renta disponible bruta	
	B.6n	*Renta disponible neta*	
	P.3	Gasto en consumo final	
	D.8	Ajuste por la variación de los derechos por pensiones	-2
	B.8b	**Ahorro bruto**	
	B.8n	***Ahorro neto***	

Cuenta de capital

III. 1.1 Cuenta de variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital

	B.8n	Ahorro neto	
	D.9r	Transferencias de capital, a cobrar	3
	D.9p	Transferencias de capital, a pagar	-5
	B.10.1	***Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital***	

III. 1.2 Cuenta de adquisiciones de activos no financieros

	B.10.1	*Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital*	
60	P.5g	Formación bruta de capital	
-40	P.51c	Consumo de capital fijo	
	B.9	**Capacidad(+)/Necesidad(-) de financiación**	



Solución

a) La elaboración de las cuentas corrientes y de capital del sector hogares las mostramos en las siguientes dos tablas que ilustran los datos del propio sector hogares que nos proporciona el problema en gris, los datos del cuadro del total de la economía que corresponden a los hogares y por tanto se pueden utilizar directamente, en amarillo, y por último los datos que se calculan específicamente para el sector, que se muestran en rojo.

Se podría avanzar en la sucesión de cuentas a través de la cuenta de producción, explotación, asignación de la renta primaria y distribución secundaria de la renta (ver figura 4.12 y figura 4.13), teniendo en cuenta algunas consideraciones, como por ejemplo que en la de explotación, el EBE de los hogares coincide con el saldo de la cuenta de explotación de la rama de servicios de alquiler de las viviendas ocupadas por sus propietarios, que es un dato del problema, y vale 80, y por eso lo ponemos en amarillo, puesto que es un dato del total de la economía que se puede usar directamente para el sector hogares. Otra apreciación que se utiliza en la cuenta de asignación de la renta primaria es que la remuneración de asalariados es la nacional y por lo tanto incluye la remuneración de los asalariados residentes por empleadores residentes y por empleadores no residentes, y por eso se opera como (655-12) + 15, es decir, 658.

En la cuenta de utilización de la renta disponible, es necesario conocer el gasto en consumo final de los hogares para poder calcular el ahorro bruto. Vamos a ver cómo podríamos calcular ese agregado. Para ello, vamos a utilizar las fórmulas del PIB por el método de la renta y de la



demanda, para, igualando, poder despejar ese gasto en consumo final deseado de los hogares.

En primer lugar sabemos que la remuneración de asalariados interior, que es la que debemos utilizar en la fórmula del PIB por el método de la renta se obtiene sumando la remuneración de asalariados pagada por residentes a residentes a la remuneración de asalariados pagada por residentes a no residentes, es decir, con los datos que tenemos sería 655. Luego:

$$PIB_{pm}(renta) = RA_{interior} + (D.2 - D.3) + (EBE + RMB) =$$
$$= 655 + 133 + 450 + RMB = 1238 + RMB$$

Pero la renta mixta bruta correspondiente al total de la economía coincide con la renta mixta bruta del sector hogares, que es el único sector que posee o del que se calcula este agregado. Ahora bien, como comentábamos antes, en la cuenta de explotación del sector hogares tenemos que el EBE+RMB vale 220, y que de ese valor, 80 corresponde al EBE. De esta forma, por diferencia, obtenemos que la renta mixta bruta del sector hogares vale 140, y por tanto:

$$PIB_{pm}(renta) = 1238 + 140 = 1378$$

Ahora conocido el valor del producto interior bruto, y teniendo en cuenta que el saldo de intercambios exteriores de bienes y servicios es -18, que es la diferencia entre exportaciones e importaciones en la economía,



podemos operar de la siguiente manera:

$$\begin{aligned} PIB_{pm}(demanda) &= GCF + FBC + X - M = \\ &= GCFE_{individual} + GCFE_{colectivo} + 294 - 18 = \\ &= GCF_{individualDeHogares} + TSE + GCFE_{colectivo} + 276 = \\ &= GCF_{individualDeHogares} + 182 + 110 + 276 = \\ &= GCF_{individualDeHogares} + 568 \end{aligned}$$

Como el $PIB_{pm}(demanda)$ debe coincidir con el obtenido por el método de la renta, que vale 1378, despejando obtenemos que:

$$GCF_{individualDeHogares} = 1378 - 568 = 810$$

que es el valor que ponemos en rojo en la cuenta de utilización de la renta disponible, necesario para poder calcular el ahorro bruto del sector.

Una vez calculado el ahorro se trabaja con las dos últimas cuentas sin problema alguno obteniendo la capacidad necesidad de financiación.



Hogares Unidad: millones de euros

Cuentas corrientes

I. Cuenta de producción

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	P.1	Producción	365
	P.11	Producción de mercado	260
	P.12	Producción para uso final propio	105
98	P.2	Consumos intermedios	
267	**B.1b/**	**Valor añadido bruto/**	
40	P.51c	Consumo de capital fijo	
227	***B.1n/***	***Valor añadido neto/***	

II. 1.1 Cuenta de explotación

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.1b/B.1*b	Valor añadido bruto / Producto interior bruto	267
	*B.1n/B.1*n*	*Valor añadido neto / Producto interior neto*	*227*
40	D.1	Remuneración de los asalariados	
10	D.2	Impuestos sobre la producción e importaciones	
-3	D.3	Subvenciones sobre la producción e importaciones	
220	***B.2b/B.3b***	**Excedente de explotación bruto + Renta mixta bruta**	
80	**B.2b**	**Excedente de explotación bruto**	
140	**B.3b**	**Renta mixta bruta**	
	B.2n/B.3n	***Excedente de explotación neto / Renta mixta neta***	

II. 1.2 Cuenta de asignación de la renta primaria

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.2b	Excedente de explotación bruto	80
	B.3b	Renta mixta bruta	140
	B.2n	*Excedente de explotación neto*	
	B.3n	*Renta mixta neta*	
	D.1	Remuneración de los asalariados (655-12+15)	658
10	D.4	Rentas de la propiedad	52
920	**B.5b/B.5*b**	**Saldo de rentas primarias bruto / Renta nacional bruta**	
880	***B.5n/B.5*n***	***Saldo de rentas primarias neto / Renta nacional neta***	

Figura 4.12: Sucesión de cuentas del sector hogares. Parte 1.



II. 2 Cuenta de distribución secundaria de la renta

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.5b/B.5*b	Saldo de rentas primarias bruto / Renta nacional bruta	920
	*B.5n/ B.5*n*	*Saldo de rentas primarias neto / Renta nacional neta*	
135	D.5	Impuestos corrientes sobre la renta, el patrimonio, etc.	
198	D.61	Cotizaciones sociales netas	1
1	D.62	Prestaciones sociales distintas de las transferencias sociales en especie	250
96	D.7	Otras transferencias corrientes	84
825	**B.6b**	**Renta disponible bruta**	
785	***B.6n***	***Renta disponible neta***	

II. 4.1 Cuenta de utilización de la renta disponible

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.6b	Renta disponible bruta	825
	B.6n	*Renta disponible neta*	
810	P.3	Gasto en consumo final	
	D.8	Ajuste por la variación de los derechos por pensiones	-2
13	**B.8b**	**Ahorro bruto**	
-27	***B.8n***	***Ahorro neto***	

Cuentas de acumulación

III. 1.1 Cuenta de variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital

Variaciones de los activos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Variaciones de pasivos S.14 Hogares
	B.8n	*Ahorro neto*	*-27*
	D.9r	Transferencias de capital, a cobrar	3
	D.9p	Transferencias de capital, a pagar	-5
-29	***B.10.1***	***Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital***	

III. 1.2 Cuenta de adquisiciones de activos no financieros

Variaciones de los activos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Variaciones de pasivos S.14 Hogares
	B.10.1	*Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital*	*-29*
60	P.5g	Formación bruta de capital	
-40	P.51c	Consumo de capital fijo	
-49	**B.9**	**Capacidad(+)/Necesidad(-) de financiación**	

	Estas celdas son datos del problema en la tabla de los hogares
	Estas celdas contienen datos del cuadro del total de la economía que corresponden a los hogares
	Estas celdas se han calculado con los datos de los hogares y el total de la economía

Figura 4.13: Sucesión de cuentas del sector hogares. Parte 2.

b.1) Para contestar a esta pregunta, una vez elaboradas las cuentas del sector hogares, se nos pregunta por la tasa de ahorro en el año T, por lo que fijándonos en la cuenta de utilización de la renta disponible vemos que dicho ahorro bruto es de 13, y también sabemos que su renta disponible bruta es 825, es decir, que la tasa de ahorro de los hogares podría estimarse en un 1.57 % de su renta disponible bruta.

b.2) En la cuenta de adquisiciones de activos no financieros puede verse que la capacidad/necesidad de financiación es -49, es decir, que registran una necesidad de financiación por valor de esa cantidad.

b.3) De una forma resumida, aunque se puede ver en las siguientes figuras 4.14 y 4.15, si ese valor aumenta de 80 a 85, entonces la producción para uso final propio en la que se engloba este concepto aumentaría en 5 y por tanto la producción también. El valor añadido de los hogares es mayor y el EBE+RMB de los hogares también aumenta en 5 unidades, pero ese aumento corresponde íntegramente a la subida del EBE de los hogares de 80 a 85, manteniéndose la RMB en el mismo valor de 140. El gasto en consumo final de los hogares también aumentaría en 5, puesto que contiene los servicios de alquiler de viviendas ocupadas por sus propietarios, por lo tanto esto implica que el PIB aumenta en 5, al igual que aumenta en los hogares el saldo de rentas primarias bruto y la renta disponible bruta. Siguiendo la sucesión puede comprobarse cómo el ahorro bruto se mantiene en 13, al igual que se mantiene el valor de la necesidad de financiación, que sigue siendo negativa.



Hogares — Unidad: millones de euros

Cuentas corrientes

I. Cuenta de producción

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	P.1	Producción	365
	P.11	Producción de mercado	260
	P.12	Producción para uso final propio	105
98	P.2	Consumos intermedios	
267	**B.1b/**	**Valor añadido bruto/**	
40	P.51c	Consumo de capital fijo	
227	***B.1n/***	***Valor añadido neto/***	

II. 1.1 Cuenta de explotación

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.1b/B.1*b	Valor añadido bruto / Producto interior bruto	267
	*B.1n/B.1*n*	*Valor añadido neto / Producto interior neto*	227
40	D.1	Remuneración de los asalariados	
10	D.2	Impuestos sobre la producción e importaciones	
-3	D.3	Subvenciones sobre la producción e importaciones	
220	***B.2b/B.3b***	**Excedente de explotación bruto + Renta mixta bruta**	
80	**B.2b**	**Excedente de explotación bruto**	
140	B.3b	Renta mixta bruta	
	B.2n/B.3n	***Excedente de explotación neto / Renta mixta neta***	

II. 1.2 Cuenta de asignación de la renta primaria

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.2b	Excedente de explotación bruto	80
	B.3b	Renta mixta bruta	140
	B.2n	*Excedente de explotación neto*	
	B.3n	*Renta mixta neta*	
	D.1	Remuneración de los asalariados (655-12+15)	658
10	D.4	Rentas de la propiedad	52
920	**B.5b/B.5*b**	**Saldo de rentas primarias bruto / Renta nacional bruta**	
880	***B.5n/B.5*n***	***Saldo de rentas primarias neto / Renta nacional neta***	

II. 2 Cuenta de distribución secundaria de la renta

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.5b/B.5*b	Saldo de rentas primarias bruto / Renta nacional bruta	920
	*B.5n/ B.5*n*	*Saldo de rentas primarias neto / Renta nacional neta*	
135	D.5	Impuestos corrientes sobre la renta, el patrimonio, etc.	
198	D.61	Cotizaciones sociales netas	1
1	D.62	Prestaciones sociales distintas de las transferencias sociales en especie	250
96	D.7	Otras transferencias corrientes	84
825	**B.6b**	**Renta disponible bruta**	
785	***B.6n***	***Renta disponible neta***	

Hogares — Unidad: millones de euros

Cuentas corrientes

I. Cuenta de producción

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	P.1	Producción	370
	P.11	Producción de mercado	260
	P.12	Producción para uso final propio (105 + 5)	110
98	P.2	Consumos intermedios	
272	**B.1b/**	**Valor añadido bruto/**	
40	P.51c	Consumo de capital fijo	
232	***B.1n/***	***Valor añadido neto/***	

II. 1.1 Cuenta de explotación

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.1b/B.1*b	Valor añadido bruto / Producto interior bruto	272
	*B.1n/B.1*n*	*Valor añadido neto / Producto interior neto*	232
40	D.1	Remuneración de los asalariados	
10	D.2	Impuestos sobre la producción e importaciones	
-3	D.3	Subvenciones sobre la producción e importaciones	
225	***B.2b/B.3b***	**Excedente de explotación bruto + Renta mixta bruta**	
85	**B.2b**	**Excedente de explotación bruto**	
140	B.3b	Renta mixta bruta	
	B.2n/B.3n	***Excedente de explotación neto / Renta mixta neta***	

II. 1.2 Cuenta de asignación de la renta primaria

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.2b	Excedente de explotación bruto	85
	B.3b	Renta mixta bruta	140
	B.2n	*Excedente de explotación neto*	
	B.3n	*Renta mixta neta*	
	D.1	Remuneración de los asalariados (655-12+15)	658
10	D.4	Rentas de la propiedad	52
925	**B.5b/B.5*b**	**Saldo de rentas primarias bruto / Renta nacional bruta**	
885	***B.5n/B.5*n***	***Saldo de rentas primarias neto / Renta nacional neta***	

II. 2 Cuenta de distribución secundaria de la renta

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.5b/B.5*b	Saldo de rentas primarias bruto / Renta nacional bruta	925
	*B.5n/ B.5*n*	*Saldo de rentas primarias neto / Renta nacional neta*	
135	D.5	Impuestos corrientes sobre la renta, el patrimonio, etc.	
198	D.61	Cotizaciones sociales netas	1
1	D.62	Prestaciones sociales distintas de las transferencias sociales en especie	250
96	D.7	Otras transferencias corrientes	84
830	**B.6b**	**Renta disponible bruta**	
790	***B.6n***	***Renta disponible neta***	

Figura 4.14: Comparativa de la sucesión de cuentas del sector hogares. Parte 1.

II. 4.1 Cuenta de utilización de la renta disponible

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.6b	Renta disponible bruta	825
	B.6n	*Renta disponible neta*	
810	P.3	Gasto en consumo final	
	D.8	Ajuste por la variación de los derechos por pensiones	-2
13	**B.8b**	**Ahorro bruto**	
-27	***B.8n***	***Ahorro neto***	

Cuentas de acumulación

III. 1.1 Cuenta de variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital

Var. de los activos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Var. Pasivos S.14 Hogares
	B.8n	*Ahorro neto*	-27
	D.9r	Transferencias de capital, a cobrar	3
	D.9p	Transferencias de capital, a pagar	-5
-29	***B.10.1***	***Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital***	

III. 1.2 Cuenta de adquisiciones de activos no financieros

Var. de los activos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Var. Pasivos S.14 Hogares
	B.10.1	*Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital*	-29
60	P.5g	Formación bruta de capital	
-40	P.51c	Consumo de capital fijo	
-49	**B.9**	**Capacidad(+)/Necesidad(-) de financiación**	

II. 4.1 Cuenta de utilización de la renta disponible

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	B.6b	Renta disponible bruta	830
	B.6n	*Renta disponible neta*	
815	P.3	Gasto en consumo final	
	D.8	Ajuste por la variación de los derechos por pensiones	-2
13	**B.8b**	**Ahorro bruto**	
-27	***B.8n***	***Ahorro neto***	

Cuentas de acumulación

III. 1.1 Cuenta de variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital

Var. de los activos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Var. Pasivos S.14 Hogares
	B.8n	*Ahorro neto*	-27
	D.9r	Transferencias de capital, a cobrar	3
	D.9p	Transferencias de capital, a pagar	-5
-29	***B.10.1***	***Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital***	

III. 1.2 Cuenta de adquisiciones de activos no financieros

Var. de los activos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Var. Pasivos S.14 Hogares
	B.10.1	*Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital*	-29
60	P.5g	Formación bruta de capital	
-40	P.51c	Consumo de capital fijo	
-49	**B.9**	**Capacidad(+)/Necesidad(-) de financiación**	

Figura 4.15: Comparativa de la sucesión de cuentas del sector hogares. Parte 2.

4.3. Soluciones del cuarto examen del Cuerpo Superior de Estadísticos del Estado de 2023. Especialidad ciencia de datos. Turno libre.

Cuestión 1. Sea un municipio con un total de N =8.000 alumnos distribuidos en $M = 16$ centros escolares de tamaños desiguales $(N_i, \forall i = 1 \ldots, 16)$. Se desea estimar la proporción P de estudiantes que practican algún deporte fuera del horario escolar para ajustar los planes deportivos municipales.

Para alcanzar estos objetivos, se utiliza un muestreo de conglomerados con submuestreo con probabilidades iguales y sin reposición en ambas etapas. En la primera etapa, se seleccionan $m = 4$ centros, y en la segunda etapa, se seleccionan los alumnos con una fracción de muestreo de $f_2 = 0,10$.

La información disponible sobre los centros muestreados es la siguiente:

Centro i	N_i	$\hat{P}_i$	$N_i\hat{P}_i$	$N_i\hat{P}_i(1-\hat{P}_i)$
1	700	0,4	280	168
2	500	0,3	150	105
3	400	0,5	200	100
4	900	0,6	540	216
Total	2.500	1,8	1.170	589

donde $\hat{P}_i$ es la estimación de la propoción de alumnos practicando algún deporte fuera del horario escolar en cada conglomerado muestreado.

Se pide:

a) Una estimación insesgada de la mencionada proporción y su error de muestreo.

b) Consideramos otro estimador de P, definido por el promedio de los porcentajes observados en los centros: $\hat{P}^* = \frac{1}{4}\sum_{i=1}^{4}\hat{P}_i$

1) Demuestre que: $Sesgo(\hat{P}^*) = -\frac{1}{N}\sum_{i=1}^{M} P_i(N_i - \frac{N}{M}) = -\frac{M}{N}Cov(N_i, P_i)$

2) Deduzca las condiciones para las que $\hat{P}^*$ sería un estimador insesgado de P.

c) Suponga que, para obtener el estimador $\hat{\theta}$ de cierta característica del alumnado, se han aplicado varios procedimientos (reponderación por falta de respuesta, calibración con variables auxiliares, ...). Por esta razón, se ha decidido estimar la varianza utilizando el método indirecto de Jackknife. Indique cómo proceder para calcular la varianza con este método en este contexto, proporcionando las expresiones necesarias para la estimación de la varianza.



Solución

(a) nota: Antes de resolver el problema y para mantener una coherencia en la notación con otros problemas de este y anteriores libros de problemas resueltos, adecuaremos la notación del problema a la que hemos utilizado en otras ocasiones.

Vamos a denotar el número de conglomerados (que son los centros escolares) de la población de estudio por $N = 16$, cada uno con tamaños diferentes que denotaremos por M_i, siendo el número de conglomerados de la muestra $n = 4$, y llamaremos al conjunto total de elementos de la población, los alumnos, $M = 8000$. A la fracción de muestreo en segunda etapa la representamos como $f_{2i} = 0{,}1$.

Sabemos que un estimador insesgado de la proporción de alumnos que practica algún deporte fuera del horario escolar es:

$$\hat{P} = \frac{\hat{A}}{M} = \frac{N \cdot \frac{\sum_{i=1}^{n} M_i \hat{P}_i}{n}}{M} = \frac{16 \frac{700 \cdot 0{,}4 + 500 \cdot 0{,}3 + 400 \cdot 0{,}5 + 900 \cdot 0{,}6}{4}}{8000} = \\ = \frac{280 + 150 + 200 + 540}{2000} = \frac{1170}{2000} = 0{,}585$$

Es decir, el 58,5 % de los alumnos hace algún deporte en horario no escolar. La expresión del estimador de la varianza de este estimador, teniendo en cuenta el tipo de muestreo, se puede escribir de la siguiente forma:

$$\hat{V}(\hat{P}) = \frac{1}{M^2} V(\hat{A}) = \left(\frac{N}{M}\right)^2 \frac{1 - f_1}{n} \frac{\sum_{i=1}^{n} \left(M_i \hat{P}_i - \frac{1}{n} \sum_{i=1}^{n} M_i \hat{P}_i\right)^2}{n - 1} + \\ + \frac{N}{nM^2} \sum_{i=1}^{n} M_i^2 \frac{(1 - f_{2i})}{m_i - 1} \hat{P}_i \hat{Q}_i$$

Con los datos que nos proporciona el problema en la tabla tenemos que

$\frac{\sum_{i=1}^{n} M_i \hat{P}_i}{n} = \frac{1170}{4} = 292{,}5$ y $f_1 = \frac{4}{16} = 0{,}25$, luego sustituyendo queda:

$$\hat{V}(\hat{P}) = \left(\frac{16}{8000}\right)^2 \frac{0{,}75}{4} \frac{(280-292{,}5)^2 + (150-292{,}5)^2 + (200-292{,}5)^2 + (540-292{,}5)^2)}{3} +$$
$$+ \frac{16}{4 \cdot 8000^2} \left(\frac{0{,}9}{69} \cdot 700 \cdot 168 + \frac{0{,}9}{49} \cdot 500 \cdot 105 + \frac{0{,}9}{39} \cdot 400 \cdot 100 + \frac{0{,}9}{89} \cdot 900 \cdot 216\right) =$$
$$= 0{,}022\,56 + 0{,}000\,336\,694 = 0{,}022\,896\,6$$

De esta forma, la estimación de su error de muestreo será la raíz de este valor, que es $\hat{\sigma}(\hat{P}) = 0{,}151\,316$

(b) Antes de calcular el sesgo del estimador, vamos a calcular su esperanza para posteriormente poder desarrollar la expresión del sesgo con mayor comodidad, haciendo uso de la variable aleatoria e_{ij}, que vale 1 si la pareja $(i, j) \in S$ está en la muestra y cero en el resto, y que cumple que

$$P(e_{ij} = 1) = P_i \frac{m_i}{M_i} = \frac{n}{N} \frac{m_i}{M_i}$$

ya que la especificación del problema dice que el muestreo es sin reposición y con probabilidades iguales en ambas etapas.

$$\mathbb{E}(\hat{P}^*) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{P}_i\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} \frac{A_{ij}}{m_i}\right) = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{A_{ij}}{m_i}\frac{n}{N}\frac{m_i}{M_i} =$$
$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{A_{ij}}{M_i} = \frac{1}{N}\sum_{i=1}^{N} \frac{A_i}{M_i} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

Podíamos haber calculado la esperanza del estimador de esta otra forma alternativa, operando para cada fase del muestreo, que quizá sea más ilustrativa:

$$\mathbb{E}[\hat{P}^*] = E_1 E_2\left[\frac{1}{n}\sum_{i=1}^{n} \hat{P}_i\right] = \frac{1}{n} E_1\left[\sum_{i=1}^{n} E_2[\hat{P}_i]\right] = \frac{1}{n} E_1\left[\sum_{i=1}^{n} P_i\right] = \frac{1}{n}\sum_{i=1}^{N} E_1[e_i] P_i$$

$$= \frac{1}{n}\frac{n}{N}\sum_{i=1}^{N} P_i = \frac{1}{N}\sum_{i=1}^{N} P_i$$



Ahora ya calculamos el sesgo, haciendo uso de que $\overline{P} = \frac{\sum_{i=1}^{N} P_i}{N}$ y que $\overline{M} = \frac{\sum_{i=1}^{N} M_i}{N}$:

$$\begin{aligned}
\text{Sesgo}(\hat{P}^*) = \mathbb{E}(\hat{P}^*) - P = \frac{1}{N}\sum_{i=1}^{N} P_i - P = \frac{1}{N}\sum_{i=1}^{N} P_i - \frac{A}{M} = \\
= \frac{1}{N}\sum_{i=1}^{N} P_i - \frac{1}{M}\sum_{i=1}^{N} M_i P_i = \frac{-1}{M}\left(\sum_{i=1}^{N} M_i P_i - \frac{\sum_{i=1}^{N} M_i \sum_{i=1}^{N} P_i}{N}\right) = \\
= \frac{-1}{M}\left(\sum_{i=1}^{N} M_i P_i - M\overline{P} - \overline{M}N\overline{P} + \overline{M}N\overline{P}\right) = \\
= \frac{-1}{M}\sum_{i=1}^{N}\left(M_i P_i - M_i\overline{P} - \overline{M}P_i + \overline{M}\overline{P}\right) = \\
= \frac{-N}{M}\frac{1}{N}\sum_{i=1}^{N}(M_i - \overline{M})(P_i - \overline{P}) = \frac{-N}{M}Cov(M_i, P_i)
\end{aligned}$$

De esta forma, las variabilidades en los tamaños de los conglomerados, es decir, la que exista entre los valores de M_i y las proporciones de realizar actividades deportivas fuera del horario lectivo afectan tanto al sesgo del estimador como a su valor esperado. Cuando la covarianza entre las variables M_i y P_i sea cero implicará que el sesgo del estimador es nulo y por tanto $\hat{P}^*$ será insesgado.

Si la correlación entre el tamaño del conglomerado y la proporción de hacer deporte fuera del horario escolar fuera muy alta el estimador tendrá un sesgo mayor, al depender de la covarianza, a tenor de la expresión obtenida, por tanto para que el sesgo no se dispare no debería haber relación entre el tamaño del conglomerado y la proporción de actividades deportivas realizadas en horario no escolar.

Por ejemplo, $\hat{P}^*$ puede subestimar P si los porcentajes altos son los de las escuelas grandes y los porcentajes bajos son los de las escuelas



pequeñas. El sesgo es nulo si la covarianza es nula, lo cual es particularmente el caso si los tamaños y/o las proporciones son similares de una escuela a otra.

(c) El método jackknife se debe a Quenouille(1949) y apareció como un método para reducir el sesgo de un estimador. Tukey(1958) demostró que el jackknife se podía utilizar para construir estimadores de la varianza. En el muestreo multietápico, este método se aplica generalmente a las unidades de muestreo en la primera etapa. La forma más sencilla de aplicar esta técnica consiste en generar una submuestra mediante la eliminación de una de estas unidades en la muestra original. Así la submuestra o muestra jackknife está formada por todas las unidades de primera etapa de la muestra original menos una (Método delete-one jackknife). Este método es aplicado en varias de las encuestas realizadas por el INE. Valliant(1993) demostró que los pesos posestratificados deben ser recalculados en cada submuestra a fin de obtener un estimador jackknife consistente en muestreo bietápico.

La explicación del método comienza partiendo la muestra s en A grupos aleatorios independientes de tamaño $m = \frac{n}{A}$. Podemos asumir que para cualquier s cada grupo es una muestra aleatoria simple sin reemplazamiento. Para cada grupo $a = 1, \ldots, A$, vamos a calcular un estimador de θ, que denotaremos por $\hat{\theta}_{(a)}$ que tiene la misma forma funcional que $\hat{\theta}$, pero calculado con los datos de la muestra una vez eliminado el grupo a. Para cada $a = 1, \ldots, A$, definimos

$$\hat{\theta}_a = A\hat{\theta} - (A-1)\hat{\theta}_{(a)}$$

que se conoce como pseudovalor a-ésimo.



El estimador *jackknife* de θ se calcula como:

$$\hat{\theta}_{JK} = \frac{1}{A}\sum_{i=1}^{A}\hat{\theta}_a$$

El estimador de la varianza *jackknife* se puede calcular con la expresión:

$$\hat{V}_{JK1} = \frac{1}{A(A-1)}\sum_{i=1}^{A}(\hat{\theta}_a - \hat{\theta}_{JK})^2$$

que se utiliza tanto para estimar $\mathbb{V}(\hat{\theta})$ como para estimar $\mathbb{V}(\hat{\theta}_{JK})$.

Si los $\hat{\theta}_a (a = 1, \ldots, A)$ fuesen variables aleatorias incorreladas con la misma esperanza puede demostrarse que $\hat{V}_{JK1}$ sería insesgado para $\mathbb{V}(\hat{\theta}_{JK})$. Sin embargo, en general están correlados y por eso no se verifica la insesgadez. No parece haber resultados concretos sobre las propiedades del estimador de la varianza jackknife cuando $\hat{\theta}$ es más complejo que el estimador de Horvitz-Thomson.

En muestreo multietápico, como es el caso de este ejercicio, este método se aplica generalmente a las unidades de primera etapa(PSU). Supongamos una muestra en primera etapa s_I que contiene n_I unidades primarias y que se ha seleccionado a partir del conjunto U_I compuesto por N_I unidades. Si dividimos s_I aleatoriamente en A grupos, con m elementos en cada grupo podemos denotar por $\hat{\theta}$ como un estimador de θ calculado con la muestra completa, y $\hat{\theta}_{(a)}$ como el estimador de θ que está basado en los datos que quedan tras eliminar el a-ésimo grupo de unidades primarias. Utilizando las expresiones anteriores podemos calcular la estimación de la varianza por el método descrito.

Para este método jackknife necesitamos fijar un número adecuado de grupos A. Para tener una buena acuracidad en los resultados del estimador de la varianza, lo ideal sería tener tantos grupos como sea

posible, lo que significa $A = n$, es decir, $m = 1$. Por otro lado, por motivos computacionales, lo ideal sería tener tan pocos grupos como fuese posible siendo el caso más extremo $A = 2$, y $m = 2$. En la práctica, la elección más frecuente parece ser $A = n$, o, cuando el coste computacional es una cuestión seria, un compromiso entre los dos extremos $A = n$ y $A = 2$.

Cuestión 2. Para la obtención de ciertas estimaciones en una población de tamaño N, se utiliza un muestreo bifásico. Responda a las cuestiones planteadas para cada caso:

a) Para estimar la media de una característica $\bar{Y}$, primero se seleccionan 2.000 elementos para determinar su estrato, mediante un muestreo aleatorio simple sin reemplazamiento. De éstas, corresponden al primer estrato 1.300 y 700 al segundo. En una segunda fase, se lleva a cabo un muestreo aleatorio simple estratificado, seleccionando 150 elementos del primer estrato y 100 en el segundo, a quienes se les solicita medir la característica de interés. Los resultados se muestran en el siguiente cuadro:

Estrato h	$\bar{y}_h$	s^2_{yh}
1	5,4	3,75
2	9,2	6,9

donde $\bar{y}_h$ representa la media muestral en el estrato h de la segunda fase y s^2_{yh} es la varianza muestral del estrato h en la segunda fase. Además, se supone que el tamaño poblacional N es mucho mayor que el tamaño muestral de primera fase.

1) Calcule una estimación insesgada de la media $\bar{Y}$.
2) Estime su varianza.

b) Se desea estimar el total de una característica Y en una población, donde las probabilidades de selección de las unidades son proporcionales a una medida de su tamaño, denotada por M_i. Sin embargo, dado que no se conocen a priori los tamaños de las unidades de la población, se extrae en una primera fase una muestra de tamaño $n^{'}$, con probabilidades iguales y sin reposición, para obtener información acerca de los tamaños $M_1, \ldots, M_i, \ldots, M_{n'}$, siendo $M^{'} = \sum_i^{n^{'}} M_i$. En la segunda fase se obtiene una submuestra de tamaño $n < n^{'}$, con probabilidades proporcionales al tamaño y con reposición. Se pide:

1) Demuestre que $\hat{Y} = \sum_{i=1}^{n} \frac{N}{n} \frac{M^{'}}{n^{'}} \frac{X_i}{M_i}$ es un estimador insesgado del total poblacional Y.
2) Obtenga de forma razonada la varianza de dicho estimador $V(\hat{Y})$.

c) ¿Sería posible utilizar el muestreo bifásico para estimar la media de una característica, $\bar{Y}$, mediante el uso de un estimador de razón, utilizando la información de una variable auxiliar X disponible para todas las unidades muestreadas en primera fase? Exponga de manera concisa cómo se llevaría a cabo cada fase para obtener el estimador de razón.



Solución

(a) La primera muestra $n' = 2000$ se usa para estratificar sus unidades, atendiendo a una o varias características que observamos, así como para estimar la proporción de unidades de la población pertenecientes al estrato h.

Tenemos que $n_1^{'} = 1300$ y $n_2^{'} = 700$ son el número de unidades muestrales de cada estrato y denotamos por $\widehat{W}_h = \frac{n_h^{'}}{n^{'}}$ las proporciones estimadas en cada estrato, que en nuestro caso valen $\widehat{W}_1 = \frac{13}{20}$ y $\widehat{W}_2 = \frac{7}{20}$.

La segunda fase consiste en tomar una submuestra aleatoria de tamaño $n_h \leq n_h^{'}$ en cada estrato independiente.

En muestreo estratificado, el estimador de la media es: $\widehat{\overline{Y}} = \sum_{h=1}^{L} W_h \overline{y}_h$, con $W_h = \frac{N_h}{N}$, pero en muestreo doble los W_h se estiman por $\widehat{W}_h$ obtenidos de la primera muestra, y con la segunda se estiman las medias tomando $\overline{y}_h = \frac{y_h}{n_h}$, quedando el estimador de la forma

$$\widehat{\overline{Y}} = \sum_{h=1}^{L} \widehat{W}_h \overline{y}_h \quad , \quad \widehat{W}_h = \frac{n_h^{'}}{n^{'}}$$

Se puede demostar que el estimador $\widehat{\overline{Y}}$ es insesgado fácilmente, y en nuestro caso valdría:

$$\widehat{\overline{Y}} = \sum_{h=1}^{2} \widehat{W}_h \overline{y}_h = \frac{13}{20} \cdot 5{,}4 + \frac{7}{20} \cdot 9{,}2 = 3{,}51 + 3{,}22 = 6{,}73$$

La varianza de este estimador es conocida y su expresión es la siguiente:

$$V\left(\widehat{\overline{Y}}\right) = \sum_h (1 - f_h) \frac{S_h^2}{n_h} \left(W_h^2 + \frac{g^{'} W_h (1 - W_h)}{n^{'}} \right) + \frac{g^{'}}{n^{'}} \sum_h W_h \left(\overline{Y}_h - \overline{Y} \right)^2$$

siendo $f_{2h} = \frac{n_h}{n_h^{'}}$ la fracción de muestreo en la segunda fase y $g^{'} = \frac{N - n^{'}}{N - 1}$.

Una expresión de la estimación de la varianza puede aproximarse (ver Särndal, Swensson, and Wretman (1992)), asumiendo que N es bastante

más grande que el tamaño muestral de primera fase, es decir, que $n^{'}$ y reemplazando $\frac{n_h^{'}-1}{n^{'}-1}$ por $\frac{n_h^{'}}{n^{'}}$, puesto que al ser las muestras de segunda fase grandes esos cocientes son muy parecidos, por:

$$\hat{V}\left(\widehat{\overline{Y}}\right) = \sum_h \left(\frac{n_h^{'}}{n^{'}}\right)^2 \frac{\hat{S}_{2h}^2}{n_h} + \frac{1}{n^{'}} \sum_h \frac{n_h^{'}}{n^{'}} \left(\overline{y}_h - \widehat{\overline{Y}}\right)^2$$

siendo $\hat{S}_{2h}^2$ la varianza de la variable de estudio en la submuestra(muestra en la segunda fase) elegida en el estrato h-ésimo.

Utilizando los datos del problema vamos a calcular esta expresión:

$$\begin{aligned}\hat{V}\left(\widehat{\overline{Y}}\right) &= \left(\frac{1300}{2000}\right)^2 \frac{3{,}75}{150} + \left(\frac{700}{2000}\right)^2 \frac{6{,}9}{100} + \\ &+ \frac{1}{2000}\left[\frac{1300}{2000}(5{,}4-6{,}73)^2 + \frac{700}{2000}(9{,}2-6{,}73)^2\right] = \\ &= 0{,}019\,015 + 0{,}000\,747\,3 = 0{,}019\,762\end{aligned}$$

El error de muestreo será la raíz de este valor que acabamos de mostrar, es decir,

$$em = \sqrt{\hat{V}\left(\widehat{\overline{Y}}\right)} = \sqrt{0{,}019\,762} = 0{,}140\,577$$

.

(b) El estimador usual del total de una característica Y con probabilidades de selección de las unidades proporcionales a una medida de su tamaño, sea M_i, viene dado por:

$$\widehat{Y} = \sum_{i=1}^{n} \frac{Y_i}{nP_i} \quad \text{, con } P_i = \frac{M_i}{M}$$

Pero al no conocerse a priori los tamaños de las unidades de la población, con la muestra de tamaño $n^{'}$ con probabilidades iguales obtenemos información de los tamaños $M_1, \ldots, M_{n'}$, con $M^{'} = \sum_{i=1}^{n^{'}} M_i$. Con la submuestra de tamaño $n < n^{'}$, basándonos en que $\frac{M_i}{\frac{N}{n^{'}}M^{'}}$ es un estimador



de $P_i = \frac{M_i}{M}$, siendo N = tamaño de la población, el estimador

$$\widehat{Y} = \sum_{i=1}^{n} \frac{N}{n} \frac{M^{'}}{n^{'}} \frac{Y_i}{M_i} = \frac{NM^{'}}{nn^{'}} \sum_{i=1}^{n} \frac{Y_i}{M_i}$$

es insesgado y es lo que vamos a comprobar a continuación.

Denotamos por E_2 a la esperanza cuando la primera muestra fija, con probabilidades proporcionales al tamaño, a $y^{'}$ al total de la muestra de la primera fase, tomada con probabilidades iguales y a $P_i = \frac{M_i}{M^{'}}$ las probabilidades en segunda fase, que son proporcionales al tamaño. Así, tenemos:

$$\begin{aligned}
\mathbb{E}\left(\widehat{Y}\right) &= E_1\left(\frac{N}{n^{'}} E_2\left(\frac{M^{'}}{n} \sum_{i=1}^{n}\left(\frac{Y_i}{M_i}\right)\right)\right) = \\
&= E_1\left(\frac{N}{n^{'}} E_2\left(\sum_{i=1}^{n} \frac{Y_i}{n\frac{M_i}{M^{'}}}\right)\right) = E_1\left(\frac{N}{n^{'}} E_2\left(\sum_{i=1}^{n} \frac{Y_i}{nP_i}\right)\right) = E_1\left(\frac{N}{n^{'}} y^{'}\right) = \\
&= NE_1\left(\frac{y^{'}}{n^{'}}\right) = N \cdot \overline{Y} = Y
\end{aligned}$$

utilizando para la segunda fase el resultado conocido de que en muestreo con probabilidades proporcionales al tamaño se cumple que $\hat{Y} = \sum_{i=1}^{n} \frac{Y_i}{nP_i}$ es un estimador insesgado del total poblacional, siendo P_i la probabilidad de seleccionar la unidad i-ésima.

De esta forma hemos demostrado la insesgadez y ahora vamos a demostrar que la expresión de la varianza de este estimador es:

$$\mathbb{V}\left(\widehat{Y}\right) = \frac{N(N-n^{'})}{n^{'}} S^2 + \frac{N}{N-1} \frac{(n^{'}-1)}{n \cdot n^{'}} \sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y\right)^2$$

En este caso también vamos a hacer uso de la expresión conocida de la varianza del estimador del total poblacional en muestreo con probabilidades proporcionales al tamaño, es decir,

$$\mathbb{V}\left(\widehat{Y}\right) = \frac{1}{n} \sum_{i=1}^{n} P_i \left(\frac{Y_i}{P_i} - Y\right)^2$$

de manera que para el estimador de la media poblacional tenemos:

$$\mathbb{V}\left(\widehat{\overline{Y}}\right) = \frac{1}{N^2}\frac{1}{n}\sum_{i=1}^{n} P_i\left(\frac{Y_i}{P_i} - Y\right)^2$$

Vamos a desarrollar la expresión de la varianza para obtener la expresión que comentábamos:

$$\begin{aligned}\mathbb{V}\left(\widehat{Y}\right) &= V_1E_2\left(\widehat{Y}|_{n'}\right) + E_1V_2\left(\widehat{Y}|_{n'}\right) = V_1\left(N\frac{y'}{n'}\right) + E_1V_2\left(\widehat{Y}|_{n'}\right) = \\ &= N^2V_1\left(\overline{y'}\right) + E_1V_2\left(\widehat{Y}|_{n'}\right) = N^2\left(\frac{1}{n'} - \frac{1}{N}\right)S^2 + E_1V_2\left(\widehat{Y}|_{n'}\right) = \\ &= \frac{N(N-n')}{n'}S^2 + E_1V_2\left(\widehat{Y}|_{n'}\right)\end{aligned}$$

Desarrollamos el segundo sumando para comprobar que el resultado es el descrito. Llamando a $Y' = \sum_{i=1}^{n'} Y_i$ tenemos:

$$\begin{aligned}E_1V_2\left(\widehat{Y}|_{n'}\right) &= E_1V_2\left(N\widehat{\overline{Y}}|_{n'}\right) = E_1\left(N^2\cdot V_2\left(\widehat{\overline{Y}}|_{n'}\right)\right) = \\ &= E_1\left(N^2\frac{1}{n'^2}\frac{1}{n}\sum_{i=1}^{n'} P_i\left(\frac{Y_i}{P_i} - Y'\right)^2\right) \overset{*_1}{=} \\ &= E_1\left(N^2\frac{1}{n'^2}\frac{1}{n}\sum_{i=1}^{n'}\sum_{j>i}^{n'} M_iM_j\left(\frac{Y_i}{M_i} - \frac{Y_j}{M_j}\right)^2\right) = \\ &= N^2E_1\left(\frac{1}{n'^2\cdot n}\sum_{i=1}^{n'}\sum_{j>i}^{n'} M_iM_j\left(\frac{Y_i}{M_i} - \frac{Y_j}{M_j}\right)^2\right) \overset{*_2}{=} \\ &= N^2\frac{1}{n'^2\cdot n}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{n'(n'-1)}{N(N-1)}M_iM_j\left(\frac{Y_i}{M_i} - \frac{Y_j}{M_j}\right)^2 = \\ &= N^2\frac{1}{n'^2\cdot n}\frac{n'(n'-1)}{N(N-1)}\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{M_i}{M}\frac{M_j}{M}\left(\frac{Y_i}{M_i} - \frac{Y_j}{M_j}\right)^2 = \\ &= \frac{N}{N-1}\frac{(n'-1)}{n\cdot n'}\sum_{i=1}^{N}\sum_{j>i}^{N} P_iP_j\left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2 = \\ &= \frac{N}{N-1}\frac{(n'-1)}{n\cdot n'}\sum_{i=1}^{N} P_i\left(\frac{Y_i}{P_i} - Y\right)^2\end{aligned}$$

De esta forma hemos comprobado el segundo sumando de la varianza y probado la fórmula que poníamos anteriormente.



Se cumple que si $n^{'}$ es grande, así como el tamaño poblacional, que según el enunciado es mucho mayor que el tamaño de muestra en primera fase ($N \gg n^{'}$), la expresión de la varianza se puede aproximar, simplificando el segundo sumando, por

$$\mathbb{V}\left(\widehat{Y}\right) \cong \frac{N(N-n^{'})}{n^{'}} S^2 + \frac{1}{n}\sum_{i=1}^{N} P_i \left(\frac{Y_i}{P_i} - Y\right)^2$$

Solo nos falta aclarar dos de los pasos del desarrollo de $E_1 V_2\left(\widehat{Y}|_{n'}\right)$. Veamos el primero de ellos.

nota $*_1$:

Conocemos que $P_i = \frac{M_i}{M'}$, $Y^{'} = \sum_{i=1}^{n^{'}} Y_i$, $M^{'} = \sum_{i=1}^{n^{'}} M_i$ y que $\sum_{i=1}^{n^{'}} P_i = 1$, por lo que

$$\begin{aligned}
\sum_{i=1}^{n^{'}} P_i \left(\frac{Y_i}{P_i} - Y^{'}\right)^2 &= \sum_{i=1}^{n^{'}} P_i \left(\frac{Y_i^2}{P_i^2} - 2\frac{Y_i Y^{'}}{P_i} + Y^{'2}\right) = \\
&= \sum_{i=1}^{n^{'}} \frac{Y_i^2}{P_i^2} P_i - 2Y^{'} \sum_{i=1}^{n^{'}} \frac{Y_i}{P_i} P_i + Y^{'} \sum_{i=1}^{n^{'}} P_i = \\
&= \sum_{i=1}^{n^{'}} \frac{Y_i^2}{P_i} - 2Y^{'2} + Y^{'2} = \sum_{i=1}^{n^{'}} \frac{Y_i^2}{P_i} - Y^{'2} = \sum_{i=1}^{n^{'}} \frac{Y_i^2}{P_i} - \left(\sum_{i=1}^{n^{'}} Y_i\right)^2 = \\
&= \sum_{i=1}^{n^{'}} \frac{Y_i^2}{P_i} - \sum_{i=1}^{n^{'}} Y_i^2 - 2\sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} Y_i Y_j = \sum_{i=1}^{n^{'}} \frac{Y_i^2}{P_i}(1-P_i) - 2\sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} Y_i Y_j = \\
&= \sum_{i=1}^{n^{'}} \frac{Y_i^2}{P_i} \left(\sum_{j\neq i}^{n^{'}} P_j\right) - 2\sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} Y_i Y_j = \sum_{i=1}^{n^{'}} \sum_{j\neq i}^{n^{'}} \frac{Y_i^2}{P_i} P_j - 2\sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} Y_i Y_j = \\
&= \sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} \left(\frac{Y_i^2}{P_i} P_j + \frac{Y_j^2}{P_j} P_i\right) - 2\sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} Y_i Y_j = \\
&= \sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} \left(\frac{Y_i^2}{P_i} P_j + \frac{Y_j^2}{P_j} P_i\right) - 2\sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} Y_i Y_j = \\
&= \sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2 P_i P_j = \sum_{i=1}^{n^{'}} \sum_{j>i}^{n^{'}} \left(\frac{Y_i}{M_i} - \frac{Y_j}{M_j}\right)^2 M_i M_j
\end{aligned}$$

nota $*_2$:

En este paso hacemos uso de la variable aleatoria e_{ij} para extender el sumatorio a todos los elementos poblacionales. Esta variable toma el valor 1 si el par $(i,j) \in S$, es decir, si está en la muestra extraída en las dos fases, y cero en caso contrario. Dadas las características de las extracciones se cumple que

$$P(e_{ij}=1) = \frac{n^{'}(n^{'}-1)}{N(N-1)}$$

al ser un muestreo aleatorio simple sin reemplazamiento en primera fase.

(c) La respuesta a este apartado es afirmativa, puesto que el muestreo doble utiliza la primera muestra de tamaño $n^{'}$ para obtener una buena estimación de $\overline{X}$ al estar la información de esta variable auxiliar disponible para todas las unidades muestreadas en primera fase, y la segunda muestra de tamaño n la utiliza para estimar $\overline{x}$ e $\overline{y}$, conformando el estimador de razón para la media en muestreo doble como:

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}}\overline{x^{'}} \ , \ \overline{x^{'}} = \text{Media de la primera muestra}$$

En el muestreo doble se podrían considerar dos formas de extraer la segunda muestra. Una primera alternativa sería que la segunda muestra fuese una muestra aleatoria de la población, seleccionada independientemente de la primer, y una segunda alternativa sería que la segunda muestra fuera una submuestra aleatoria de la primera.

Se tiene que en cualquier caso, la esperanza del estimador es:

$$\mathbb{E}\left(\widehat{\overline{Y}}_R\right) = \overline{X}\mathbb{E}(\hat{R})$$

que será insesgado si lo es $\hat{R} = \frac{\overline{y}}{\overline{x}}$.



Las dos alternativas comentadas afectan al cálculo de la varianza resultando expresiones diferentes, que no son objeto de este ejercicio y que pueden consultarse en cualquier referencia bibliográfica de muestreo.

Ejercicio 3

Las siguientes estimaciones se calcularon a partir de una muestra de 7.634 mujeres encuestadas en una Encuesta General de Hogares. La variable dependiente toma el valor 1 si la mujer tiene un empleo remunerado, y 0 en caso contrario.

	MCO	Logit	Probit
ALTO	0.093	0.423	0.259
	(0.015)	(0.071)	(0.043)
NOCUAL	-0.210	-0.898	-0.554
	(0.013)	(0.056)	(0.035)
EDAD	0.038	0.173	0.107
	(0.003)	(0.124)	(0.008)
EDAD2	-0.051	-0.230	-0.142
	(0.003)	(0.069)	(0.009)
CAS	0.024	0.103	0.063
	(0.009)	(0.057)	(0.035)
Constante	-0.068	-2.587	-1.593
	(0.049)	(0.225)	(0.137)

Donde ALTO es uno si el encuestado tiene una cualificación educativa superior, cero en caso contrario; NOCUAL es uno si el encuestado no tiene cualificaciones, cero en caso contrario; EDAD es la edad en años; EDAD2 es

(edad x edad)/100; CAS es uno si está casado, cero en caso contrario. Los errores estándar calculados de forma convencional se indican entre paréntesis para los resultados de mínimos cuadrados ordinarios (MCO) y los errores estándar asintóticos entre paréntesis en los demás casos.

a) En el contexto de los modelos GLM, indique qué papel desempeña la función de enlace y cuáles son las funciones de enlace teóricas para cada uno de los modelos estimados, es decir, para MCO, Logit, Probit.

b) Explique brevemente cómo se calculan las estimaciones Probit cuando el modelo no tiene intercepto y solo tiene una variable explicativa. Indique qué propiedades generales tienen las correspondientes estimaciones.

c) Para los tres conjuntos de estimaciones, pruebe la hipótesis nula de que el coeficiente de CAS es cero. ¿Qué estadístico de contraste consideraría más fiable? Explíquelo.

d) Utilizando las estimaciones MCO y Probit, calcule las probabilidades estimadas de estar desempleada para una mujer casada de 40 años con estudios superiores.

e) Compruebe la hipótesis de significación conjunta del modelo probit, sabiendo que:

$$\ln L_R = -416,01$$
$$\ln L_{NR} = -321,25$$

donde $\ln L_R$ y $\ln L_{NR}$ son los logaritmos de la verosimilitud de los modelos probit restringido y no restringido, respectivamente. Utilice un nivel de significatividad del 5 por ciento.



Solución

a) El uso de los modelos GLM amplía el modelo lineal general al relacionar la variable dependiente con los factores y las covariables mediante una función de enlace g. Esta función conecta el valor esperado de la variable de respuesta $\mu = \mathbb{E}[Y]$, con la predicción lineal $\eta = X\beta$, donde X son las variables independientes del modelo y β los respectivos coeficientes. De esta forma:

$$g(\mu) = \eta \Leftrightarrow \mu = g^{-1}(\eta)$$

siendo g la función de enlace que además es una función invertible.

El rol principal de la función g es permitir que la relación entre la media de la variable dependiente μ y la combinación lineal de distribuciones η sea lineal, incluso si la naturaleza de la variable dependiente Y no sigue una distribución normal o no tiene un comportamiento lineal.

Por ejemplo, si se están modelando probabilidades (como en una regresión logística), estas están restringidas al intervalo $[0, 1]$ y una combinación lineal de predictores ($\eta = X\beta$) puede tomar valores fuera de este rango. La función de enlace g transforma la media de la variable dependiente ($\mu = \mathbb{E}[Y]$) para que pueda relacionarse linealmente con η. Así, la función g asegura que los valores predichos respeten las propiedades de Y y exiende el modelo lineal a un marco más general.

Las funciones de enlace teóricas para los modelos de Mínimos Cuadrados Ordinarios (MCO), Logit y Probit, que son casos específicos dentro del marco de los Modelos Lieales Generalizados (GLM), son las siguientes:

- **Mínimos Cuadrados Ordinarios (MCO)**. El modelo es de la

forma:

$$Y = \beta_0 + \beta_1 ALTO + \beta_2 NOCUAL + \beta_3 EDAD + \beta_4 EDAD2 + \beta_5 CAS + \varepsilon$$

Se utiliza en regresión lineal para modelar variables de respuesta continua, la variable de respuesta (Y) se supone que sigue una distribución normal. La función de enlace teórica es:

$$g(\mu) = \mu$$

es decir, la función identidad. Esto significa que no se aplica ninguna transformación a la media esperada $(\mu = \mathbb{E}[Y])$. La relación entre la variable de respuesta (Y) los predictores (X) es directamente lineal

$$\mu = \eta = X\beta$$

- **Logit**. El modelo es de la forma:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 ALTO + \beta_2 NOCUAL + \beta_3 EDAD + \beta_4 EDAD2 + \beta_5 CAS)}}$$

Equivalente a

$$\ln\left(\frac{P(Y=1|X)}{1 - P(Y=1|X)}\right) = \beta_0 + \beta_1 ALTO + \beta_2 NOCUAL + \beta_3 EDAD + \\ + \beta_4 EDAD2 + \beta_5 CAS$$

Se utiliza en regresión logística para modelar variables de respuesta categórica binaria $(Y \in \{0, 1\})$, que se supone siguen una distribución de Bernoulli. La función de enlace es el logit definido como:

$$g(\mu) = \ln\left(\frac{\mu}{1 - \mu}\right)$$

donde $\mu = \mathbb{E}[Y]$ es la probabilidad de éxito $P(Y = 1)$. La predicción lineal $\eta = X\beta$ está en el rango $(-\infty, \infty)$, mientras que μ



está en el rago $[0, 1]$. La transformación logit asegura esta correspondencia:

$$\mu = g^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

- **Probit**. El modelo es de la forma:

$$P(Y = 1 \mid X) = \Phi(\beta_0 + \beta_1 ALTO + \beta_2 NOCUAL + \beta_3 EDAD + \beta_4 EDAD2 + \beta_5 CAS)$$

donde $\Phi(\cdot)$ es la función de distribución acumulativa de la normal estándar $\mathcal{N}(0, 1)$.

Al igual que en el caso anterior, también se utiliza para modelar variables binarias, sin embargo asume que la variable latente subyacente sigue una distribución normal estándar. Así, la función de enlace es la inversa de la función de distribución acumulativa, $\Phi(\cdot)$, de la normal estándar $\mathcal{N}(0, 1)$:

$$g(\mu) = \Phi^{-1}(\mu)$$

Siendo la predicción lineal $\eta = X\beta$ una distribución normal estándar, la probabilidad μ de que $Y = 1$ se obtiene como:

$$\mu = g^{-1}(\eta) = \Phi(\eta).$$

A diferencia del modelo Logit, aquí la relación entre η y μ está determinada por la forma de la curva normal estándar.

b) En un modelo Probit sin intercepto, la probabilidad de que la variable dependiente tome el valor 1 está dada por:

$$P(Y = 1 \mid X) = \Phi(\beta X),$$

donde $\Phi(\cdot)$ es la función de distribución acumulativa de la normal estándar $\mathcal{N}(0, 1)$, β es el coeficiente que queremos estimar, X es la

única variable explicativa. La probabilidad de que $Y = 0$ es:

$$P(Y = 0 \mid X) = 1 - \Phi(\beta X).$$

Las estimaciones Probit se calcularán a partir de la función de verosimilitud:

$$L(\beta) = \prod_{i=1}^{n} \Phi(\beta X_i)^{Y_i} \left[1 - \Phi(\beta X_i)\right]^{1-Y_i}.$$

Tomando el logaritmo para simplificar el cálculo:

$$\ell(\beta) = \sum_{i=1}^{n} \left[Y_i \ln\left(\Phi(\beta X_i)\right) + (1 - Y_i) \ln\left(1 - \Phi(\beta X_i)\right)\right].$$

Para encontrar la estimación de β, maximizamos la log-verosimilitud. La derivada de la log-verosimilitud con respecto a β es

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{Y_i}{\Phi(\beta X_i)} - \frac{1 - Y_i}{1 - \Phi(\beta X_i)}\right] \cdot \phi(\beta X_i) X_i,$$

donde $\phi(\cdot)$ es la función de densidad de la normal estándar. Dado que la ecuación de primer orden $\frac{\partial \ell(\beta)}{\partial \beta} = 0$ no tiene una solución analítica cerrada, se utilizan métodos numéricos para encontrar la estimación de β que maximiza la log-verosimilitud.

Las estimaciones obtenidas en el modelo Probit sin intercepto poseen las siguientes propiedades generales:

1. Consistencia: la estimación de β, denotada como $\hat{\beta}$, es consistente bajo condiciones estándar de máxima verosimilitud. Esto significa que conforme el tamaño de la muestra n crece, $\hat{\beta}$ converge en probabilidad al verdadero valor de β.

2. Normalidad asintótica: bajo condiciones regulares, $\hat{\beta}$ es asintóticamente normal, es decir:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\beta)),$$



donde $I(\beta)$ es la información de Fisher, que en este caso se calcula como el valor esperado de la segunda derivada negativa de la log-verosimilitud.

3. Eficiencia asintótica: la estimación obtenida por máxima verosimilitud es eficiente en el sentido de que alcanza la cota de Cramér-Rao en muestras grandes, es decir, tiene la menor varianza posible entre los estimadores insesgados asintóticamente normales.

4. Sesgo en muestras pequeñas: en muestras pequeñas, los estimadores de máxima verosimilitud pueden estar sesgados, lo que significa que $E[\hat{\beta}] \neq \beta$. Sin embargo, el sesgo tiende a desaparecer a medida que el tamaño muestral aumenta.

5. Identificación y escala: en un modelo Probit sin intercepto, la interpretación de β es relativa a la escala de la variable explicativa X. A diferencia de un modelo con intercepto, no hay un término independiente que ajuste el nivel base de la función Probit, lo que puede afectar la interpretación de los coeficientes.

c) Dados los tres conjuntos de estimaciones para cada uno de los modelos definidos en el apartado a), las hipótesis a contrastar para probar la hipótesis nula de que el coeficiente de CAS es cero son:

- **Hipótesis nula** (H_0): El coeficiente de la variable es cero, es decir, la variable no tiene efecto sobre la variable dependiente.

$$H_0 : \beta_5 = 0$$

- **Hipótesis alternativa** (H_A): El coeficiente de la variable es distinto de cero, es decir, la variable tiene un efecto sobre la variable

dependiente.

$$H_A : \beta_5 \neq 0$$

El estadístico de prueba que se utiliza en los modelos econométricos para contrastar la hipótesis nula H_0 de que el coeficiente de una variable es igual a cero es el estadístico t,

$$t = \frac{\hat{\beta}_i - 0}{\text{SE}(\hat{\beta}_i)},$$

donde $\hat{\beta}_i$ es el valor estimado del coeficiente de la variable y $\text{SE}(\hat{\beta}_i)$ es la estimación de la desviación estándar de $\hat{\beta}_i$. Se utiliza dicho estadístico porque bajo H_0, si los errores son normales, el numerador $\hat{\beta}_i - 0$ tiene una distribución normal. El estadístico t sigue una distribución t de Student con $n - k$ grados de libertad, donde n es el número de observaciones en el modelo y k es el número de parámetros estimados incluyendo el intercepto, si lo hay.

El valor del estadístico t se compara con el valor crítico de la distribución t para un nivel de significación α y $n - k$ grados de libertad. Si el valor absoluto de t es mayor que el valor crítico, se rechaza la hipótesis nula. Si es menor, no se puede rechazar la hipótesis nula.

	MCO	Logit	Probit
$\hat{\beta}$	0.024	0.103	0.063
$SE(\hat{\beta})$	0.009	0.057	0.035
t	2.67	1.81	1.80

Para un modelo con 6 variables, $k = 6$ y $n = 7634$, los grados de libertad $n - k = 7634 - 6 = 7628$. Con 7682 grados de libertad, el valor



crítico para una prueba bilateral al nivel de significación $\alpha = 0{,}05$ sería aproximadamente

$$t_{\alpha,n-k} = t_{0{,}025,7628} \simeq 1{,}96$$

Dado que el tamaño de muestra es grande, el valor obtenido se aproxima bastante al que se obtiene con una distribución normal estándar.

Por lo tanto, al comparar los estadísticos t obtenidos con el valor crítico $t_{0{,}025,7628}$, tenemos la siguiente tabla:

MCO	Logit	Probit
$t_{MCO} > t_{\alpha,n-k}$	$t_{Logit} < t_{\alpha,n-k}$	$t_{Probit} < t_{\alpha,n-k}$

En el caso de MCO, dado que el valor t es mayor que 1.96, se rechaza la hipótesis nula ($H_0 : \beta = 0$). En el caso de Logit y Probit, como los valores t son menores que 1.96, no se rechaza la hipótesis nula ($H_0 : \beta = 0$).

El estadístico t utilizado en el caso de MCO no es fiable ya que, en modelos con variable dependiente binaria, el término de error no tiene varianza constante, lo que viola uno de los supuestos clave de MCO. Consecuentemente, en presencia de heterocedassticiadad, los errores estándar de los coeficientes pueden estar mal estimados y llevar a inferencias incorrectas. Por el contrario, los modelos Logit y Probit son preferidos ya que están diseñados específicamente para variables dependientes binarias y manejan de manera más adecuada la no linealidad en la relación entre las variables.

d) La información conocida para el cálculo de esta probabilidad es:

- ALTO = 1 (tiene estudios superiores)

- NOCUAL = 0 (no está en el grupo sin cualificaciones)
- EDAD = 40
- EDAD2 = $\frac{40*40}{100} = 16$

Para la probabilidad de estar desempleada, lo que necesitamos calcular es la probabilidad de que la variable dependiente (tener empleo remunerado) tome el valor 0. Esto es equivalente a calcular $1 - P(Y = 1)$. De esta forma, las probabilidades para MCO y Probit serán:

MCO:

$$\begin{aligned}\hat{Y} =&\hat{\beta}_0 + \hat{\beta}_1 ALTO + \hat{\beta}_2 NOCUAL + \hat{\beta}_3 EDAD + \hat{\beta}_4 EDAD2 + \hat{\beta}_5 CAS \\ =& - 0{,}068 + 0{,}093 \cdot 1 - 0{,}210 \cdot 0 + 0{,}038 \cdot 40 - 0{,}051 \cdot 16 + 0{,}024 \cdot 1 \\ =&0{,}753\end{aligned}$$

Por tanto, la probabilidad de estar desempleada según el modelo MCO es $1 - \hat{Y}_{MCO} = 1 - 0{,}753 = 0{,}247$

PROBIT: tal y como se exponía en el apartado a, en este caso la variable dependiente es binaria y se utiliza una transformación de la función acumulada de la distribución normal estándar para modelar la probabilidad de estar desempleada.

$$\begin{aligned}\Phi^{-1}\left(P(Y = 1)\right) =&\hat{\beta}_0 + \hat{\beta}_1 ALTO + \hat{\beta}_2 NOCUAL + \hat{\beta}_3 EDAD + \\ &+\hat{\beta}_4 EDAD2 + \hat{\beta}_5 CAS = \\ &= - 1{,}593 + 0{,}259 \cdot 1 - 0{,}554 \cdot 0 + 0{,}107 \cdot 40 + \\ &-0{,}142 \cdot 16 + 0{,}063 \cdot 1 = \\ &=0{,}737\end{aligned}$$

El valor lineal de $\Phi^{-1}\left(P(Y = 1)\right)$ es 0.737. Ahora, usando la función



acumulativa normal estándar para calcular la probabilidad estimada

$$(P(Y = 1)) = \Phi(0{,}737) \simeq 0{,}769.$$

De modo que la probabilidad de estar desempleada es $1 - P(Y = 1) = 1 - 0{,}769 = 0{,}231$

Por lo tanto, ambos modelos predicen probabilidades similares de estar desempleada en este caso.

e) Las hipótesis nula y alternativa para la prueba de razón de verosimilitudes son:

- H_0: Los coeficientes del modelo son conjuntamente iguales a cero (es decir, las variables explicativas no tienen un efecto significativo sobre la variable dependiente).
- H_1: Al menos uno de los coeficientes es distinto de cero.

Matemáticamente:

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : \beta_i \neq 0 \text{ para algún } i = 0, \ldots, 5 \end{cases}$$

El estadístico de la prueba de razón de verosimilitudes (LR) se calcula como:

$$LR = -(2 \ln L_R - \ln L_{NR})$$

donde: $\ln L_R$ es el logaritmo de la verosimilitud del modelo restringido y $\ln L_{NR}$ es el logaritmo de la verosimilitud del modelo no restringido. Sustituyendo con los valores dados:

$$LR = -(2 \ln L_R - \ln L_{NR}) = -2 \cdot (-416{,}01) - (-321{,}25) = 189{,}52$$

El estadístico LR sigue una distribución χ^2 con k grados de libertad, donde k es el número de restricciones, que corresponde al número de



coeficientes en el modelo restringido. En este caso, el modelo tiene 5 variables explicativas (ALTO, NOCUAL, EDAD, EDAD2, CAS) por lo que $k = 5$. Para un nivel de significación del 5 % y $k = 5$ grados de libertad, el valor crítico $\chi^2_{0,05,5}$ se puede obtener de tablas de la distribución chi-cuadrado o con un software estadístico.

$$\chi^2_{0,05,5} \simeq 11{,}07$$

Comparamos el valor calculado de LR con el valor crítico. Dado que

$$LR > \chi^2_{0,05,5},$$

rechazamos la hipótesis nula H_0. Conclusión, hay evidencia suficiente, al nivel de significancia del 5 %, para concluir que las variables explicativas son conjuntamente significativas en el modelo Probit.

Ejercicio 4

Para promover la vida social y turística de un municipio, determinada Asociación Cultural y Turística ofrece desde 2023 un programa de excursiones anuales para sus miembros de dos tipos, unas de carácter nacional y otras de carácter internacional. La implementación de este programa se realiza siguiendo las pautas que se establecen a continuación:

- Cada año, los asociados que aún no participan en el programa de excursiones tienen una probabilidad de 0,4 de unirse a las excursiones nacionales y de 0,3 a las internacionales, el año siguiente. Si no se integran al programa, seguirán sin participar en él.

- El 60 % de los asociados que han participado en excursiones nacionales un determinado año continuarán el próximo año participando en las



mismas, mientras que el 70 % de los que realizaron las excursiones internacionales repetirá en éstas. El resto de participantes, cambiará al otro tipo de excursión, manteniéndose para siempre en el programa de la asociación.

Considerando que en 2023 ningún miembro participaba en estas excursiones programadas, se pide

a) Modelar el proceso estocástico asociado utilizando una cadena de Markov, describiendo sus principales componentes y propiedades.

b) Establecer la proporción de miembros de la asociación que durante el año 2025 disfrutarán de cada uno de los tipos de excursiones.

c) Determinar el comportamiento a largo plazo de la proporción de miembros que realizará cada uno de los tipos de excursiones. Asimismo, calcular el tiempo medio que tarda en estar en el programa un asociado que no estaba en él de partida.

Solución

La dinámica descrita se ajusta a un proceso discreto de Markov. Hay tres estados $= \{e_F, e_N, e_I\}$, que codifican el estar aún (F)uera del programa, estar en las excursiones (N)acionales, y estar en las (I)nternacionales, respectivamente.

Una representación del sistema de transición entre estados es la mostrada en la Figura 4.16.

La matriz de transición M (en su versión derecha) es aquella en la que $M_{i,j}$ es la probabilidad de transicionar en un paso desde e_i a e_j. En nuestro

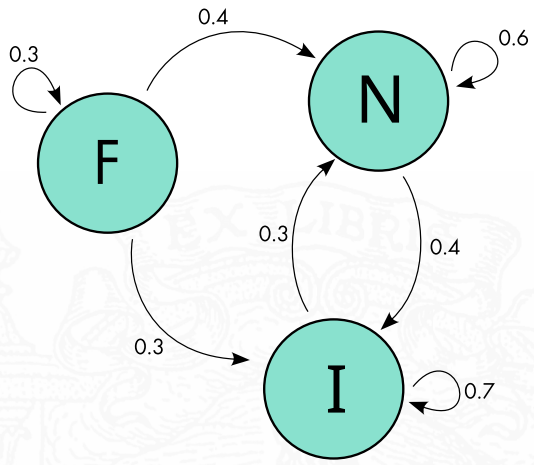


Figura 4.16: Diagrama del proceso de Markov asociado al problema.

caso:

$$M = \begin{pmatrix} 0{,}3 & 0{,}4 & 0{,}3 \\ 0 & 0{,}6 & 0{,}4 \\ 0 & 0{,}3 & 0{,}7 \end{pmatrix}$$

Si v (vector fila) codifica las probabilidades de que un individuo esté en un momento dado en los estados e_F, e_N y e_I, respectivamente, entonces $v \cdot M$ contiene las probabilidades tras una transición, en el paso siguiente.

Puesto que la probabilidad de las transiciones solo depende del estado actual, y no de cómo se ha llegado a él, la matriz de transición describe por completo la dinámica del proceso de Markov.

En este caso concreto no hay ningún estado absorbente (son aquellos en los que la probabilidad de salir de ellos es cero).

El estado e_F es transitorio, de hecho una vez se abandona es imposible volver a él.



Apartado b)

Se nos aclara que en 2023 ningún individuo está participando en un programa. Es decir, dado un individuo, su vector de probabilidades es $v_{2023} = (1, 0, 0)$. Dado que el sistema es de Markov con matriz de transición M, resulta

$$v_{2024} = v_{2023} \cdot M = (0{,}3, 0{,}4, 0{,}3)$$

y

$$v_{2025} = v_{2024} \cdot M = (0{,}3, 0{,}4, 0{,}3) \cdot M = (0{,}09, 0{,}45, 0{,}46).$$

Esto se interpreta como que hay una probabilidad del 9 % de que en 2025 siga fuera del programa, de un 45 % de que esté haciendo viajes nacionales y de un 46 % de que esté haciendo viajes internacionales.

Apartado c)

El comportamiento a largo plazo viene dado por $\lim_{k\to\infty} M^k$, cuando existe. En nuestro caso, una vez se abandona el estado e_F no puede regresarse a él, pero desde e_N y e_I hay una probabilidad no nula tanto de quedarse como de marchar al otro en un solo paso. Eso implica que restringida a $\{e_N, e_I\}$ la cadena es irreducible y aperiódica, con lo que es de tipo ergódico: converge a una distribución estacionaria única sin importar el estado inicial. Es decir, la sucesión de vectores $v_{2023}, v_{2024}, v_{2025} \ldots$ converge. La convergencia se alcanza necesariamente en un punto fijo de M:

$$w \cdot M = w,$$

donde por ser w un vector de probabilidades, todas las componentes son no negativas y suman 1. Escrito de otra forma,

$$w \cdot (M - I) = 0$$

donde I es la matriz identidad. Es decir, w es un vector del *kernel* (por la izquierda) de $M - I$. Vamos a calcular $w = (w_1, w_2, w_3)$ resolviendo las ecuaciones.

$$w \cdot (M - I) = 0 \iff (w_1 w_2 w_3) \cdot \begin{pmatrix} -0{,}7 & 0{,}4 & 0{,}3 \\ 0 & -0{,}4 & 0{,}4 \\ 0 & 0{,}3 & -0{,}3 \end{pmatrix} = (0, 0, 0)$$

de donde $w_1 = 0$ y $-0{,}4w_2 + 0{,}3w_3 = 0$. Además sabemos que $w_1 + w_2 + w_3 = 1$. Entonces resulta $4w_2 = 3w_3 = 3(1 - w_2) = 3 - 3w_2$, de donde $w_2 = \frac{3}{7}$ y $w_3 = \frac{4}{7}$. Por tanto $w = (0, \frac{3}{7}, \frac{4}{7})$ son las probabilidades estacionarias. Aproximadamente un 42,85 % de estar en viajes nacionales y un 57,14 % en internacionales. Es el único punto fijo, así que se converge necesariamente a él.

En cuanto al tiempo (en años) que se tarda en entrar en el programa, es una distribución geométrica $X \sim \text{Geom}(p = 0{,}7)$[1]. Sabemos que su esperanza es $\mathbb{E}[X] = \frac{1}{p} = \frac{1}{0{,}7} = \frac{10}{7}$.

Ejercicio 5

Suponga que le piden implementar un código que incluya los algoritmos Lasso y random forest para un caso de clasificación. Escriba los pasos a seguir, preferiblemente en código Python, R o pseudocódigo, para llevar a cabo el proyecto completo. Las características del proyecto son las siguientes:

- Objetivo: Utilizar las características del empleado/a para predecir si abandonarán la compañía. Datos disponibles en el archivo "abandono.csv"

[1] Esto es así porque se tiene una probabilidad en cada paso de un 0,7 de entrar en el programa de viajes, y una vez se entra nunca se abandona.



- Variable target: Attrition (valores “ Yes” , “No”)
- Features (covariables): 30 variables (var1 a var30)
- Número de observaciones: 1470

En el código debe tener en cuenta lo siguiente:

a) Teniendo en cuenta que la prevalencia de abandono general está en torno al 16 %, ¿qué métricas de clasificación considera adecuadas para evaluar el desempeño de los algoritmos?

b) Entre los features (covariables) hay variables de carácter cuantitativo y otras de carácter cualitativo (tanto nominal como ordinal). Algunas variables de carácter cuantitativo pueden contener outliers, tener fuerte asimetría y, habitualmente, estarán en unidades distintas. ¿Qué transformaciones considera adecuadas en las variables de carácter cuantitativo para lograr que los algoritmos puedan tener mejores resultados? Suponga que las variables cuantitativas son las 15 primeras (var1 a var15).

c) En las variables de carácter nominal (var16 a var25), ¿es necesario crear dummies por categorías en ambos algoritmos? Justifique su respuesta.

d) Suponga que le informan de que algunas de las categorías de las variables cualitativas var17 y var18 tienen muy pocas observaciones. ¿Qué efectos tendría esta situación sobre el entrenamiento y validación de algoritmos y cómo podríamos resolverlo?

e) ¿Cómo se pueden codificar las variables ordinales var26 a var30 en ambos algoritmos?



f) Su supervisora le dice que debe realizarse un proceso de tuneado sobre los hiperparámetros de cada algoritmo. ¿Por qué cree que le pide esto?. Describa detalladamente los pasospara llevar a cabo este proceso teniendo en cuenta que se debe utilizar validación cruzada con la muestra de training. Tenga en cuenta, además, que todos los resultados deben ser reproducibles.

g) ¿Cómo evaluaría la posible existencia de overfitting en los algoritmos planteados?

Solución

A continuación se escriben en código Python los pasos a seguir para implementar el proyecto que incluye el uso de los algoritmos Lasso para la selección de características y Random Forest para clasificación.

El primer paso sería cargar y explorar los datos. Para ello se importa el archivo `abandono.csv` y se inspecciona su contenido:

```
import pandas as pd
data = pd.read_csv("abandono.csv")
print(data.info())
print(data.describe())
print(data.head())
```

Una vez cargados los datos se pasa al preprocesamiento de los datos:

1. **Transformaciones para variables cuantitativas (`var1` a `var15`).**

   Las siguientes transformaciones mejorarán el rendimiento del modelo:

   - **Imputación de valores faltantes:** Sustituir valores faltantes con la mediana, que es más robusta frente a *outliers*.



- **Escalado:** Normalizar las variables para que tengan media cero y desviación estándar uno. Se hace a la vez que el ajuste de la regresión Lasso para hacerlo de forma ordenada y en un solo paso. Para el Random Forest no es necesario pues el algoritmo funciona de forma adecuada con las variables sin escalar.
- **Transformaciones para outliers y asimetría:** Cuando las variables no presenten valores negativos, se aplican transformaciones como logarítmica o Box-Cox según la distribución de las variables. En caso de que la variable no normal presente valores negativos, se utilizaría la transformación de Yeo-Johnson.

2. **Codificación de variables nominales (`var16` a `var25`).**

   - Para Lasso, es necesario codificar las variables nominales en formato numérico, como *One-Hot Encoding*, estrategia que consiste en crear una columna binaria (que solo puede contener los valores 0 o 1) para cada valor único que exista en la variable categórica que se está codificando. Para evitar colinealidad, es necesario eliminar una categoria por variable.

   - Para Random Forest, aunque puede manejar variables categóricas directamente, el uso de codificaciones puede mejorar el rendimiento. En casos con categorías poco frecuentes, se recomienda agruparlas en una categoría *Otros* o utilizar técnicas de *Target Encoding*.

3. **Manejo de categorías con pocas observaciones (`var17` y `var18`).**

   Las categorías con pocas observaciones pueden causar sobreajuste, ya que el modelo podría memorizar patrones específicos de estas categorías. Para evitarlo:

- En el caso de Lasso, se pueden agrupar categorías poco frecuentes en una categoría *Otros* y a continuación proceder a la codificación.
- Realizar pruebas adicionales para validar el impacto de estas categorías en el desempeño del modelo.

4. **Codificación de variables ordinales (`var26` a `var30`).**

   Para ambas técnicas, como comentábamos, las variables ordinales deben codificarse asignando valores numéricos que respeten su orden y si es necesario se agrupan previamente en función de los umbrales que consideremos. Por ejemplo, una variable con categorías "Bajo", "Medio" y "Alto" se puede codificar como 1, 2 y 3 respectivamente.

```
#importacion de librerias
from sklearn.preprocessing import StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder
import matplotlib.pyplot as plt
import pandas as pd


# Se crea un objeto imputer que reemplaza los valores faltantes en las columnas seleccionadas utilizando la mediana de cada columna
imputer = SimpleImputer(strategy='median')
# se seleccionan las primeras 15 columnas del conjunto de datos y se aplica el imputer a esas columnas
data.iloc[:, :15] = imputer.fit_transform(data.iloc[:, :15])  # Cuantitativas

#A continuación se estudia la distribución de las variables 1 a 15 para comprobar si es necesario
```




```
    aplicar alguna transformación. Se pone el caso de var1
     pero sería analogo para el resto:
data['var1'].hist() #se inspecciona la distribución
plt.title('Distribucion de var1')
plt.show()
#Caso 1: Distribución sesgada a la derecha a la que se
   aplica una transformación logarítmica
data['var1'] = np.log(data['var1'])
#Caso 2: Transfromación BoxCox para hacer que la variable
    siga una distribución más cercana a la normal
data['var1'], _ = boxcox(data['var1'])
#Caso 3: Distribución aproximadamente normal, no es
   necesaria la transformación.

#Codificación de las variables nominales mediante One-Hot
    Encoding
data = pd.get_dummies(data,
columns=[f'var{i}' for i in range(16, 26)],
drop_first=True) #elimina una categoría por variable para
    evitar colinealidad

# Codificar ordinales (var26 a var30) y nominales (var16
   a var25)
for col in ["var26", "var27", "var28", "var29", "var30"]:
le = LabelEncoder()
data[col] = le.fit_transform(data[col])
```

Tras el preprocesamiento de los datos, se pasa a aplicar el algoritmo Lasso que se nos pide en el enunciado. El algoritmo Lasso, Least Absolute Shrinkage and Selection Operator, es una técnica de regularización que se utiliza para selección de características y reducción de la complejidad del modelo. Lasso es un tipo de regresión lineal regularizada en la que se penaliza la magnitud de



los coeficientes de las variables en el modelo. La regularización L1 utilizada por Lasso agrega una penalización proporcional a la suma de los valores absolutos de los coeficientes. La fórmula general de Lasso para el modelo de regresión es: ser útil de la siguiente manera:

$$\text{mín} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

donde y_i es el valor observado (si un empleado abandonó o no), $\hat{y}_i$ es la predicción del modelo, β_j son los coeficientes de las características y λ es el parámetro de regularización que controla la cantidad de penalización.

En el contexto del problema, se desea predecir si un empleado abandonará la compañía utilizando 30 variables, Lasso servirá para construir un modelo más sencillo, eficiente y fácil de interpretar al seleccionar las variables más relevantes para predecir si un empleado abandonará la empresa mientras penaliza aquellas características menos importantes.

```
#importación de librerías
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV
#se estandarizan las variables del conjunto de datos
scaler = StandardScaler()
#se define el modelo Lasso especificando que se usa
    regularización L1, el solucionador liblinear que es
    adecuado para pequeños conjuntos de datos y funciona
    bien con regularización L1 y se establece una semilla
    para la aleatoriedad:
lasso = LogisticRegression(penalty='l1', solver='
    liblinear', random_state=42)
```




```
12  # se crea un pipeline que combina el escalado de las
        características primero y después se ajusta el modelo
        de regresión logística con Lasso a los datos escalados
        :
13  pipeline = Pipeline([('scaler', scaler), ('lasso', lasso)
        ])
14  #se define la búsqueda de hiperparametros mediante
        GridSearchCV:
15  X = data.drop(columns=["Attrition"])  # O la variable
        objetivo segun el dataset
16  y = data["Attrition"].map({"Yes": 1, "No": 0})  #
        Convertir a binario si es categórica
17  #se generan los valores del parámetro lambda en escala
        logarítmica para C (que es el inverso de lambda)
18  C_values = np.logspace(0, -2, num=30)  # 30 valores de 1
        a 0.01 en escala logarítmica
19  param_grid = {'lasso__C': C_values}
20  grid_search = GridSearchCV(pipeline, param_grid, cv=5,
        scoring='f1')
21  #y se ajusta el modelo con los datos:
22  grid_search.fit(X, y)
23  #una vez que se ha completado la búsqueda, se selecciona
        el mejor estimador
24  best_lasso = grid_search.best_estimator_.named_steps['
        lasso']
25  # Hacer predicciones con el mejor modelo
26  y_pred = best_lasso.predict(X)
27  #se visualiza la importancia de las variables
28  coefs = best_lasso.coef_.flatten()
29  plt.figure(figsize=(10, 5))
30  plt.barh(X.columns, coefs)
31  plt.xlabel("Peso del coeficiente")
32  plt.ylabel("Variable")
```




```
plt.title("Importancia de las variables segun Lasso")
plt.show()
```

Otra opción, siguiendo las indicaciones del enunciado, es utilizar el algoritmo de Random Forest para la clasificación. La principal diferencia respecto al modelo Lasso que acabamos de entrenar es que Random Forest es no lineal (el anterior sí lo es). Podría ocurrir que las variables que Lasso considera poco explicativos en el caso lineal, tengán un poder predictivo no lineal que sea captado por el modelo de Random Forest. Se entrena el modelo Random Forest.

```
#importación de librerias
from sklearn.ensemble import RandomForestClassifier
#se crea una instancia del modelo RandomForestClassifier
    y se almacena en la variable rf, se usa una semilla
    aleatoria con un valor específico  para que los
    resultados del modelo sean reproducibles
rf = RandomForestClassifier(random_state=42)
rf.fit(X, y)
```

En el enunciado se dice que la prevalencia de abandono general es de en torno al 16 %. Esto quiere decir que el 16 % de empleados del conjunto de datos han abandonado la compañía. En otras palabras, de las 1470 observaciones de los empleados, aproximadamente el 16 % de esos empleados tienen como valor en la variable "Attrition" el valor "Yes". El restante 84 % tiene el valor "No", que indica que no abandonaron la empresa.

A la vista de lo expuesto, se puede concluir que las clases están desbalanceadas: 16 % de abandono frente al 84 % que no abandonan. Por ello, es fundamental usar métricas que sean sensibles a este desbalance del modelo. Cuando se tiene un problema de clasificación con clases desbalanceadas, las métricas tradicionales como la precisión pueden no ser la mejor opción para



evaluar el desempeño del modelo ya que podría ocurrir que un modelo con una precisión alta esté simplemente prediciendo la clase mayoritaria la mayor parte del tiempo. En consecuencia es fundamental usar métricas que sean sensibles a este desbalance: F1-Score, ROC-AUC y Curva Precision-Recall.

- **F1-Score:** es la media armónica entre la precisión y la recall (sensibilidad)

$$\text{precisión} = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos positivos}}$$

$$\text{recall} = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos negativos}}$$

  Es decir, la precisión mide de todas las instancias que el modelo predice como positivas, ¿cuántas son realmente positivas? Mientras que recall, de todas las instancias que son realmente positivas, ¿cuántas han sido correctamente identificadas por el modelo? F1-Score alto significa que el modelo está haciendo un buen trabajo tanto identificando correctamente las instancias positivas como evitando etiquetar instancias negativas como positivas. F1-Score bajo indica que el modelo tiene dificultades en identificar correctamente las instancias positivas, lo cual es un problema en clases desbalanceadas, donde se quiere priorizar la detección de la clase minoritaria.

- **ROC-AUC:** El ROC (Receiver Operating Characteristic) es una curva que muestra la tasa de verdaderos positivos (recall) frente a la tasa de falsos positivos para diferentes umbrales de clasificación. El ROC-AUC (Área bajo la curva ROC) mide el rendimiento global del modelo sin importar el umbral de clasificación, es decir, se tiene en cuenta la tasa de falsos positivos y la tasa de verdaderos positivos para todos los posibles valores de umbral. El ROC-AUC es útil en clases desbalanceadas porque te da una visión más general del desempeño del modelo, al



considerar todos los posibles umbrales. La clase minoritaria puede ser mejor evaluada con esta métrica, ya que no se ve tan afectada por la clase mayoritaria.

- **Curva Precision-Recall:** muestra cómo varía la precisión frente a la recall para diferentes umbrales. Es especialmente útil cuando tenemos clases desbalanceadas porque, al igual que el F1-Score, esta curva se enfoca más en la clase positiva (la minoritaria en problemas desbalanceados). En escenarios desbalanceados, la clase mayoritaria puede dominar la precisión, y la recall de la clase positiva (minoritario) podría quedar rezagada. En este contexto, la curva Precision-Recall es más informativa que la curva ROC, ya que se enfoca directamente en cómo se está tratando la clase minoritaria. Una curva Precision-Recall más alta indica que el modelo tiene una buena capacidad para predecir correctamente la clase positiva sin hacer demasiados errores (falsos positivos).

```
#importación de librerias
from sklearn.metrics import classification_report,
    roc_auc_score, precision_recall_curve
from sklearn.model_selection import train_test_split

# División del conjunto de datos en entrenamiento y
    prueba (80% entrenamiento, 20% prueba)
X_train, X_test, y_train, y_test = train_test_split(X, y,
     test_size=0.2, random_state=42)
# Predicción de la clase para el conjunto de prueba
    usando Random Forest
y_pred = rf.predict(X_test)
# Predicción de probabilidades para el conjunto de prueba
y_proba = rf.predict_proba(X_test)[:, 1]
```




```
# informe de métricas de clasificación (precision, recall, F1-score) para cada clase
print(classification_report(y_test, y_pred))
# calcula e imprime el puntaje ROC-AUC del modelo
print(f"ROC-AUC: {roc_auc_score(y_test, y_proba):.2f}")
# calcula las curvas de precisión y recall basadas en las probabilidades predichas
precision, recall, _ = precision_recall_curve(y_test, y_proba)
```

Finalmente, una vez evaluado el desempeño del modelo, se procede al tuneado de hiperparámetros. El tuneado de hiperparámetros es importante para optimizar el desempeño de los modelos. Ambos algoritmos (Lasso y Random Forest) tienen hiperparámetros clave que afectan significativamente su rendimiento. A continuación, se detalla el proceso:

1. Lasso (Regresión Logística con penalización L1):
   - C: Inverso de la fuerza de regularización. Valores bajos de C generan mayor penalización, eliminando más características.
   - solver: Asegurarse de usar un solver compatible con la penalización L1 (como liblinear o saga).
2. Random Forest:
   - n_estimators: Número de árboles en el bosque. Un número más alto puede mejorar la estabilidad, pero incrementa el costo computacional.
   - max_depth: Profundidad máxima de los árboles. Controla la complejidad del modelo.



- min_samples_split y min_samples_leaf: Controlan el número mínimo de muestras necesarias para dividir un nodo o mantener una hoja.
- max_features: Número de características consideradas al buscar el mejor split.

Existen diferentes métodos de optimización para evaluar el desempeño de las combinaciones de hiperparámetros como son la búsqueda en malla o Grid Search o la búsqueda aleatoria o Randomized Search. En este caso, se usan los modelos con validación cruzada.

```
# Importación de librerias
from sklearn.model_selection import train_test_split,
    GridSearchCV, RandomizedSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.pipeline import Pipeline
import pandas as pd

# División del conjunto de datos en entrenamiento y
    prueba (80% entrenamiento, 20% prueba)
X_train, X_test, y_train, y_test = train_test_split(X, y,
    test_size=0.2, random_state=42, stratify=y)

# Lasso: ajuste de hiperparámetro C
lasso_param_grid = {'lasso__C': [0.01, 0.1, 1, 10, 100]}

# Se define un pipeline que incluye la estandarización y
    la regresión logística con penalización L1 (Lasso)
lasso_pipeline = Pipeline([
('scaler', StandardScaler()),
```




```
18  ('lasso', LogisticRegression(penalty='l1', solver='
        liblinear', random_state=42))
19  ])
20
21  # Se usa GridSearchCV para encontrar el mejor valor de C
        mediante validación cruzada
22  grid_search_lasso = GridSearchCV(lasso_pipeline,
        param_grid=lasso_param_grid, cv=5, scoring='f1')
23
24  # Se ajusta el modelo Lasso a los datos sin aplicar SMOTE
25  grid_search_lasso.fit(X_train, y_train)
26
27  # Obtener las características seleccionadas por Lasso
28  best_lasso = grid_search_lasso.best_estimator_.
        named_steps['lasso']
29  selected_features = X_train.columns[(best_lasso.coef_[0]
        != 0)]
30
31  X_train_selected = X_train[selected_features]
32  X_test_selected = X_test[selected_features]
33
34  # En cuanto al Random Forest, aplicar StandardScaler
35  scaler = StandardScaler()
36  X_train_scaled = scaler.fit_transform(X_train)
37  X_test_scaled = scaler.transform(X_test)
38
39  # Random Forest: ajuste de hiperparámetros
40  n_features = X.shape[1]  # Número total de caracterí
        sticas en el dataset, se usara a continuación
41
42  rf_param_grid = {
43    'n_estimators':[50, 100, 200, 500, 1000],        # Nú
          mero de árboles en el bosque
```




```
    'max_depth': [5, 10, 20, 30, None],           # Profundidad máxima de los árboles
    'min_samples_split': [2, 4, 6, 10, 20],       # Mínimo de muestras necesarias para dividir un nodo
    'max_features': (np.linspace(1, n_features, num=5, dtype=int)).tolist() # Máximo numero de características consideradas por división
}

# Se inicializa un modelo Random Forest
random_forest = RandomForestClassifier(random_state=42)

# Se usa RandomizedSearchCV para buscar los mejores hiperparámetros
random_search_rf = RandomizedSearchCV(
random_forest, param_distributions=rf_param_grid,
n_iter=50, cv=5, scoring='f1', random_state=42
)

# Se entrena el modelo Random Forest con las características seleccionadas y normalizadas
random_search_rf.fit(X_train_scaled, y_train)

# Resultados finales
print("Mejores hiperparámetros para Lasso:", grid_search_lasso.best_params_)
print("Mejores hiperparámetros para Random Forest:", random_search_rf.best_params_)
```

Por último, el overfitting ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, perdiendo capacidad de generalización. Para evaluarlo:



1. Diferencia significativa entre las métricas de entrenamiento y prueba: un alto desempeño en el conjunto de entrenamiento pero bajo en el de prueba es un claro signo de sobreajuste.

2. Curvas de aprendizaje: si el desempeño en entrenamiento se mantiene alto mientras que en validación es consistentemente bajo, indica sobreajuste.

3. Comportamiento de las métricas: una alta precisión combinada con un bajo recall puede ser un síntoma de que el modelo está ajustándose demasiado a patrones específicos.

```
#importación de librerias
from sklearn.metrics import classification_report,
    roc_auc_score
from sklearn.model_selection import learning_curve
import matplotlib.pyplot as plt
# Métricas de entrenamiento: predicciones sobre los datos
    de entrenamiento balanceados
train_pred = random_search_rf.best_estimator_.predict(X)
# probabilidades de clase positiva (1) para calcular mé
    tricas como ROC-AUC
train_proba = random_search_rf.best_estimator_.
    predict_proba(X)[:, 1]
# reporte de clasificación con métricas detalladas (
    Precision, Recall, F1, etc.)
print("Entrenamiento:")
print(classification_report(y, train_pred))  # Se
    comparan las predicciones con las etiquetas
    balanceadas
# se calcula el ROC-AUC en el conjunto de entrenamiento
print(f"ROC-AUC (train): {roc_auc_score(y, train_proba)
    :.2f}")
```




```
14
15  # métricas de prueba: predicciones sobre el conjunto de
        prueba
16  test_pred = random_search_rf.best_estimator_.predict(
        X_test)
17  # probabilidades de clase positiva (1) para calcular mé
        tricas
18  test_proba = random_search_rf.best_estimator_.
        predict_proba(X_test)[:, 1]
19  # reporte de clasificación para el conjunto de prueba
20  print("Prueba:")
21  print(classification_report(y_test, test_pred))  # Se
        comparan las predicciones con las etiquetas reales del
         test
22  # se calcula el ROC-AUC en el conjunto de prueba
23  print(f"ROC-AUC (test): {roc_auc_score(y_test, test_proba
        ):.2f}")
24
25  # Curvas de aprendizaje: se evalua el desempeño del
        modelo con diferentes tamaños de datos de
        entrenamiento.
26  # train_sizes: Números de ejemplos de entrenamiento
        evaluados (crecen de menor a mayor).
27  # train_scores y test_scores: Metricas (F1-Score)
        obtenidas para cada tamaño de datos.
28  train_sizes, train_scores, test_scores = learning_curve(
29  random_search_rf.best_estimator_,  # Modelo Random Forest
         ajustado con los mejores parámetros
30  X,  # Datos balanceados seleccionados
31  y,  # Etiquetas balanceadas
32  cv=5,  # Validación cruzada con 5 particiones
33  scoring='f1'  # Métrica de evaluacion (F1-Score)
34  )
```




```
35
36  # Graficar curvas de aprendizaje
37  plt.plot(train_sizes, train_scores.mean(axis=1), label='
        Train')  # Curva para entrenamiento
38  plt.plot(train_sizes, test_scores.mean(axis=1), label='
        Validation')  # Curva para validación
39  plt.legend()  # Añadimos una leyenda para distinguir las
        curvas
40  plt.xlabel('Training Size')
41  plt.ylabel('F1-Score')
42  plt.title('Learning Curve')
43  plt.show()
```

Consideramos interesante aportar una versión de este ejercicio en el lenguaje de programación R, puesto que puede ayudar a todos aquellas personas que estén más familiarizados con este lenguaje que con python o cualquier otro. Nuestra intención es presentar herramientas a los opositores para que puedan comprender de la mejor forma posible los ejercicios propuestos, y como hemos comentado con anterioridad, es evidente que no se pretende que los aspirantes puedan contestar en el limitado tiempo que tienen una solución como estas, sino que el objetivo de este texto va más en la dirección de poder comprender bien cada ejercicio.

Hemos comentado que como la prevalencia de positivos es baja, la acuracidad no es una medida adecuada para evaluar el performance de los modelos propuestos. En su lugar, se pueden considerar la curva ROC o la f-measure (ó f1) que mide la media armónica entre la sensibilidad y la precisión.

En la codificación en lenguaje R se seguirá la implementación propuesta en el metapaquete **tidymodels** (ver el textbook disponible en `https://www.tmwr.org/`). Esta metodología incluye por separado las etapas de preprocesamiento (en objetos del tipo *recipe*) y el algoritmo a implementar en



un objeto de tipo *workflow* (que incluye recipe + algoritmo). Sobre este objeto workflow se pueden realizar procesos de validación cruzada con datos de training para evaluar el performance predictivo. Adicionalmente, se puede entrenar el objeto workflow con el conjunto de training completo y evaluar el performance sobre el conjunto de test. Ver los detalles en el textbook disponible en `https://www.tmwr.org/`.

```
## Lectura de Datos
attrition <- readr::read_csv("abandono.csv")
## Carga paquete tidymodels
library(tidymodels)
## En caso de conflicto con el nombre de funciones,
# elige las disponibles en tidymodels
tidymodels_prefer()
## Establece semilla para reproducibilidad de resultados
set.seed(123890)
## División de datos en training y test
attr_split <- initial_split(attrition, prop = 0.7)
attr_train <- training(attr_split)
attr_test <- testing(attr_split)
```

- Transformaciones para las variables cuantitativas para reducir asimetría y efecto outliers: *Box-Cox* (no negativas) o *Yeo-Johnson* (similar a Box-Cox para variables con valores negativos). El valor del parámetro λ para la transformación Box-Cox se estima a partir de los datos (y, en caso de no necesitar transformar la variable, corresponde a un valor de $\lambda = 1$ que mantiene la forma de la distribución).

- Las categorías con pocos valores en las variables nominales (por ejemplo: menor de un umbral de 2 % de observaciones) se *agruparán en una categoría única* para impedir el sobreajuste.



- La codificación de las nominales con variables dummies es necesaria para el algoritmo Lasso pero no es estrictamente necesaria para el Random Forest (aunque puede hacerse).

- Para las variables ordinales, se escoge una *codificación numérica que mantenga el orden de las categorías correspondientes* (habitualmente se realiza con enteros o polinomios que garanticen el mantenimiento del orden de las categorías ordinales). No es necesaria en el caso de Random Forest aunque puede hacerse.

En el siguiente paso podemos declarar un objeto de tipo recipe con steps de preprocesamiento para el algoritmo Lasso de la siguiente forma:

```
rec_lasso <- recipe(Attrition ~ . , data = attr_train) |>
# Elimina variables near-zero-variance
  step_nzv(all_predictors()) |>
# Codifica variables ordinales con valores polinomiales que respeten el orden
  step_ordinalscore(all_ordered_predictors()) |>
# Agrupa categorías que no llegan al 2% en las variables nominales
  step_other(all_nominal_predictors(), threshold = 0.02)  |>
# En el algoritmo Lasso, se codifican como dummies los features cualitativos
# eliminando la primera dummy (one-hot FALSE).
  step_dummy(all_nominal_predictors(), one_hot = FALSE) |>
# Se realiza transformación BoxCox para reducir asimetría y outliers
# en predictores numéricos (si hay valores
# negativos se utilizaría step_YeoJohnson)
  step_BoxCox(all_numeric_predictors()) |>
# En el algoritmo Lasso es necesario normalizar los predictores numéricos
#  (centrando y escalando)
  step_normalize(all_numeric_predictors())
```

De la misma forma hacemos algo similar para el algoritmo Random Forest, que en este caso necesita menos preprocesamiento que Lasso.

```
rec_rf <- recipe(Attrition ~ . ,
                 data = attr_train) |>
  step_nzv(all_predictors()) |>
  step_other(all_nominal_predictors(), threshold = 0.02)
```



- Creación de modelos para cada algoritmo (Lasso y Random Forest) así como objetos workflow que integren el objeto recipe y el modelo.

```
# Se tuneará el parámetro de penalización Lasso.
# El argumento mixture = 1 indica Lasso puro.
# El paquete (engine) utilizado será glmnet
lasso_model <- logistic_reg(penalty = tune(), mixture = 1) |>
  set_engine("glmnet")  |>
  set_mode("classification")
# Se añade recipe y algoritmo en el workflow
lasso_wf <- workflow() |>
  add_recipe(rec_lasso) |>
  add_model(lasso_model)
```

```
# Se tunearán los argumentos mtry (número de features seleccionados en cada división),
# trees (número de árboles) y min_n (número mínimo de observaciones en hojas finales)
# El paquete (engine) utilizado será randomForest
rf_model <-rand_forest(mtry = tune(), trees = tune(), min_n = tune()) |>
  set_engine("randomForest") |>
  set_mode("classification")
# Se añade recipe y algoritmo en el objeto workflow
rf_wf <-  workflow() |>
  add_model(rf_model) |>
  add_recipe(rec_rf)
```

Respecto a la pregunta relativa a que la supervisora dice que debe realizarse un proceso de tuneado sobre los hiperparámetros de cada algoritmo y el por qué cree que se pide esto, podemos decir que el tuneado se hace para elegir de forma óptima los hiperparámetros de cada modelo (especificados ya en el propio objeto workflow) de manera que se optimiza el rendimiento de los modelos, mejorando su capacidad de generalización. Se elegirán mediante un proceso de validación cruzada con los datos de training para impedir el posible overfitting.

Lo primero será crear un objeto cross-validation (fijando semilla) que se pueda utilizar con ambos workflows (lasso y random forest).




```
# Se fija semilla para reproducibilidad
set.seed(101112)
# Número de folds en CV
k <- 5
# Se crean dos objetos CV: uno para el tuneado y
# otro para que, una vez elegidos los hiperparámetros, se evalúe el performance
# con otro proceso de CV distinto del utilizado para el tuneado.
attrcv_tun <- vfold_cv(attr_train, v = k)
attrcv_asse <- vfold_cv(attr_train, v = k)
```

Creación de un grid para los hiperparámetros a tunear en cada workflow (lasso y random forest).

```
# Inicializa parámetros en lasso_wf
lasso_param <- extract_parameter_set_dials(lasso_wf)
# Crea grid para el lambda de Lasso con
# 100 valores en escala logarítmica entre (-15 y 1)
```

```
grid_lasso_tune <- data.frame(
  penalty = penalty() |>
  range_set(c(-15, 1)) |>
  value_seq(n = 100, original = TRUE))
# Inicializa parámetros en rf_wf
rf_param <- extract_parameter_set_dials(rf_wf)
# Crea grid con combinaciones de valores para mtry, trees y min_n (4x2x2)
grid_rf <- expand.grid(mtry = c(5, 10, 15, 25),
                       trees = c(500, 1000),
                       min_n = c(10, 50))
```

Utilizando los objetos workflow, el objeto CV generado y los grids se puede realizar el tuneado con la función `tune_grid()`




```
# Vector métricas a calcular en cada etapa de CV
metr_cl <- metric_set(roc_auc, accuracy, sensitivity, precision, f_meas, specificity)
# Tuneado hiperparámetros lasso con CV
# El argumento de control es para que, al calcular las métricas,
# el positivo sea el valor "Yes" o "TRUE" (segundo valor)
lasso_reg_tune <- tune_grid(lasso_wf,
                            attrcv_tun,
                            grid = grid_lasso_tune,
                            metrics = metr_cl,
                            control = control_grid(event_level = "second"))
# Selección mejores hiperparámetros y actualización workflow
best_roc_lasso <- select_best(lasso_reg_tune, metric = "roc_auc")
best_lasso_wf <- finalize_workflow(lasso_wf, best_roc_lasso)
# Tuneado hiperparámetros random forest con CV
rf_tune_grid <- tune_grid(rf_wf,
                          attrcv_tun,
                          grid = grid_rf,
                          metrics = metr_cl,
                          control = control_grid(event_level = "second"))
best_roc_rf <- select_best(rf_tune_grid, metric = "roc_auc")
best_rf_wf <- finalize_workflow(rf_wf, best_roc_rf)
```

Una vez tuneados los hiperparámetros y actualizados los workflows con los mejores valores del tuneado, se puede realizar otro proceso de validación cruzada (utilizando el segundo objeto CV generado) para realizar un assessment predictivo de Lasso y random forest. Para ello se puede utilizar la función `fit_resamples()` con los workflows actualizados:

```
ctrl_assess <- control_resamples(event_level = "second")
asses_best_lasso_wf <- fit_resamples(best_lasso_wf,
                                     resamples = attrcv_asse,
                                     metrics = metr_cl,
                                     control = ctrl_assess)
asses_best_rf_wf <- fit_resamples(best_rf_wf,
                                  resamples = attrcv_asse,
                                  metrics = metr_cl,
                                  control = ctrl_assess)
# Muestra las métricas de ambos assessment
collect_metrics(asses_best_lasso_wf)
collect_metrics(asses_best_rf_wf)
```

La pregunta relativa a cómo evaluar la posible existencia de overfitting en los algoritmos planteados, una vez realizado el assessment con validación cruzada anterior, se pueden estimar los modelos tuneados con toda la mues-



tra de training y evaluar sobre los datos de test. Esto es hecho de forma automática con la función `last_fit()`. Las métricas en el test deberían ser similares a las obtenidas en el assessment de validación cruzada anterior.

```
# Estimación con todo el training y evaluación sobre test para Lasso
ctrl_lastfit <- control_last_fit(event_level = "second")
# Se incluye el objeto attr_split que
# proporcionaba el splitting inicial (incluye training y test)
final_fit_lasso <- best_lasso_wf |>
                     last_fit(attr_split,
                              metrics = metr_cl,
                              control = ctrl_lastfit)
collect_metrics(final_fit_lasso)
# Estimación con todo el training y evaluación sobre test para Random Forest
final_fit_rf <- best_rf_wf |>
                     last_fit(attr_split,
                              metrics = metr_cl,
                              control = ctrl_lastfit)
collect_metrics(final_fit_rf)
```

Si no hay overfitting las métricas obtenidas aquí deberían ser similares a las del assessment anterior con validación cruzada.

Para terminar con la resolución de este ejercicio, se comenta brevemente una diferencia clave entre ambas propuestas, la de Python y la de R, que se basan en esencia en el mismo razonamiento, y que radica en la manera en que funciona la fase de preprocesamiento. En R se lleva a cabo en cada etapa de validación cruzada, mientras que en la resolución en Python, se hace una vez antes del entrenamiento del modelo y no en cada etapa. De esta forma, en la propuesta de R la transformación step_BoxCox(all_numeric_predictors()) se ejecuta en cada subconjunto de entrenamiento dentro de la validación cruzada. Como el parámetro λ se ajusta según los datos de entrada, este valor puede cambiar en cada pliegue porque cada conjunto de entrenamiento es ligeramente diferente. Análogamente, step_normalize(all_numeric_predictors()) se recalcula en cada pliegue, lo que significa que la media y la desviación estándar usadas para centrar y escalar los datos pueden variar en cada itera-



ción. Lo mismo ocurre con la codificación de variables categóricas (step_dummy(), step_other()) y la eliminación de variables de baja varianza (step_nzv()).

Por el contrario, siguiendo la propuesta del código Python, la transformación Box-Cox, la normalización y la codificación se realizan en la etapa de preprocesamiento antes del entrenamiento del modelo. Esto significa que el valor λ de la transformación Box-Cox se calcula una sola vez y permanece fijo durante la posterior validación cruzada. También permanecen fijas la media y la desviación estandar usadas para entrar y escalar los datos y la codificación de las variables categóricas.

El hecho de que en Python el preprocesamiento se haga una vez antes de entrenar el modelo puede hacer que el mismo sea más consistente entre pliegues, pero menos adaptable a cambios en los datos. Por otro lado, en R, al realizar el preprocesimiento en cada iteración, se obtiene una transformación más flexible, aunque podría introducir mayor variabilidad en los coeficientes finales.

Ejercicio 6

Considere las siguientes relaciones correspondientes a una base de datos de una determinada universidad:

- `Student(snum: integer, sname: string, major: string, level: string, age: integer)`
- `Class(name: string, meets at: string, room: string, fid: integer)`
- `Enrolled(snum: integer, cname: string)`
- `Faculty(fid: integer, fname: string, deptid: integer)`



La relación `Enrolled` tiene un registro por cada par estudiante-clase tal que el estudiante está inscrito en la clase. El atributo major en la relación `Student` se refiere al grado que cursa el estudiante mientras que el atributo `fname` en la relación `Faculty` se refiere al nombre del profesor que imparte el curso. No se deben imprimir duplicados en ninguna de las respuestas correspondientes a consultas SQL.

a) Crear las tablas mediante comandos SQL incluyendo las claves y las referencias de integridad necesarias.

b) Escribir código SQL para imprimir los nombres de todos los estudiantes de level = `'JR'` que están inscritos en una clase impartida por `'I. Teach'`

c) Escribir código SQL para imprimir la edad del estudiante de más edad que esté inscrito en `'Historia'` o en un curso impartido por `'I. Teach'`

d) Escribir código SQL para imprimir los nombres de los estudiantes inscritos en el mayor número de clases.

Solución

Antes de nada, conviene representar el esquema con las tablas y las relaciones entre ellas, ver Figura 4.17.

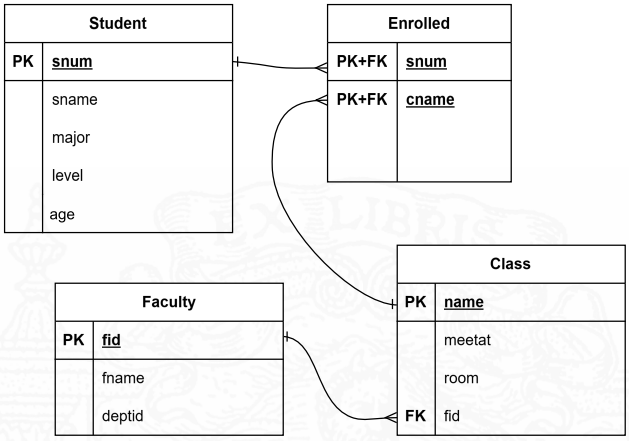


Figura 4.17: Esquema para la base de datos de la universidad.

Apartado a)

Los comandos SQL serían:

```
CREATE TABLE Faculty (
    fid INTEGER,
    fname TEXT,
    deptid INTEGER,
    PRIMARY KEY(fid)
)

CREATE TABLE Student (
    snum INTEGER,
    sname TEXT,
    major TEXT,
    level TEXT,
    age INTEGER,
    PRIMARY KEY(snum)
)
```




```
CREATE TABLE Class (
    name TEXT,
    meetsat TEXT,
    room BLOB,
    fid INTEGER,
    PRIMARY KEY(name),
    FOREIGN KEY(fid) REFERENCES Faculty(fid)
)

CREATE TABLE Enrolled (
    snum INTEGER,
    cname TEXT,
    PRIMARY KEY(snum, cname),
    FOREIGN KEY(snum) REFERENCES Student(snum),
    FOREIGN KEY(cname) REFERENCES Class(name)
)
```

El campo `deptid` de `Faculty` parece pensado como clave ajena a otra tabla de departamentos. Pero como esa tabla no figura en el esquema, lo ignoramos.

Apartado b)

```
SELECT DISTINCT S.sname
FROM Faculty F
JOIN Class C ON (F.fname='I.Teach' AND F.fid = C.fid)
JOIN Enrolled E ON (E.cname = C.name)
JOIN Student S ON (S.snum = E.snum AND S.level='JR')
```

Los filtrados de `F.fname='I.Teach'` y `S.level='JR'` se podrían mover a una cláusula `WHERE` tras los joins, pero resulta más eficiente la solución del



texto porque se cruzan menos filas durante la búsqueda.

Apartado c)

```sql
SELECT MAX(S.age) as Maxedad
FROM Faculty F
JOIN Class C ON F.fid = C.fid
JOIN Enrolled E ON E.cname = C.name
JOIN Student S ON S.snum = E.snum
WHERE C.name = 'Historia' or F.fname='I.Teach'
```

Apartado d)

Esta query es lo suficientemente compleja como para usar una cláusula `WITH`, que permite hacer depender la query principal de subqueries auxiliares que reciben un nombre.

```sql
WITH StudentClassCounts AS (
    SELECT E.snum, S.sname, COUNT(E.cname) AS classcount
    FROM Enrolled E
    JOIN Student S ON E.snum = S.snum
    GROUP BY E.snum, S.sname
),
MaxClassCount AS (
    SELECT MAX(classcount) AS maxcount
    FROM StudentClassCounts
)
SELECT sname
FROM StudentClassCounts
WHERE classcount = (SELECT maxcount FROM MaxClassCount);
```



Ejercicio 7

Ciertos mensajes llegan a una instalación de comunicaciones de acuerdo con un proceso Poisson a razón de 2 por hora. La instalación consta de tres canales, cada mensaje llega a un canal libre, si los tres están libres, o se pierde si todos los canales están ocupados. El tiempo que los mensajes permanecen en un canal es una variable aleatoria que depende de las condiciones meteorológicas al momento de llegada. Si el mensaje llega cuando las condiciones son «buenas», su tiempo de procesamiento es una variable aleatoria con función de distribución:

$$F(x) = x, \quad 0 < x < 1.$$

Mientras que si las condiciones son «malas», su tiempo de procesamiento tiene la función de distribución:

$$F(x) = x^3, \quad 0 < x < 1.$$

Al principio las condiciones son buenas y, posteriormente, se van alternando en períodos buenos y malos: los buenos tienen una duración fija de 2 horas y los malos de una hora.

a) Realice un esbozo de cuáles serían los pasos para simular la distribución del número de mensajes perdidos hasta el instante $T = 100$.

b) Escriba el pseudocódigo, o el código en algún lenguaje imperativo (C++, Java, Python o R), que permita obtener, mediante simulación, la distribución del número de mensajes perdidos hasta el instante $T = 100$. Se puede utilizar una función `random()` que genere valores con distribución $U(0, 1)$.



Solución

Apartado a)

La simulación se puede estructurar en los siguientes pasos:

1. Simulación de los tiempos de llegada de mensajes. Partiendo de $T = 0$, y entendiendo que las unidades son horas, cada tiempo de llegada resulta de incrementar el anterior con el sorteo de una variable exponencial de $\lambda = 2$, todas ellas independientes entre sí. Esto se debe a que en un proceso de Poisson los tiempos entre eventos siguen dicha distribución.

   Se generan tiempos sucesivamente hasta obtener uno posterior a $T = 100$. De esta manera se obtiene la lista ordenada de los tiempos de llegada de mensajes. Nótese que son tiempos absolutos, es decir el número de horas desde el inicio del proceso.

2. Simulación del tiempo de resolución de cada mensaje a partir del de llegada, si es que fuera procesado. Los tiempos de salida se obtienen añadiendo un incremento al de llegada según una simulación de la variable aleatoria correspondiente a la climatología. Con buena climatología dicho incremento sigue una $U(0, 1)$, y con mala por el método de la transformada inversa, la raíz cúbica de una $U(0, 1)$.

   Con esto podemos emparejar cada tiempo de llegada de un mensaje con su tiempo de finalizado (si es que fuera procesado). Nótese que también son tiempos absolutos, es decir el número de horas desde el inicio del proceso, y que no tienen por qué estar ordenados entre sí.

3. Dada la lista `[(t_lleg1, t_sal1), (t_lleg2, t_sal2),...]` de tiempos de llegada y finalizado simulados, veamos cómo computar el número de rechazos. Se mantiene una lista (en el código será `cola_sals`) con



los tiempos de finalizado de los procesos en canales ocupados. Como al principio todos está libres, se inicializa vacía. Luego se van procesando iterativamente los mensajes `(t_llegk, t_salk)` en su orden de llegada:

a) Primeramente, se considera si hay canales que estuvieran ocupados al procesar el mensaje anterior, pero que han acabado ya su procesamiento y se encuentran por tanto libres al llegar el k-ésimo. Esos canales deberían actualizarse como disponibles, y para ello de la lista `cola_sals` se eliminan los tiempos que sean menores que `t_llegk`, si los hubiera.

   Por ejemplo, si el siguiente mensaje a tener en cuenta es `t_llegk=23.4`, `t_salk=24.9` y teníamos `cola_sals=[21.8, 22.9, 25.2]`, deberíamos actualizar esa lista reseteándola como `cola_sals=[25.2]`.

b) Si la lista `cola_sals` sigue teniendo longitud tres, eso significa que todos los canales permanecen ocupados en el momento de recepción de este k-ésimo mensaje, y por tanto no se puede aceptar, con lo que se incrementa el conteo de rechazados. En otro caso, debemos reflejar que uno de los canales se ocupa de él, y esto se consigue incluyendo su tiempo de finalizado `t_salk` en la lista `cola_sals`.

   En el ejemplo anterior, el mensaje se acepta y resetearíamos `cola_sals=[25.2]` como `cola_sals=[25.2, 24.9]`.

Lo anterior puede encapsularse en una función `simul_suelta`, que devuelva finalmente ese número de rechazos. Es decir, esta función es capaz de obtener una única realización de la variable aleatoria *número de rechazos* cuya distribución nos piden.

4. Para conocer la distribución de la variable *número de rechazos* pedida,



vamos a obtener una muestra suficientemente grande. Sea `nsims` su tamaño. Programamos una función que llama `nsims` veces a la función `simul_suelta`, y almacena los distintos resultados.

Apartado b)

Una posible implementación en Python:

```
import numpy as np
import numpy.random as ra

def sorteo_tiempos_llegada(lam=2, horiz=100):
t = 0
ts = []
while True:
    t += -np.log(ra.uniform()) / lam
    if t > horiz:
        return ts
    ts.append(t)

def sorteo_tiempo_salida(t_lleg):
    res = t_lleg % 3
    if res > 2:  # Condiciones malas
        return t_lleg + ra.uniform() ** (1/3)
    else:  # Condiciones buenas
        return t_lleg + ra.uniform()

def sorteo_intervalos(lam=2, horiz=100):
    return [(t_lleg, sorteo_tiempo_salida(t_lleg))
                for t_lleg in sorteo_tiempos_llegada(lam, horiz)]

def simul_suelta(intervs, ncolas=3):
    cola_sals = []
```




```
        nrechaz = 0
        for (t_lleg, t_sal) in intervs:
            cola_sals = [t_s for t_s in cola_sals if t_s > t_lleg]
            if len(cola_sals) < ncolas:
                cola_sals.append(t_sal)
            else:
                nrechaz += 1
        return nrechaz

    def simul_distribucion(nsims, ncolas=3):
        return [simul_suelta(sorteo_intervalos(), ncolas)
                                                    for _ in range(nsims)]

    # Codigo para probar el muestreo:
    dist = simul_distribucion(nsims=100, ncolas=3)
    print(f"Una muestra de la distribucion: {dist}")
```

Una representación de la simulación de la recepción de mensajes es la que aparece en la Figura 4.18.

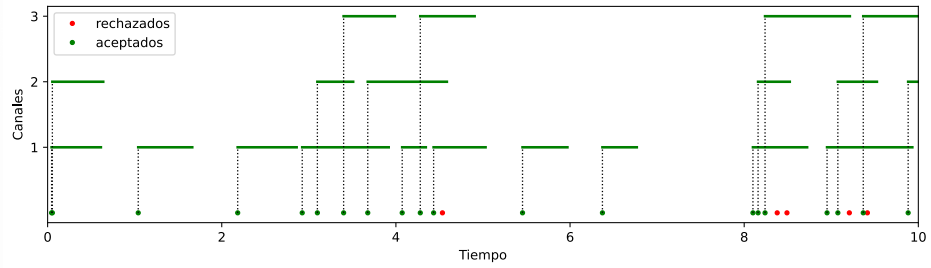


Figura 4.18: Simulación de la recepción de mensajes. En verde, los mensajes aceptados y el intervalo temporal en el canal que los procesa. En rojo, los tiempos de llegada de los rechazados. Solo aparecen las 10 primeras horas.

En cuanto a la distribución pedida (y obtenida por la función `simul_distribucion`), una posible distribución de los rechazos es la mostrada en la Figura 4.19.

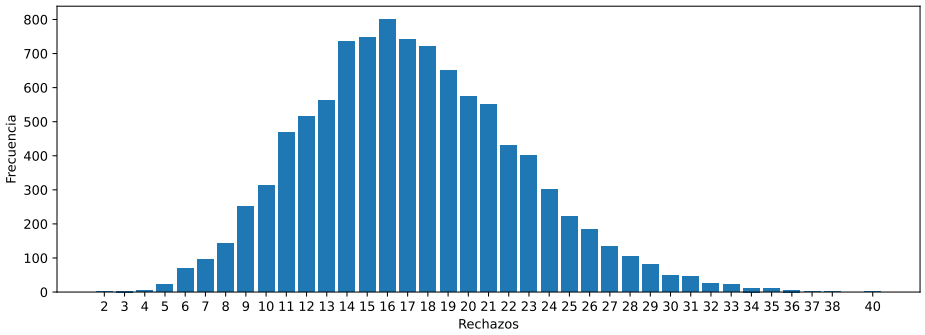


Figura 4.19: Distribución obtenida del número de rechazos al realizar la simulación 10,000 veces.

Ejercicio 8

Escriba el pseudocódigo, o el código en algún lenguaje imperativo (C++, Java, Python o R), de un algoritmo para calcular la media recortada (*trimmed mean*). El algoritmo debe aceptar como parámetros de entrada una secuencia de números y el parámetro alpha de la media recortada. No se puede utilizar ninguna función tipo *sort* predefinida sino que la ordenación debe estar incluida en el algoritmo y ser eficiente. Justificar la complejidad algorítmica de todos los pasos del procedimiento descrito.

Solución

La media recortada de una serie de datos con parámetro α consiste en eliminar de la serie los datos cuyo percentil es menor que α o mayor que $1-\alpha$, y posteriormente calcular la media de los restantes.

Suponemos de ahora en adelante que n denota el tamaño de la secuencia cuya media recortada se desea calcular.



Estrategia naive

Un primer enfoque para este problema podría ser ordenar los valores, y luego promediar aquellos que están entre la posición $n\alpha$ y la $n(1-\alpha)$. En términos de complejidad, la ordenación podría hacerse con un *quicksort* o un *mergesort* con un orden asintótico de $O(n \ln n)$ operaciones, luego simplemente habría que recorrer la sublista entre las posiciones descritas (para sumar los valores), lo que supone $O(n)$, con lo que la complejidad del algoritmo completo seguiría siendo $O(n \ln n)$.

Estrategia mejorada

Sin embargo, hay un enfoque mejor que no requiere la ordenación de la lista, sino que es capaz de obtener la media recortada con $O(n)$ operaciones. Es el que vamos a presentar en esta sección. En el problema de la mediana (Ejercicio 8, cuarto examen A1 del turno libre de ciencia de datos, 2.2) ya presentamos un algoritmo lineal capaz de encontrar, dentro de una lista posiblemente desordenada, el valor con ranking `k` para un `k` cualquiera dado. En concreto, lo implementamos con la función `busca_kesimo`. Dada la lista `nums` de valores y el parámetro `alpha`, el problema de la media recortada puede descomponerse en tres fases:

1. Buscar en la lista `nums` el valor cuyo ranking es el entero más próximo a $n\alpha$. Aquí n representa la cardinalidad de `nums`. Esto se hace a través de `busca_kesimo`, en tiempo $O(n)$, y da lugar al umbral μ_{izq}.

2. Buscar en la lista `nums` el valor cuyo ranking es el entero más próximo a $n(1-\alpha)$. Esto se hace a través de `busca_kesimo`, en tiempo $O(n)$, y da lugar al umbral μ_{der}.

3. Se recorren todos los valores en `nums`, llevando actualizada la suma



parcial de aquellos que están entre μ_{izq} y μ_{der}, y su conteo. Esto naturalmente se hace en tiempo $O(n)$.

4. La media recortada se devuelve como esa suma parcial una vez finalizado el recorrido de `nums`, dividiendo por el número de valores sumados. Esta operación es $O(1)$.

Al ser una secuencia de operaciones, la complejidad de la secuencia es el máximo de las complejidades de sus pasos, es decir, resulta un algoritmo $O(n)$.

Una posible implementación en Python:

```python
import numpy.random as ra


def trocear(nums):
    pivote = ra.choice(nums)
    menores, iguales, mayores = [], [], []
    for e in nums:
        if e < pivote:
            menores.append(e)
        elif e == pivote:
            iguales.append(e)
        else:
            mayores.append(e)
    trozos = [menores, iguales, mayores]
    return [tr for tr in trozos if tr] # Devuelve los trozos no vacios


def busca_kesimo(nums, k):
    """Devuelve el k-esimo elemento en rango de nums. 0 para el primero,
    1 para el segundo..."""
    # Caso base 1 de la recursion: todos los numeros son iguales
    if all(nums[0] == e for e in nums):
        return nums[0]
```




```
    # Caso base 2 de la recursion: el dato buscado es un extremo
    if k == 0:
        return min(nums)
    if k == len(nums) - 1:
        return max(nums)
    # Caso recursivo
    for tr in trocear(nums):
        if len(tr) > k: # El numero buscado esta en el trozo tr
            return busca_kesimo(tr, k)
        else: # El numero buscado esta en alguno de los trozos restantes
            k -= len(tr)


def media_recortada(nums, alpha):
    """Calcula la media recortada, que elimina el porcentaje alpha de datos m^^c3^^a1s
    # Busca los umbrales del intervalo que se selecciona
    umbral_izq = busca_kesimo(nums=nums, k=int(len(nums) * alpha))
    umbral_der = busca_kesimo(nums=nums, k=int(len(nums) * (1 - alpha)))
    # Filtra los elementos de nums entre los umbrales, y los promedia
    total = 0
    ndatos = 0
    for e in nums:
        if umbral_izq <= e <= umbral_der:
            total += e
            ndatos += 1
    return total / ndatos
```



EX LIBRIS
INSTITUTO NACIONAL
DE
ESTADISTICA



Capítulo 5

Año 2023. Promoción interna.

"If the people are doubting how far you can go,
go so far that you can't hear them anymore."

—Michele Ruiz



Antoni Gaudí (1852-1926) se ha convertido en una figura cumbre de la arquitectura internacional. A lo largo del siglo XX su obra ha pasado del olvido más absoluto al reconocimiento más entusiasta, y hoy es reconocido como un innovador y como un gran artista. Actualmente, su obra constituye una de las aportaciones más relevantes de la arquitectura catalana de los siglos XIX y XX. La Casa Batlló se sitúa en el número 43 de Paseo de Gracia, en Barcelona. D. Josep Batlló concedió total libertad creativa a Antoni Gaudí, encargándole unas obras que en principio consistían en derribar el edificio. Sin embargo, gracias la audacia de Gaudí, se descartó el derrumbe de la Casa, llevando a cabo una reforma integral entre 1904 y 1906. El arquitecto cambió completamente la fachada, redistribuyendo la tabiquería interior, ampliando el patio de luces y haciendo de su interior una auténtica obra de arte.

$2023 = 17^2 \cdot 7$

2023 is not a de Polygnac number because $2023 - 2^9 = 1511$ is a prime.



5.1. Soluciones del segundo examen del Cuerpo Superior de Estadísticos del Estado de 2023. Promoción interna.

Este examen es idéntico al examen de turno libre del año 2023 salvo el ejercicio número 4, o cuestión 4, relativa a muestreo. El ejercicio 1 es el 1 de libre, el 2 coincide con el 2 de libre, el 3 es igual al 3 de libre, el 4 es el que comentamos que es diferente, que resolvemos en esta sección, el ejercicio 5 es el mismo que el 7 de libre y el ejercicio 6 es el 6 de libre.

Cuestión 4. Una empresa especializada en servicios de seguridad busca estimar el total de viviendas equipadas con sistemas de seguridad en una determinada área urbana, con el propósito de desarrollar estrategias de marketing dirigidas a promover sus servicios. Para lo cual se seleccionan 80 conglomerados utilizando muestreo aleatorio simple sin reposición y probabilidades iguales con una fracción de muestreo de $\frac{1}{50}$. El tamaño de cada conglomerado es de 10 viviendas. Asimismo, para cada conglomerado i seleccionado en la muestra, se denota por A_i al número total de viviendas que tienen contratadas servicios de seguridad. Los resultados de la investigación son los siguientes:

$$\sum_{i=1}^{80} A_i = 370 \quad \sum_{i=1}^{80} A_i^2 = 2536$$

Se pide:

a) Una estimación insesgada del total de viviendas con servicios de seguridad contratados en el área urbana y su error de muestreo relativo o coeficiente de variación.

b) Una estimación del efecto de diseño del muestreo descrito comparado con el muestreo aleatorio simple sin reposición.

c) Estimar el coeficiente de correlación intraconglomerados o coeficiente de homogeneidad de los conglomerados.

Solución.

a) Tenemos que el número de conglomerados de la muestra son $n = 80$ y como la fracción de muestreo es de $f = \frac{n}{N} = \frac{1}{50}$ entonces despejando obtenemos que el número de conglomerados de la población es $N = 50 \cdot 80 = 4000$.



Además sabemos que en cada conglomerado hay $\overline{M} = 10$ viviendas, siendo A_i el número de viviendas con seguridad en el conglomerado i-ésimo.

Para obtener una estimación insesgada del total de viviendas con servicios de seguridad contratados en el área urbana, $\widehat{A}$, utilizaremos las fórmulas habituales para conglomerados con el mismo tamaño $\overline{M} = 10$. Sabiendo que se obtiene una muestra aleatoria simple (sin reemplazamiento) de $n = 80$ conglomerados con probabilidades iguales, un estimador insesgado del total de clase vendrá dado por:

$$\widehat{A}_{MCU} = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{\overline{M}} A_{ij} = \frac{N}{n}\sum_{i=1}^{n} A_i = \frac{4000}{80}\cdot 370 = 50\cdot 370 = 18\,500$$

donde el subíndice MCU significa muestreo de conglomerados unietápico o sin submuestreo.

Para el cálculo de la estimación insesgada de su error de muestreo se obtiene de la raíz cuadrada de la estimación insesgada de la varianza de $\widehat{A}_{MCU}$, que se calcula , teniendo en cuenta que:

$$P_i = \frac{\sum_{j=1}^{\overline{M}} A_{ij}}{\overline{M}} = \frac{A_i}{\overline{M}}$$

$$\overline{P} = \frac{1}{n}\sum_{i=1}^{n} P_i = \frac{1}{n\overline{M}}\sum_{i=1}^{n} A_i = \frac{1}{80\cdot 10}\cdot 370 = \frac{37}{80}$$

$$\sum_{i=1}^{n} A_i = n\cdot\overline{M}\cdot\overline{P}$$

como:

$$\begin{aligned}\widehat{Var}_{MCU}\left(\widehat{A}_{MCU}\right) &= N^2\overline{M}^2\widehat{Var}\left(\widehat{P}\right) = \\ &= \frac{N^2\overline{M}^2(1-f)}{n(n-1)}\sum_{i=1}^{n}\left(P_i-\overline{P}\right)^2 = \\ &= \frac{N^2\overline{M}^2(1-f)}{n(n-1)}\left[\frac{1}{\overline{M}}\right]^2\left(\sum_{i=1}^{n}A_i^2 - n\overline{M}^2\cdot\overline{P}^2\right) = \\ &= \frac{N^2(1-f)}{n(n-1)}\left(\sum_{i=1}^{n}A_i^2 - n\overline{M}^2\cdot\overline{P}^2\right) = \end{aligned}$$

$$= \frac{4000^2\left(1-\frac{1}{50}\right)}{80\cdot(79)}\left(2536-80\cdot 10^2\cdot\left(\frac{37}{80}\right)^2\right) =$$
$$= 2\,046\,215{,}1898$$

Así, la estimación insesgada del error de muestreo del total de la población estimado será:

$$\widehat{\sigma}_{MCU}\left(\widehat{A}_{MCU}\right) = \sqrt{\widehat{Var}_{MCU}\left(\widehat{A}_{MCU}\right)} = \sqrt{2\,046\,215{,}1898} = 1430{,}459.$$

con lo cual el error de muestreo relativo o coeficiente de variación será:

$$cv = \frac{\widehat{\sigma}_{MCU}\left(\widehat{A}_{MCU}\right)}{\widehat{A}_{MCU}} = \frac{1430{,}459}{18\,500} = 0{,}07732$$

es decir, un $7{,}732\,\%$

b) Para llevar a cabo la comparación de la eficiencia del muestreo de conglomerados y el muestreo aleatorio simple una posibilidad es calcular la varianza en el caso de realizar muestreo aleatorio simple y compararla con la varianza obtenida en el apartado anterior. Esta comparación es posible ya que ambos estimadores son insesgados. Para un muestreo aleatorio simple el estimador del total se calcula como:

$$\hat{A}_{MAS} = N\cdot\overline{M}\cdot\widehat{P}_{MAS} = 4000\cdot 10\cdot\frac{\sum_{i=1}^{80}A_i}{n\cdot\overline{M}} = 4000\cdot 10\cdot\frac{370}{80\cdot 10} =$$
$$= \frac{40\,000}{800}\cdot 370 = 50\cdot 370 = 18\,500$$

y la varianza sería con la fórmula habitual, pero interpretando de forma correcta los datos de la misma según la información que nos proporciona el



enunciado:

$$\begin{aligned}\widehat{Var}_{MAS}(\hat{A}_{MAS}) &= N'^2\left(1-\frac{n'}{N'}\right)\frac{\hat{P}\hat{Q}}{n'-1} = \\ &= N^2\overline{M}^2\left(1-\frac{n\cdot\overline{M}}{N\overline{M}}\right)\frac{\hat{P}\hat{Q}}{n\cdot\overline{M}-1} = \\ &= 4000^2\cdot 10^2\cdot\left(1-\frac{80\cdot 10}{4000\cdot 10}\right)\frac{\frac{37}{80}\cdot\frac{43}{80}}{80\cdot 10-1} = \\ &= 16\cdot 10^8\cdot\frac{3920}{4000}\frac{37\cdot 43}{799\cdot 80^2} = \\ &= 487\,853{,}5669\end{aligned}$$

La eficiencia relativa del muestreo por conglomerados respecto al aleatorio simple se calcula como el cociente entre sus varianzas:

$$eff(\hat{A}_{MCU},\hat{A}_{MAS}) = \frac{\widehat{Var}_{MCU}(\hat{A}_{MCU})}{\widehat{Var}_{MAS}(\hat{A}_{MAS})} = \frac{2\,046\,215{,}1898}{487\,853{,}5669} = 4{,}194$$

lo que muestra que el error muestral en muestreo de conglomerados es mayor que en el muestreo aleatorio simple, debido a que el efecto del diseño es mayor que uno. La razón de este comportamiento está en la propia estructura de este tipo de muestreo en el que las unidades muestrales están formadas por conglomerados de unidades, mientras que en el aleatorio simple las unidades son 'individuos' que, para un número dado de elementos de la muestra, aportan, generalmente mayor precisión.

Los errores de la estimación en el muestreo por conglomerados suelen ser mayores que en el muestreo aleatorio simple y se pueden analizar en comparación con los que se derivarían del mismo a través de la fórmula que detallamos en el siguiente apartado del ejercicio en el que aparece el coeficiente de correlación intraconglomerados.

c) La eficiencia del muestreo por conglomerados respecto a la del muestreo aleatorio simple también se puede estudiar calculando el coeficiente de correlación intraconglomerados. Este se define como el coeficiente de correlación

lineal entre todos los pares de valores posibles de la variable en estudio medidos sobre unidades dentro del mismo conglomerado. Es decir, es una medida de la homogeneidad en el interior de los conglomerados de forma que el muestreo por conglomerados será preferido al muestreo aleatorio simple cuando el coeficiente de homogeneidad intraconglomerados sea lo más pequeño posible, ya que lo ideal es la heterogeneidad dentro de los conglomerados. La estimación de este coeficiente, $\widehat{\delta}$, viene dada por:

$$\widehat{\delta} = \frac{\widehat{S}_b^2 - \widehat{S}_0^2}{\left(\overline{M} - 1\right)\widehat{S}_0^2}$$

donde

$$\widehat{S}_0^2 = \frac{(N-1)\,\widehat{S}_b^2 + \left(N\overline{M} - N\right)\widehat{S}_W^2}{N\overline{M} - 1}$$

$$\widehat{S}_b^2 = \frac{\overline{M}}{n-1}\sum_{i=1}^{n}\left(P_i - \overline{P}\right)^2$$

$$\widehat{S}_W^2 = \frac{\sum_{i=1}^{n}\overline{M}P_i\left(1-P_i\right)}{n\left(\overline{M}-1\right)}$$

La varianza del muestreo monoetápico de conglomerados se puede expresar en función de la varianza del muestreo aleatorio simple y el coeficiente de correlación intraconglomerado:

$$\widehat{Var}_{MCU}\left(\widehat{P}\right) = \widehat{Var}_{MAS}\left(\widehat{P}\right)\left[1 + \left(\overline{M}-1\right)\widehat{\delta}\right]$$

y la comparación de la eficiencia entre muestreo monoetápico de conglomerados y muestreo aleatorio simple podría resumirse de la siguiente forma:

- si $\widehat{\delta} > 0 \Rightarrow$ conglomerados peor que aleatorio simple
- si $\widehat{\delta} = 0 \Rightarrow$ conglomerados igual que aleatorio simple
- si $\widehat{\delta} < 0 \Rightarrow$ conglomerados mejor que aleatorio simple



Veamos,

$$\widehat{S}_b^2 = \frac{\overline{M}}{n-1}\sum_{i=1}^{n}\left(P_i-\overline{P}\right)^2 = \frac{\overline{M}}{n-1}\left(\frac{1}{\overline{M}}\right)^2\left[\sum_{i=1}^{n}A_i^2 - n\overline{M}^2\overline{P}^2\right] =$$

$$= \frac{1}{(n-1)\,\overline{M}}\left[\sum_{i=1}^{n}A_i^2 - n\overline{M}^2\overline{P}^2\right] =$$

$$= \frac{1}{(n-1)\,\overline{M}}\left[\sum_{i=1}^{n}A_i^2 - n\overline{M}^2\left(\frac{1}{n\overline{M}}\sum_{i=1}^{n}A_i\right)^2\right] =$$

$$= \frac{1}{79\cdot 10}\left[2536 - 80\cdot 10^2\cdot\left(\frac{1}{80\cdot 10}\cdot 370\right)^2\right] = \frac{3299}{3160} = 1,0439$$

$$\widehat{S}_W^2 = \frac{1}{n\left(\overline{M}-1\right)\overline{M}}\left[\overline{M}\sum_{i=1}^{n}A_i - \sum_{i=1}^{n}A_i^2\right] =$$

$$= \frac{1}{80\cdot 10\cdot(10-1)}\left[10\cdot 370 - 2536\right] = \frac{1164}{7200} = 0,16161616$$

$$\widehat{S}_0^2 = \frac{(N-1)\,\widehat{S}_b^2 + \left(N\overline{M}-N\right)\widehat{S}_W^2}{N\overline{M}-1} =$$

$$= \frac{(4000-1)\,1,0439 + (4000\cdot 10 - 4000)\,0,16161616}{4000\cdot 10 - 1} =$$

$$= 0,24988$$

Y por tanto, la estimación del coeficiente de correlación intraconglomerados es:

$$\widehat{\delta} = \frac{\widehat{S}_b^2 - \widehat{S}_0^2}{\left(\overline{M}-1\right)\widehat{S}_0^2} = \frac{1,0439 - 0,24988}{(10-1)\,0,24988} = 0,3531.$$

Como se cumple que $\widehat{\delta} > 0$, esto confirma que el muestreo por conglomerados no es preferible al muestreo aleatorio simple.

Sabemos que:

$$\widehat{Var}_{MCU}\left(\widehat{P}\right) = \widehat{Var}_{MAS}\left(\widehat{P}\right)\left[1 + \left(\overline{M}-1\right)\widehat{\delta}\right]$$

y por tanto

$$\widehat{Var}_{MCU}\left(\widehat{A}\right) = N^2\cdot\overline{M}^2\cdot\widehat{Var}_{MCU}\left(\widehat{P}\right) =$$



$$= N^2 \cdot \overline{M}^2 \cdot \widehat{Var}_{MAS}\left(\widehat{P}\right)\left[1 + \left(\overline{M} - 1\right)\widehat{\delta}\right] =$$
$$= \widehat{Var}_{MAS}\left(\widehat{A}\right)\left[1 + \left(\overline{M} - 1\right)\widehat{\delta}\right] =$$
$$= \widehat{Var}_{MAS}\left(\widehat{A}\right)\left[1 + (10 - 1) \cdot 0,3531\right] =$$
$$= \widehat{Var}_{MAS}\left(\widehat{A}\right) \cdot 4,177$$

lo que confirma lo que indicábamos en el anterior apartado respecto al efecto del diseño.



5.2. Soluciones del cuarto examen del Cuerpo Superior de Estadísticos del Estado de 2023. Especialidad ciencia de datos. Promoción interna.

Este examen es el mismo que se ha puesto para el turno libre, por lo que las soluciones del mismo ya se han podido desarrollar en la sección 4.3.



EX LIBRIS
INSTITUTO NACIONAL
DE
ESTADISTICA



Bibliografía

Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley.

Agresti, A. (2015). Foundations of Linear and Generalized Linear Models. John Wiley.

Ash, R.B. (1970). Basic Probability Theory. John Wiley.

Aznar, G,A., García, A, y Martín, A.(1994). Ejercicios de econometría I y II. Pirámide

Casella, G. y Berger, R.L. (1992). Statistical Inference (2nd ed.). Duxbury Advanced Series.

Castro, J. (2009). Statistical disclosure control in tabular data (Report DR 2009-11), Universitat Politécnica de Catalunya.

Celestino Rey, F. (2010). Historia de los cuerpos especiales de estadística de la Administración General del Estado, 1860-2010. Visión Libros.

Cuadras, C.M. (2002). Problemas de Probabilidad y Estadística. PPU Barcelona.

Cuadras, C.M. (2012). Nuevos métodos de análisis multivariante.

García, A., Quesada, V. (1998). Lecciones de Cálculo de Probabilidades. Díaz de Santos.

Feller, W. (1966). An Introduction to Probability Theory and Its Applications. Vol. 1. John Wiley.

Feller, W. (1966). An Introduction to Probability Theory and Its Appli-



cations. Vol.2. John Wiley.

Fernández, C.C., y Fuentes, F. (1995). Curso de Estadística Descriptiva. Teoría y Práctica. Madrid: Ariel.

Fernández, F.R. y Mayor Gallego, J. A. (1995). Ejercicios y prácticas de muestreo en poblaciones finitas. EUB. Ediciones Universitarias de Barcelona.

Gibbons, J. and S. Chakraborti (2011). Nonparametric Statistical Inference. Chapman and Hall/CRC; 5ª ed.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Nordholt, E. S., Seri, G. and De Woldf, P. P. (2010). Handbook on statistical disclosure control. ESSNet SDC, Eurostat.

Ibarrola, P., Pardo, L., Quesada, V. (1997). Teoría de la Probabilidad. Síntesis.

Leguina, J. (1989). Fundamentos de Demografía.

Keppel, G. y Zedeck, S. (1989). Data Analysis for Research. Analysis of variance and multiple regression/ correlation approaches. New York: Freeman and Company.

Lequiller, F. y Derek Blades, (2014). Understanding National Accounts, OECD.

Livi Bacci, M. (1993). Introducción a la demografía, Ariel.

Martín-Guzmán, P. y Martín Pliego, F. (1985). Curso Básico de Estadística Económica. Madrid: AC.

Martín Pliego, F.J. (1990). Curso práctico de estadística económica.

Matilla, M; Pérez, P y Sanz, B. (2017). Econometría y Predicción. McGraw-Hill .

Merediz, A. (2004). Historia de la estadística oficial como institución pública en España.

Montero, J., Pardo, L, Morales, D., Quesada, V. (1988). Ejercicios y Pro-



blemas de Cálculo de Probabilidades. Díaz de Santos.

Peña, D. (1987). Estadística. Modelos y métodos 1: Fundamentos. Madrid: Alianza Universidad.

Peña, D. (1987). Estadística. Modelos y métodos 2. Modelos Lineales y Series Temporales. Madrid: Alianza Universidad.

Peña, D. (2002). Analisis de datos multivariantes. McGraw-Hill Interamericana de España/UNED.

Pérez L. C. (2005). Muestreo estadístico: conceptos y problemás resueltos. Alhambra.

Pérez L. C. (2014). Modelos Econometricos Con Datos de Panel: Conceptos y Ejercicios Resueltos. Createspace.

Pfeffermann, D. y Rao, C.R. (2009). Handbook od statisitics. Vol. 29A Sample surveys: design, methods and applications, North-Holland.

Pressat, Roland (1987). "Diccionario de la Demografía". Oikos-Tau.

Pressat, Roland (1983). "El análisis demográfico". Fondo de cultura económica.

Queouille, M. (1949). "Approximation tests of correlation in time series". Journal of the Royal statistical Society, B, 11, 18-84.

Ruíz Maya, L. y Martín Pliego, F.J. (2004). Fundamentos de inferencia estadística Ed. Thompson-Paraninfo.

Särndal, C.-E, Swensson, B. y Wretman, J. (1992). Model assisted survey sampling. Springer.

Särndal, C.-E, y Lündstrom, S. (2005). Estimation in surveys with non-response. Wiley.

Schattschneider, D., Escher, M. C. (2004). M.C. Escher: Visions of symmetry. New York: H.N. Abrams.

Sistema Europeo de Cuentas, SEC 2010, (2013). EUROSTAT (Comisión



Europea).

Statistics Canada (2010). Survey methods and practices. Statistics Canada.

Tukey, J. (1958). "Bias and confidence in not quite large samples". Annals of Mathematical Statistics, 29, 614.

Uriel, E. y Peiró, A. (2000. Introducción al análisis de series temporales. Madrid: AC.

Valliant, R., Dorfmann, A.H. y Royall, R.M. (2000). Finite population sampling and inference: a prediction approach. Wiley.

Valliant, R. (1993). "Post-Stratification and Conditional Variance Estimation". Journal of the American Statistical Association, 88, 89-96.

Vélez, R., Hernández, V. (1995). Cálculo de Probabilidades 1. UNED

Vélez, R., y Garcia, A. (2012). Principios de Inferencia Estadistica. UNED.

Vinuesa, J. (1994). Demografia, análisis y proyecciones (Espacios y sociedades) .

Vinuesa y Puga (2007). Técnicas y ejercicios de demografía.

Wooldridge, J.B. (2001). Econometric Analysis of Cross-Section and Panel Data. MIT Press.

Wooldridge, J.B. (2010). Introducción a la econometría. Un enfoque moderno. Cengage Learning.



Apéndices



EX LIBRIS
INSTITUTO NACIONAL
DE
ESTADISTICA



Apéndice A

Exámenes de problemas de 2022. Turno libre.

A.1. Segundo examen del Cuerpo Superior de Estadísticos del Estado. Parte común. Turno libre.

CUESTIÓN 1

El marco muestral contiene 70 empresas dedicadas al comercio, agrupadas en 2 estratos de tamaño, medido por el número de asalariados.

(a) Calcule el tamaño muestral en cada estrato, necesario para estimar el total de asalariados, aplicando afijación óptima, con un error de muestreo relativo prefijado del 1,5 %, usando el estimador estratificado de expansión simple sin reemplazamiento.

Para ello, se dispone de la siguiente información:

Estrato de tamaño	Tramo de asalariados	Tamaño de la población (N_h)	X= Total de asalariados	Desviación típica de X (S_h)
1	[1,9]	50	250	1,1
2	[10,500]	20	850	8,5
Total		70	1100	

(b) La muestra, estratificada aleatoria simple sin reemplazamiento, obtenida en el apartado anterior se investiga y responden 4 empresas del estrato de tamaño 1 y 7 del tamaño 2; el resto de unidades investigadas se han negada a colaborar. Suponiendo que la falta de respuesta es aleatoria, se repondera.

Calcule el estimador del total de la cifra de negocios (Y), usando el estimador estratificado de expansión simple sin reemplazamiento reponderado por falta de respuesta y su error de muestreo relativo o coeficiente de elevación estimado. Los datos necesarios son los siguientes:

valores de la variable **y** de las unidades que responde

	Estrato de tamaño h	
unidades	h=1	h=2
i	y_{1i}	y_{2i}
1	40	1700
2	70	2000
3	60	1800
4	30	1750
5		2500
6		1700
7		1800

INSTITUTO NACIONAL DE ESTADISTICA



CUESTIÓN 2

Proporcione un estimador puntual del parámetro $\theta > 0$ a partir de una observación X con función de densidad

$$f(x|\theta) = \begin{cases} 1/\theta, \text{ si } 0 < x < \theta \\ 0, \text{ en otro caso} \end{cases}$$

siendo la distribución a priori

$$g(\theta) = \begin{cases} \theta \exp(-\theta), \text{ si } \theta > 0 \\ 0, \text{ en otro caso} \end{cases}$$

cuando se emplea como función de pérdida

$$l(\theta, a) = 5(\theta - a)^2, a > 0.$$

Aplíquese al caso en que la observación es $6,1$.



CUESTIÓN 3

Consideramos un sistema de representación de números enteros en complemento a dos, con 8 bits. Llamemos Entero8bits a este tipo de datos.

(a) ¿Cuántos enteros distintos pueden representarse en este sistema? ¿Cuál es el entero mínimo y el máximo que se pueden representar en dicho sistema?

(b) Represente, en este sistema, las siguientes cantidades enteras:

- 100
- -100

(c) Observe este fragmento de programa en seudocódigo:

```
n: Entero8bits   (1)
n = 100          (2)
print(n)         (3)
n = n*2          (4)
print(n)         (5)
```

Es decir, paso a paso, este código expresa lo siguiente:

(1) Se declara la variable n como un entero representado internamente como un Entero8bits.
(2) Se asigna la cantidad 100 a dicha variable n.
(3) Se imprime en efecto el valor 100 asignado.
(4) Se duplica el valor de dicha variable, almacenando dicho valor duplicado en la misma.
(5) Se imprime en efecto el nuevo valor.

Al ejecutarlo, esperamos ver en el output estándar, primero, la cantidad 100, y luego la cantidad 200, pero no es así: vemos la cantidad 100 en primer lugar, pero en segundo, un número negativo: "menos cincuenta y tantos". ¿Qué número es éste? ¿A qué se debe este resultado?



CUESTIÓN 4

El Producto Interior Bruto a precios de mercado (PIB) de una economía en el año t-1, medido a precios corrientes, se cifró en 525.000 unidades monetarias (u.m)

(a) Se sabe que, en el año t, el excedente bruto de explotación y renta mixta bruta generados en dicha economía se estimaron en 230.000 u.m.

Así mismo, la remuneración a recibir por los asalariados residentes en el conjunto de dicho año t se cifró en 275.000 u.m. (de ellas, 1.500 provenían de empleadores no residentes). Por otro lado, los empleadores residentes debían abonar 1900 u.m a asalariados no residentes en la economía nacional.

Dados además los datos que figuran en la siguiente tabla sobre impuestos devengados y subvenciones a pagar durante el año en dicha economía, determine el crecimiento económico de la misma en el año t respecto al año t-1 en términos nominales.

IVA	4100
Impuestos y derechos sobre las importaciones (excluido el IVA)	200
Otros impuestos sobre los productos	1700
Otros impuestos sobre las producción	1100
Impuesto corriente sobre la renta	5350
Impuestos sobre el capital	410
Subvenciones a las importaciones	50
Otras subvenciones a los productos	600
Otras subvenciones a la producción	1350

(b) Por otro lado, el índice de volumen encadenado del PIB de dicha economía en los años t-1 y t se ha estimado en 110,7 y 108,1, respectivamente. ¿Cuál fue el crecimiento real de dicha economía en el año t respecto a t-1 (es decir, en términos de volumen)? ¿qué podríamos decir sobre la evolución de los precios de los bienes y servicios producidos en dicha economía? ¿a qué tasa evolucionó el deflactor implícito del PIB?

INSTITUTO NACIONAL DE ESTADISTICA

CUESTIÓN 5

De una población se conocen las siguientes tasas de mortalidad referidas al año 2019.

Grupos de edad	${}_nm_x$
Menos de 1 año	0,00431
De 1 a 4 años	0,00015
De 5 a 19 años	0,00013
De 20 a 34 años	0,00024
De 35 a 49 años	0,00087
De 50 a 64 años	0,00408
De 65 a 79 años	0,01557
De 80 a 94 años	0,10733
De 95 años y más	0,29376

Calcule las siguientes series:

(a) Probabilidades de muerte, ${}_nq_x$ (suponga $a_0 = 0,1$ y ${}_na_x = 0,5$ para los demás casos).

(b) Supervivientes, l_x, y defunciones de la tabla, d_x.

(c) Población estacionaria, L_x (suponga $a_0 = 0,1$ y ${}_na_x = 0,5$ para los demás casos).

(d) Tiempo vivido, ${}_nT_x$.

(e) ¿Cuál fue la esperanza de vida al nacer de esta población en 2019?



CUESTIÓN 6

Sea una población de tamaño N, $U = u_1, u_1, \ldots u_N$ y sean N valores $\pi_1, \pi_2, \ldots, \pi_N$, tal que $0 < \pi_k < 1$ para todo k=1,2...k. Se generan un conjunto de N realizaciones independientes $\alpha_1, \alpha_2, \ldots, \alpha_N$ de una distribución uniforme sobre el intervalo (0,1).

El elemento u_k es seleccionado en la muestra 's' si $\alpha_k < \pi_k$, en caso contrario no se selecciona.

Se pide:

(a) Calcular el valor esperado y varianza del tamaño muestral n_s.

(b) Calcular la probabilidad de que la muestra s tenga al menos 1 elemento.

(c) Sea una población formada por 10 elementos, para la cual se dispone de la información de una variable auxiliar x que toma los siguientes valores:

unidad	x
1	10
2	40
3	15
4	20
5	10
6	15
7	50
8	25
9	60
10	40

Se desea seleccionar una muestra de tamaño 5 mediante el método de selección sistemática con probabilidades proporcionales al tamaño. Suponiendo que el arranque aleatorio seleccionado en el intervalo de muestreo es 35, determine las unidades que entran en la muestra.

INSTITUTO NACIONAL DE ESTADISTICA

CUESTIÓN 7

En un estudio sobre una enfermedad, se eligen al azar 220 personas adultas que no padecen la misma; se comprueba que 60 de ellas no consumen alcohol. De entre 150 personas adultas con la enfermedad, también elegidas al azar, se observa que 15 de ellas no consumen alcohol.

Discuta a partir de qué valor del nivel de significación se rechazaría la hipótesis nula de que el consumo de alcohol no tiene influencia sobre la presencia de la enfermedad, si se dispone de un software con un comando de sintaxis $qnorm(p)$ que proporciona, para una probabilidad p, el correspondiente cuantil de la distribución normal estándar.

INSTITUTO NACIONAL DE ESTADISTICA

7



CUESTIÓN 8

Con la información que se refleja en la tabla siguiente sobre las relaciones exteriores de la economía española en el año 2020, elabore las cuentas del resto del mundo referidas a dicho año.

Desde el año 2012, la economía española ha presentado capacidad de financiación respecto al resto del mundo. ¿Se mantuvo también esa tendencia en el año 2020?

Si dicho saldo se calculara como resultado de la cuenta corriente y de capital de la balanza de pagos, conforme al 6º Manual de Balanza de Pagos y Posición de Inversión Internacional, ¿la conclusión sería la misma?

Operaciones con el resto del mundo	**2020**
Exportaciones de bienes	265450
Importaciones de bienes (CIF)	281576
Importaciones de bienes (FOB)	274076
Exportaciones de servicios	78890
- De la cuales, gasto en consumo de no residentes en el territorio económico	14055
Importaciones de servicios	53978
- De las cuales, gasto en consumo de residentes fuera del territorio económico	61161
Impuestos menos subvenciones sobre la producción	-3346
Remuneración de asalariados residentes con empleador no residente	2615
Remuneración de asalariados no residentes con empleador residente	311
Rentas de la propiedad a recibir	43210
Rentas de la propiedad a pagar	46124
Impuestos corrientes sobre la renta y el patrimonio a recibir	1379
Impuestos corrientes sobre la renta y el patrimonio a pagar	1544
Cotizaciones sociales netas a recibir	71
Cotizaciones sociales netas a pagar	819
Prestaciones sociales distintas de transferencias sociales en especie a recibir	4055
Prestaciones sociales distintas de transferencias sociales en especie a pagar	563
Otras transferencias corrientes a recibir	11921
Otras transferencias corrientes a pagar	26733
Transferencias de capital a recibir	4550
Transferencias de capital a pagar	636
Adquisiciones menos cesiones de activos no financieros no producidos	1220

Nota: en millones de euros.

INSTITUTO NACIONAL DE ESTADISTICA



A.2. Cuarto examen del Cuerpo Superior de Estadísticos del Estado. Especialidad ciencia de datos. Turno libre.



CUESTIÓN 1

Se quiere estimar el total de una característica Y en una población de tamaño N. Para ello, se tiene información de dos variables auxiliares X_1, total poblacional de la variable X_1 y X_2, total poblacional de la variable X_2.

Se pide:

A) Sea el estimador sintético para el total de Y dado por la siguiente expresión:

$$\hat{Y}_R = \alpha \hat{Y}_{R1} + \beta \hat{Y}_{R2}$$

Donde $\alpha + \beta = 1$ e $\hat{Y}_{Ri}$ es el estimador del total de Y por el método de la razón usando la variable auxiliar X_i, i=1,2.

(a) ¿Es el estimador $\hat{Y}_R$ insesgado?

(b) Calcule el valor de α que minimice la varianza de $\hat{Y}_R$

B) Se quiere estimar la superficie media de 100 explotaciones agrarias de una comarca en 2023, media en hectáreas, que se denota por $\bar{Y} = Y/N$. De censos anteriores se conoce la superficie media en 2020 ($\hat{X}_1 = 1100$) y 2010 ($\hat{X}_2 = 900$).

Para ello se extrae una muestra aleatoria simple sin reemplazamiento de tamaño n=40. Los datos disponibles son:

- Estimación de las medias usando un estimador de expansión simple:

$$\hat{\bar{Y}} = 1050; \qquad \hat{\bar{X}}_1 = 990; \qquad \hat{\bar{X}}_2 = 910;$$

- Cuasidesviaciones muestrales:

$$s_y = 975; \qquad s_{x1} = 901; \qquad s_{x2} = 820;$$

$$s_{x1y} = 905; \qquad s_{x2y} = 850; \qquad s_{x1x2} = 800;$$

(c) Calcule la estimación de la media de Y usando el estimador de razón $\hat{Y}_{R1}$ junto a su error de muestreo.

(d) Calcule la estimación de la media de Y usando el estimador de razón $\hat{Y}_{R2}$ junto a su error de muestreo.

(e) Calcule la estimación de la media de Y usando el estimador de razón $\hat{Y}_R$ cuando α toma el valor obtenido a partir de la expresión del apartado A.2), junto a su error de muestreo.

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA

CUESTIÓN 2

Suponga que las observaciones (x, y) satisfacen el modelo $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ donde $\epsilon_1, ..., \epsilon_n$ se distribuyen independientemente según una $N(0, \sigma^2)$.

(a) Escriba la verosimilitud de los datos.

(b) Suponga que a priori $(\beta_1, ..., \beta_p)$ son independientes e idénticamente distribuidas según una distribución de densidad $p(\beta) = \frac{1}{2b}\exp(-|\beta|/b)$, con parámetro b común. Argumente que el estimador lasso surge como moda de la correspondiente distribución a posteriori.

(c) Suponga, ahora, que a priori $(\beta_1, ..., \beta_p)$ son independientes e idénticamente distribuidas según una distribución normal de media 0 y varianza c. Argumente que el estimador de regresión *ridge* surge como moda de la correspondiente distribución a posteriori.



CUESTIÓN 3

Tenemos una base de datos con datos geográficos del mundo: contiene información sobre países, con su código correspondiente, la población total del país, la superficie y la renta per cápita. También contiene información sobre ciudades, con su nombre, un código, su población y el país al que pertenecen. Es posible que nuestra base de datos no contenga la población de alguna ciudad. En cambio, los demás datos (nombres de las ciudades y países, códigos, etc.) son obligatorios.

(a) Proporcione un ejemplo de un documento XML bien formado para este caso, e indique las reglas que cumple para que se considere un documento bien formado.

(b) Dé instrucciones SQL de creación de dichas tablas, sin olvidar vincularlas con las claves externas necesarias, en su caso.

(c) Elabore una consulta SQL para averiguar cuáles son las ciudades (5 máx.) más pobladas del país del mundo con la renta per cápita más alta registrado en nuestra base de datos.

(d) Elabore una consulta SQL para eliminar de nuestra base de datos el país de menor superficie. Esta operación debe evitar dejar inconsistencias en la base de datos con respecto a las ciudades del país eliminado.

INSTITUTO NACIONAL DE ESTADÍSTICA

CUESTIÓN 4

Diseñe una función que simule una v. a. exponencial, cuya función de densidad es la siguiente:

$$f(x) = \lambda\, e^{-\lambda x}, \text{ para } x \geq 0$$

Para la simulación, podemos suponer definida una función $U(0,1)$, esto es, una función que simula una v. a. real, uniformemente distribuida en el intervalo [0, 1].

El diseño de este algoritmo puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R.

INSTITUTO NACIONAL DE ESTADISTICA



CUESTIÓN 5

Una población está formada por $M = 3000$ viviendas (unidades elementales) agrupadas en 30 municipios de tamaños desiguales $(M_i, i = 1, 2, \ldots, 30)$. Se quiere estimar la proporción de viviendas que pertenecen a una cierta clase y para ello se usa un muestreo de conglomerados con submuestreo.

En ambas etapas se emplea un procedimiento de selección con probabilidades iguales sin reposición.

En la primera etapa se seleccionan 3 municipios con los siguientes valores de M_i: 10; 50; 100.

En la segunda etapa, realizada con una fracción de muestreo de $f_{2i} = 4/M_i$, se obtienen en los 3 municipios muestrales los valores $2; 2; 3$ respectivamente para el número de viviendas que pertenecen a la clase.

Se pide:

(a) Una estimación insesgada de la mencionada proporción y su error de muestreo.

(b) Una estimación, usando el estimador de la razón al tamaño, de la proporción mencionada y su error de muestreo.

(c) Comentar ventajas e inconvenientes del estimador de (b) frente al estimador de (a).

CUESTIÓN 6

Se modeliza la ocurrencia diaria de lluvia en otoño en cierta estación meteorológica mediante una cadena de Markov con dos estados (1) no llueve y (2) llueve. Se han recogido los siguientes datos de lluvia en 49 días:

2222222222 1121111111 2211112221 2111112111 111111111

Se pide:

(a) Defina una matriz de transición para la cadena.

(b) Estime mediante máxima verosimilitud las entradas de la matriz de la cadena.

(c) Suponga distribuciones a priori uniformes para los elementos de la diagonal de la matriz.

- Calcule las distribuciones a posteriori para los elementos de la cadena.
- Calcule las probabilidades de transición esperadas a posteriori.
- Calcule la probabilidad predictiva a posteriori de lluvia para el día 50.

(d) Calcule la probabilidad predictiva a posteriori de que no llueva los días 50 y 51.



CUESTIÓN 7

Una agencia de viajes necesita desarrollar un sistema para gestionar información sobre pisos turísticos, incluyendo el país, ciudad y dirección en que se ubican, así como otras características tales como número de dormitorios y baños, disponibilidad de cocina, etc.

(a) Defina XML, diga qué reglas debe cumplir un documento bien formado y proporcione un ejemplo de un documento XML bien formado para el ejemplo anterior.

(b) Se desea diseñar una aplicación web para gestionar la información de nuestra agencia. Describa los conceptos de frontend, backend, lado del cliente y lado del servidor, apoyándose en nuestra agencia de viajes. En cada uno de estos conceptos, mencione alguna tecnología que tendría sentido usar.

INSTITUTO NACIONAL DE ESTADISTICA



CUESTIÓN 8

Deseamos obtener la mediana de una secuencia de números enteros. Los números enteros no están ordenados. La secuencia puede ser un vector o una lista, posiblemente de gran tamaño, por lo que nuestro algoritmo ha de ser lo más eficiente posible.

El diseño de este algoritmo puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R. No estará permitido usar ninguna función predefinida que ordene dicha secuencia. Adicionalmente, se pide discutir la complejidad en tiempo de dicho algoritmo, o al menos un boceto de dicho cálculo.

INSTITUTO NACIONAL DE ESTADISTICA

8



A.3. Segundo examen del Cuerpo de Diplomados en Estadística del Estado. Turno libre.

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA

Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

Cuestionario de examen

Segundo ejercicio de la fase de oposición de Diplomados en Estadística del Estado 2022. Turno libre.

Convocatoria de la oferta pública de empleo 2020, 2021 y 2022.

Resolución de 26 de septiembre de 2022, de la Subsecretaría del Ministerio de Asuntos Económicos y Transformación Digital (BOE 3 de octubre de 2022)



Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

CUESTIÓN PRÁCTICA 1

Una variable aleatoria continua tiene la siguiente función de densidad:

$$f(x)=\begin{cases}\frac{1}{2}e^{x} & si\ x \leq 0\\ \frac{1}{2}e^{-x} & si\ x > 0\end{cases}$$

Obtenga:

1) **La función de distribución de X**
2) **La probabilidad de que X sea menor o igual que 1**
3) **La probabilidad del suceso** $\{|X| \leq 2 \text{ U } (X > 0)\}$
4) **Si sabemos que X es mayor que 1, la probabilidad de que X sea menor que 2**
5) **La esperanza de X**

CUESTIÓN PRÁCTICA 2

Una fábrica produce determinados artículos que se agrupan en cajas de 15 unidades. En un control de defectuosos de una partida de 140 cajas se obtienen los siguientes datos

Nº artículos defectuosos	Nº de cajas
0	57
1	57
2	18
3	5
4	3
5 o más	0

Calcule:

1) **El número medio de defectuosos por caja**
2) **La varianza y la desviación típica del número de defectuosos por caja**
3) **La mediana, el primer y tercer cuartil y el percentil del 95% del número de defectuosos por caja**
4) **La probabilidad de que un artículo sea defectuoso**

1



Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

CUESTIÓN PRÁCTICA 3

Sea $X_1, \ldots, X_n$, una muestra aleatoria simple de una población que sigue una distribución Normal $N(0, \sigma^2)$ siendo σ^2 desconocido. Calcular:

1) **El estimador máximo verosímil de σ^2**
2) **¿Es un estimador insesgado?**
3) **Intervalo de confianza de nivel 0,95 para σ^2. Indique el intervalo en función de los valores de la distribución del estimador.**

CUESTIÓN PRÁCTICA 4

En una federación deportiva el número de jugadores y el de entrenadores se distribuye según se indica en esta tabla de frecuencias:

Años	Jugadores (en miles) X_i	Entrenadores (en miles) Y_i
2017	80	3,1
2018	85	3,3
2019	93	5,2
2020	105	6,3
2021	116	7,1

Se pide:

1) **Ecuación de la recta de regresión de Y sobre X**
2) **Interpretación del coeficiente de regresión de Y sobre X**
3) **Calcular el coeficiente de correlación lineal de Pearson. Interpretación.**
4) **Calcular el coeficiente de determinación. Interpretación.**
5) **Predicción del número de entrenadores en 2022 para un número de jugadores estimados en ese año de 125.000.**

INSTITUTO NACIONAL DE ESTADÍSTICA

2

Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

CUESTIÓN PRÁCTICA 5

El mercado de altramuces en el país B tiene las siguientes curvas de oferta y demanda:

X_{Oferta}=1.000 + 400*P

$X_{Demanda}$=2.000 – 200*P

Nota: Los altramuces son legumbres características de la zona Mediterránea. Estas semillas comestibles se tomaban habitualmente como aperitivo.

1) **Razone si el paradigma del mercado de altramuces es un monopolio, oligopolio o un mercado de competencia perfecta.**
2) **Determine la expresión matemática de la curva de demanda individual de cada consumidor y la oferta individual de cada productor, sabiendo que en el mercado hay 200 consumidores idénticos y 50 productores con las mismas características.**
3) **Calcule el punto de equilibrio del mercado.**
4) **Si el precio de fija en 2 unidades monetarias indique que ocurrirá en el mercado.**
5) **Represente en un solo gráfico las respuestas a los dos apartados anteriores.**

CUESTIÓN PRÁCTICA 6

A partir de los datos de la siguiente tabla de mortalidad referida a 2020, completar las celdas sombreadas(*). Las tasas de mortalidad y los riesgos de muerte en la tabla están expresados en tantos por mil.

(*) Nota: Expresar los resultados redondeados a dos decimales

Año 2020	Tasa de mortali-dad	Promedio de años vividos el último año de vida	Riesgo de muerte	Supervi-vientes	Defuncion es teóricas	Población estaciona-ria	Tiempo por vivir	Esperanza de vida
Ambos sexos								
0 años	2,52	0,13	2,51	(1)	251,40	(5)	8.232.951,82	82,33
1 año	0,20	0,45	0,20	(2)	(3)	99.737,61	8.133.171,13	81,54
2 años	0,10	0,49	0,10	99.728,70	10,39	99.723,43	(4)	80,55
3 años	0,11	0,52	0,11	99.718,31	10,73	99.713,18	7.933.710,08	(6)
4 años	(7)	(9)	0,05	99.707,59	5,03	99.705,15	7.833.996,90	78,57
5 años	0,07	0,53	(8)	99.702,55	6,49	99.699,51	7.734.291,75	77,57
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
100 y más años	490,28	1,97	(10)	2.177,59	(11)	4.282,44	(12)	1,97



Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

CUESTION PRACTICA 7

A partir de la información de las rúbricas SEC2010 de una economía imaginaria que se indican en la siguiente tabla:

Rúbrica SEC2010	Unidades Monetarias. Precios corrientes de 2021
Consumo individual efectivo	1.340
Formación neta de capital	195
Exportaciones de bienes y servicios	500
Excedente Bruto de Explotación / Renta Mixta Bruta	790
VAB a precios básicos	2.150
Consumo colectivo efectivo	660
Consumo de capital fijo	50
Otros impuestos netos sobre la producción	290
Variación de existencias	15
Impuestos sobre productos e importaciones	455
Adquisiciones menos cesiones de objetos valiosos	25
Producción	3.280
Importaciones de bienes y servicios	200
Consumos intermedios	1.130

Calcule:

1) **El gasto en consumo final**
2) **La formación bruta de capital fijo**
3) **El saldo neto exterior de bienes y servicios**
4) **Sabiendo que el PIB a precios de mercado de 2021 valorado a precios constantes de 2015 es de 2.120 unidades monetarias, cuál sería el deflactor implícito del PIB para 2021 en base 2015=100**
5) **El valor añadido bruto a precios de mercado (o precios de adquisición)**
6) **Los impuestos netos sobre productos e importaciones**
7) **Las subvenciones a los productos**
8) **La remuneración de los factores de producción**
9) **La remuneración de los asalariados**
10) **Compruebe que es posible determinar el PIB a precios de mercado valorado a precios corrientes de 2021 por las tres vías: producción, gasto y renta.**

INSTITUTO NACIONAL DE ESTADISTICA

Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

CUESTIÓN PRÁCTICA 8

Una población de tamaño N=500 se divide en dos estratos. Obtenemos una muestra estratificada sin reemplazamiento de tamaño n=50 con los siguientes datos:

Estrato	N_h	n_h	$\bar{x}_h$	S_h
1	100	15	30	15
2	400	35	10	8

Obtener:

1) **La estimación de la media poblacional**
2) **La estimación de la varianza de la media anterior**
3) **¿Son los tamaños muestrales seleccionados en cada estrato compatibles con la afijación óptima?**
4) **Si consideramos además los costes de entrevista por unidad en cada estrato C1=25 y C2=4, ¿Cuál sería la afijación optima para el tamaño muestral total de n=50?**

CUESTIÓN PRÁCTICA 9

Se dispone de los siguientes datos de una población

Grupos de edad	Nacimientos en 2020 por edad de la madre		Población residente a 1 de enero de 2020		Población residente a 1 de enero de 2021	
	Hombres	Mujeres	Hombres	Mujeres	Hombres	Mujeres
< 15	0	0	6720	6800	6800	6512
15-19	16	10	3244	3150	3250	3165
20-24	35	30	3192	2914	4000	2956
25-29	32	40	2800	2880	2865	2889
30-34	45	47	3096	2853	3098	2855
35-39	58	55	2950	2900	2876	2942
40-44	38	32	2970	3013	2983	3000
45-49	30	25	3050	3140	3013	3150
50-54	0	0	3200	3220	3390	3267
55-59	0	0	3091	3131	3035	3100
60+	0	0	4400	4600	4400	4600
TOTAL	254	239	38713	38601	39710	38436

Calcular:

1) **La tasa bruta de natalidad**
2) **La tasa global de fecundidad**
3) **La tasa específica de fecundidad para el grupo 35-39 años**
4) **La tasa específica de fecundidad de niñas para el grupo 20-24**
5) **El ratio de masculinidad al nacimiento**



Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

CUESTIÓN PRÁCTICA 10

El índice al Comercio al por Menor (ICM) es un índice tipo Laspeyres encadenado a diciembre. La serie en base 2015 para los años 2018 y 2019 es la siguiente:

Periodo	ICM precios corrientes	Periodo	ICM precios corrientes
2018M01	107,932	2019M01	110,592
2018M02	92,094	2019M02	94,749
2018M03	103,085	2019M03	104,867
2018M04	101,836	2019M04	105,410
2018M05	106,685	2019M05	110,619
2018M06	111,649	2019M06	111,750
2018M07	115,606	2019M07	120,931
2018M08	106,530	2019M08	109,390
2018M09	104,006	2019M09	106,667
2018M10	110,587	2019M10	112,339
2018M11	110,459	2019M11	113,286
2018M12	127,888	2019M12	130,688

La muestra utilizada para el cálculo del índice en el año 2020 fue de 12.006 empresas. Conocemos los siguientes datos adicionales:

	Facturación en miles de euros
Volumen de facturación en comercio al por menor en enero de 2020 con una muestra de 12.006 empresas	185.776,98
Volumen de facturación en comercio al por menor en diciembre de 2019 con una muestra de 12.006 empresas	215.964,12
Volumen de facturación en comercio al por menor en diciembre de 2019 con una muestra de 12.158 empresas	221.476,38

1) **Calcular el índice de Comercio al por Menor (ICM) en base 2015 para el periodo de enero de 2020. Justifica su elaboración. Justifica cómo lo has elaborado.**
2) **¿Cómo calcularías el coeficiente de enlace para pasar esta serie a base 2019? ¿Qué condición tiene que cumplir el año base?**
3) **Sabiendo que el IPC en base 2015 asociado a los productos que recoge el ICM tiene como valor índice 103,813 en diciembre de 2019, calcula el índice deflactado. ¿Qué conseguimos con la deflación del índice a precios corrientes?**
4) **Calcule la tasa de variación mensual del índice de precios corrientes de diciembre de 2019 y explique su significado.**

Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

7

Ejercicio segundo. Convocatoria 2022 del turno libre de Diplomados en Estadística del Estado

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA

8



EX LIBRIS
INSTITUTO NACIONAL
DE
ESTADISTICA



Apéndice B

Exámenes de problemas de 2022. Promoción interna.

B.1. Segundo examen del Cuerpo Superior de Estadísticos del Estado. Parte común. Promoción interna.

CUESTIÓN 1

Se quiere estimar la superficie total de olivar de una provincia en el 2023. Se dispone como marco de muestreo el listado de explotaciones agrarias del Censo Agrario 2020. Se forman tres estratos de tamaño y en cada uno de ellos se obtiene una muestra aleatoria simple sin reemplazamiento. Se dispone de la siguiente información:

Estratos	Número de explotaciones Población (N_h)	Muestra (n_h)	Superficie media muestral (Ha.)	Desviacíón típica muestral (Ha)
1	10200	100	2,5	1
2	5000	1200	10	4,5
3	500	250	50	15

Se pide:

- Para cada uno de los estratos:

(a) Estime de manera insesgada el total de hectáreas (Ha) de Olivar junto a su error de muestreo relativo.

(b) Calcule los tamaños muestrales necesarios para realizar las estimaciones anteriores con unos errores de muestreo estimados inferiores a 1000 Ha en el estrato 1, a 500 Ha en el estrato 2, y a 10 Ha en el estrato 3.

- Para el total provincial:

(c) Estime de manera insesgada el total de hectáreas (Ha) de Olivar junto a su error de muestreo relativo.

(d) Calcule los tamaños muestrales por estrato, necesarios para realizar la estimación anterior con un error de muestreo estimado inferior a 500 Ha para el total provincial.

INSTITUTO NACIONAL DE ESTADISTICA



CUESTIÓN 2

Considere tres variables aleatorias A, B y C tales que su distribución conjunta tiene la forma $p(a, b, c) = p(a) \times p(c) \times p(b|a, c)$. A y C tienen distribuciones de Bernoulli de parámetros θ_A y θ_C, respectivamente. Dados a y c, B tiene distribución de Bernoulli.

(a) Dadas n observaciones independientes idénticamente distribuidas según A, derive el estimador de máxima verosimilitud de θ_A.

(b) Dadas las 5 observaciones de (A, C, B) siguientes ((0,1,1),(0,0,0),(1,0,1),(0,0,0),(0,1,0)), calcule el estimador de máxima verosimilitud de θ_A, el de θ_C y, finalmente, el del parámetro de Bernoulli de la distribución de B, dados los antecesores a y c, siempre que sea posible. Comente el resultado.

Suponga que se dispone de información a priori para cada uno de los seis parámetros antes mencionados, modelizada en los seis casos mediante distribuciones uniformes $U(0, 1)$ (e independientes):

(c) Dadas n observaciones independientes idénticamente distribuidas según A, derive la media de la distribución a posteriori de θ_A.

(d) Dadas las mismas 5 observaciones de (A, C, B), calcule las medias de las distribuciones a posteriori de θ_A, de θ_C y, finalmente, del parámetro de Bernoulli de la distribución de B, dados los antecesores a y c. Comente el resultado.

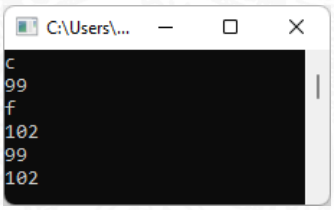
INSTITUTO NACIONAL DE ESTADISTICA

CUESTIÓN 3

Observe este fragmento de programa en C/C++. Consiste simplemente en una secuencia de instrucciones. Hemos añadido un breve comentario a su derecha explicando su funcionamiento:

```
char car = 'c';          // Sea la variable car, de tipo carácter, con el valor inicial 'c'
printf(" %c \n", car);   // Escribir el valor de car como un carácter
printf(" %d \n", car);   // Escribir el valor de car como un entero
car = car + 3;           // Aumentar el valor de car en 3 unidades
printf(" %c \n", car);   // Escribir el valor de car como un carácter
printf(" %d \n", car);   // Escribir el valor de car como un entero
int n = int('c');        // Sea la variable entera n, que toma el valor inicial del
                         // código ASCII de 'c'
printf(" %d \n", n);     // Escribir el valor de n como un entero
n = n + 3;               // Aumentar el valor de n en 3 unidades
printf(" %d \n", n);     // Escribir el valor de n como un entero
```

Cada instrucción printf contiene el símbolo '\n', que indica un salto de fin de línea, de manera que la ejecución de este fragmento de programa genera la salida siguiente, una línea por cada instrucción printf:



Ahora, vamos con la representación interna de la información que registra caracteres y enteros. Sabemos que la variable 'car' se está almacenando en una palabra (1 byte) y que el código ASCII del carácter 'c' es 99.

(a) Dé la representación, en binario y en hexadecimal, del entero 99.

(b) Dé la representación, en binario y en hexadecimal, del carácter 'c'.

(c) ¿Por qué la instrucción siguiente no genera un error?

car = car + 3;

(d) En una máquina cuyas palabras de memoria son de 1 byte, tenemos una palabra con el siguiente valor almacenado: "01100001". ¿Qué puede representar?



CUESTIÓN 4

Dada la población:

u_i	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
x_i	2	7	2	3	5	6	4	1

se desea obtener una muestra sistemática de tamaño 3 para lo cual se determina un periodo $k = N/n$, redondeado al entero más próximo y un arranque aleatorio entre 1 y K.

Se pide:

(a) Determine el espacio muestral y la función de probabilidad.

(b) Calcule la probabilidad π_i que tiene la unidad u_i de pertenecer a la muestra.

(c) Calcule la $V(\hat{X})$ siendo $\hat{X}$ el estimador insesgado del total.

(d) A partir de la distribución en el muestreo de $\hat{X}$, calcule $E(\hat{X})$ y $V(\hat{X})$.

INSTITUTO NACIONAL DE ESTADISTICA

CUESTIÓN 5

Se tienen datos $\{x_1, x_2, ..., x_n\}$ independientes idénticamente distribuidos de una normal de media μ y varianza conocida e igual a 9. Se determina que, a priori, μ se distribuye según una distribución normal de media 10 y varianza 4.

(a) Calcule la verosimilitud del modelo.

(b) Calcule el estimador de máxima verosimilitud de μ.

(c) Calcule la distribución a posteriori de μ.

(d) Calcule la moda a posteriori de μ.

(e) Proporcione la expresión de un intervalo de probabilidad 0.8 a posteriori para μ, suponiendo que se tiene acceso a una rutina $qnorm(p)$ que, dada una probabilidad p, nos da el correspondiente cuantil de la distribución normal estándar.



CUESTIÓN 6

En una localidad turística hay 12 empresas de alojamiento dedicadas a la hostelería, de las que se conoce su cifra de negocio del último mes, expresada en u.m.

Empresa	*Cifra de negocios*	*Grupos de actividad CNAE*
A	25.000	551 Hoteles y alojamientos similares
B	30.000	552 Alojamientos turísticos y otros alojamientos de corta estancia
C	15.000	553 Campings y aparcamientos para caravanas
D	12.000	559 Otros alojamientos
E	18.000	551 Hoteles y alojamientos similares
F	13.000	551 Hoteles y alojamientos similares
G	3.000	552 Alojamientos turísticos y otros alojamientos de corta estancia
H	2.000	552 Alojamientos turísticos y otros alojamientos de corta estancia
I	11.000	559 Otros alojamientos
J	2.000	552 Alojamientos turísticos y otros alojamientos de corta estancia
K	2.000	552 Alojamientos turísticos y otros alojamientos de corta estancia

Se pretende publicar una tabla que refleje la cifra de negocio de dichas empresas atendiendo al grupo CNAE en el que se engloba su actividad y respetando la confidencialidad de los informantes. La tabla propuesta es la siguiente:

Grupos de actividad CNAE	*Cifra de negocios*
551 Hoteles y alojamientos similares	56.000
552 Alojamientos turísticos y otros alojamientos de corta estancia	39.000
553 Campings y aparcamientos para caravanas	15.000
559 Otros alojamientos	23.000

Justifique razonadamente si la tabla propuesta resulta segura en un contexto de control de secreto estadístico atendiendo a los siguientes criterios:

(a) Según la regla de la frecuencia (o valor límite), para k=2.

(b) Según la regla de dominancia, para n=1, k=75 %.

(c) Según la regla del porcentaje, para p=10 %.

(d) Proponga alguna alternativa concreta para solucionar los problemas de identificación encontrados.

B.2. Cuarto examen del Cuerpo Superior de Estadísticos del Estado. Especialidad ciencia de datos. Promoción interna.



CUESTIÓN 1

Para estimar el número medio de portátiles por vivienda en una determinada zona rural se realiza un diseño en dos etapas:

Primera etapa: Se obtiene una muestra aleatoria simple con reemplazamiento de m=3 pueblos entre un total de M=300. El tamaño de cada pueblo (N_i) se mide por el número de viviendas.

Segunda etapa: Se selecciona una muestra aleatoria simple de 5 viviendas en cada uno de los pueblos seleccionados en primera etapa.

El número total de viviendas es de N=9000 .

Los datos disponibles son:

Unidades primarias (pueblos)	Tamaños (N_i)	Número de portátiles observados en las viviendas de la muestra m_i=5
1	50	1;0;3;2;2
2	60	2;2;1;2;3
3	65	2;4;3;2;1

Se pide:

(a) Calcular un estimador insesgado para el número medio de portátiles por vivienda en los siguientes casos:

- Muestreo con probabilidades iguales en las dos etapas.

- Muestreo con probabilidades proporcionales al tamaño en la primera etapa.

(b) Estimar el error de muestreo en ambos casos. ¿Cuál de los dos estimadores es preferible?

CUESTIÓN 2

Se formula un modelo (de regresión no lineal) de red neuronal de dos capas ocultas completamente conectada con las siguientes características:

i) entradas $\mathbf{x} \in \mathbf{R}^6$;
ii) salida $y \in \mathbf{R}$;
iii) primera capa oculta con cuatro nodos y función de activación logística;
iv) segunda capa oculta con dos nodos y función de activación ReLU.

Se pide:

(a) Representar el modelo gráficamente.

(b) Formular la transformación asociada a un nodo de la primera capa oculta. Formule la transformación asociada a un nodo de la segunda capa oculta.

(c) Representar la relación entre la respuesta y y la salida de la segunda capa oculta.

(d) Se dispone de datos $(\mathbf{x}_i, y_i)_{i=1}^n$ para entrenamiento de la red. Se desea ajustar la red mediante mínimos cuadrados con regularizador de caída de pesos (weight decay). Formule la función objetivo del problema que debería resolverse para ajustar el modelo.

(e) Describir analíticamente una iteración genérica del algoritmo del descenso del gradiente para ese problema. Describa qué rol tendría el método de retropropagación en tal algoritmo.

(f) Describir qué problemas se dan cuando el tamaño n de los datos es grande y el modelo de red neuronal es profundo. Describa cómo pueden mitigarse esos problemas con un método de descenso del gradiente estocástico.

(g) Se dispone de datos $(\mathbf{x}_i, y_i)_{i=1}^n$ para ajustar la red. Explique cómo procedería, y con qué propósito, a dividir el conjunto en uno de entrenamiento y otro de contraste para un mejor ajuste del modelo.

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA



CUESTIÓN 3

Tenemos el siguiente esquema relacional, con información de una pequeña biblioteca:

Socio(cód_socio, nombre, dirección. email)
Libro(núm_registro, título, autor, isbn, ubicación)
EnPréstamo(núm_registro, cód_socio, fecha_préstamo)

Se pide diseñar consultas de álgebra relacional para averiguar la siguiente información:

(a) Qué títulos están actualmente en préstamo.

(b) Qué socios tienen más de tres ejemplares en préstamo.

(c) Nombre y direcciones de email de los socios que tienen libros desde hace más de 30 días. (Admitimos la operación "fecha + 30" o "fecha – 30" para indicar fechas posteriores o anteriores en 30 días a una fecha dada y las operaciones fecha1 < fecha2 para expresar que una fecha es anterior a otra.)

(d) Un socio devuelve un libro. ¿Cuál será la operación de álgebra relacional que actualiza nuestro esquema?

CUESTIÓN 4

Tenemos un archivo de texto plano, llamado "población.csv", con datos de poblaciones del mundo. He aquí el inicio de su contenido:

```
7.975.105.155
China,1.425.887.337
India,1.417.173.173
Estados Unidos,338.289.857
Indonesia,275.501.339
```

La primera línea es la población mundial total; las siguientes, las de cada país. El nombre del país puede contener más de una palabra (ejemplos: Corea del Norte, Islas Vírgenes de los Estados Unidos). Las cifras de poblaciones vienen en el formado mostrado, con puntos separando los miles, millones, etc. Los nombres de los países y las cifras vienen separados con una coma.

Deseamos elegir el nombre de un país aleatoriamente, con probabilidades proporcionales a sus poblaciones. Se pide diseñar un programa o algoritmo para ello.

El diseño pedido puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R. En la simulación de la variable aleatoria que se necesite, se podrá asumir que se tiene definida y disponible una función básica que simule una variable aleatoria uniforme continua U(0, 1). No se podrá usar ninguna otra función aleatoria.



CUESTIÓN 5

Sea una población formada por N=40 empresas. Se quiere estimar el total de su cifra de negocios (CN) y para ello se obtiene una muestra aleatoria simple (sin reemplazamiento) de n=4 empresas.

Se conoce el número de ocupados (x) de cada una de las empresas de la población. El total de ocupados en la población es X=160.

Los datos disponibles son los siguientes:

Unidades de la Muestra (i)	Y (CN, miles de euros)	X (ocupados)
1	100	3
2	200	5
3	500	7
4	300	5

Se pide:

(a) Estimar el total de la Y (CN) por el método de la razón.

(b) Estimar el sesgo del estimador del apartado anterior.

(c) Estimar su varianza.

(d) Explicar si es o no relevante el efecto del sesgo en el intervalo de confianza a partir de los resultados de los apartados anteriores.

INSTITUTO NACIONAL DE ESTADISTICA

CUESTIÓN 6

Suponga que las observaciones (x, y) de un problema pueden modelizarse mediante regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, donde $\epsilon_1, ..., \epsilon_n$ se distribuyen independientemente según una $N(0, \sigma^2)$. Para simplificar, suponga que σ es conocida e igual a 1.

Se pide:

(a) Escribir la log-verosimilitud de los datos.

(b) Suponga que, a priori, (β_0, β_1) son independientes e idénticamente distribuidas según una distribución normal de media 0 y varianza c^2. Calcule la distribución a posteriori de los parámetros.

(c) Basándose en tal distribución a posteriori, explique cómo contrastaría la hipótesis nula $H_0 : \beta_1 = 0$.

(d) Calcular la distribución predictiva de y para una nueva observación x_{n+1}.

(e) Dar un intervalo predictivo de probabilidad 0.88 para y dado x_{n+1}, basado en la salida de una rutina `rnorm(s)` que permite generar `s` observaciones de una normal estándar.



CUESTIÓN 7

Un supermercado necesita diseñar una base de datos para gestionar su inventario de productos y controlar las ventas. Cada producto se registra con su código, una descripción sucinta, su precio y las cantidades de dicho producto mínima (para reponer existencias cuando baje de dicha cantidad), real (la que hay realmente en los estantes) y máxima (la que cabe en los estantes). Cada cliente se registra con su email, su nombre y su domicilio). Cada vez que un cliente acude al supermercado, al pagar se genera un único tique, con la fecha, la lista de productos y la cantidad de cada uno, y el importe total de la factura. El supermercado registra, además, lógicamente, el cliente que ha realizado dicha compra.

Se pide:

(a) Realizar un diseño E/R de dicha base de datos.

(b) Definir, con este diseño, los siguientes conceptos, si están presentes, o inventa un ejemplo sencillo pero apropiado para cada uno de ellos:
- Entidades fuertes, entidades débiles.
- Atributos, claves.
- Relación uno a uno, uno a muchos, muchos a muchos.

(c) Participación total o parcial en una relación.

(d) Atributos de una relación.

(e) Atributo multivalorado, atributo calculado.

INSTITUTO NACIONAL DE ESTADISTICA

CUESTIÓN 8

Tenemos un archivo de texto plano, llamado "pps.csv" (abreviando país, población, superficie, separados por comas), con datos de poblaciones y superficies de países del mundo. He aquí el inicio de su contenido:

```
China,1.425.887.337,9.562.900
India,1.417.173,173 3,287.300
Estados Unidos,338.289.857,9.831.500
Indonesia,275.501.339,1.913.600
```

Cada línea consigna el nombre de cada país, seguido de su población y su superficie en kilómetros cuadrados. El nombre del país puede contener más de una palabra (ejemplos: Corea del Norte, Islas Vírgenes de los Estados Unidos). Las cifras de poblaciones vienen en el formado mostrado, con puntos separando los miles, millones, etc. Los nombres de los países y las cifras vienen separados con una coma. En las cantidades, se usa el punto para separar miles, millones, etc.

Se pide:

(a) Diseñar un programa o algoritmo para obtener el coeficiente de correlación entre ambas variables, superficie y población.

(b) Como sabes, es posible realizar esto recorriendo el archivo dos veces, una para calcular las medias y otra para recopilar el resto de la información. Pero la lectura de archivos es bastante costosa computacionalmente. Por eso, en este segundo apartado se pide diseñar un programa o algoritmo para obtener obligatoriamente este coeficiente leyendo el archivo en una sola pasada. Si ya lo has hecho así en tu primer apartado, basta con que en éste lo menciones.

El diseño puede realizarse en seudocódigo o en algún lenguaje imperativo como C/C++, Java, Python o R. No está permitido usar ninguna función predefinida que calcule la media ni otras medidas estadísticas.



B.3. Segundo examen del Cuerpo de Diplomados en Estadística del Estado. Promoción interna grupo C1.

Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

Cuestionario de examen

Segundo ejercicio de la fase de oposición de Diplomados en Estadística del Estado 2022. Promoción interna grupo C1.

Convocatoria de la oferta pública de empleo 2020, 2021 y 2022.

Resolución de 13 de diciembre de 2022, de la Subsecretaría del Ministerio de Asuntos Económicos y Transformación Digital (BOE 28 de diciembre de 2022)



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 1

Un colectivo de personas ha recibido las rentas anuales que se presentan en la siguiente distribución de frecuencias:

Rentas (en 10^4 €)	Nº de personas
4-6	12
6-8	13
8-10	35
10-12	25
12-14	15
	100

Calcular:

1) **La renta media**
2) **La moda de la distribución de renta**
3) **La mediana de la distribución de renta**
4) **La concentración de renta en el colectivo (índice de Gini)**

CUESTIÓN PRÁCTICA 2

Dado el vector aleatorio (X,Y) con función de densidad conjunta:

$$f(x,y)=\begin{cases}\frac{3}{8}(x^2+y^2) \rightarrow -1 \le x \le 1; -1 \le y \le 1 \\ 0 \quad \rightarrow \quad \text{resto}\end{cases}$$

Se pide:

1) **Funciones de densidad marginales de X y de Y**
2) **Coeficiente de correlación lineal entre X e Y**
3) **Indique si X e Y son variables aleatorias independientes**
4) **Varianza de la variable Z=2X-3Y**



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 3

En una distribución bidimensional de frecuencias se han calculado la recta de regresión de Y sobre X:

Y=1,8 X -5

Y la de X sobre Y:

X= 0,5 Y + 3

Determinar los valores medios de X y de Y. Deducir si la relación de las variables X e Y es alta o baja.

CUESTIÓN PRÁCTICA 4

Se tiene una variable aleatoria X que sigue una distribución Binomial(m,p) con parámetro 'p' desconocido. Se extrae una muestra de tamaño 'n'. Calcular:

1) **un estimador de 'p' por el método de los momentos**
2) **calcular el error cuadrático medio del estimador**
3) **Hallar la región crítica proporcionada por el test de razón de verosimilitud basada en este estimador para el contraste: H0: p<= p0 ; H1: p>p0**

INSTITUTO NACIONAL DE ESTADISTICA

2

Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTION PRACTICA 5

El índice de Precios de un determinado sector presenta las siguientes características técnicas:

- **Tipo Laspeyres**
- **Año Base 2018 con media anual =100**

El índice General de Precios de ese sector consta de 7 agrupaciones diferenciadas, y conocemos los siguientes datos sobre una de esas agrupaciones, la agrupación A1, que consta a su vez de cuatro subgrupos.

Datos sobre la Agrupación A1:

Agrupación A1	Ponderación (tanto por 1)	Índice enero 2022	Ìndice febrero 2022
Subgrupo A11	0,015	121,3	122,5
Subgrupo A21	0,024	117,4	116,3
Subgrupo A31	0,007	106,0	106,0
Subgrupo A31	0,036	115,8	118,3

Conocemos que el Índice General de Precios del sector en febrero 2022 es 111,5. Se pide calcular:

1) La variación mensual (%) del Índice de Precios de la Agrupación A1 en el mes de febrero de 2022.

2) Repercusión que se produciría en el Índice de la Agrupación A1 en el mes de marzo 2022 si el subgrupo A31 sufriera un incremento del 3,5% en sus precios dicho mes. ¿Cuál sería, además, la repercusión en el Índice General de Precios del Sector?

INSTITUTO NACIONAL DE ESTADISTICA

3



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 6

Utilizando los datos siguientes relativos a una economía imaginaria,

Gasto en consumo final	2300
Formación bruta de capital fijo	215
Deflactor implícito del PIB para 2021 en base 2015=100	115,1
Impuestos netos sobre productos	415
Subvenciones a los productos	70
Remuneración de los asalariados	1135
Exportaciones de bienes y servicios	545
Consumo colectivo efectivo	692
Consumo de capital fijo	70
Otros impuestos netos sobre la producción	335
Variación de existencias	16
Adquisiciones menos cesiones de objetos valiosos	27
Producción	3380
Importación de bienes y servicios	210
Consumos intermedios	902

calcule en términos SEC2010:

1) **El consumo individual efectivo**
2) **La formación neta de capital**
3) **El saldo neto exterior de bienes servicios**
4) **El PIBpm a precios constantes de 2015**
5) **El valor añadido bruto a precios de mercado (o precios de adquisición)**
6) **El VAB a precios básicos**
7) **Los impuestos sobre los productos**
8) **La remuneración de los factores de producción**
9) **El Excedente Bruto de Explotación**
10) **Comprueba que es posible determinar el PIBpm a precios corrientes por las tres vías: producción, gasto y renta**

INSTITUTO NACIONAL DE ESTADISTICA

4

Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 7

Se quiere comparar la estructura de mortalidad de dos poblaciones distintas A y B. Se dispone de una tercera población C que servirá como referencia. A partir de los datos de la tabla:

Grupos de edad	Población de C en cada grupo de edad	Tasas específicas de mortalidad de la población A	Tasas específicas de mortalidad de la población B
De 0 a 4 años	2152	0,34	0,15
De 5 a 24 años	9943	0,19	0,14
De 25 a 44 años	12840	3,53	1,15
De 45 a 64 años	16942	9,81	12,78
De 65 a 84 años	23354	40,73	33,76
De 85 y más	15004	93,65	98,83
TOTAL	80235		

Calcule las tasas estandarizadas de mortalidad de las poblaciones A y B mediante el método directo de estandarización o 'población tipo' utilizando la población C e indique qué población, A o B, es la que tiene una menor tasa estandarizada.

CUESTIÓN PRÁCTICA 8

Se considera una población de N=100 hogares todos ellos constituidos por 4 personas. Se quiere estimar la proporción de personas con cierto atributo. Para ello se obtiene una muestra aleatoria sin reemplazamiento de 5 hogares obteniéndose los siguientes valores:

Hogar de la muestra	Nª de personas con atributo
1	2
2	3
3	1
4	2
5	1

1) **Obtenga una estimación de la proporción del atributo en la población**
2) **Una estimación de la varianza de la estimación anterior**



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 9

Una aldea cuenta con 50 personas que todos los días se dedican a la avicultura o a la agricultura. Si utilizan todos sus recursos y teniendo en cuenta la tecnología que poseen, son capaces de llegar a las siguientes combinaciones de patatas y aves.

	Aves	kg de patatas
A	0	105
B	15	95
C	25	85
D	30	75
E	45	55
F	55	0

1. Dibuja la frontera de posibilidades de producción (FPP)

2. ¿Cuál es el significado de la FPP de esta aldea?

3. ¿Cuál es el coste de oportunidad para la aldea de pasar de la combinación E a la C?

4. Clasifica las siguientes seis combinaciones de bienes: (25 aves, 100 Kg patatas) ; (30 aves, 75 Kg patatas) ; (45 aves, 55 Kg patatas) ; (15 aves, 95 kg patatas) ; (45 aves, 10 Kg patatas) ; (35 aves, 75 Kg patatas) como ineficientes, inalcanzables o eficientes ¿Qué quieren decir esos términos?

INSTITUTO NACIONAL DE ESTADÍSTICA

Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 10

A partir de la información de la tabla

Grupos de edad	Mujeres residentes (a fecha 1 de julio)	Nacimientos por edad de la madre
De 15 a 19 años	1.188.022	5.470
De 20 a 24 años	1.165.886	23.781
De 25 a 29 años	1.237.778	56.916
De 30 a 34 años	1.357.305	112.874
De 35 a 39 años	1.574.520	102.129
De 40 a 44 años	1.905.822	32.832
De 45 a 49 años	1.958.973	3.094
TOTAL	10.388.306	337.096

Calcule:

1) **Las tasas específicas de fecundidad por grupo de edad de la madre**
2) **El índice sintético de fecundidad**
3) **La edad media a la maternidad**

INSTITUTO NACIONAL DE ESTADISTICA



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA

8

B.4. Segundo examen del Cuerpo de Diplomados en Estadística del Estado. Promoción interna grupo C1 especialidad estadística.



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

Cuestionario de examen

Segundo ejercicio de la fase de oposición de Diplomados en Estadística del Estado 2022. Promoción interna grupo C1 Especialidad Estadística.

Convocatoria de la oferta pública de empleo 2020, 2021 y 2022.

Resolución de 13 de diciembre de 2022, de la Subsecretaría del Ministerio de Asuntos Económicos y Transformación Digital (BOE 28 de diciembre de 2022)

Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 1

Un colectivo de personas ha recibido las rentas anuales que se presentan en la siguiente distribución de frecuencias:

Rentas (en 10^4 €)	Nº de personas
4-6	12
6-8	13
8-10	35
10-12	25
12-14	15
	100

Calcular:

1) **La renta media**
2) **La moda de la distribución de renta**
3) **La mediana de la distribución de renta**
4) **La concentración de renta en el colectivo (índice de Gini)**

CUESTIÓN PRÁCTICA 2

Dado el vector aleatorio (X,Y) con función de densidad conjunta:

$$f(x,y) = \begin{cases} \frac{3}{8}(x^2+y^2) \rightarrow -1 \leq x \leq 1; -1 \leq y \leq 1 \\ 0 \quad \rightarrow \quad \text{resto} \end{cases}$$

Se pide:

1) **Funciones de densidad marginales de X y de Y**
2) **Coeficiente de correlación lineal entre X e Y**
3) **Indique si X e Y son variables aleatorias independientes**
4) **Varianza de la variable Z=2X-3Y**

1

CUESTIÓN PRÁCTICA 3

Se tiene una variable aleatoria X que sigue una distribución Binomial(m,p) con parámetro 'p' desconocido. Se extrae una muestra de tamaño 'n'. Calcular:

1) **un estimador de 'p' por el método de los momentos**
2) **calcular el error cuadrático medio del estimador**
3) **Hallar la región crítica proporcionada por el test de razón de verosimilitud basada en este estimador para el contraste: H0: p<= p0 ; H1: p>p0**

CUESTIÓN PRÁCTICA 4

Se quiere comparar la estructura de mortalidad de dos poblaciones distintas A y B. Se dispone de una tercera población C que servirá como referencia. A partir de los datos de la tabla:

Grupos de edad	Población de C en cada grupo de edad	Tasas específicas de mortalidad de la población A	Tasas específicas de mortalidad de la población B
De 0 a 4 años	2152	0,34	0,15
De 5 a 24 años	9943	0,19	0,14
De 25 a 44 años	12840	3,53	1,15
De 45 a 64 años	16942	9,81	12,78
De 65 a 84 años	23354	40,73	33,76
De 85 y más	15004	93,65	98,83
TOTAL	80235		

Calcule las tasas estandarizadas de mortalidad de las poblaciones A y B mediante el método directo de estandarización o 'población tipo' utilizando la población C e indique qué población, A o B, es la que tiene una menor tasa estandarizada.



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 5

Utilizando los datos siguientes relativos a una economía imaginaria,

Gasto en consumo final	2300
Formación bruta de capital fijo	215
Deflactor implícito del PIB para 2021 en base 2015=100	115,1
Impuestos netos sobre productos	415
Subvenciones a los productos	70
Remuneración de los asalariados	1135
Exportaciones de bienes y servicios	545
Consumo colectivo efectivo	692
Consumo de capital fijo	70
Otros impuestos netos sobre la producción	335
Variación de existencias	16
Adquisiciones menos cesiones de objetos valiosos	27
Producción	3380
Importación de bienes y servicios	210
Consumos intermedios	902

calcule en términos SEC2010:

1) **El consumo individual efectivo**
2) **La formación neta de capital**
3) **El saldo neto exterior de bienes servicios**
4) **El PIBpm a precios constantes de 2015**
5) **El valor añadido bruto a precios de mercado (o precios de adquisición)**
6) **El VAB a precios básicos**
7) **Los impuestos sobre los productos**
8) **La remuneración de los factores de producción**
9) **El Excedente Bruto de Explotación**
10) **Comprueba que es posible determinar el PIBpm a precios corrientes por las tres vías: producción, gasto y renta**

INSTITUTO NACIONAL DE ESTADÍSTICA

3

Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 6

Se considera una población de N=100 hogares todos ellos constituidos por 4 personas. Se quiere estimar la proporción de personas con cierto atributo. Para ello se obtiene una muestra aleatoria sin reemplazamiento de 5 hogares obteniéndose los siguientes valores:

Hogar de la muestra	Nª de personas con atributo
1	2
2	3
3	1
4	2
5	1

1) **Obtenga una estimación de la proporción del atributo en la población**
2) **Una estimación de la varianza de la estimación anterior**

CUESTIÓN PRÁCTICA 7

A partir de la información de la tabla

Grupos de edad	Mujeres residentes (a fecha 1 de julio)	Nacimientos por edad de la madre
De 15 a 19 años	1.188.022	5.470
De 20 a 24 años	1.165.886	23.781
De 25 a 29 años	1.237.778	56.916
De 30 a 34 años	1.357.305	112.874
De 35 a 39 años	1.574.520	102.129
De 40 a 44 años	1.905.822	32.832
De 45 a 49 años	1.958.973	3.094
TOTAL	10.388.306	337.096

Calcule:

1) **Las tasas específicas de fecundidad por grupo de edad de la madre**
2) **El índice sintético de fecundidad**
3) **La edad media a la maternidad**



Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.

CUESTIÓN PRÁCTICA 8

Una aldea cuenta con 50 personas que todos los días se dedican a la avicultura o a la agricultura. Si utilizan todos sus recursos y teniendo en cuenta la tecnología que poseen, son capaces de llegar a las siguientes combinaciones de patatas y aves.

	Aves	kg de patatas
A	0	105
B	15	95
C	25	85
D	30	75
E	45	55
F	55	0

1. Dibuja la frontera de posibilidades de producción (FPP)

2. ¿Cuál es el significado de la FPP de esta aldea?

3. ¿Cuál es el coste de oportunidad para la aldea de pasar de la combinación E a la C?

4. Clasifica las siguientes seis combinaciones de bienes: (25 aves, 100 Kg patatas) ; (30 aves, 75 Kg patatas) ; (45 aves, 55 Kg patatas) ; (15 aves, 95 kg patatas) ; (45 aves, 10 Kg patatas) ; (35 aves, 75 Kg patatas) como ineficientes, inalcanzables o eficientes ¿Qué quieren decir esos términos?

INSTITUTO NACIONAL DE ESTADÍSTICA

5

Ejercicio segundo. Convocatoria 2022 de Diplomados en Estadística del Estado. Promoción interna.



Apéndice C

Exámenes de problemas de 2023. Turno libre.

C.1. Segundo examen del Cuerpo Superior de Estadísticos del Estado. Parte común. Turno libre.



MINISTERIO DE ECONOMÍA, COMERCIO Y EMPRESA

INE Instituto Nacional de Estadística

Oposición al Cuerpo Superior de Estadísticos de Estado

Segundo Ejercicio

Convocatoria de la oferta pública de empleo de 2023

Resolución de 14 de diciembre de 2023, de la Subsecretaría, por la que se convocan procesos selectivos para ingreso, por el sistema de acceso libre y promoción interna, en el Cuerpo Superior de Estadísticos del Estado. (BOE 27 de Diciembre de 2023)



Cuestión 1. En una población formada por $N = 8$ unidades $\{1, 2, 3, 4, 5, 6, 7, 8\}$, se considera un diseño muestral que genera las siguientes muestras posibles:

$$p(\{1,3,5,6\}) = \frac{1}{8}; \quad p(\{1,4,6,8\}) = \frac{1}{8}; \quad p(\{4,5,7,8\}) = \frac{1}{8}$$
$$p(\{2,3,7,8\}) = \frac{1}{4}; \quad p(\{2,4,6,8\}) = \frac{3}{8}$$

Se pide:

a) Calcular las probabilidades de inclusión de primer orden, π_i, de todos los elementos de la población.

b) El diseño de la muestra, ¿es aleatorio simple sin reemplazamiento, aleatorio simple estratificado o ninguno de ellos? Razonar la respuesta.

c) Si los tamaños de estas 8 unidades son $\{100, 40, 40, 20, 20, 10, 5, 5\}$ respectivamente, ¿podría aplicarse un diseño muestral sin reposición, con tamaño $n = 4$, de manera que las π_i sean proporcionales a sus tamaños? Justifique su respuesta.

Cuestión 2. La función de densidad conjunta de las variables (X, Y) es la siguiente:

$$f_a(x, y) = a^2 e^{a(x+y)}, x \geq 0, y \geq 0.$$

Se toma una muestra aleatoria simple de tamaño n.

a) Calcule el estimador máximo verosímil del parámetro a.

b) Utilice dicho estimador para obtener $\hat{a}$ a partir de la información muestral recogida en la siguiente tabla:

x	1	2	3	4	5
y	2	0	1	0	1



Cuestión 3. .

a) Sea un sistema de representación numérica de enteros con 8 bits.

- Expresar el número decimal -39, en representación binaria utilizando formatos signo-magnitud (con primer dígito para signo), complemento a 1 y complemento a 2. El dígito de signo se representa como 1 negativo y 0 positivo.
- Si un número binario en complemento a 2 se representa como 10101010, ¿cuál es su valor decimal?.
- ¿Cuál es el entero mínimo y máximo representables en complemento a 2 en este sistema (que incluye bit de signo)?.

b) En un sistema binario (sin bit para signo), ¿a qué valor decimal exacto equivale el número 110101,111?

c) ¿Cuál es el valor decimal exacto del número binario $0,10101101101\ldots = 0,1\widehat{0101}$?

Cuestión 4. Complete las diez casillas marcadas en la siguiente tabla de mortalidad, indicando la fórmula utilizada en cada caso o razonando la respuesta si no es necesario aplicar ninguna fórmula.

	$_nq_x$	$_nl_x$	$_nd_x$	$_nL_x$	$_nT_x$	$_ne_x$	$_na_x$
0 años	1,8246	100.000,00	182,46	99.859,10	g	83,23	i
De 1 a 4 años	0,4937	99.817,54	49,28	e	8.223.335,74	82,38	0,2768
De 5 a 9 años	a	99.768,26	33,45	498.759,90	7.824.208,14	78,42	0,5133
De 10 a 14 años	0,3820	99.734,81	38,10	498.585,83	7.325.448,24	73,45	0,5369
De 15 a 19 años	1,0772	99.696,71	107,39	498.197,46	6.826.862,41	68,48	0,4672
De 20 a 24 años	1,6008	99.589,32	159,42	497.563,59	6.328.664,95	63,55	0,5195
De 25 a 29 años	1,3157	99.429,90	130,82	496.837,63	5.831.101,36	58,65	0,5232
De 30 a 34 años	2,3522	99.299,08	233,57	495.948,73	5.334.263,73	53,72	0,5319
De 35 a 39 años	3,4211	99.065,51	338,91	494.551,62	4.838.315,00	48,84	0,5421
De 40 a 44 años	4,5478	98.726,60	448,99	492.563,95	4.343.763,38	44,00	0,5238
De 45 a 49 años	8,0801	98.277,61	794,09	f	3.851.199,43	39,19	0,5401
De 50 a 54 años	13,7566	97.483,52	1.341,04	484.447,87	3.361.637,39	34,48	0,5571
De 55 a 59 años	23,8322	96.142,48	2.291,29	475.372,55	2.877.189,52	29,93	0,5339
De 60 a 64 años	34,7539	93.851,19	3.261,69	461.721,45	2.401.816,97	25,59	0,5380
De 65 a 69 años	50,7729	90.589,50	4.599,49	442.359,47	1.940.095,52	h	0,5396
De 70 a 74 años	73,6006	85.990,01	d	415.013,80	1.497.736,05	17,42	0,5280
De 75 a 79 años	121,5686	c	9.684,29	376.070,32	1.082.722,25	13,59	0,5408
De 80 a 84 años	204,4377	69.976,80	14.305,90	317.073,42	706.651,93	10,10	0,5413
De 85 a 89 años	374,6874	55.670,90	20.859,18	228.407,19	389.578,51	7,00	0,5211
De 90 a 94 años	592,7031	34.811,72	20.633,01	119.639,04	161.171,32	4,63	0,4725
95 y más años	b	14.178,71	14.178,71	41.532,28	41.532,28	2,93	j

INSTITUTO NACIONAL DE ESTADISTICA



Cuestión 5. Teniendo en cuenta la siguiente información contable correspondiente a una economía para un determinado año T (datos en miles de millones EUR):

Remuneración de los asalarados residentes	660
siendo de empleadores no residentes	18
Remuneración de los asalariados no residentes por empleadores residentes	15
Otros impuestos menos subvenciones sobre la producción	13
Saldo de intercambios exteriores de bienes y servicios	-20
Consumo colectivo efectivo	112
Transferencias sociales en especie	159
Renta disponible ajustada bruta del Sector Hogares	1.032
Ahorro bruto del Sector Hogares	75

a) Elabore el siguiente cuadro macroeconómico (en hoja aparte):

(miles de millones EUR)	**AÑO T**
PRODUCTO INTERIOR BRUTO A PRECIOS DE MERCADO	
Demanda	
Gasto en consumo final	
Gasto en consumo final de los hogares	
Gasto en consumo final de las ISFLSH	15
Gasto en consumo final de las AAPP	
Formación bruta de capital	294
Exportaciones de bienes y servicios	
Importaciones de bienes y servicios	540
Oferta	
Agricultura, ganadería, silvicultura y pesca	33
Industria	158
Construcción	
Servicios	922
Impuestos menos subvenciones sobre los productos	
Rentas	
Remuneración de los asalariados	
Excedente de explotación bruto / Renta mixta bruta	
Impuestos menos subvenciones sobre la producción y las importaciones	133

Suponiendo que, en ese año, el ajuste por la variación de los derechos por pensiones no es significativo.

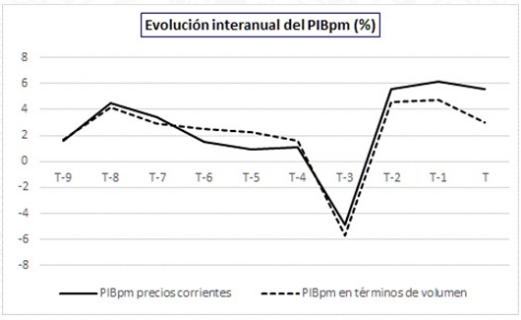
b) Indique cómo se vería afectado el cuadro anterior si, además, se producen conjuntamente las siguientes transacciones económicas:

1) Reducción del tipo del IVA en productos energéticos, que provoca un descenso de 2 miles de millones EUR en los ingresos de las AAPP.

2) Aumento del servicio producido por personal doméstico remunerado en 3 miles de millones EUR.

3) Disminución de la aportación del hogar al pago de medicamentos con receta en 5 miles de millones EUR.

c) Atendiendo al siguiente gráfico, ¿qué se podría decir de la variación del deflactor implícito del PIB en los últimos 10 años en esta economía?





Cuestión 6. Para el estudio del comportamiento medio de una determinada característica cuantitativa Y en una población, se propone aplicar un muestreo aleatorio estratificado.

La población ha sido agrupada en tres estratos de acuerdo a su edad: menos de 20 años, de 20 a 49 años y de 50 o más años. La información de la que se dispone de los estratos es la siguiente:

Estrato	Tramos edad	N_h	$\bar{Y}_h$	S_h^2
1	De 0 a 20 años	150	20	36
2	De 20 a 49 años	250	35	144
3	De 50 o más años	200	60	256

Donde :

- N_h= Número de personas en el estrato h.
- $\bar{Y}_h$= Valor medio de la característica en el estrato h.
- S_h^2= Cuasivarianza poblacional de Y en el estrato h.

Se realiza un muestreo sin reemplazamiento y con probabilidades iguales en cada estrato. Se pide:

a) Obtener la cuasivarianza poblacional de Y. ¿Que proporción supone la variación interestrato de la variación total?

b) Comentar a la vista de los resultados obtenidos en el apartado anterior, si la variable de estratificación es apropiada en términos de ganancia en precisión.

c) Calcular el tamaño muestral necesario, para que siendo una muestra autoponderada, se obtenga una varianza para el estimador de la media de 1,45. Asimismo, calcular los tamaños muestrales en los estratos.

Cuestión 7. Considere que la esperanza condicional de una variable dependiente y es la siguiente

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

donde las variables son continuas.

a) Calcule qué efecto esperado (efecto parcial) sobre y tiene un cambio de una unidad en la variable x_1, considerando que x_2 permanece constante.

b) Si dispusiera de una muestra de tamaño n, y quisiera estimar el siguiente modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + u_i, i = 1, ..., n,$$

donde los errores u_i no son observables, ¿cómo se deduciría y cuál sería la expresión del estimador de mínimos cuadrados ordinarios de $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$?



Cuestión 8. .

a) Elabore la sucesión de cuentas corrientes y de capital del sector Hogares de la página siguiente, que corresponde a una economía para un año T, teniendo en cuenta la información (en miles de millones EUR) para el total de la economía que aparece en esta tabla:

Formación bruta de capital	294
Consumo colectivo efectivo	110
Transferencias sociales en especie (TSE)	182
Remuneración de los asalariados por empleadores residentes	655
de la cual, a asalariados no residentes	12
Remuneración de los asalariados residentes por empeadores no residentes	15
Excedente de Explotación Bruto	450
Saldo de intercambios exteriores de bienes y servicios	-18
Impuestos menos subvenciones sobre la producción y las importaciones	133
Cotizaciones Sociales netas (total empleos)	198
Prestaciones Sociales distintas de las TSE (total recursos)	250
Saldo de la Cuenta de Explotación de la rama *servicios de alquiler de las viviendas ocupadas por sus propietarios*	80

Se considera que, para esta economía, las adquisiciones menos cesiones de activos no producidos son despreciables en el año T.

b) Responda a las siguientes cuestiones:

1) ¿Cuál sería la tasa de ahorro de los hogares en el año T? Interprete el resultado.
2) ¿Qué se podría concluir de su Capacidad / necesidad de financiación?
3) Si para dicha economía se realiza una nueva estimación del importe de los servicios de alquileres imputados de las viviendas ocupadas por sus propietarios para ese año, y el resultado supera en 5 miles de millones EUR a la actual, ¿qué saldos y operaciones se verían modificados a la luz de esta nueva información?

INSTITUTO NACIONAL DE ESTADISTICA

Cuentas corrientes

I. Cuenta de producción

Empleos S.14 Hogares	Código	Operaciones y otros flujos y saldos contables	Recursos S.14 Hogares
	P.1	Producción	365
	P.11	Producción de mercado	260
	P.12	Producción para uso final propio	105
98	P.2	Consumos intermedios	
	B.1b/	**Valor añadido bruto/**	
	P.51c	Consumo de capital fijo	
	B.1n/	***Valor añadido neto/***	

II. 1.1 Cuenta de explotación

	B.1b/B.1*b	Valor añadido bruto / Producto interior bruto	
	*B.1n/B.1*n*	*Valor añadido neto / Producto interior neto*	
40	D.1		
10	D.2		
-3	D.3		
	B.2b	**Excedente de explotación bruto**	
	B.3b	**Renta mixta bruta**	
	B.2n/B.3n	***Excedente de explotación neto / Renta mixta neta***	

II. 1.2 Cuenta de asignación de la renta primaria

	B.2b	Excedente de explotación bruto	
	B.3b	Renta mixta bruta	
	B.2n/B.3n	*Excedente de explotación neto / Renta mixta neta*	
	D.1	Remuneración de los asalariados	
10	D.4	Rentas de la propiedad	52
	B.5b/B.5*b	**Saldo de rentas primarias bruto**	
	B.5n/B.5*n	***Saldo de rentas primarias neto***	

II. 2 Cuenta de distribución secundaria de la renta

	B.5b/B.5*b	Saldo de rentas primarias bruto / Renta nacional bruta	
	*B.5n/B.5*n*	*Saldo de rentas primarias neto / Renta nacional neta*	
135	D.5	Impuestos corrientes sobre la renta, el patrimonio, etc.	
	D.61	Cotizaciones sociales netas	1
1	D.62	Prestaciones sociales distintas de las TSE	
96	D.7	Otras transferencias corrientes	84
	B.6b	**Renta disponible bruta**	
	B.6n	***Renta disponible neta***	

II. 4.1 Cuenta de utilización de la renta disponible

	B.6b	Renta disponible bruta	
	B.6n	*Renta disponible neta*	
	P.3	Gasto en consumo final	
	D.8	Ajuste por la variación de los derechos por pensiones	-2
	B.8b	**Ahorro bruto**	
	B.8n	***Ahorro neto***	

Cuenta de capital

III. 1.1 Cuenta de variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital

	B.8n	Ahorro neto	
	D.9r	Transferencias de capital, a cobrar	3
	D.9p	Transferencias de capital, a pagar	-5
	B.10.1	***Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital***	

III. 1.2 Cuenta de adquisiciones de activos no financieros

	B.10.1	*Variaciones del patrimonio neto debidas al ahorro y a las transferencias de capital*	
60	P.5g	Formación bruta de capital	
-40	P.51c	Consumo de capital fijo	
	B.9	**Capacidad(+)/Necesidad(-) de financiación**	



C.2. Cuarto examen del Cuerpo Superior de Estadísticos del Estado. Especialidad ciencia de datos. Turno libre.

MINISTERIO
DE ECONOMÍA, COMERCIO
Y EMPRESA

INE
Instituto Nacional de Estadística

Oposición al Cuerpo Superior de Estadísticos de Estado

Cuarto Ejercicio

Convocatoria de la oferta pública de empleo de 2023

Resolución de 14 de diciembre de 2023, de la Subsecretaría, por la que se convocan procesos selectivos para ingreso, por el sistema de acceso libre y promoción interna, en el Cuerpo Superior de Estadísticos del Estado. (BOE 27 de Diciembre de 2023)

Especialidad I: Estadística-Ciencia de Datos



Cuestión 1. Sea un municipio con un total de N =8.000 alumnos distribuidos en $M = 16$ centros escolares de tamaños desiguales $(N_i, \forall i = 1 \ldots, 16)$. Se desea estimar la proporción P de estudiantes que practican algún deporte fuera del horario escolar para ajustar los planes deportivos municipales.

Para alcanzar estos objetivos, se utiliza un muestreo de conglomerados con submuestreo con probabilidades iguales y sin reposición en ambas etapas. En la primera etapa, se seleccionan $m = 4$ centros, y en la segunda etapa, se seleccionan los alumnos con una fracción de muestreo de $f_2 = 0,10$.

La información disponible sobre los centros muestreados es la siguiente:

Centro i	N_i	$\hat{P}_i$	$N_i\hat{P}_i$	$N_i\hat{P}_i(1-\hat{P}_i)$
1	700	0,4	280	168
2	500	0,3	150	105
3	400	0,5	200	100
4	900	0,6	540	216
Total	2.500	1,8	1.170	589

donde $\hat{P}_i$ es la estimación de la propoción de alumnos practicando algún deporte fuera del horario escolar en cada conglomerado muestreado.

Se pide:

a) Una estimación insesgada de la mencionada proporción y su error de muestreo.

b) Consideramos otro estimador de P, definido por el promedio de los porcentajes observados en los centros: $\hat{P}^* = \frac{1}{4}\sum_{i=1}^{4}\hat{P}_i$

1) Demuestre que: $Sesgo(\hat{P}^*) = -\frac{1}{N}\sum_{i=1}^{M} P_i(N_i - \frac{N}{M}) = -\frac{M}{N}Cov(N_i, P_i)$

2) Deduzca las condiciones para las que $\hat{P}^*$ sería un estimador insesgado de P.

c) Suponga que, para obtener el estimador $\hat{\theta}$ de cierta característica del alumnado, se han aplicado varios procedimientos (reponderación por falta de respuesta, calibración con variables auxiliares, ...). Por esta razón, se ha decidido estimar la varianza utilizando el método indirecto de Jackknife. Indique cómo proceder para calcular la varianza con este método en este contexto, proporcionando las expresiones necesarias para la estimación de la varianza.

Cuestión 2. Para la obtención de ciertas estimaciones en una población de tamaño N, se utiliza un muestreo bifásico. Responda a las cuestiones planteadas para cada caso:

a) Para estimar la media de una característica $\bar{Y}$, primero se seleccionan 2.000 elementos para determinar su estrato, mediante un muestreo aleatorio simple sin reemplazamiento. De éstas, corresponden al primer estrato 1.300 y 700 al segundo. En una segunda fase, se lleva a cabo un muestreo aleatorio simple estratificado, seleccionando 150 elementos del primer estrato y 100 en el segundo, a quienes se les solicita medir la característica de interés. Los resultados se muestran en el siguiente cuadro:

Estrato h	$\bar{y}_h$	s^2_{yh}
1	5,4	3,75
2	9,2	6,9

donde $\bar{y}_h$ representa la media muestral en el estrato h de la segunda fase y s^2_{yh} es la varianza muestral del estrato h en la segunda fase. Además, se supone que el tamaño poblacional N es mucho mayor que el tamaño muestral de primera fase.

1) Calcule una estimación insesgada de la media $\bar{Y}$.
2) Estime su varianza.

b) Se desea estimar el total de una característica Y en una población, donde las probabilidades de selección de las unidades son proporcionales a una medida de su tamaño, denotada por M_i. Sin embargo, dado que no se conocen a priori los tamaños de las unidades de la población, se extrae en una primera fase una muestra de tamaño $n^{'}$, con probabilidades iguales y sin reposición, para obtener información acerca de los tamaños $M_1, \ldots, M_i, \ldots, M_{n'}$, siendo $M^{'} = \sum_i^{n^{'}} M_i$. En la segunda fase se obtiene una submuestra de tamaño $n < n^{'}$, con probabilidades proporcionales al tamaño y con reposición. Se pide:

1) Demuestre que $\hat{Y} = \sum_{i=1}^{n} \frac{N}{n} \frac{M^{'}}{n^{'}} \frac{X_i}{M_i}$ es un estimador insesgado del total poblacional Y.
2) Obtenga de forma razonada la varianza de dicho estimador $V(\hat{Y})$.

c) ¿Sería posible utilizar el muestreo bifásico para estimar la media de una característica, $\bar{Y}$, mediante el uso de un estimador de razón, utilizando la información de una variable auxiliar X disponible para todas las unidades muestreadas en primera fase? Exponga de manera concisa cómo se llevaría a cabo cada fase para obtener el estimador de razón.

INSTITUTO NACIONAL DE ESTADÍSTICA



Cuestión 3. Las siguientes estimaciones se calcularon a partir de una muestra de 7.634 mujeres encuestadas en una Encuesta General de Hogares. La variable dependiente toma el valor 1 si la mujer tiene un empleo remunerado, y 0 en caso contrario.

	MCO	Logit	Probit
ALTO	0.093	0.423	0.259
	(0.015)	(0.071)	(0.043)
NOCUAL	–0.210	–0.898	–0.554
	(0.013)	(0.056)	(0.035)
EDAD	0.038	0.173	0.107
	(0.003)	(0.124)	(0.008)
EDAD2	–0.051	–0.230	–0.142
	(0.003)	(0.069)	(0.009)
CAS	0,024	0.103	0.063
	(0.009)	(0.057)	(0.035)
Constante	-0.068	-2.587	-1.593
	(0.049)	(0.225)	(0.137)

Donde ALTO es uno si el encuestado tiene una cualificación educativa superior, cero en caso contrario; NOCUAL es uno si el encuestado no tiene cualificaciones, cero en caso contrario; EDAD es la edad en años; EDAD2 es (edad x edad)/100; CAS es uno si está casado, cero en caso contrario.

Los errores estándar calculados de forma convencional se indican entre paréntesis para los resultados de mínimos cuadrados ordinarios (MCO) y los errores estándar asintóticos entre paréntesis en los demás casos.

a) En el contexto de los modelos GLM, indique qué papel desempeña la función de enlace y cuáles son las funciones de enlace teóricas para cada uno de los modelos estimados, es decir, para MCO, Logit, Probit.

b) Explique brevemente cómo se calculan las estimaciones Probit cuando el modelo no tiene intercepto y sólo tiene una variable explicativa. Indique qué propiedades generales tienen las correspondientes estimaciones.

c) Para los tres conjuntos de estimaciones, pruebe la hipótesis nula de que el coeficiente de CAS es cero. ¿Qué estadístico de contraste consideraría más fiable? Explíquelo.

d) Utilizando las estimaciones MCO y Probit, calcule las probabilidades estimadas de estar desempleada para una mujer casada de 40 años con estudios superiores.

e) Compruebe la hipótesis de significación conjunta del modelo probit, sabiendo que:

$$lnL_R = -416{,}01$$

$$lnL_{NR} = -321{,}25$$

donde lnL_R y lnL_{NR} son los logaritmos de la verosimilitud de los modelos probit restringido y no restringido, respectivamente. Utilice un nivel de significatividad del 5 por ciento.

Cuestión 4. Para promover la vida social y turística de un municipio, determinada Asociación Cultural y Turística ofrece desde 2023 un programa de excursiones anuales para sus miembros de dos tipos, unas de carácter nacional y otras de carácter internacional.

La implementación de este programa se realiza siguiendo las pautas que se establecen a continuación:

- Cada año, los asociados que aún no participan en el programa de excursiones tienen una probabilidad de 0,4 de unirse a las excursiones nacionales y de 0,3 a las internacionales, el año siguiente. Si no se integran al programa, seguirán sin participar en él.
- El 60 % de los asociados que han participado en excursiones nacionales un determinado año continuarán el próximo año participando en las mismas, mientras que el 70 % de los que realizaron las excursiones internacionales repetirá en éstas. El resto de participantes, cambiará al otro tipo de excursión, manteniéndose para siempre en el programa de la asociación.

Considerando que en 2023 ningún miembro participaba en estas excursiones programadas, se pide:

a) Modelar el proceso estocástico asociado utilizando una cadena de Markov, describiendo sus principales componentes y propiedades.

b) Establecer la proporción de miembros de la asociación que durante el año 2025 disfrutarán de cada uno de los tipos de excursiones.

c) Determinar el comportamiento a largo plazo de la proporción de miembros que realizará cada uno de los tipos de excursiones. Asimismo, calcular el tiempo medio que tarda en estar en el programa un asociado que no estaba en él de partida.

INSTITUTO NACIONAL DE ESTADISTICA



Cuestión 5. Suponga que le piden implementar un código que incluya los algoritmos Lasso y random forest para un caso de clasificación. Escriba los pasos a seguir, preferiblemente en código Python, R o pseudocódigo, para llevar a cabo el proyecto completo. Las características del proyecto son las siguientes:

- Objetivo: Utilizar las características del empleado/a para predecir si abandonarán la compañía. Datos disponibles en el archivo "abandono.csv"
- Variable target: Attrition (valores " Yes" , "No")
- Features (covariables): 30 variables (var1 a var30)
- Número de observaciones: 1470

En el código debe tener en cuenta lo siguiente:

a) Teniendo en cuenta que la prevalencia de abandono general está en torno al 16 %, ¿qué métricas de clasificación considera adecuadas para evaluar el desempeño de los algoritmos?.

b) Entre los features (covariables) hay variables de carácter cuantitativo y otras de carácter cualitativo (tanto nominal como ordinal). Algunas variables de carácter cuantitativo pueden contener outliers, tener fuerte asimetría y, habitualmente, estarán en unidades distintas. ¿Qué transformaciones considera adecuadas en las variables de carácter cuantitativo para lograr que los algoritmos puedan tener mejores resultados? Suponga que las variables cuantitativas son las 15 primeras (var1 a var15).

c) En las variables de carácter nominal (var16 a var25), ¿es necesario crear dummies por categorías en ambos algoritmos?. Justifique su respuesta.

d) Suponga que le informan de que algunas de las categorías de las variables cualitativas var17 y var18 tienen muy pocas observaciones. ¿Qué efectos tendría esta situación sobre el entrenamiento y validación de algoritmos y cómo podríamos resolverlo?

e) ¿Cómo se pueden codificar las variables ordinales var26 a var30 en ambos algoritmos?

f) Su supervisora le dice que debe realizarse un proceso de tuneado sobre los hiperparámetros de cada algoritmo. ¿Por qué cree que le pide esto?. Describa detalladamente los pasos para llevar a cabo este proceso teniendo en cuenta que se debe utilizar validación cruzada con la muestra de training. Tenga en cuenta, además, que todos los resultados deben ser reproducibles.

g) ¿Cómo evaluaría la posible existencia de overfitting en los algoritmos planteados?

Cuestión 6. Considere las siguientes relaciones correspondientes a una base de datos de una determinada universidad:

- Student(snum: integer, sname: string, major: string, level: string, age: integer)
- Class(name: string, meets at: string, room: string, fid: integer)
- Enrolled(snum: integer, cname: string)
- Faculty(fid: integer, fname: string, deptid: integer)

La relación Enrolled tiene un registro por cada par estudiante-clase tal que el estudiante está inscrito en la clase. El atributo major en la relación Student se refiere al grado que cursa el estudiante mientras que el atributo fname en la relación Faculty se refiere al nombre del profesor que imparte el curso. No se deben imprimir duplicados en ninguna de las respuestas correspondientes a consultas SQL.

a) Crear las tablas mediante comandos SQL incluyendo las claves y las referencias de integridad necesarias.

b) Escribir código SQL para imprimir los nombres de todos los estudiantes de level = 'JR' que están inscritos en una clase impartida por 'I. Teach'

c) Escribir código SQL para imprimir la edad del estudiante de más edad que esté inscrito en 'Historia' o en un curso impartido por 'I. Teach'

d) Escribir código SQL para imprimir los nombres de los estudiantes inscritos en el mayor número de clases.



Cuestión 7. Ciertos mensajes llegan a una instalación de comunicaciones de acuerdo con un proceso Poisson a razón de 2 por hora. La instalación consta de tres canales, cada mensaje llega a un canal libre, si los tres están libres, o se pierde si todos los canales están ocupados. El tiempo que los mensajes permanecen en un canal es una variable aleatoria que depende de las condiciones meteorológicas al momento de llegada. Si el mensaje llega cuando las condiciones son "buenas", su tiempo de procesamiento es una variable aleatoria con función de distribución:

$$F(x) = x, \qquad 0 < x < 1$$

Mientras que si las condiciones son "malas", su tiempo de procesamiento tiene la función de distribución:

$$F(x) = x^3, \qquad 0 < x < 1$$

Al principio las condiciones son buenas y, posteriormente, se van alternando en períodos buenos y malos: los buenos tienen una duración fija de 2 horas y los malos de una hora.

a) Realice un esbozo de cuáles serían los pasos para simular la distribución del número de mensajes perdidos hasta el instante $T = 100$.

b) Escriba el pseudocódigo, o el código en algún lenguaje imperativo (C++, Java, Python ó R), que permita obtener, mediante simulación, la distribución del número de mensajes perdidos hasta el instante T = 100. Se puede utilizar una función random() que genere valores con distribución $U(0, 1)$.

Cuestión 8. Escriba el pseudocódigo, o el código en algún lenguaje imperativo (C++, Java, Python ó R), de un algoritmo para calcular la media recortada (trimmed mean). El algoritmo debe aceptar como parámetros de entrada una secuencia de números y el parámetro alpha de la media recortada. No se puede utilizar ninguna función tipo sort predefinida sino que la ordenación debe estar incluida en el algoritmo y ser eficiente. Justificar la complejidad algorítmica de todos los pasos del procedimiento descrito.

INSTITUTO NACIONAL DE ESTADISTICA



C.3. Segundo examen del Cuerpo de Diplomados en Estadística del Estado. Turno libre.

EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA

PROCESO SELECTIVO CONVOCADO POR RESOLUCIÓN DE 14 DE DICIEMBRE DE 2023, DE LA SUBSECRETARÍA DE ECONOMÍA, COMERCIO Y EMPRESA (BOE 27 DE DICIEMBRE DE 2023)

CONVOCATORIA DE LA OFERTA PÚBLICA DE EMPLEO 2021, 2022 Y 2023

SEGUNDO EJERCICIO DE LA FASE DE OPOSICIÓN DE
DIPLOMADOS EN ESTADÍSTICA DEL ESTADO
INGRESO LIBRE

CUESTIONARIO 2024/DEE-L

AVISO IMPORTANTE PARA LOS OPOSITORES

1. De conformidad con lo establecido en las bases de la convocatoria, **deberá elegir y responder a siete cuestiones prácticas (7), de las diez (10) propuestas por el Tribunal**, relacionadas con todo el programa del temario excepto el apartado de 'Derecho Público y Organización del Estado'.

   El número de cuestiones prácticas respondidas no podrá ser superior al número de problemas a elegir en este ejercicio. En caso de responder una pregunta que NO quiere que se le corrija, táchela claramente para que el Tribunal lo tenga en cuenta.

2. El **tiempo máximo** para la realización de este ejercicio es de **150 minutos.**



EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA

CUESTIONARIO DE PREGUNTAS CON RESPUESTAS MULTIPLES BASADO EN EL PROGRAMA DEL ANEXO II

CUESTIÓN PRÁCTICA 1

Sea **X** variable aleatoria que sigue una distribución exponencial de parámetro **λ > 0**.
Sea **g** función de **X** tal que $\mathbf{g(X) = X^2}$.

1) Halle **E(g(X))**.
2) Calcule **P(g(X) > 8 | X > 2)**.
3) Sea Y otra variable aleatoria con la misma distribución que X y sea:

$$f(x,y) = \begin{cases} \lambda^2 e^{-\lambda(x+y)} & si\ x \geq 0, y \geq 0 \\ 0 & en\ otro\ caso \end{cases}$$

la función de densidad conjunta de (X, Y).
¿Son X e Y independientes? ¿Por qué?

4) Calcule **E(g(X)Y)**.
5) Sean **U =** $\frac{X}{X+Y}$ y **V = X + Y**. Calcule la función de densidad conjunta de **(U, V)**.

CUESTIÓN PRÁCTICA 2

De una población con *N*=6 unidades, se obtiene una muestra de *n*=2 unidades sin reposición y con probabilidades iguales.

Posteriormente se agregan las unidades en dos estratos:

Estratos *(h)*	X_{hi}
I	0, 1, 3
II	5, 6, 9

Se pide:

1) Varianza del estimador de la media, $V(\bar{x})$, sin estratificar.
2) Varianza de la media con estratificación, $V(\bar{x}_{st})$, y fracción de muestreo $f_h = {}^1/_3$, en cada estrato.

3) Si comparamos las varianzas obtenidas en los apartados a) y b).
¿Qué conclusión se podría sacar?



CUESTIÓN PRÁCTICA 3

Sea X_1, X_2, …, X_n una muestra aleatoria simple de una distribución de Bernoulli de parámetro p con $0 < p < 1$.

1) Utilice el método de los momentos para hallar un estimador de p. También calcule el estimador de p máximo verosímil. ¿Son consistentes?
2) Calcule la Cota de Cramér-Rao y el Estimador Insesgado de Mínima Varianza para estimar el parámetro p.
3) Calcule el error cuadrático medio del estimador obtenido en el apartado 2).
4) Se dispone de una muestra aleatoria simple de viviendas de una urbanización en el 73 A de una calle con datos sobre el número de animales de compañía que conviven en la vivienda almacenada en la siguiente tabla:

NUMER	CALIF	BLOQUE	PORTAL	ESCALERA	PLANTA	PUERTA	NUM_ANI_CO
73	A		1		3	D	2
73	A		1		5	B	0
73	A		2		2	C	0
73	A		1		2	A	0
73	A		2		1	C	2
73	A		2		3	A	1
73	A		2		1	D	1
73	A		2		1	B	0
73	A		1		2	C	0
73	A		2		4	B	3
73	A		1		4	D	0
73	A		1		4	E	1
73	A		2		3	C	0
73	A		1		1	E	0
73	A		2		4	C	1
73	A		1		4	B	1
73	A		2		5	E	0
73	A		1		5	A	1
73	A		1		5	D	2

Denotando p como la probabilidad de que en una vivienda habite algún animal de compañía, calcule la estimación de p para el 73 A, portal 1 usando el estimador obtenido en el apartado 2).

CUESTIÓN PRÁCTICA 4

Se conocen los siguientes datos de las cuentas nacionales de una economía para un año t:

Rúbrica SEC2010	**Unidades Monetarias. Precios corrientes de t**
Formación bruta de capital (P.5g)	300
Otros impuestos (-) subvenciones s/ la producción (D.29 -D.39)	10
Producción (P.1)	2.600
Adquisiciones (-) cesiones de activos no producidos (NP)	5
Consumos intermedios (P.2)	1.400
Exportaciones de bienes y servicios (P.6)	510
Remuneración de los asalariados (D.1)	650
Variación de existencias (P.52)	8
Ahorro nacional neto (B.8n)	110
Transferencias de capital netas del exterior (D.9r -D.9p)	15
Renta disponible ajustada bruta (B.7g)	1.316
Saldo de intercambios exteriores de bienes y servicios (B.11)	26
Consumo de capital fijo (P.51c)	206
Excedente de explotación bruto (B.2g)	410

Para esa economía en el año t se pide:

1) Calcular el PIB a precios de mercado por la perspectiva de demanda.

2) Calcular la cuenta de bienes y servicios y verificar que se cumplen las propiedades de esa cuenta con los datos del problema.

3) Calcular la Renta Mixta Bruta (B.3g). ¿En qué cuentas y para qué sectores institucionales del sistema SEC2010 se calcula la Renta Mixta Bruta?

4) Con los datos que ha obtenido para esta economía ¿se puede calcular el PIB por tres ópticas alternativas? Justificar la respuesta conceptual y numéricamente.

5) Si el PIB pm de esa economía en el año precedente (t-1) era 1250 (a precios corrientes de ese año) y se sabe que la variación del deflactor implícito del PIB entre los dos periodos fue del 2%, calcular cuál fue la tasa de crecimiento real del PIB de esa economía entre t-1 y t.



CUESTIÓN PRÁCTICA 5

Sea una determinada población que a 1 de enero del año 2023 cuenta con 4.111.555 individuos y supongamos que durante el año 2023 se producen 41.198 nacimientos, 37.760 defunciones, 20.084 inmigraciones y 908 emigraciones.

Considere los siguientes datos:

$e^{0,0135} = 1,0135$	$\sqrt{1,0055} = 1,0027$	$(1,0027)^5 = 1,0135$

1) Calcule el crecimiento natural, el saldo migratorio y el crecimiento total que dicha población registra durante el año 2023.
2) Utilizando el crecimiento experimentado por esta población en 2023 y suponiendo un crecimiento aritmético de la misma, ¿qué tamaño alcanzaría a 1 de enero de 2029?
3) Utilizando el crecimiento experimentado por esta población en 2023, ¿qué tamaño alcanzaría a 1 de enero de 2029 suponiendo un crecimiento geométrico?
4) Por último, utilizando el crecimiento experimentado por esta población en 2023, realizar el mismo cálculo suponiendo un crecimiento exponencial.

CUESTIÓN PRÁCTICA 6

En el mercado de un determinado bien, la función de oferta responde a la forma: $Q_o = 40\ p$ mientras que la de demanda se estima que es: $Q_d = 4000 - 10\ p$.

Donde Q_d es la cantidad demandada, Q_o es la cantidad ofertada y p es el precio mostrado en unidades monetarias genéricas.

1) Calcule el precio y la cantidad de equilibrio.
2) Represéntelo gráficamente.
3) Halle el excedente del consumidor y del productor.
4) Suponga ahora que el Gobierno fija un impuesto a las empresas por el que la función de oferta varía pasando a ser: $Q'_o = 20\ p - 200$. Calcule el nuevo precio y la nueva cantidad de equilibrio y represéntelo gráficamente.
5) Calcule la perdida irrecuperable de eficiencia que se ha producido por la imposición del impuesto.

CUESTIÓN PRÁCTICA 7

Un establecimiento registra el número de unidades vendidas de un producto (*X*) y el precio unitario en euros (*Y*), para cada uno de los trimestres de tres años consecutivos:

Año	Trimestre	X	Y
2015	Trimestre 1	80	20
	Trimestre 2	100	30
	Trimestre 3	100	20
	Trimestre 4	110	30
2016	Trimestre 1	90	40
	Trimestre 2	120	20
	Trimestre 3	110	30
	Trimestre 4	90	50
2017	Trimestre 1	100	20
	Trimestre 2	100	40
	Trimestre 3	100	20
	Trimestre 4	100	40

Donde[1]:

$$\sum_i \sum_j x_i * n_{ij} = 1.200,\ \sum_i \sum_j {x_i}^2 * n_{ij} = 121.200$$

$$\sum_i \sum_j y_j * n_{ij} = 120,\ \sum_i \sum_j {y_i}^2 * n_{ij} = 400$$

$$\sum_i \sum_j x_i * y_j * n_{ij} = 35.700$$

Calcule:

1) La media aritmética y mediana de las ventas trimestrales.
2) En los trimestres en los que se producen 100 unidades, calcule el precio medio pagado por unidad.
3) El coeficiente de variación de las ventas trimestrales.
4) Calcule la moda y la varianza de la variable Z=3X+2.
5) La covarianza de las variables X e Y.

[1] En el desarrollo del examen se detectó una incoherencia en los datos de la cuestión práctica 7. En la corrección se valorará tanto el uso de los resultados del enunciado en el texto como los derivados de la tabla.

CUESTIÓN PRÁCTICA 8

Un determinado IPC sectorial con Base 2019 (media anual=100), tipo Laspeyres, consta de 21 artículos clasificados en siete grupos.

1) Conocidos los siguientes datos en marzo 2024:

Agrupación	**Índice (período *t*)**	**Ponderación (tanto por uno)**
IPC General	114,9	1,000
Grupo 6 (telecom.)	119,3	0,180
Artículo 6.1 (telefonía móvil)	132,9	0,045

Se pide calcular la repercusión que supondría en el grupo 6 (telecomunicaciones) un incremento de precios del 7% en el artículo 6.1 (telefonía móvil) en el mes de abril-24.

También se pide calcular la repercusión en el IPC General de esa misma variación del 7% en el artículo 6.1 en abril-24.

2) Conocido ese IPC General = **114,9** en el mes de marzo-24, se pide obtener el nuevo IPC General resultante para marzo-24 tras excluir del cálculo del Índice tres artículos de los veintiunos iniciales. En la tabla de más abajo se indican los datos de los tres artículos a excluir.

Artículo	**Índice (período *t*)**	**Ponderación (tanto por uno)**
Art. 3.1	120	0,1
Art. 4.2	107	0,2
Art. 7.1	115	0,1

Calcular el nuevo Índice general en el período *t (marzo-24)*

CUESTIÓN PRÁCTICA 9

Los datos de producción de energía eléctrica y renta per cápita de un determinado país en el período 2015 a 2019 fueron:

Años	Energía Eléctrica (miles mill. kW/h)	Renta Per Cápita (miles €)
2015	32	35
2016	38	40
2017	41	43
2018	46	47
2019	52	51

Se pide:

1) Hallar las ecuaciones de las rectas de regresión y dibujarlas
2) Significado de los coeficientes de regresión
3) Calcular el valor del coeficiente de correlación
4) Previsión de la renta per cápita cuando la producción de energía eléctrica sea de 60.000 millones de kW/h
5) ¿Qué confianza merecen las predicciones?

CUESTIÓN PRÁCTICA 10

Dado el proceso:

$Y_t=1,4Y_{t-1} - 0,5Y_{t-2} + a_t$

Donde a_t es un proceso de ruido blanco con varianza δ_a^2 se pide:

1) ¿De qué tipo de proceso estamos hablando? Escribir el proceso en notación de retardos.
2) ¿Qué significa que un proceso es estacionario?
3) ¿Cuáles son las condiciones que se deben cumplir para que este proceso sea estacionario?
4) ¿Este proceso es estacionario? Razone la respuesta.



EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA



EX LIBRIS
INSTITUTO NACIONAL
DE
ESTADISTICA



Apéndice D

Exámenes de problemas de 2023. Promoción interna.

D.1. Segundo examen del Cuerpo Superior de Estadísticos del Estado. Parte común. Promoción interna.

MINISTERIO DE ECONOMÍA, COMERCIO Y EMPRESA

INE Instituto Nacional de Estadística

Oposición al Cuerpo Superior de Estadísticos de Estado

Segundo Ejercicio

Convocatoria de la oferta pública de empleo de 2023

Resolución de 14 de diciembre de 2023, de la Subsecretaría, por la que se convocan procesos selectivos para ingreso, por el sistema de acceso libre y promoción interna, en el Cuerpo Superior de Estadísticos del Estado. (BOE 27 de Diciembre de 2023)

Acceso Promoción Interna, CDEE



Cuestión 1. En una población formada por $N = 8$ unidades $\{1, 2, 3, 4, 5, 6, 7, 8\}$, se considera un diseño muestral que genera las siguientes muestras posibles:

$$p(\{1,3,5,6\}) = \frac{1}{8}; \quad p(\{1,4,6,8\}) = \frac{1}{8}; \quad p(\{4,5,7,8\}) = \frac{1}{8}$$
$$p(\{2,3,7,8\}) = \frac{1}{4}; \quad p(\{2,4,6,8\}) = \frac{3}{8}$$

Se pide:

a) Calcular las probabilidades de inclusión de primer orden, π_i, de todos los elementos de la población.

b) El diseño de la muestra, ¿es aleatorio simple sin reemplazamiento, aleatorio simple estratificado o ninguno de ellos? Razonar la respuesta.

c) Si los tamaños de estas 8 unidades son $\{100, 40, 40, 20, 20, 10, 5, 5\}$ respectivamente, ¿podría aplicarse un diseño muestral sin reposición, con tamaño $n = 4$, de manera que las π_i sean proporcionales a sus tamaños? Justifique su respuesta.

INSTITUTO NACIONAL DE ESTADISTICA

Cuestión 2. La función de densidad conjunta de las variables (X, Y) es la siguiente:

$$f_a(x, y) = a^2 e^{a(x+y)}, x \geq 0, y \geq 0.$$

Se toma una muestra aleatoria simple de tamaño n.

a) Calcule el estimador máximo verosímil del parámetro a.

b) Utilice dicho estimador para obtener $\hat{a}$ a partir de la información muestral recogida en la siguiente tabla:

x	1	2	3	4	5
y	2	0	1	0	1

CUERPOS DE NUMERUS CLAUSUS INSTITUTO NACIONAL DE ESTADISTICA



Cuestión 3. .

a) Sea un sistema de representación numérica de enteros con 8 bits.

- Expresar el número decimal -39, en representación binaria utilizando formatos signo-magnitud (con primer dígito para signo), complemento a 1 y complemento a 2. El dígito de signo se representa como 1 negativo y 0 positivo.
- Si un número binario en complemento a 2 se representa como 10101010, ¿cuál es su valor decimal?.
- ¿Cuál es el entero mínimo y máximo representables en complemento a 2 en este sistema (que incluye bit de signo)?.

b) En un sistema binario (sin bit para signo), ¿a qué valor decimal exacto equivale el número 110101,111?

c) ¿Cuál es el valor decimal exacto del número binario $0,10101101101\ldots = 0,1\widehat{0101}$?

Cuestión 4. Una empresa especializada en servicios de seguridad busca estimar el total de viviendas equipadas con sistemas de seguridad en una determinada área urbana, con el propósito de desarrollar estrategias de marketing dirigidas a promover sus servicios. Para lo cual se seleccionan 80 conglomerados utilizando muestreo aleatorio simple sin reposición y probabilidades iguales con una fracción de muestreo de $\frac{1}{50}$. El tamaño de cada conglomerado es de 10 viviendas. Asimismo, para cada conglomerado i seleccionado en la muestra, se denota por A_i al número total de viviendas que tienen contratadas servicios de seguridad. Los resultados de la investigación son los siguientes:

$$\sum_{i=1}^{80} A_i = 370 \quad \sum_{i=1}^{80} A_i^2 = 2536$$

Se pide:

a) Una estimación insesgada del total de viviendas con servicios de seguridad contratados en el área urbana y su error de muestreo relativo o coeficiente de variación.

b) Una estimación del efecto de diseño del muestreo descrito comparado con el muestreo aleatorio simple sin reposición.

c) Estimar el coeficiente de correlación intraconglomerados o coeficiente de homogeneidad de los conglomerados.



Cuestión 5. Considere que la esperanza condicional de una variable dependiente y es la siguiente

$$E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

donde las variables son continuas.

a) Calcule qué efecto esperado (efecto parcial) sobre y tiene un cambio de una unidad en la variable x_1, considerando que x_2 permanece constante.

b) Si dispusiera de una muestra de tamaño n, y quisiera estimar el siguiente modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + u_i, i = 1, ..., n,$$

donde los errores u_i no son observables, ¿cómo se deduciría y cuál sería la expresión del estimador de mínimos cuadrados ordinarios de $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$?

Cuestión 6. Para el estudio del comportamiento medio de una determinada caracterísica cuantitativa Y en una población, se propone aplicar un muestreo aleatorio estratificado.

La población ha sido agrupada en tres estratos de acuerdo a su edad: menos de 20 años, de 20 a 49 años y de 50 o más años. La información de la que se dispone de los estratos es la siguiente:

Estrato	Tramos edad	N_h	$\bar{Y}_h$	S_h^2
1	De 0 a 20 años	150	20	36
2	De 20 a 49 años	250	35	144
3	De 50 o más años	200	60	256

Donde :

- N_h= Número de personas en el estrato h.
- $\bar{Y}_h$= Valor medio de la característica en el estrato h.
- S_h^2= Cuasivarianza poblacional de Y en el estrato h.

Se realiza un muestreo sin reemplazamiento y con probabilidades iguales en cada estrato. Se pide:

a) Obtener la cuasivarianza poblacional de Y. ¿Que proporción supone la variación interestrato de la variación total?

b) Comentar a la vista de los resultados obtenidos en el apartado anterior, si la variable de estratificación es apropiada en términos de ganancia en precisión.

c) Calcular el tamaño muestral necesario, para que siendo una muestra autoponderada, se obtenga una varianza para el estimador de la media de 1,45. Asimismo, calcular los tamaños muestrales en los estratos.



D.2. Cuarto examen del Cuerpo Superior de Estadísticos del Estado. Especialidad ciencia de datos. Promoción interna.

MINISTERIO
DE ECONOMÍA, COMERCIO
Y EMPRESA

INE
Instituto Nacional de Estadística

Oposición al Cuerpo Superior de Estadísticos de Estado

Cuarto Ejercicio

Convocatoria de la oferta pública de empleo de 2023

Resolución de 14 de diciembre de 2023, de la Subsecretaría, por la que se convocan procesos selectivos para ingreso, por el sistema de acceso libre y promoción interna, en el Cuerpo Superior de Estadísticos del Estado. (BOE 27 de Diciembre de 2023)

Especialidad I: Estadística-Ciencia de Datos



Cuestión 1. Sea un municipio con un total de N =8.000 alumnos distribuidos en $M = 16$ centros escolares de tamaños desiguales $(N_i, \forall i = 1 \ldots, 16)$. Se desea estimar la proporción P de estudiantes que practican algún deporte fuera del horario escolar para ajustar los planes deportivos municipales.

Para alcanzar estos objetivos, se utiliza un muestreo de conglomerados con submuestreo con probabilidades iguales y sin reposición en ambas etapas. En la primera etapa, se seleccionan $m = 4$ centros, y en la segunda etapa, se seleccionan los alumnos con una fracción de muestreo de $f_2 = 0,10$.

La información disponible sobre los centros muestreados es la siguiente:

Centro i	N_i	$\hat{P}_i$	$N_i\hat{P}_i$	$N_i\hat{P}_i(1-\hat{P}_i)$
1	700	0,4	280	168
2	500	0,3	150	105
3	400	0,5	200	100
4	900	0,6	540	216
Total	2.500	1,8	1.170	589

donde $\hat{P}_i$ es la estimación de la propoción de alumnos practicando algún deporte fuera del horario escolar en cada conglomerado muestreado.

Se pide:

a) Una estimación insesgada de la mencionada proporción y su error de muestreo.

b) Consideramos otro estimador de P, definido por el promedio de los porcentajes observados en los centros: $\hat{P}^* = \frac{1}{4}\sum_{i=1}^{4}\hat{P}_i$

1) Demuestre que: $Sesgo(\hat{P}^*) = -\frac{1}{N}\sum_{i=1}^{M} P_i(N_i - \frac{N}{M}) = -\frac{M}{N}Cov(N_i, P_i)$

2) Deduzca las condiciones para las que $\hat{P}^*$ sería un estimador insesgado de P.

c) Suponga que, para obtener el estimador $\hat{\theta}$ de cierta característica del alumnado, se han aplicado varios procedimientos (reponderación por falta de respuesta, calibración con variables auxiliares, ...). Por esta razón, se ha decidido estimar la varianza utilizando el método indirecto de Jackknife. Indique cómo proceder para calcular la varianza con este método en este contexto, proporcionando las expresiones necesarias para la estimación de la varianza.

Cuestión 2. Para la obtención de ciertas estimaciones en una población de tamaño N, se utiliza un muestreo bifásico. Responda a las cuestiones planteadas para cada caso:

a) Para estimar la media de una característica $\bar{Y}$, primero se seleccionan 2.000 elementos para determinar su estrato, mediante un muestreo aleatorio simple sin reemplazamiento. De éstas, corresponden al primer estrato 1.300 y 700 al segundo. En una segunda fase, se lleva a cabo un muestreo aleatorio simple estratificado, seleccionando 150 elementos del primer estrato y 100 en el segundo, a quienes se les solicita medir la característica de interés. Los resultados se muestran en el siguiente cuadro:

Estrato h	$\bar{y}_h$	s^2_{yh}
1	5,4	3,75
2	9,2	6,9

donde $\bar{y}_h$ representa la media muestral en el estrato h de la segunda fase y s^2_{yh} es la varianza muestral del estrato h en la segunda fase. Además, se supone que el tamaño poblacional N es mucho mayor que el tamaño muestral de primera fase.

1) Calcule una estimación insesgada de la media $\bar{Y}$.
2) Estime su varianza.

b) Se desea estimar el total de una característica Y en una población, donde las probabilidades de selección de las unidades son proporcionales a una medida de su tamaño, denotada por M_i. Sin embargo, dado que no se conocen a priori los tamaños de las unidades de la población, se extrae en una primera fase una muestra de tamaño $n^{'}$, con probabilidades iguales y sin reposición, para obtener información acerca de los tamaños $M_1, \ldots, M_i, \ldots, M_{n'}$, siendo $M^{'} = \sum_i^{n^{'}} M_i$. En la segunda fase se obtiene una submuestra de tamaño $n < n^{'}$, con probabilidades proporcionales al tamaño y con reposición. Se pide:

1) Demuestre que $\hat{Y} = \sum_{i=1}^{n} \frac{N}{n} \frac{M^{'}}{n^{'}} \frac{X_i}{M_i}$ es un estimador insesgado del total poblacional Y.
2) Obtenga de forma razonada la varianza de dicho estimador $V(\hat{Y})$.

c) ¿Sería posible utilizar el muestreo bifásico para estimar la media de una característica, $\bar{Y}$, mediante el uso de un estimador de razón, utilizando la información de una variable auxiliar X disponible para todas las unidades muestreadas en primera fase? Exponga de manera concisa cómo se llevaría a cabo cada fase para obtener el estimador de razón.

INSTITUTO NACIONAL DE ESTADISTICA



Cuestión 3. Las siguientes estimaciones se calcularon a partir de una muestra de 7.634 mujeres encuestadas en una Encuesta General de Hogares. La variable dependiente toma el valor 1 si la mujer tiene un empleo remunerado, y 0 en caso contrario.

	MCO	Logit	Probit
ALTO	0.093	0.423	0.259
	(0.015)	(0.071)	(0.043)
NOCUAL	–0.210	–0.898	–0.554
	(0.013)	(0.056)	(0.035)
EDAD	0.038	0.173	0.107
	(0.003)	(0.124)	(0.008)
EDAD2	–0.051	–0.230	–0.142
	(0.003)	(0.069)	(0.009)
CAS	0,024	0.103	0.063
	(0.009)	(0.057)	(0.035)
Constante	-0.068	-2.587	-1.593
	(0.049)	(0.225)	(0.137)

Donde ALTO es uno si el encuestado tiene una cualificación educativa superior, cero en caso contrario; NOCUAL es uno si el encuestado no tiene cualificaciones, cero en caso contrario; EDAD es la edad en años; EDAD2 es (edad x edad)/100; CAS es uno si está casado, cero en caso contrario.

Los errores estándar calculados de forma convencional se indican entre paréntesis para los resultados de mínimos cuadrados ordinarios (MCO) y los errores estándar asintóticos entre paréntesis en los demás casos.

a) En el contexto de los modelos GLM, indique qué papel desempeña la función de enlace y cuáles son las funciones de enlace teóricas para cada uno de los modelos estimados, es decir, para MCO, Logit, Probit.

b) Explique brevemente cómo se calculan las estimaciones Probit cuando el modelo no tiene intercepto y sólo tiene una variable explicativa. Indique qué propiedades generales tienen las correspondientes estimaciones.

c) Para los tres conjuntos de estimaciones, pruebe la hipótesis nula de que el coeficiente de CAS es cero. ¿Qué estadístico de contraste consideraría más fiable? Explíquelo.

d) Utilizando las estimaciones MCO y Probit, calcule las probabilidades estimadas de estar desempleada para una mujer casada de 40 años con estudios superiores.

e) Compruebe la hipótesis de significación conjunta del modelo probit, sabiendo que:

$$lnL_R = -416{,}01$$

$$lnL_{NR} = -321{,}25$$

donde lnL_R y lnL_{NR} son los logaritmos de la verosimilitud de los modelos probit restringido y no restringido, respectivamente. Utilice un nivel de significatividad del 5 por ciento.

Cuestión 4. Para promover la vida social y turística de un municipio, determinada Asociación Cultural y Turística ofrece desde 2023 un programa de excursiones anuales para sus miembros de dos tipos, unas de carácter nacional y otras de carácter internacional.

La implementación de este programa se realiza siguiendo las pautas que se establecen a continuación:

- Cada año, los asociados que aún no participan en el programa de excursiones tienen una probabilidad de 0,4 de unirse a las excursiones nacionales y de 0,3 a las internacionales, el año siguiente. Si no se integran al programa, seguirán sin participar en él.
- El 60 % de los asociados que han participado en excursiones nacionales un determinado año continuarán el próximo año participando en las mismas, mientras que el 70 % de los que realizaron las excursiones internacionales repetirá en éstas. El resto de participantes, cambiará al otro tipo de excursión, manteniéndose para siempre en el programa de la asociación.

Considerando que en 2023 ningún miembro participaba en estas excursiones programadas, se pide:

a) Modelar el proceso estocástico asociado utilizando una cadena de Markov, describiendo sus principales componentes y propiedades.

b) Establecer la proporción de miembros de la asociación que durante el año 2025 disfrutarán de cada uno de los tipos de excursiones.

c) Determinar el comportamiento a largo plazo de la proporción de miembros que realizará cada uno de los tipos de excursiones. Asimismo, calcular el tiempo medio que tarda en estar en el programa un asociado que no estaba en él de partida.

INSTITUTO NACIONAL DE ESTADISTICA



Cuestión 5. Suponga que le piden implementar un código que incluya los algoritmos Lasso y random forest para un caso de clasificación. Escriba los pasos a seguir, preferiblemente en código Python, R o pseudocódigo, para llevar a cabo el proyecto completo. Las características del proyecto son las siguientes:

- Objetivo: Utilizar las características del empleado/a para predecir si abandonarán la compañía. Datos disponibles en el archivo "abandono.csv"
- Variable target: Attrition (valores " Yes" , "No")
- Features (covariables): 30 variables (var1 a var30)
- Número de observaciones: 1470

En el código debe tener en cuenta lo siguiente:

a) Teniendo en cuenta que la prevalencia de abandono general está en torno al 16 %, ¿qué métricas de clasificación considera adecuadas para evaluar el desempeño de los algoritmos?.

b) Entre los features (covariables) hay variables de carácter cuantitativo y otras de carácter cualitativo (tanto nominal como ordinal). Algunas variables de carácter cuantitativo pueden contener outliers, tener fuerte asimetría y, habitualmente, estarán en unidades distintas. ¿Qué transformaciones considera adecuadas en las variables de carácter cuantitativo para lograr que los algoritmos puedan tener mejores resultados? Suponga que las variables cuantitativas son las 15 primeras (var1 a var15).

c) En las variables de carácter nominal (var16 a var25), ¿es necesario crear dummies por categorías en ambos algoritmos?. Justifique su respuesta.

d) Suponga que le informan de que algunas de las categorías de las variables cualitativas var17 y var18 tienen muy pocas observaciones. ¿Qué efectos tendría esta situación sobre el entrenamiento y validación de algoritmos y cómo podríamos resolverlo?

e) ¿Cómo se pueden codificar las variables ordinales var26 a var30 en ambos algoritmos?

f) Su supervisora le dice que debe realizarse un proceso de tuneado sobre los hiperparámetros de cada algoritmo. ¿Por qué cree que le pide esto?. Describa detalladamente los pasos para llevar a cabo este proceso teniendo en cuenta que se debe utilizar validación cruzada con la muestra de training. Tenga en cuenta, además, que todos los resultados deben ser reproducibles.

g) ¿Cómo evaluaría la posible existencia de overfitting en los algoritmos planteados?

Cuestión 6. Considere las siguientes relaciones correspondientes a una base de datos de una determinada universidad:

- Student(snum: integer, sname: string, major: string, level: string, age: integer)
- Class(name: string, meets at: string, room: string, fid: integer)
- Enrolled(snum: integer, cname: string)
- Faculty(fid: integer, fname: string, deptid: integer)

La relación Enrolled tiene un registro por cada par estudiante-clase tal que el estudiante está inscrito en la clase. El atributo major en la relación Student se refiere al grado que cursa el estudiante mientras que el atributo fname en la relación Faculty se refiere al nombre del profesor que imparte el curso. No se deben imprimir duplicados en ninguna de las respuestas correspondientes a consultas SQL.

a) Crear las tablas mediante comandos SQL incluyendo las claves y las referencias de integridad necesarias.

b) Escribir código SQL para imprimir los nombres de todos los estudiantes de level = 'JR' que están inscritos en una clase impartida por 'I. Teach'

c) Escribir código SQL para imprimir la edad del estudiante de más edad que esté inscrito en 'Historia' o en un curso impartido por 'I. Teach'

d) Escribir código SQL para imprimir los nombres de los estudiantes inscritos en el mayor número de clases.



Cuestión 7. Ciertos mensajes llegan a una instalación de comunicaciones de acuerdo con un proceso Poisson a razón de 2 por hora. La instalación consta de tres canales, cada mensaje llega a un canal libre, si los tres están libres, o se pierde si todos los canales están ocupados. El tiempo que los mensajes permanecen en un canal es una variable aleatoria que depende de las condiciones meteorológicas al momento de llegada. Si el mensaje llega cuando las condiciones son "buenas", su tiempo de procesamiento es una variable aleatoria con función de distribución:

$$F(x) = x, \qquad 0 < x < 1$$

Mientras que si las condiciones son "malas", su tiempo de procesamiento tiene la función de distribución:

$$F(x) = x^3, \qquad 0 < x < 1$$

Al principio las condiciones son buenas y, posteriormente, se van alternando en períodos buenos y malos: los buenos tienen una duración fija de 2 horas y los malos de una hora.

a) Realice un esbozo de cuáles serían los pasos para simular la distribución del número de mensajes perdidos hasta el instante $T = 100$.

b) Escriba el pseudocódigo, o el código en algún lenguaje imperativo (C++, Java, Python ó R), que permita obtener, mediante simulación, la distribución del número de mensajes perdidos hasta el instante T = 100. Se puede utilizar una función random() que genere valores con distribución $U(0, 1)$.

Cuestión 8. Escriba el pseudocódigo, o el código en algún lenguaje imperativo (C++, Java, Python ó R), de un algoritmo para calcular la media recortada (trimmed mean). El algoritmo debe aceptar como parámetros de entrada una secuencia de números y el parámetro alpha de la media recortada. No se puede utilizar ninguna función tipo sort predefinida sino que la ordenación debe estar incluida en el algoritmo y ser eficiente. Justificar la complejidad algorítmica de todos los pasos del procedimiento descrito.

INSTITUTO NACIONAL DE ESTADISTICA



Apéndice E

Galeria de imágenes

E.1. Autoría y referencias bibliográficas de las imágenes portada de cada capítulo

1. Imágenes de la portada del capítulo 2 (Año 2022):
   - Dibujo del número pi con los dígitos coloreados:
   https://www.reddit.com/r/dataisbeautiful/comments/7apa1y/2500_digits_of_pi_coloured_oc/
   - Imágenes de los simpsons:
   1) Imagen (2,2): https://experimentalmath.info/blog/category/disclaimer-and-copyright/ El copyright es de David H. Bailey ©2023.
   2) Imagen (1,2): http://www.numberphile.com/
   3) Imagen (2,1): https://t.me/Crypto_Trendingz
   4) Imagen (1,1): https://www.pinterest.es/pin/859483910111063672/ drawthesimpsons.tumblr.com

2. Imagen de la portada del capítulo 3 (Año 2022):
   La imagen la hemos tomado de la web:

- https://www.revistadearte.com/2022/01/23/los-maestros-antiguos-cara-a-cara-en-el-museo-picasso-malaga/

Los derechos de esta imagen son de:

Pablo Picasso (1881-1973)

Busto de hombre

Mougins, 8 noviembre 1970

Óleo sobre contrachapado

96,5 x 57,5 cm /

Fundación Almine y Bernard Ruiz-Picasso para el Arte, Madrid

©FABA Foto: Soko-studio

©Sucesión Pablo Picasso, VEGAP, Madrid, 2021

3. Imagen de la portada del capítulo 4 (Año 2023):

   La imagen la hemos tomado de la siguiente web, de acuerdo al nombre de la pintura M-Maybe:

   - https://www.lichtensteincatalogue.org/catalogue/entry.php?id=578e

   Es una pintura que está en el Museo Ludwig, en Colonia, Alemania.

4. Imagen de la portada del capítulo 5 (Año 2023):

   Hemos cogido la imagen de la web:

   - https://www.getyourguide.es/barcelona-l45/barcelona-casa-batllo-experiencia-10d-t398519/?ranking_uuid=6be3ae24-2deb-4d91-9345-0ec212ff6fd5/

   Desconocemos quien es el autor de la fotografía.



EX LIBRIS
INSTITUTO NACIONAL DE ESTADISTICA



9 788409 557516