

**FACULTAD DE INGENIERÍA**

**CARRERA DE INGENIERÍA DE SISTEMAS**

**ESTADÍSTICA APLICADA**

**2023 – 2**

**Guía de ejercicios**

**Temas: Introducción a los datos, Asociación entre variables,**

**Estimación de parámetros**

1. Se encuestó a 245 estudiantes de la universidad “Santander”. Se sabe que en esta universidad el 72% del alumnado pertenece al área de Humanidades y el resto al área de Ciencias. Después de procesar la información muestral se obtuvo que:

* 56% eran mujeres,
* el 33.2% prefiere el rock como género musical,
* el 45.7% gusta de la emisora 104.7 VIVA F.M.
* La media del número de veces que un oyente llamó a la radio en el último mes fue de 2.86 llamadas y la moda de esa misma variable fue igual a uno.

Para la información identifique las variables de estudio, el tipo de cada variable y su gráfico respectivo.

1. Determinar si los siguientes enunciados son V o F.
2. La característica método de pago (débito, transferencia, efectivo) es de escala nominal.
3. Si al realizar un diagrama de dispersión entre dos variables cuantitativas se observa una relación exponencial, entonces es mejor calcular el coeficiente de correlación de Spearman.
4. Al calcular la media y mediana de la variable ingresos de un empleado se tiene que la media es mucho mayor que la mediana. Entonces, los ingresos de un empleado presentan una distribución asimétrica negativa.
5. Para relacionar el método de pago (débito, transferencia, efectivo) con la zona donde reside el cliente (norte, sur, este, oeste) una de las medidas estadísticas más adecuadas de usar es el estadístico Cramer’s V.
6. Para comparar el grado de homogeneidad entre el ingreso de un empleado y sus respectivas edades se emplea la varianza.
7. El encargado de una tienda registró el tiempo (minutos) que los clientes estuvieron en la tienda desde el momento en que tomaron un número hasta que se fueron. Se tienen los siguientes datos para una muestra de 16 clientes:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 15 | 14 | 16 | 14 | 14 | 14 | 13 | 8 |
| 12 | 9 | 7 | 17 | 10 | 15 | 16 | 16 |

1. ¿Qué gráfico o gráficos exploratorios permiten describir estos datos?
2. Calcular e interpretar las medidas de tendencia central (media, mediana y moda).
3. Calcular e interpretar las medidas de dispersión rango intercuartil y desviación estándar.
4. ¿Cuánto debería ser el tiempo mínimo que pasa un cliente en la tienda para pertenecer al 18% de los clientes que pasan mayor tiempo en la tienda?
5. El tiempo máximo que pasa un cliente en la tienda para pertenecer al 12% de los clientes que pasan menor tiempo en la tienda será:
6. ¿Qué tipo de distribución presenta esta característica? *Considerar calcular el coeficiente de asimetría y curtosis.*
7. El área de recursos humanos de una pequeña empresa considera que existe una relación entre el ausentismo (días) y la edad (años) del empleado. Para verificar esto, selecciona una muestra aleatoria de 8 empleados y obtiene los siguientes datos:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Días ausente** | 15 | 6 | 10 | 18 | 9 | 7 | 14 | 11 |
| **Edad** | 27 | 61 | 37 | 23 | 46 | 58 | 29 | 36 |

¿Qué se puede comentar al respecto? Responder de manera visual y cuantitativa.

1. Se tiene la siguiente matriz de correlación de 5 variables cuantitativas:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **A** | 1 | 0.58 | -0.34 | 0.22 | 0.06 |
| **B** | 0.58 | 1 | 0.10 | -0.12 | -0.65 |
| **C** | -0.34 | 0.10 | 1 | 0.79 | -0.02 |
| **D** | 0.22 | -0.12 | 0.79 | 1 | 0.18 |
| **E** | 0.06 | -0.65 | -0.02 | 0.18 | 1 |

1. ¿Entre qué variables existe una mayor asociación?
2. ¿Entre qué variables existe una menor asociación?
3. ¿Cuáles variables presentan la mayor asociación positiva?
4. ¿Cuáles variables presentan la mayor asociación negativa?
5. Una pizzería local desea conocer un poco más sobre el comportamiento de sus pedidos. Para esto, selecciona una muestra aleatoria de 9 pedidos realizados la última semana.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Orden** | App | Web | Web | Web | App | App | App | Web | App |
| **Distancia de traslado (km)** | 7.5 | 13.3 | 4.7 | 14.6 | 8.4 | 12.6 | 6.2 | 10.5 | 9.8 |
| **Tiempo de entrega (min)** | 25 | 32 | 26 | 38 | 34 | 41 | 36 | 37 | 31 |
| **Satisfacción de entrega** | 5 | 3 | 2 | 2 | 2 | 3 | 1 | 4 | 4 |

1. Identificar el tipo de cada una de las características recopiladas.
2. ¿Qué tipo de distribución presenta la distancia de traslado del pedido? *Considerar calcular el coeficiente de asimetría.*
3. ¿Cuál debe ser el tiempo de entrega máximo del pedido para pertenecer al 10% de los pedidos que se entregan más rápido?
4. ¿La distancia de traslado presenta datos más homogéneos que el tiempo de entrega?
5. Cuantificar e interpretar la asociación entre el tiempo de entrega del pedido y la distancia de traslado.
6. ¿El nivel de satisfacción de la entrega del pedido está relacionado a la forma de orden del pedido? Sustentar su respuesta.

Los mismos datos se encuentran en el archivo “Pizzeria.xlsx” o “Pizzeria.csv”.

1. Los datos presentados a continuación corresponden a los montos y el tipo de pago realizados por una muestra de 100 clientes de una cadena de supermercados. Los directivos de esta empresa pidieron tomar la muestra para conocer el comportamiento de los **pagos de los clientes**.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cheque** | | | | | | | | | |
| 5.00 | 17.87 | 18.77 | 21.11 | 22.67 | 25.96 | 30.60 | 31.07 | 31.74 | 34.67 |
| 35.38 | 36.09 | 36.48 | 37.20 | 37.60 | 37.94 | 38.58 | 39.55 | 40.51 | 41.10 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Efectivo** | | | | | | | | | |
| 1.09 | 1.27 | 1.85 | 2.44 | 2.87 | 2.96 | 3.31 | 3.65 | 4.34 | 4.75 |
| 5.08 | 5.15 | 5.91 | 5.98 | 6.93 | 7.02 | 7.17 | 7.22 | 7.40 | 7.41 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tarjeta** | | | | | | | | | |
| 14.44 | 19.78 | 22.59 | 25.57 | 26.57 | 26.91 | 27.66 | 27.89 | 33.76 | 44.53 |
| 46.13 | 46.24 | 48.11 | 50.30 | 52.35 | 52.63 | 53.32 | 54.19 | 55.21 | 57.55 |

* 1. Para las tres formas de pago halle e interprete la media, la mediana y la moda.
  2. ¿Cuál de las tres formas de pago presenta mayor homogeneidad?
  3. Para las tres formas de pago halle e interprete el coeficiente de asimetría

1. Novecientos cincuenta escolares se clasificaron de acuerdo a sus hábitos alimenticios y a su coeficiente intelectual:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Estado de nutrición** | Coeficiente intelectual | | | |
| < 80 | 80 - 90 | 90 - 99 | ≥100 |
| Nutrición buena | 245 | 228 | 177 | 219 |
| Nutrición pobre | 31 | 27 | 13 | 10 |

Indique si existe asociación entre las variables y la fuerza de asociación

1. Un investigador está interesado en verificar si el rendimiento en una planta de un grupo de personas está asociado con la antigüedad en la planta. En el estudio se eligieron al azar 10 trabajadores y se registraron los siguientes datos:

|  |  |
| --- | --- |
| **Rendimiento** | **Antigüedad** |
| Muy bajo | 10 |
| Muy alto | 15 |
| Bajo | 18 |
| Medio | 22 |
| Bajo | 18 |
| Muy bajo | 12 |
| Medio | 15 |
| Medio | 14 |
| Alto | 25 |
| Medio | 22 |

Calcule e interprete la medida de asociación más adecuada entre el rendimiento y la antigüedad en la planta.

1. Al gerente de comercialización de una cadena grane de supermercados le gustaría determinar si hay correlación entre el espacio (en pies) en estantes sobre las ventas (en cientos de soles) de comida para mascotas. Se selecciona una muestra aleatoria de 12 supermercados de igual tamaño y los resultados se presentan a continuación:

|  |  |  |
| --- | --- | --- |
| **Tienda** | **Espacio** | **Ventas** |
| 1 | 5 | 1.6 |
| 2 | 5 | 2.2 |
| 3 | 5 | 1.4 |
| 4 | 10 | 1.9 |
| 5 | 10 | 2.4 |
| 6 | 10 | 2.6 |
| 7 | 15 | 2.3 |
| 8 | 15 | 2.7 |
| 9 | 15 | 2.8 |
| 10 | 20 | 2.6 |
| 11 | 20 | 2.9 |
| 12 | 20 | 3.1 |

Halle e interprete el coeficiente de correlación para las variables en estudio

1. Se cuenta con una muestra de 200 empleados de una empresa, en la que el salario promedio fue de 75 mil soles anuales y una desviación estándar de 10 mil nuevos soles.
2. Determine los límites de confianza del 90% para la estimación del salario medio verdadero.
3. ¿Con qué nivel de confianza se puede decir que la media verdadera es 75 ± 1.5 miles de soles?
4. ¿Se puede afirmar que con un nivel de confianza del 95% se obtiene un intervalo más amplio? Justifique
5. Un sondeo realizado con 100 votantes tomados en forma aleatoria de todos los votantes de determinado distrito indica que, el 55% están a favor de cierto candidato. Encontrar los límites de confianza del 95%, 90% y 99% para la proporción de todos los votantes a favor del candidato.
6. Se desea realizar una investigación sobre el nivel de satisfacción de los alumnos de la Universidad de Lima con respecto a las clases híbridas. Para ello se tomó una muestra aleatoria de 100 alumnos seleccionados al azar, encontrándose que 10 de ellos estaban insatisfechos con las clases híbridas. Hallar un intervalo de confianza al 99% sobre el nivel de insatisfacción de los alumnos.
7. Se toma una muestra de 34 personas que escuchen radio y se determina la cantidad de horas, por semana que escuchan la radio. Los datos se muestran a continuación:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 9 | 8 | 7 | 4 | 8 | 6 | 8 | 8 | 7 | 10 | 8 | 10 | 6 | 7 | 7 | 8 | 9 |
| 6 | 5 | 8 | 5 | 6 | 8 | 7 | 8 | 5 | 5 | 8 | 7 | 6 | 6 | 4 | 5 | 6 |

1. Determinar el intervalo de confianza a un nivel del 90% de confianza para la verdadera media de la cantidad de horas por semana que escuchan radio y ¿se podrá afirmar que el valor de 5 cae dentro de este intervalo?
2. ¿Se puede afirmar que la media de la cantidad de horas por semana que escuchan radio es superior a 5 horas por semana? Evalúe a un nivel de confianza del 90%.
3. ¿Se puede afirmar que la media de la cantidad de horas por semana que escuchan radio es diferente a 6.3 horas por semana? Evalúe a un nivel de confianza del 95%.
4. Cierta fábrica de ventanas y puertas comerciales cuenta con una maquinaria que está configurada para producir una sección de acero de 0.65 cm de espesor. Se seleccionó una muestra aleatoria de 14 secciones de acero y se registró el espesor de cada una en la siguiente tabla:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0.541 | 0.757 | 0.599 | 0.823 | 0.645 | 0.688 | 0.518 |
| 0.640 | 0.780 | 0.754 | 0.765 | 0.739 | 0.564 | 0.625 |

Suponga que la distribución del espesor de la sección es normal.

1. Realizar una estimación puntual de la verdadera dispersión del 50% central de los valores de espesor de la sección de acero.
2. Estimar un intervalo de confianza del 99% para el verdadero espesor promedio de la sección de acero.
3. A medida que los rodillos se erosionan, la máquina comienza a producir secciones de acero demasiado gruesas. Usando su respuesta de la parte a), ¿hay alguna evidencia que sugiera que el verdadero espesor medio de la sección de acero es más de 0.65 centímetros? Justifique su respuesta.
4. Según los registros históricos de supervisión de la maquinaria, se sabe que la desviación estándar poblacional de los espesores de las secciones de acero es 0.097 cm. ¿Cuánto debería ser el tamaño de muestra que se necesita seleccionar si se quiere estimar el espesor medio de la sección de acero con un margen de error máximo de 0.057 centímetros y una confianza del 97%?
5. Si la máquina está produciendo secciones de aceros con tamaños de espesor muy variados, la producción debe detenerse y realizar una inmediata calibración a la maquinaria. Al 95% de confianza, ¿hay alguna evidencia que sugiera que la desviación estándar de los espesores de las secciones de acero sea superior a 0.20 centímetros y que deba detenerse la producción?
6. La cámara de comercio de una ciudad se encuentra interesada en estimar la cantidad promedio de dinero que gasta la gente que asiste a convenciones que se llevan a cabo en la ciudad, se seleccionaron 60 personas y se les preguntó la cantidad que gastaban por día. Se obtuvo la siguiente información en dólares:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 150 | 175 | 163 | 148 | 142 | 189 | 135 | 174 |
| 168 | 152 | 158 | 184 | 134 | 146 | 155 | 163 |

Si se supone que la cantidad de dinero gastado en un día es una variable aleatoria con distribución Normal.

1. Obtenga los intervalos de confianza del 95% y 98% para la cantidad promedio real.
2. Obtener una estimación bootstrap para la cantidad promedio real y el intervalo a un nivel de confianza del 95%. Considere 1000 muestras bootstrap, use set.seed(10) y desconocida la distribución de la variable cantidad de dinero gastado en un día.
3. Algunos docentes creen que el éxito escolar está directamente relacionado con la cantidad de tiempo dedicado a completar las tareas escolares. El archivo **“*Tareas.xlsx*”** contiene información sobre los puntajes de progreso obtenidos por una muestra de estudiantes de nivel secundaria y el tiempo empleado en el desarrollo de sus tareas secundarias (Menos de una hora, Más de dos horas).
4. Realizar una estimación puntual del verdadero puntaje de progreso promedio de los estudiantes de nivel secundario para cada una de las categorías de tiempo.
5. Realizar una estimación por intervalo de confianza del verdadero puntaje de progreso promedio de los estudiantes de nivel secundario para cada una de las categorías de tiempo. Usar un nivel de confianza del 95%.
6. Estimar un intervalo de confianza bootstrap del 95% del verdadero puntaje de progreso mediano de los estudiantes de nivel secundario para cada una de las categorías de tiempo.
7. Una empresa que desarrolla software de inteligencia artificial tiene interés en estimar la proporción de personas que reconocen el logo de la manzana mordida de Apple. De los 952 consumidores seleccionados al azar y encuestados, 132 pudieron identificar el logo asociado a Apple.
8. ¿Cuánto es la estimación puntual de la verdadera proporción de consumidores que reconocen el logo de Apple?
9. Encontrar un intervalo de confianza del 94% para la verdadera proporción de consumidores que reconocen el logo de Apple.
10. La empresa comercializará una versión Apple de su nuevo software si la verdadera proporción de personas que reconocen el logo es superior a 0.10. Al 94% de confianza o 6% de significación, ¿hay alguna evidencia que sugiera que la verdadera proporción de personas que reconocen el logo es mayor que 0.10? Justifique su respuesta.
11. Suponga que la proporción de consumidores que reconocen el logo de Apple se estima en 0.18. ¿Cuántos consumidores se deberían seleccionar para la encuesta si se permitiera un error máximo de 0.021 con un 98% de confianza?
12. En un estudio para evaluar la eficacia de un producto que previene la caries dental, participaron 50 personas que usaron este producto durante seis meses y una vez concluido este periodo, se registró el número de caries que presentaron.

| Número de caries | Número de personas |
| --- | --- |
| 0 | 36 |
| 1 | 9 |
| 2 | 3 |
| 3 | 2 |
|  |  |

1. Estime, con un nivel de confianza del , la proporción de personas que no tienen caries luego de usar el producto por 6 meses. Interprete.
2. Estime, con un nivel de confianza del , la media del número de caries de las personas que usan el producto. A partir del intervalo hallado comente si hay evidencia suficiente para concluir que, la media del número de caries de las personas que usan el producto es menor a 0.50 caries por persona. Justifique su respuesta.
3. Considere que en el futuro se desea repetir este estudio. En este caso, se tiene como objetivo estimar la proporción de personas que no presentan caries luego de usar el producto 6 meses con un margen de error no mayor de 0.04 a un nivel de confianza del y estimar la media del número de caries por persona con un margen de error de 0.1 a un nivel de confianza del 99%. ¿Qué tamaño de muestra recomendaría para cumplir con los objetivos de este nuevo estudio? Considere la información presentada como una muestra piloto de este futuro estudio.
4. El archivo “**MarketingDirecto.csv”** contiene datos de 1000 clientes de un vendedor de marketing directo el cuál vende sus productos sólo a través de correos electrónicos personalizados. Considerando las variables Monto (gasto en dólares) e Hijos (número de hijos):
5. Calcule e interprete un intervalo de confianza del 95% para el gasto medio de un cliente.
6. Calcule e interprete un intervalo de confianza del 92% para la desviación estándar del monto de préstamo.
7. Calcule e interprete un intervalo de confianza del 90% para la proporción de clientes que tienen más de un hijo. De acuerdo con el intervalo hallado, ¿es posible concluir que la mayoría de los clientes tienen más de un hijo?
8. Si se desease tomar una muestra para la siguiente campaña, halle el tamaño de muestra a un 95% de confianza para el gasto medio si se desea obtener un error de estimación de 3000 soles.
9. Si se desease tomar una muestra para la siguiente campaña, halle el tamaño de muestra a un 95% de confianza para la proporcición de personas que tienen casa alquilada si se desea obtener un error de estimación de 4%.