



MINISTERIO
DE ECONOMÍA, COMERCIO
Y EMPRESA

INE
Instituto Nacional de Estadística

Oposición al Cuerpo Superior de Estadísticos de Estado

Cuarto Ejercicio

Especialidad I: Estadística y Ciencia de Datos

Convocatoria de la oferta pública de empleo de 2025

Resolución de 22 de diciembre de 2025, 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 29 de Diciembre de 2025)

Cuestión 1. Sea $U = \{1, \dots, N\}$ una población finita de tamaño $N = 1500$ y sea $Y = \sum_{k \in U} y_k$ el total poblacional que se desea estimar. Asimismo, se considera una medida de tamaño $x_k > 0$ asociada a cada unidad k , cuyos valores no se conocen para todas las unidades de U .

Se selecciona mediante muestreo aleatorio simple sin reemplazamiento una muestra s_a de tamaño $n_a = 150$, en cuyas unidades se observan los valores x_k . A continuación, entre las unidades de s_a se realizan $n = 4$ extracciones independientes con reemplazamiento, con probabilidades de selección proporcionales a x_k . Los valores y_k se observan únicamente para las unidades seleccionadas en estas extracciones.

Para la muestra s_a seleccionada se obtiene $X_{s_a} = \sum_{k \in s_a} x_k = 9000$, y las cuatro extracciones posteriores proporcionan los siguientes valores:

Extracción	x_k	y_k
1	45	72
2	90	126
3	60	102
4	45	72

- Construya, a partir del procedimiento de selección descrito, un estimador de expansión del total poblacional Y , especificando las probabilidades de selección consideradas, y calcule su valor para la muestra observada.
- Demuestre que el estimador obtenido en el apartado anterior es insesgado para el total poblacional Y .
- Deduzca la expresión de la varianza del estimador $\hat{Y}$ a partir del procedimiento de selección descrito, identificando las contribuciones debidas a cada una de sus dos etapas.

Cuestión 2. Un equipo de investigación biomédica ha recopilado información sobre pacientes con distintos perfiles clínicos. El objetivo es desarrollar un modelo predictivo robusto que permita clasificar automáticamente a nuevos pacientes en uno de los tres grupos diagnósticos posibles: Tipo 1, Tipo 2 y Tipo 3. Para ello, se dispone del archivo `perfil_pacientes.csv`, el cual contiene las siguientes características:

- Variable a predecir (target): Diagnóstico, con niveles “Tipo 1”, “Tipo 2” y “Tipo 3”, cuyas frecuencias relativas en el fichero son, respectivamente, del 40 %, el 50 % y el 10 %.
- Total de observaciones: 3.000.
- Variables predictoras (35 en total):
 - `var_1` a `var_20`: cuantitativas, con escalas heterogéneas y presencia de asimetrías marcadas y valores atípicos.
 - `var_21` a `var_30`: cualitativas nominales.
 - `var_31` a `var_35`: cualitativas ordinales con entre 3 y 5 niveles por variable.

Se solicita el desarrollo de un procedimiento automático de clasificación utilizando Máquinas de Vector Soporte (SVM) mediante kernel de base radial (RBF). Escriba, con este fin, un script detallado (en R, Python, lenguaje similar o pseudocódigo estructurado) que aborde los aspectos enumerados en los siguientes apartados.

Indicación: Para llevar a cabo el procedimiento, inicialmente divida el fichero en un conjunto de training (70%) y un conjunto de test (30%) mediante muestreo estratificado por la variable target.

- (a) Describa qué transformaciones de preprocesamiento considera necesarias para preparar las variables cuantitativas, las cualitativas nominales y las ordinales, teniendo en cuenta el algoritmo utilizado. Encapsule todos los pasos en un único flujo de preprocesamiento (por ejemplo, un pipeline de scikit-learn, un objeto recipe de tidymodels o su equivalente descrito en pseudocódigo).
- (b) Proponga y justifique una métrica adecuada, teniendo en cuenta el algoritmo elegido, para este problema de clasificación multiclase, en el que las clases están desbalanceadas con frecuencias del 40 %, el 50 % y el 10 % para Tipo 1, Tipo 2 y Tipo 3, respectivamente.
- (c) Realice, mediante validación cruzada utilizando exclusivamente datos de training, un tuneado del modelo SVM con kernel radial. Indique qué hiperparámetros deben ajustarse y cuál es el rango o rejilla para cada hiperparámetro. Describa el esquema de validación cruzada para elegir los mejores valores del tuneado garantizando la reproducibilidad de los resultados. Indique, asimismo, cómo validaría finalmente el procedimiento elegido antes de su puesta en producción.

Cuestión 3. Una compañía aseguradora dispone de una reserva inicial de 100 unidades monetarias para hacer frente a 20 siniestros. Las cuantías $X_1, \dots, X_{20}$ son independientes y siguen una distribución Pareto con parámetros $x_m = 2$ y $\alpha = 3$, cuya función de distribución es

$$F(x) = 1 - \left(\frac{2}{x}\right)^3, \quad x \geq 2.$$

Se define el importe agregado de los siniestros como:

$$S = X_1 + \dots + X_{20}$$

Se considera que se produce la ruina si el importe agregado excede la reserva, es decir, la probabilidad de ruina viene dada por:

$$p = P(S > 100)$$

- (a) Diseñe un algoritmo en Python, R, otro lenguaje similar o pseudocódigo para estimar p mediante M simulaciones independientes. Tenga en cuenta que, en el entorno de simulación, únicamente pueden generarse números aleatorios independientes con distribución $Uniforme(0, 1)$. Se supondrá que el generador no devuelve el valor 1. Indique la complejidad temporal del algoritmo en función de M . Obtenga un estimador del error estándar para el estimador propuesto y construya un intervalo de confianza asintótico bilateral del 95 % para p basado en la aproximación normal.
- (b) Para aumentar la frecuencia de las simulaciones que producen ruina, se propone generar los siniestros bajo una distribución Pareto alternativa con el mismo parámetro de escala ($x_m = 2$) y parámetro de forma $\alpha^* = 2$, cuya función de densidad es

$$f^*(x) = \frac{2 \cdot 2^2}{x^3} = \frac{8}{x^3}, \quad x \geq 2.$$

La función de densidad de la distribución original es

$$f(x) = \frac{3 \cdot 2^3}{x^4} = \frac{24}{x^4}, \quad x \geq 2.$$

Obtenga la razón de verosimilitudes (o peso de importancia) de una muestra $x_1, \dots, x_{20}$ simulada bajo la distribución alternativa (simplifique la expresión obtenida cuando sea posible).

- (c) Adapte el algoritmo del apartado 1 para calcular de nuevo el estimador de p por muestreo de importancia (utilizando los pesos obtenidos en el apartado b). Estime, a partir de las M simulaciones, el error estándar del nuevo estimador y explique cómo compararía la eficiencia de los estimadores obtenidos (mediante simulación directa y muestreo de importancia).

Cuestión 4. Se estudia el número anual de incidencias registrado en 100 centros sanitarios con distintos niveles de exposición. Para analizar la influencia de un nuevo protocolo se considera inicialmente el modelo

$$Y_i \sim \text{Poisson}(\mu_i), \qquad \log(\mu_i) = \log(E_i) + \beta_0 + \beta_1 X_i,$$

donde E_i es el número de miles de pacientes atendidos y $X_i = 1$ si el centro aplica el nuevo protocolo, siendo $X_i = 0$ en caso contrario.

La distribución marginal de los recuentos observados es:

Número de incidencias y	0	1	2	3	4	5
Número de centros	55	20	12	7	4	2

Se han ajustado cuatro modelos, obteniéndose las siguientes salidas:

Modelo	$\widehat{\beta}_1$	$\text{se}(\widehat{\beta}_1)$	p -valor
Poisson	−0,405	0,150	0,007
Cuasi-Poisson	−0,405	0,201	0,044
Binomial negativa	−0,350	0,190	0,065
Poisson cero-inflado	−0,365	0,200	0,068

En el modelo Poisson cero-inflado, β_1 pertenece al componente de recuento.

Para el modelo cuasi-Poisson se ha obtenido la estimación

$$\widehat{\phi} = 1,80.$$

Los cocientes de dispersión de Pearson, calculados utilizando la función de varianza correspondiente a cada modelo, son:

Modelo	χ^2_P/gl
Poisson	1,80
Binomial negativa	1,06
Poisson cero-inflado	1,02

Para la binomial negativa se utiliza la parametrización

$$\text{Var}(Y_i \mid X_i, E_i) = \mu_i + \frac{\mu_i^2}{k}, \qquad \widehat{k} = 1,13.$$

La comparación del ajuste proporciona:

Modelo	Log-verosimilitud	AIC	Número esperado de ceros
Poisson	−137,6	279,2	40,3
Binomial negativa	−119,4	244,8	52,0
Poisson cero-inflado	−121,5	249,0	54,2

Los números esperados de ceros se han obtenido sumando las probabilidades ajustadas para los 100 centros, teniendo en cuenta sus exposiciones y covariables.

El componente de inflación del modelo Poisson cero-inflado contiene únicamente una constante. Por tanto, la probabilidad estimada de pertenecer al estado de cero estructural es común a todos los centros y vale

$$\hat{\pi} = 0,25.$$

Puede utilizar, si lo considera necesario,

$$e^{-0,405} \simeq 0,667, \quad e^{-0,350} \simeq 0,705, \quad e^{-0,365} \simeq 0,694,$$

y

$$\text{AIC} = -2\ell + 2q,$$

donde ℓ es la log-verosimilitud maximizada y q es el número de parámetros estimados. Responda razonadamente:

- a) Explique la función de $\log(E_i)$ en el modelo. Calcule e interprete la razón de tasas asociada al protocolo en el modelo de Poisson.
- b) Compare los resultados de Poisson y cuasi-Poisson. Explique por qué la estimación de β_1 no cambia, mientras que sí cambian su error estándar y el resultado del contraste. ¿Qué conclusión sobre el protocolo se obtendría con cada modelo al nivel del 5 %?
- c) Interprete el parámetro k de la binomial negativa y compare la forma en que cuasi-Poisson y binomial negativa modelizan la sobredispersión. Indique también por qué no aparece un AIC para el modelo cuasi-Poisson.
- d) Un investigador sostiene que los 55 ceros observados demuestran que debe utilizarse un modelo Poisson cero-inflado. Valore esta afirmación empleando toda la información suministrada y seleccione razonadamente entre Poisson, binomial negativa y Poisson cero-inflado.
- e) Evalúe la adecuación de las siguientes deducciones justificando la respuesta:
 - 1) “Como el coeficiente es significativo en el modelo Poisson, entonces el protocolo reduce la tasa de incidencias”.
 - 2) “El parámetro $\pi = 0,25$ significa que el 25 % de todos los ceros observados son ceros estructurales”.
 - 3) “Si se selecciona la binomial negativa, el efecto estimado del protocolo es una reducción aproximada del 29,5 % en la tasa de incidencias”.

Cuestión 5. Se considera una variable espacial $Z(s)$, donde s representa la localización, observada en nueve áreas geográficas contiguas $A_1, \dots, A_9$ dispuestas según la cuadrícula de la figura, siendo unitario el lado de cada área. Se denota por z_i el valor observado en el área A_i y se consideran vecinas dos áreas que comparten una frontera común.

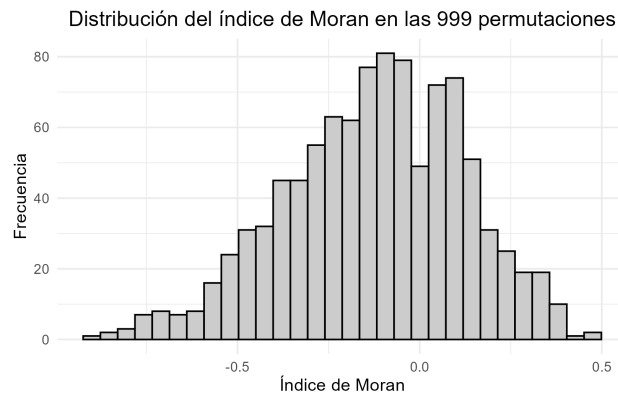
A_7 $z_7 = 3$	A_8 $z_8 = 5$	A_9 $z_9 = 10$
A_4 $z_4 = 4$	A_5 $z_5 = 6$	A_6 $z_6 = 8$
A_1 $z_1 = 2$	A_2 $z_2 = 7$	A_3 $z_3 = 9$

- (a) Considérese el estadístico I (índice de Moran), que permite estudiar la autocorrelación espacial de n áreas:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2},$$

donde $w_{ij} = \begin{cases} 1, & \text{si } A_i \text{ y } A_j \text{ son vecinas,} \\ 0, & \text{en otro caso,} \end{cases} \quad w_{ii} = 0, \quad S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}.$

- (I) Razone cómo intervienen la vecindad y las desviaciones respecto de la media en el signo de I . Calcule e interprete el estadístico para los datos de la figura. Puede utilizar $\bar{z} = 6$ y $\sum_{i=1}^n (z_i - \bar{z})^2 = 60$.
- (II) Suponga ahora que I se calcula con los pesos normalizados por filas $w_{ij}^* = w_{ij} / \sum_{j=1}^n w_{ij}$. Interprete los nuevos pesos y analice las consecuencias de esta normalización.
- (b) Para evaluar el comportamiento del estadístico I , se realizan 999 permutaciones aleatorias de los valores z_i entre las áreas, sin modificar las relaciones de vecindad, obteniéndose la distribución que se muestra en la figura.



Explique la finalidad de este procedimiento y utilice la distribución obtenida para evaluar el resultado del apartado anterior. Indique, además, cómo obtendría el p -valor a partir de las permutaciones realizadas.

- (c) Considérese como localización de cada área su centro. Supóngase que $E[Z(s)] = \mu$ y que su función de covarianza es

$$C(h) = 9 \exp \left\{ -\sqrt{h_1^2 + 4h_2^2} \right\}, \quad h = (h_1, h_2).$$

- (I) Obtenga la función de autocorrelación espacial. Analice la estacionariedad de segundo orden y la isotropía del modelo, justificando sus conclusiones con áreas de la figura.
- (II) Determine el lugar geométrico de los desplazamientos h para los que la autocorrelación es igual a e^{-2} e interprete el resultado en términos de la estructura de dependencia espacial.

Cuestión 6. De una población finita $U = \{1, \dots, N\}$ se selecciona, mediante muestreo aleatorio simple sin reemplazamiento, una muestra s de tamaño n , con objeto de estimar el total

$$Y = \sum_{k \in U} y_k.$$

Para cada unidad seleccionada se observa, además de y_k , una característica que permite clasificarla de forma unívoca en uno de H grupos. Esta característica no ha intervenido en la selección de la muestra. A partir de una fuente externa se conocen los totales poblacionales correspondientes a dichos grupos,

$$N_1, \dots, N_H, \quad \sum_{h=1}^H N_h = N.$$

Se denota por s_h el conjunto de unidades de la muestra clasificadas en el grupo h y por $n_h = |s_h|$ su tamaño, suponiendo $n_h > 0$, $\forall h = 1, \dots, H$.

- (a) Utilizando los efectivos poblacionales N_h , proponga un estimador del total Y basado en la estimación separada dentro de los H grupos. Expréselo como $\hat{Y} = \sum_{k \in s} w_k y_k$ y determine los pesos w_k . Indique qué propiedad satisfacen estos pesos respecto a los N_h y comente su relación con el estimador del total bajo muestreo estratificado.
- (b) Para cada unidad k de la muestra, considere el vector de variables indicadoras

$$\mathbf{x}_k = (I(k \text{ pertenece al grupo } 1), \dots, I(k \text{ pertenece al grupo } H))', \quad \mathbf{X} = (N_1, \dots, N_H)'$$

Denótense por d_k los pesos de diseño. Entre todos los conjuntos de pesos que satisfacen

$$\sum_{k \in s} w_k \mathbf{x}_k = \mathbf{X},$$

se consideran aquellos que minimizan

$$\Phi(w, d) = \frac{1}{2} \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k}.$$

Obtenga justificadamente la expresión de los pesos resultantes y compruebe que, para las variables indicadoras anteriores, se recuperan los pesos obtenidos en el apartado (a). Identifique el procedimiento de ajuste de pesos que se está aplicando e interprete las restricciones impuestas.

- (c) Considere ahora un vector general de variables auxiliares $\mathbf{x}_k$, cuyo total poblacional $\mathbf{X}$ es conocido. Utilizando los pesos obtenidos en el apartado anterior, demuestre que el correspondiente estimador puede expresarse como

$$\hat{Y} = \hat{Y}_{HT} + (\mathbf{X} - \hat{\mathbf{X}}_{HT})' \hat{\mathbf{B}},$$

donde $\hat{Y}_{HT}$ y $\hat{\mathbf{X}}_{HT}$ son los estimadores de Horvitz–Thompson de Y y $\mathbf{X}$, respectivamente. Obtenga la expresión de $\hat{\mathbf{B}}$ e identifique el estimador resultante.

Cuestión 7. Un instituto de estadística recibe un fichero de n observaciones numéricas que se leen secuencialmente, una a una, y que no pueden almacenarse en memoria por su tamaño. Se desea calcular la media y la varianza muestral del conjunto leyendo el fichero una sola vez.

Se emplean las definiciones siguientes: la media es $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ y la varianza muestral es $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, extendiéndose ambos sumatorios a $i = 1, \dots, n$, con $n \geq 2$. Para cada $k = 1, \dots, n$ se denotan por M_k la media de las k primeras observaciones y por S_k la suma de cuadrados de las desviaciones respecto de M_k , es decir, $S_k = \sum_{i=1}^k (x_i - M_k)^2$, de modo que $s^2 = \frac{S_n}{n-1}$.

- (a) Escriba, en Python, R, otro lenguaje similar o pseudocódigo estructurado, un algoritmo que recorra el fichero una única vez y devuelva $\bar{x}$ y s^2 , actualizando en cada nueva observación x_k las cantidades M_k y S_k mediante las siguientes recurrencias, que deberá utilizar tal como se indican:

$$\begin{aligned} k = 1 : \quad & M_1 = x_1, \quad S_1 = 0, \\ k \geq 2 : \quad & M_k = M_{k-1} + \frac{x_k - M_{k-1}}{k}, \\ & S_k = S_{k-1} + (x_k - M_{k-1})(x_k - M_k). \end{aligned}$$

Indique la complejidad temporal del algoritmo en notación Θ en función de n , así como el número exacto de operaciones aritméticas —sumas y restas consideradas conjuntamente, multiplicaciones y divisiones— que se ejecutan por cada observación leída a partir de la segunda. Para este recuento, supóngase que la diferencia entre la observación en curso y la media anterior se calcula una sola vez y se reutiliza en ambas recurrencias, sin reescribir algebraicamente las expresiones. Además, no se cuentan las comparaciones, el incremento del contador k , las operaciones de lectura ni las operaciones finales necesarias para obtener $s^2 = \frac{S_n}{n-1}$.

Indique, asimismo, la complejidad espacial auxiliar, entendida como la memoria adicional a la ocupada por la observación en curso. Compare ambas complejidades con las del procedimiento clásico de dos pasadas, que calcula primero $\bar{x}$ y recorre después los datos para acumular $\sum_{i=1}^n (x_i - \bar{x})^2$. (a efectos exclusivos de esta comparación, se admite que el fichero pueda releerse íntegramente desde disco una segunda vez). Precise si ambos procedimientos coinciden o difieren en el orden de sus complejidades temporal y espacial auxiliar, y cuántos recorridos completos del fichero requiere cada uno.

- (b) El fichero se reparte en dos bloques disjuntos y no vacíos A y B , con $n_A \geq 1$ y $n_B \geq 1$, de los que se han obtenido por separado los tríos

$$(n_A, M_A, S_A) \quad \text{y} \quad (n_B, M_B, S_B)$$

mediante el algoritmo del apartado a. Escriba las fórmulas que permiten obtener el trío (n, M, S) del conjunto de observaciones formado por $A \cup B$ a partir únicamente de esos seis valores, sin volver a leer los datos. La fórmula de S deberá expresarse utilizando explícitamente la diferencia de medias $M_B - M_A$; no se admitirá como respuesta la reconstrucción de las sumas de cuadrados brutas

$$\sum_i x_i^2.$$

- (c) El fichero se procesa ahora en P bloques de igual tamaño n/P , siendo P un divisor de n , cada uno en un procesador distinto que trabaja simultáneamente con los demás, combinándose después secuencialmente los P tríos en un único procesador mediante $P - 1$ aplicaciones de las fórmulas del apartado b.

Suponiendo que los P bloques pueden procesarse simultáneamente y sin contención de acceso a los datos, y que el coste de comunicación entre procesadores es nulo, indique, en notación Θ , la complejidad temporal del tiempo de ejecución del procedimiento completo — no del trabajo total agregado de los P procesadores— en función de n y de P , distinguiendo la fase de recorrido de la fase de combinación.

Permitiendo que P sea función de n , exprese la condición asintótica sobre P que debe cumplirse para que el coste de la fase de combinación sea despreciable frente al de la fase de recorrido.

Cuestión 8. El siguiente documento XML representa la tasa de paro, expresada en porcentaje, por comunidad autónoma y año:

```
<tasaParo>
  <dato id="D001">
    <comunidad>Comunidad de Madrid</comunidad>
    <anyo>2024</anyo>
    <porcentaje>9.2</porcentaje>
  </dato>
  <dato id="D002">
    <comunidad>Andalucia</comunidad>
    <anyo>2024</anyo>
    <porcentaje>16.3</porcentaje>
  </dato>
</tasaParo>
```

- (a) Elabore un esquema XSD 1.0 que permita validar el documento anterior y cumpla simultáneamente las siguientes condiciones:
- El elemento raíz `<tasaParo>` debe contener uno o más elementos `<dato>`.
 - Cada elemento `<dato>` debe tener un atributo obligatorio `id`, de tipo `xs:ID`, y debe contener, en el siguiente orden y exactamente una vez cada uno, los elementos `<comunidad>`, `<anyo>` y `<porcentaje>`.
 - `<comunidad>` debe ser de tipo `xs:string`.
 - `<anyo>` debe ser de un tipo simple derivado por restricción de `xs:integer` y estar comprendido entre 2000 y 2100, ambos inclusive.
 - `<porcentaje>` debe ser de un tipo simple derivado por restricción de `xs:decimal` y estar comprendido entre 0 y 100, ambos inclusive.
 - El esquema no debe incorporar comodines (`xs:any` ni `xs:anyAttribute`), por lo que no se admitirán elementos ni atributos distintos de los declarados expresamente. A estos efectos, no se tendrá en cuenta la posible presencia de los atributos previstos por la especificación en el espacio de nombres `xsi`.
 - El esquema no debe declarar un espacio de nombres de destino (atributo `targetNamespace`). El documento XML anterior no está cualificado y debe validar tal como se proporciona.
- (b) Escriba una hoja de estilo XSLT 1.0 que genere un documento HTML cuyo elemento `<body>` contenga, como únicos elementos hijos y en este orden:
- Un encabezado `<h1>` con el texto exacto *Tasa de paro por comunidad autónoma*.
 - Una lista no ordenada `<ul>` con un elemento `<li>` por cada elemento `<dato>` hijo de `<tasaParo>`, respetando el orden de aparición de dichos elementos en el documento XML.

El contenido textual de cada elemento `<li>` debe ser exactamente: *nombre de la comunidad: porcentaje %*, con un único espacio tras los dos puntos y sin espacio entre la cifra y el signo de porcentaje. El valor del porcentaje se reproducirá literalmente tal como figura en el documento XML, sin reformatearlo ni redondearlo (por ejemplo, 9.2, y no 9,2 ni 9.20).

No se exige la declaración DOCTYPE ni la inclusión de los elementos `<head>` y `<title>`; su presencia, no obstante, será admisible y no se penalizará. Tampoco es obligatorio declarar `<xsl:output method="html"/>`.

Para el documento proporcionado, la lista deberá contener exactamente los elementos:

```
<li>Comunidad de Madrid: 9.2%</li>
```

```
<li>Andalucia: 16.3%</li>
```