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 2024

Resolución de 22 de diciembre de 2024, 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 31 de Diciembre de 2024)

Cuestión 1. En una población finita $U = \{1, \ldots, N\}$, las unidades están agrupadas en M conglomerados de tamaños N_i para $i = 1, \ldots, M$, cumpliendo $N = \sum_{i=1}^{M} N_i$. Con el objetivo de estimar la media poblacional $\bar{Y}$ de cierta característica de la población, se propone un plan de muestreo bietápico y se consideran tres escenarios, cada uno con sus cuestiones específicas.

1) En la primera etapa se seleccionan m unidades primarias con reemplazamiento y con probabilidades proporcionales a su tamaño (número de unidades secundarias). En la segunda etapa se seleccionan n_i unidades secundarias de la unidad primaria i sin reemplazamiento y con probabilidades iguales. Se supone que los valores N_i son conocidos $\forall i = 1, \ldots M$.
   a) Justificar que la probabilidad de inclusión de una unidad primaria i bajo estas condiciones es aproximadamente $\pi_i \approx mp_i$, donde p_i es la probabilidad de que, en cada extracción, el conglomerado i sea seleccionado. Suponer que p_i es pequeño.
   b) Calcular la probabilidad de inclusión π_{ij} de una unidad secundaria j, $j = 1, \ldots, N_i$, perteneciente a la unidad primaria i, determinar la condición para que todas las π_{ij} sean iguales y comentar si en tal caso el diseño sería autoponderado.

2) Ahora suponemos que los conglomerados son de igual tamaño y que en la primera etapa se seleccionan m unidades primarias y, dentro de cada una, n unidades secundarias, siempre sin reemplazamiento y con probabilidades iguales.

   Incluir una unidad primaria en la muestra tiene un coste $c_1 > 0$, mientras que entrevistar a una unidad secundaria supone un coste $c_2 > 0$. El presupuesto total disponible es C_0 y, a partir de un estudio previo, se conoce el coeficiente de correlación intraconglomerado ρ de la variable de interés.
   a) Plantee el problema de elección de m y n que minimice la varianza del estimador insesgado de la media poblacional, bajo la restricción de coste.
   b) Resuelva el problema anterior y obtenga los valores óptimos de m y n para las unidades de primera y segunda etapa.

3) En la primera etapa se seleccionan m conglomerados sin reemplazamiento y con probabilidades iguales. En cada conglomerado i se eligen n_i unidades secundarias también sin reemplazamiento y con probabilidades iguales. Ahora bien, no se conoce el número total de unidades secundarias $N = \sum_{i=1}^{M} N_i$.
   a) Proponga un estimador de la media poblacional $\bar{Y}$ que no requiera conocer N, y justifique su construcción.
   b) Estudie el sesgo del estimador propuesto.

Cuestión 2. Con el fin de obtener estimaciones de determinados parámetros en una población finita de tamaño N, se recurre a un esquema de muestreo en ocasiones sucesivas. Se pide contestar a las cuestiones que se formulan a continuación.

1) Para estimar el ingreso mensual promedio de los hogares de un cierto territorio, se propone realizar una encuesta en dos periodos consecutivos que denotaremos por t y $t+1$.

   En t se selecciona mediante muestreo aleatorio simple sin reemplazamiento una muestra s_1 de tamaño n. En el periodo $t+1$, volvemos a entrevistar a m unidades que formaban parte de la muestra s_1 y se seleccionan u nuevas mediante muestreo aleatorio simple de manera que el tamaño total permanezca igual a n.

   Sea z_i el valor de la variable en la unidad i en t e y_k el valor de la variable en la uniadd k en $t+1$.

   a) Para estimar el cambio medio poblacional , se propone como estimador la diferencia de medias muestrales en las dos ocasiones, $\hat{D} = \bar{y} - \bar{x}$.
      - Obtenga de manera razonada la expresión de la varianza del estimador en términos de la proporción de emparejamiento, μ, y del coeficiente de correlación lineal entre los valores de la variable en una y otra ocasión, ρ.
      - Interprete el efecto de ρ y de μ sobre la precisión de $\hat{D}$.

   b) Discutir qué cambiaría en la estrategia solapamiento con el fin de minimizar la varianza, si el objetivo fuese estimar un ingreso promedio durante el período combinado de t y $t+1$ mediante el estimador $\hat{M} = \frac{\bar{y}+\bar{z}}{2}$

   Nota. Se puede suponer que N es muy grande en relación con n.

2) Se le encarga asesorar sobre aspectos relacionados con el diseño muestral y la recogida de datos de una encuesta destinada a estudiar la evolución de una característica a lo largo del tiempo.

   a) Usted plantea, entre otros elementos relacionados con el diseño muestral, utilizar un panel rotante. Describa de manera clara en qué consiste y exponga, como si se lo explicara al responsable del proyecto, los aspectos metodológicos que deben considerarse, así como sus principales ventajas e inconvenientes para valorar su implementación.

   b) Si uno de los métodos de recogida por los que se opta es un cuestionario autocumplimentado a través de internet, indique qué pruebas de usabilidad recomendaría llevar a cabo antes de su implantación.

Cuestión 3. Un grupo de investigadores está desarrollando un sistema para predecir si un paciente recibirá un diagnóstico positivo de una enfermedad (variable binaria `Diagnóstico`) en función de factores clínicos y hábitos de vida. Se recogen los siguientes datos de 200 pacientes:

- `Diagnóstico` (1 = positivo, 0 = negativo)
- `Edad` (años)
- `Fuma` (1 = sí, 0 = no)
- `Ejercicio` (1 = realiza ejercicio regular, 0 = no)
- `Categoría_dieta` (1 = buena, 2 = regular, 3 = mala)
- `Visitas_médico` (número de visitas al médico en el último año)
- `Tipo_enfermedad` (1 = cardiovascular, 2 = respiratoria, 3 = metabólica, 4 = sin diagnóstico)

Se propone el siguiente modelo de regresión logística para `Diagnóstico`:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \cdot \text{Edad}_i + \beta_2 \cdot \text{Fuma}_i + \beta_3 \cdot \text{Ejercicio}_i \tag{1}$$

donde $\pi_i = P(\text{Diagnóstico}_i = 1)$. Se obtienen las siguientes estimaciones:

$$\hat{\beta}_1 = 0{,}03, \quad \hat{\beta}_2 = 0{,}85, \quad \hat{\beta}_3 = -0{,}60 \tag{2}$$

Preguntas

Para responder a estas cuestiones puede ser de utilidad conocer los siguiente valores de la función exponencial. Úselos si lo considera conveniente.

x_1	e^{x_1}	x_2	e^{x_2}	x_3	e^{x_3}	x_4	e^{x_4}
-0.9	0.41	-0.4	0.67	0.1	1.11	0.6	1.82
-0.85	0.43	-0.35	0.7	0.15	1.16	0.65	1.92
-0.8	0.45	-0.3	0.74	0.2	1.22	0.7	2.01
-0.75	0.47	-0.25	0.78	0.25	1.28	0.75	2.12
-0.7	0.5	-0.2	0.82	0.3	1.35	0.8	2.23
-0.65	0.52	-0.15	0.86	0.35	1.42	0.85	2.34
-0.6	0.55	-0.1	0.9	0.4	1.49	0.9	2.46
-0.55	0.58	-0.05	0.95	0.45	1.57		
-0.5	0.61	0	1	0.5	1.65		
-0.45	0.64	0.05	1.05	0.55	1.73		

1) a) Interprete el efecto de las variables `Fuma` y `Ejercicio` sobre la probabilidad de diagnóstico positivo.
   b) ¿Cuál es el odds ratio asociado a `Ejercicio`?
   c) Inteprete el odds ratio anterior para el grupo que practica ejercicio.
   d) Si un paciente de 50 años, fumador y sin ejercicio tiene $\hat{\pi} = 0{,}45$, calcule el valor de $\hat{\pi}$ para un paciente idéntico que sí hace ejercicio.

2) Se ajusta un modelo de categoría base con `Tipo_enfermedad` como variable respuesta y `Edad` y `Categoría_dieta` como explicativas. Se toma como categoría base "sin diagnóstico" (categoría 4):

$$\log\left(\frac{P(Y_i = j)}{P(Y_i = 4)}\right) = \alpha_{j0} + \alpha_{j1} \cdot \text{Edad}_i + \alpha_{j2} \cdot \text{Dieta}_i, \quad j = 1, 2, 3 \tag{3}$$

   a) ¿Qué representa α_{21} en este modelo?

   b) ¿Qué indica que $\alpha_{32} > 0$ para la categoría "metabólica"?

3) Se desea modelizar el número de visitas al médico (`Visitas_médico`) usando un modelo de Poisson:

$$\log(\mu_i) = \gamma_0 + \gamma_1 \cdot \text{Edad}_i + \gamma_2 \cdot \text{Fuma}_i \tag{4}$$

   a) ¿Cuál es la interpretación de γ_1?

   b) Si $\hat{\mu} = 4{,}5$, ¿cuántas visitas se esperaría para un paciente 10 años mayor, manteniendo constantes las demás variables? Considere que $\hat{\gamma_1} = 0{,}02$.

4) Al ajustar el modelo anterior, se encuentra que el estadístico de devianza sobre los grados de libertad da un valor de 2,4.

   a) ¿Qué indica este valor respecto al ajuste del modelo?

   b) Proponga dos soluciones metodológicas para abordar el problema detectado.

Cuestión 4. Un modelo de red neuronal profunda incluye dos capas ocultas: la primera capa oculta contiene 3 unidades y la segunda capa oculta contiene 2 unidades. La red genera una única salida continua (variable cuantitativa). Las variables de entrada están compuestas por 2 variables cuantitativas, 1 variable cualitativa binaria y 1 variable cualitativa con 3 categorías.

1) Describa con precisión la arquitectura de la red neuronal, especificando el número total de entradas, la disposición de las capas, así como las funciones de activación utilizadas en cada capa.
2) Escriba una expresión explícita para la función $f(X)$ estimada por el modelo, suponiendo funciones de activación tipo ReLU (Rectified Linear Unit) en las capas ocultas. La expresión debe ser lo más detallada posible.
3) Calcule el número total de parámetros (pesos y sesgos) que deben ajustarse durante el entrenamiento del modelo.
4) Explique de forma rigurosa los pasos del algoritmo de retropropagación (backpropagation) que se utilizan para actualizar los parámetros del modelo, incluyendo las fórmulas necesarias para el cálculo de los gradientes.
5) ¿Qué modificaciones serían necesarias en la arquitectura y en la función de pérdida, si la variable objetivo fuera una variable categórica con 5 clases mutuamente excluyentes?.

Cuestión 5. Sea $\mathbf{X}$ un vector aleatorio, $\mathbf{X} = (X_1, X_2, X_3)'$ de media cero, cuya matriz de varianzas-covarianzas poblacionales es la siguiente:

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

1) Obtenga los autovalores y autovectores de Σ.
2) Determine el vector $\mathbf{Y} = (Y_1, Y_2, Y_3)'$ de componentes principales y la proporción de la varianza total explicada por cada componente principal.
3) Calcúlese la correlación entre la primera componente principal Y_1 y la variable original X_2.
4) ¿Varían las componentes principales si se cambia la unidad de medida de las variables respecto a las componentes obtenidas con las unidades originales? Razone su respuesta.

Cuestión 6. Una empresa de ingeniería está interesada en modelizar la vida útil (en cientos de horas) de un componente electrónico que opera en entornos industriales de alta exigencia. Según estudios empíricos previos, la vida útil del componente puede aproximarse mediante una distribución Weibull con los siguientes parámetros:

- Parámetro de forma: $k = 2$
- Parámetro de escala: $\lambda = 500$

La función de distribución acumulada (CDF) de una variable aleatoria$X \sim \text{Weibull}(k, \lambda)$ viene dada por:

$$F(x) = 1 - \exp\Big(-\big(\tfrac{x}{\lambda}\big)^k\Big), \quad x \geq 0$$

1) Diseñar un algoritmo de simulación, utilizando el método de inversión, para estimar un intervalo de confianza al 95 % de la mediana del tiempo de vida útil. El algoritmo de simulación debe expresarse en pseudocódigo o, preferentemente, implementarse en un lenguaje de programación como Python ó R. El algoritmo debe ser computacionalmente eficiente teniendo en cuenta que, en el entorno de simulación disponible, únicamente se pueden generar números aleatorios con distribución Uniforme(0,1) y no existen funciones directas ni para realizar ordenaciones de números, ni para calcular directamente la mediana de un conjunto de valores. Justificar detalladamente cuál sería la complejidad computacional del algoritmo propuesto (denomine n al tamaño de cada muestra generada y M al número de simulaciones).
2) Describir cómo incorporar alguna técnica de reducción de varianza en el proceso de simulación para mejorar la precisión del estimador del tiempo mediano de vida útil.

Cuestión 7. Una universidad está desarrollando una base de datos para gestionar los proyectos de investigación (pid), sus investigadores (rid) y los fondos asociados (fid). Se proporcionan las siguientes relaciones:

- **Researcher** (*rid: integer*, *rname: varchar*, *department: varchar*, *email: varchar*)
- **Project** (*pid: integer*, *pname: varchar*, *budget: decimal*, *start_date: date*, *end_date: date*)
- **WorksOn** (*rid: integer*, *pid: integer*, *role: varchar*, *hours_per_week: integer*)
- **Funding** (*fid: integer*, *pid: integer*, *sponsor: varchar*, *amount: decimal*)

Información adicional:

- Cada investigador puede trabajar en varios proyectos, y cada proyecto puede tener varios investigadores.
- Un investigador puede tener distintos roles en distintos proyectos, pero únicamente un rol por proyecto.
- Cada proyecto puede tener varias fuentes de financiación y un mismo sponsor puede financiar varios proyectos.

Se pide:

1) Construir detalladamente el diagrama Entidad-Relación que represente el diseño conceptual de la base de datos, utilizando preferentemente la notación de Chen. El diagrama debe cumplir obligatoriamente con los siguientes requisitos:
   a) Representar todas las entidades y todas las relaciones que se derivan de la descripción anterior.
   b) Incluir todos los atributos correspondientes a cada entidad y a cada relación (cuando proceda).
   c) Utilizar rectángulos para entidades, rombos para las relaciones y óvalos para los atributos (según la notación de Chen).
   d) Incluir los atributos propios de las relaciones, si existen.
   e) Señalar de forma clara y diferenciada las claves primarias (PK) y las claves foráneas (FK) de cada entidad o relación, utilizando subrayado o notación explícita.
   f) No se deben utilizar atributos multivaluados ni compuestos.
   g) Representar explícitamente las restricciones de integridad referencial, incluyendo:
      - Participación total o parcial de cada entidad en sus respectivas relaciones.
      - Cardinalidades mínimas y máximas.
2) Escriba una consulta SQL que, para cada investigador, devuelva su nombre junto con el nombre de aquel o aquellos proyectos en los que trabaja el mayor número de horas semanales. En caso de empate, deben incluirse todos los proyectos que alcanzan el máximo de horas. Elija la solución que proporcione el código más eficiente que sea compatible con cualquier variante de SQL.

3) Calcular, para cada investigador, la media de financiación por proyecto en los que ha trabajado, y listar, ordenados por sponsor, aquellos proyectos cuya media supera la media general del total. Elija la solución que proporcione el código más eficiente que sea compatible con cualquier variante de SQL.

Cuestión 8. Se proporciona el siguiente fragmento de un archivo XML:

```
<produccion>
  <investigador id="I001">
    <nombre>María García</nombre>
    <publicaciones>
      <publicacion>
        <titulo>Inteligencia Artificial en Medicina</titulo>
        <anyo>2021</anyo>
        <tipo>Articulo</tipo>
        <revista>Journal of Medical AI</revista>
      </publicacion>
    </publicaciones>
  </investigador>
</produccion>
```

Y un esquema incompleto:

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
 <xs:element name="produccion">
  <!-- COMPLETAR -->
 </xs:element>
</xs:schema>
```

1) Elabore el esquema XML (XSD) que permita validar correctamente el documento anterior, asegurando que se representen de forma adecuada la estructura jerárquica y la secuencia de los elementos, así como el tipo de dato correspondiente al elemento <anyo>y al atributo id.
2) Escriba una expresión XPath que recupere los títulos (<titulo>) de todas las publicaciones cuyo tipo sea "Articulo" y cuyo año sea posterior a 2020.
3) Modifique el XML inicial para que incluya una segunda publicación del mismo investigador, pero de tipo "Libro", con un nuevo campo <editorial >. Debe mantenerse la estructura válida respecto al esquema definido en el apartado 1.