

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

Optimizando la Experiencia del Cliente en el E-commerce: Un Enfoque Basado en Datos

En la era de la transformación digital, el comercio electrónico (e-commerce) se ha convertido en un pilar fundamental de la economía global. La competencia es intensa y las empresas que logran prosperar son aquellas que entienden profundamente a sus clientes y utilizan los datos para tomar decisiones estratégicas. Un análisis riguroso del comportamiento del cliente es esencial para:

- **Marketing y Ventas:** Personalizar ofertas, optimizar la inversión publicitaria (ROI) y segmentar a los clientes para campañas más efectivas.
- **Operaciones y Logística:** Prever la demanda, gestionar el inventario y optimizar la cadena de suministro para garantizar entregas rápidas y eficientes.
- **Finanzas:** Comprender los impulsores clave de los ingresos y la rentabilidad, y analizar el Valor de Vida del Cliente (Customer Lifetime Value).
- **Estrategia Corporativa:** Identificar nuevas oportunidades de mercado, mejorar la experiencia del usuario (UX) y construir una ventaja competitiva sostenible.

1.1. Descripción de la Base de Datos

Ha sido contratado(a) como Analista de Datos para “OmniRetail”, una de las principales empresas de venta minorista en línea. Su misión es analizar los datos de los clientes para proporcionar insights accionables que impulsen el crecimiento de la empresa. La base de datos ecommerce_data contiene información transaccional, de comportamiento y demográfica de una muestra de clientes.

La base de datos contiene las siguientes variables:

Variables Cuantitativas

- **Age:** (Numérica) Edad del cliente en años.
- **CustomerSince_days:** (Numérica) Antigüedad del cliente, en días.
- **TimeOnSite_min:** (Numérica) Tiempo total de la sesión en el sitio web, en minutos.
- **PagesViewed:** (Conteo) Número de páginas distintas visitadas durante la sesión.
- **LastLogin_days:** (Conteo) Días transcurridos desde la última visita del cliente.
- **NumPurchases_lifetime:** (Conteo) Número total de compras realizadas por el cliente a lo largo de su historia.
- **CustomerSupportInteractions:** (Conteo) Número de veces que el cliente ha contactado a soporte.
- **ReturnedItems_last_purchase:** (Conteo) Número de artículos devueltos de su compra anterior.
- **ItemsInCart:** (Conteo) Número de artículos en el carrito para la transacción actual.
- **AvgItemValue_USD:** (Numérica) Valor promedio por artículo en el carrito ($\text{TotalPurchase_USD} / \text{ItemsInCart}$).
- **TotalPurchase_USD:** (Numérica - Variable de Respuesta) Monto total de la compra en USD.

Variables Cualitativas

- **CustomerID:** (ID) Identificador numérico único del cliente.
- **Gender:** (Nominal) Género del cliente.
- **Region:** (Nominal) Región geográfica del cliente.
- **MembershipLevel:** (Ordinal) Nivel de lealtad del cliente (Bronce, Plata, Oro, Platino).
- **AcquisitionSource:** (Nominal) Canal a través del cual el cliente llegó al sitio web.
- **DeviceType:** (Nominal) Dispositivo usado para la compra (Móvil, Desktop, Tablet).
- **UsedDiscount:** (Binaria) Indica si se aplicó un cupón de descuento ('Sí'/'No').
- **ShippingType:** (Nominal) Tipo de envío seleccionado por el cliente.

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 gerencia de OmniRetail desea investigar si existe una diferencia significativa en el indicador clave **TimeOnSite_min** entre los distintos segmentos de clientes definidos por la categoría **AcquisitionSource** (Seleccione las dos categorías con mayor número de registros).

Filtros obligatorios:

- Incluir únicamente clientes donde **UsedDiscount** = "Sí"
- Incluir únicamente clientes donde **PagesViewed** ≤ 8

Ejercicio: Con un nivel de significancia del 3 %, pruebe si hay diferencias significativas entre la proporción de **TimeOnSite_min** que superan los 14.14 entre los grupos más poblados en la variable **AcquisitionSource**. Interprete los resultados en el contexto del problema.

- b) (1 punto) El grupo de Marketing están interesados en conocer el rango de valores bajo los cuales se puede encontrar la verdadera diferencia de promedios de **TimeOnSite_min**, cuando se compara entre los grupos definidos por **Gender** (Seleccione las dos categorías con mayor número de registros).

Filtros obligatorios:

- Incluir únicamente clientes donde `ShippingType` = “Dia Siguiente”
- Incluir únicamente clientes donde `ItemsInCart` ≥ 5

Ejercicio: Para verificar esto, emplee un nivel de confianza del 92 % para la verdadera diferencia de promedios de las variables de interés. Interprete los resultados en el contexto del problema.

- c) **(1 punto)** Basado en un estándar de la industria (o un objetivo interno de la compañía), se ha establecido que el valor promedio real de la métrica `LastLogin_days` es de 14.97.

Filtros obligatorios:

- Incluir únicamente clientes donde `Region` = “Este”
- Incluir únicamente clientes donde `MembershipLevel` = “Bronce”

Ejercicio: OmniRetail necesita verificar si el comportamiento promedio de un subgrupo específico de sus clientes en la métrica `LastLogin_days` se alinea con el valor de 14.97. Para probar esto, construya un intervalo de confianza del 91 % y verifique si dicho valor es viable. Concluya en el contexto de la población.

- d) **(1 punto)** Si se toma como cierto el valor 14.97 para el valor promedio de la métrica `LastLogin_days` (punto anterior), y se define una región crítica preestablecida dada por:

$$RC : \{\bar{X} | \bar{X} < 11.58 \quad \text{ó} \quad \bar{X} > 18.29\}$$

Ejercicio: Emplee un nivel de significancia del 11 % para calcular la probabilidad de cometer un Error Tipo I (α) y un Error Tipo II (β). Para el Error Tipo II emplee cómo hipótesis alternativa específica el valor $\mu_1 = 17.36$. Interprete los resultados 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 clientes donde `MembershipLevel` = “Bronce”
- Incluir únicamente clientes donde `NumPurchases_lifetime` ≥ 7

Ejercicio:

- **Prueba 1:** Evalúe si la variable `LastLogin_days` sigue una distribución lognormal.
- **Prueba 2:** Evalúe si la variable `TimeOnSite_min` sigue una distribución exponencial.

Para cada prueba, emplee un nivel de significancia del 10 % y utilice la prueba de Cramer-Von Mises o Anderson-Darling. Presente el juego de hipótesis, P-Valor y conclusión.

****Nota**** Si el ajuste de alguna distribución les saca error, pueden probar ajustando la distribución Poisson u otra de su elección cómo alternativa.