

Departamento de Estadística y Matemáticas
Facultad de Ciencias Económicas
Estadística II
Parcial III

Nombre: _____ Cédula:_____

Antes de iniciar con el desarrollo del trabajo **LEA ATENTAMENTE:**

- La calificación de cada punto desarrollado estará conformado por tres partes:
 1. El planteamiento del ejercicio, que abarca el correcto planteamiento del juego de hipótesis, y/o intervalo de confianza, la correcta evaluación de las condiciones que conducen al empleo de una u otra ecuación.
 2. El desarrollo del ejercicio planteado junto a sus códigos.
 3. Las interpretaciones resultantes en el contexto de los datos y del enunciado planteado.
- Si se presenta una situación en donde al momento de la calificación, se encuentra que los códigos generados son realizados con IA Generativa, se **calificará con 0 en ese punto.** >:c
- Enviar al correo **jivan.perez@udea.edu.co**, los archivos correspondientes a los códigos generados, una vez finalice con su desarrollo. :D

1. Contexto del Problema

El Costo Económico de las Enfermedades Cardiovasculares: Un Análisis de Salud Pública

Las enfermedades cardiovasculares (ECV) representan uno de los mayores desafíos para los sistemas de salud a nivel mundial, siendo la principal causa de muerte y generando costos significativos tanto para los sistemas de salud como para la economía en general. Según la Organización Mundial de la Salud (OMS), las ECV son responsables de aproximadamente 17.9 millones de muertes anuales, lo que representa el 31 % de todas las muertes globales.

Desde una perspectiva económica y de gestión empresarial, comprender los factores de riesgo asociados con las ECV es crucial para:

- **Compañías de seguros:** Evaluar riesgos y establecer primas adecuadas para seguros de salud y vida.
- **Empresas empleadoras:** Diseñar programas de bienestar corporativo que reduzcan el ausentismo laboral y mejoren la productividad.
- **Sistemas de salud públicos y privados:** Optimizar la asignación de recursos y desarrollar estrategias preventivas costo-efectivas.
- **Planificadores de políticas públicas:** Diseñar intervenciones en salud pública que maximicen el impacto con recursos limitados.

1.1. Descripción de la Base de Datos

Ha sido contratado(a) como analista de datos por una importante compañía de seguros de salud que busca desarrollar un modelo de evaluación de riesgo cardiovascular para sus asegurados. La empresa ha recopilado información en la base de datos entregada sobre pacientes que incluye datos demográficos, clínicos y de estilo de vida.

La base de datos contiene las siguientes variables:

Variables Demográficas:

- **id**: Identificador único del paciente
- **edad_años**: Edad del paciente en años
- **genero**: Género del paciente (“Mujer” / “Hombre”)

Variables Clínicas:

- **altura_cm**: Altura en centímetros
- **peso_kg**: Peso en kilogramos
- **IMC**: Índice de Masa Corporal (calculado automáticamente)
- **presion_sistolica**: Presión arterial sistólica en mmHg
- **presion_diastolica**: Presión arterial diastólica en mmHg
- **colesterol**: Nivel de colesterol (“Normal” / “Por encima de lo normal” / “Muy por encima de lo normal”)
- **glucosa**: Nivel de glucosa en sangre (“Normal” / “Por encima de lo normal” / “Muy por encima de lo normal”)
- **enfermedad_cardiovascular**: Indica si el paciente ha sido diagnosticado con enfermedad cardiovascular (“No” / “Sí”)

Variables de Estilo de Vida:

- **fumador**: Indica si el paciente es fumador (“No” / “Sí”)
- **consume_alcohol**: Indica si consume alcohol regularmente (“No” / “Sí”)
- **actividad_fisica**: Indica si realiza actividad física regular (“No” / “Sí”)

2. Preguntas del Examen

Basándose en la base de datos asignada, responda las siguientes preguntas. **Cada pregunta incluye filtros específicos que debe aplicar antes de realizar el análisis correspondiente.**

- a) **(1 punto)** La compañía de seguros desea comparar la variable **presion_diastolica** entre dos grupos de asegurados definidos por **genero**.

Filtros obligatorios:

- Incluir únicamente pacientes donde **enfermedad_cardiovascular** = "Sí"
- Incluir únicamente pacientes donde **edad_años** > 47

Ejercicio: Construya un intervalo de confianza del 99 % para la diferencia de medias de **presion_diastolica** entre los grupos definidos por **genero**. Interprete los resultados en el contexto del problema.

- b) **(1 punto)** La literatura médica sugiere que el valor promedio de **edad_años** en la población general es de 48.

Filtros obligatorios:

- Incluir únicamente pacientes donde **actividad_fisica** = "No"
- Incluir únicamente pacientes donde **glucosa** = "Normal"

Ejercicio: Plantee y ejecute una prueba de hipótesis bilateral para verificar si la media de **edad_años** en su muestra filtrada es igual a 48, empleando un nivel de significancia del 1 %. Incluya todos los pasos formales de la prueba.

- c) **(1 punto)** Basándose en la prueba de hipótesis del punto anterior (punto b), suponga que el verdadero valor poblacional de **edad_años** es $\mu_1 = 50$.

Ejercicio: Calcule el Error Tipo II (β) para esta hipótesis alternativa específica. Qué significa este valor en el contexto del problema.

d) **(1 punto)** La compañía desea saber si existe relación entre `actividad_fisica` y `enfermedad_cardiovascular`.

Filtros obligatorios:

- Incluir únicamente pacientes donde `colesterol` = "Por encima de lo normal"
- Incluir únicamente pacientes donde `edad_años` >50

Tarea: Plantee y ejecute una prueba de independencia empleando un nivel de significancia del 8 %. Calcule la tabla de frecuencias observadas, frecuencias esperadas, y presente el correspondiente juego de hipótesis, estadístico de prueba, P-Valor y conclusión en el contexto del problema.

e) **(1 punto)** Se desea evaluar el comportamiento distribucional de dos variables clave en el portafolio de asegurados.

Filtros obligatorios (aplicar a ambas pruebas):

- Incluir únicamente pacientes donde `consume_alcohol` = "Sí"
- Incluir únicamente pacientes donde `edad_años` >40

Ejercicio:

- **Prueba 1:** Evalúe si la variable `edad_años` sigue una distribución gamma.
- **Prueba 2:** Evalúe si la variable `presion_sistolica` sigue una distribución log-normal.

Para cada prueba, emplee un nivel de significancia del 4 % y utilice la prueba de Cramer-Von Mises o Anderson-Darling. Presente el juego de hipótesis, P-Valor y conclusión.