Inferencia Estadística

Ejercicio: Introducción a la Inferencia Estadística

DSLab

noviembre, 2024

Ejercicio 1

Pregunta: Describe en tus propias palabras qué es la Ciencia de Datos y su importancia en el análisis de grandes volúmenes de datos.

Solución: La Ciencia de Datos es una disciplina interdisciplinaria que se centra en la extracción de conocimiento significativo a partir de grandes conjuntos de datos. Es crucial en un mundo impulsado por los datos, ya que permite tomar decisiones informadas y hacer predicciones basadas en el análisis de datos. La Ciencia de Datos combina elementos de estadística, informática y conocimiento específico del dominio para interpretar datos y aplicar este conocimiento en diversas áreas como la medicina, las finanzas y la tecnología. Su importancia radica en su capacidad para transformar datos crudos en información valiosa que puede impulsar la innovación y la eficiencia en múltiples campos.

Ejercicio 2

Pregunta: Enumera las herramientas estadísticas que se utilizan en la inferencia estadística y explica brevemente su propósito.

Solución: En la inferencia estadística, se utilizan diversas herramientas para analizar datos y hacer generalizaciones sobre una población a partir de muestras:

- **Pruebas de Hipótesis**: Se utilizan para determinar si existe suficiente evidencia en una muestra de datos para inferir que una cierta condición es verdadera para toda la población.

- **Intervalos de Confianza**: Proporcionan un rango estimado que es probable que contenga el valor de un parámetro desconocido de la población, con un cierto nivel de confianza.
- **Análisis de Varianza (ANOVA)**: Permite comparar tres o más medias de grupos para determinar si al menos una de las medias es diferente de las demás.
- **Chi-cuadrado (χ^2)**: Es una prueba que mide la discrepancia entre los datos observados y los datos que se esperarían según un modelo específico.
- **T-test**: Evalúa si las medias de dos grupos son estadísticamente diferentes entre sí.
- **Correlación**: Mide la relación entre dos variables y la fuerza de esta relación.
- **Estadística Bayesiana**: Utiliza la probabilidad para representar la incertidumbre sobre los parámetros del modelo y actualiza esta incertidumbre a medida que se obtienen más datos.
- **Métodos de Muestreo**: Incluyen técnicas para seleccionar muestras representativas de la población para realizar inferencias estadísticas.
- **Métodos No Paramétricos**: Son técnicas que no asumen una distribución específica de los datos y son útiles cuando no se cumplen los supuestos de los métodos paramétricos.

Ejercicio 3

Pregunta: Define los términos “población” y “muestra” y explica la diferencia entre ambos.

Solución:

- **Población**: Se refiere al conjunto completo de elementos o resultados que se están estudiando, del cual se desean obtener conclusiones. La población incluye a todos los individuos, mediciones, objetos o eventos que cumplen con un conjunto de especificaciones previamente definidas. Por ejemplo, si estamos estudiando la altura de los estudiantes de una universidad, la población sería la altura de **todos** los estudiantes de esa universidad.

- **Muestra**: Es un subconjunto de la población que se selecciona para representarla. La muestra debe ser representativa de la población para que las inferencias hechas a partir de ella sean válidas. Por ejemplo, si elegimos a 100 estudiantes al azar de la universidad mencionada anteriormente, esos 100 estudiantes constituirían una muestra de la población.

La **diferencia principal** entre ambos términos es el alcance. Mientras que la población es el grupo completo que se quiere estudiar, la muestra es solo una parte de ese grupo. Las muestras se utilizan porque a menudo es impracticable o imposible estudiar toda la población debido a limitaciones de tiempo, costo o logística. Por lo tanto, se selecciona una muestra para obtener estimaciones o pruebas sobre la población completa.

Ejercicio 4

Pregunta: ¿Qué es una distribución de probabilidad y cómo se relaciona con las variables cualitativas y cuantitativas?

Solución: Una **distribución de probabilidad** es una función matemática que describe la probabilidad de ocurrencia de los diferentes posibles resultados en un experimento. En otras palabras, asigna probabilidades a cada posible resultado de una variable aleatoria. Las distribuciones de probabilidad se relacionan con las variables cualitativas y cuantitativas de la siguiente manera:

- **Variables Cualitativas (o Categóricas)**: Son aquellas que describen una cualidad o categoría y no tienen un orden o medida numérica inherente. Las distribuciones de probabilidad para estas variables son conocidas como **distribuciones discretas** y asignan probabilidades a resultados específicos. Un ejemplo es la **distribución binomial**, que puede modelar eventos como el lanzamiento de una moneda, donde los resultados son categóricos (cara o cruz).
- **Variables Cuantitativas**: Son variables que se pueden medir en una escala numérica y tienen sentido hablar de valores mayores o menores. Las distribuciones de probabilidad asociadas a estas variables son **distribuciones continuas** y asignan probabilidades a intervalos de números. Por ejemplo, la **distribución normal** es una distribución continua que se utiliza comúnmente para modelar fenómenos naturales como la altura o el peso de individuos.

Ejercicio 5

Pregunta: Realiza un resumen descriptivo de un conjunto de datos utilizando medidas de tendencia central y dispersión.

Solución: Imaginemos un conjunto de datos que representa las calificaciones de un grupo de estudiantes en un examen:

```
datos=c(72, 85, 90, 68, 88, 76, 95, 89, 75, 80)
datos
```

```
##  [1] 72 85 90 68 88 76 95 89 75 80
```

Media:

$$\text{Media} = \frac{72 + 85 + 90 + 68 + 88 + 76 + 95 + 89 + 75 + 80}{10} = \frac{818}{10} = 81.8$$

Mediana: Ordenamos los datos: 68, 72, 75, 76, 80, 85, 88, 89, 90, 95. Como hay 10 valores,

la mediana es el promedio de los dos valores centrales:

$$\text{Mediana} = \frac{80 + 85}{2} = \frac{165}{2} = 82.5$$

Moda: No hay un valor que se repita más de una vez, por lo que no hay moda en este conjunto de datos.

Rango:

$$\text{Rango} = 95 - 68 = 27$$

Varianza: Primero, calculamos la media de los cuadrados de las desviaciones respecto a la media:

$$\sigma^2 = \frac{(72 - 81.8)^2 + (85 - 81.8)^2 + ... + (80 - 81.8)^2}{10} = 71.4$$

Desviación Estándar:

$$\sigma = \sqrt{71.4} \approx 8.45$$

Cuartiles: - **Primer Cuartil (Q1)**: Valor en el percentil 25 (primera mitad inferior de los datos):

$$Q1 = 75$$

- **Tercer Cuartil (Q3)**: Valor en el percentil 75 (primera mitad superior de los datos):

$$Q3 = 89$$

Resumen Descriptivo

- **Media**: 81.8
- **Mediana**: 82.5
- **Moda**: No hay moda
- **Rango**: 27
- **Varianza**: 71.4
- **Desviación Estándar**: 8.45
- **Primer Cuartil (Q1)**: 75
- **Tercer Cuartil (Q3)**: 89

Este resumen descriptivo proporciona una visión clara y concisa de las características principales del conjunto de datos, permitiendo una mejor comprensión de su distribución y variabilidad.

Ejercicio 6

Pregunta: Explica la diferencia entre estadística descriptiva e inferencial y proporciona un ejemplo de cómo se utiliza cada una.

Solución: La **estadística descriptiva** y la **estadística inferencial** son dos ramas principales de la estadística que tienen propósitos y métodos diferentes:

- **Estadística Descriptiva**: Se centra en resumir y describir las características de un conjunto de datos. Utiliza medidas como la media, mediana, moda, rango y desviación estándar para dar una visión general de los datos. Por ejemplo, si tenemos los resultados de una prueba de matemáticas de una clase, la estadística descriptiva podría incluir el cálculo de la media de las calificaciones, la calificación más alta, la más baja y la variabilidad de las calificaciones.
- **Estadística Inferencial**: Va más allá de la descripción de los datos y busca hacer predicciones o generalizaciones sobre una población basándose en una muestra de datos. Utiliza herramientas como pruebas de hipótesis, intervalos de confianza y regresión para inferir patrones y tomar decisiones. Por ejemplo, si queremos saber si un nuevo método de enseñanza es efectivo, podríamos aplicarlo a una muestra de estudiantes y usar la estadística inferencial para determinar si los resultados observados en la muestra pueden generalizarse a todos los estudiantes.

Solución:

Ejercicio 7

Pregunta: Diseña un experimento para ilustrar cómo el muestreo aleatorio simple puede ser utilizado para estimar una característica de una población.

Solución: Para ilustrar cómo el muestreo aleatorio simple puede ser utilizado para estimar una característica de una población, consideremos el siguiente experimento:

Objetivo del Experimento: Estimar la proporción de personas en una ciudad que prefieren el transporte público sobre otros medios de transporte.

Población: Todos los residentes de la ciudad que son mayores de edad y utilizan algún medio de transporte para desplazarse.

Característica de Interés: Preferencia por el transporte público.

Procedimiento: 1. **Definición de la Población**: Identificar a todos los residentes de la ciudad que son mayores de edad y utilizan algún medio de transporte.

2. **Selección de la Muestra**:

- Utilizar un registro actualizado de la población, como el padrón municipal, para obtener una lista de individuos.
- Seleccionar una muestra aleatoria de individuos utilizando un generador de números aleatorios.
- Determinar el tamaño de la muestra necesario para obtener resultados con un nivel de confianza y un margen de error deseado.

3. **Recolección de Datos**:
   - Contactar a los individuos seleccionados y preguntarles si prefieren el transporte público sobre otros medios de transporte.
   - Registrar las respuestas afirmativas y negativas.
4. **Análisis de Datos**:
   - Calcular la proporción de respuestas afirmativas en la muestra.
   - Utilizar esta proporción como una estimación puntual de la preferencia en la población total.
5. **Estimación de la Población**:
   - Calcular un intervalo de confianza para la proporción estimada, lo que proporcionará un rango dentro del cual se espera que se encuentre la verdadera proporción de la población con un cierto nivel de confianza.
6. **Conclusión**:
   - Presentar la proporción estimada y el intervalo de confianza como la estimación de la preferencia por el transporte público en la población.
   - Discutir las limitaciones del estudio y la posibilidad de sesgo si la muestra no fue perfectamente aleatoria o si hubo una tasa de respuesta baja.

Ejercicio 8

Pregunta: Explica el Teorema Central del Límite y su relevancia en la inferencia estadística.

Solución: ### Teorema Central del Límite (TCL)

El Teorema Central del Límite es un principio fundamental en estadística que establece que, bajo ciertas condiciones, la distribución de la suma de un gran número de variables aleatorias independientes y identicamente distribuidas (i.i.d.) tiende a aproximarse a una distribución normal (gaussiana), independientemente de la forma de la distribución original de las variables.

Enunciado del Teorema Central del Límite

Para una muestra de tamaño n tomada de una población con cualquier distribución de probabilidad con media μ y desviación estándar σ, la distribución de la media muestral $\bar{X}$ se aproximará a una distribución normal a medida que n se haga grande. Matemática-

mente, si $X_1, X_2, ..., X_n$ son variables aleatorias independientes e idénticamente distribuidas con media μ y desviación estándar σ, entonces la media muestral $\bar{X}$ se distribuye aproximadamente como:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Importancia y Relevancia en la Inferencia Estadística

1. **Justificación del Uso de la Distribución Normal**: El TCL permite justificar el uso de la distribución normal en la inferencia estadística. Incluso si la población original no está distribuida normalmente, la distribución de las medias muestrales se aproximará a una distribución normal si el tamaño de la muestra es suficientemente grande.
2. **Construcción de Intervalos de Confianza**: Permite construir intervalos de confianza para estimar parámetros poblacionales. Por ejemplo, al estimar la media poblacional, podemos utilizar la distribución normal para determinar un rango donde probablemente se encuentra la media verdadera.
3. **Pruebas de Hipótesis**: Facilita la realización de pruebas de hipótesis. Dado que la distribución de la media muestral es aproximadamente normal, se pueden aplicar métodos estadísticos basados en la normalidad para decidir si rechazar o no una hipótesis nula.
4. **Simplificación de Cálculos**: El TCL simplifica el análisis de datos, ya que permite trabajar con la distribución normal, que tiene propiedades bien definidas y es ampliamente comprendida y tabulada, facilitando los cálculos y la interpretación de los resultados.
5. **Aplicación Universal**: Es aplicable en una amplia gama de situaciones, desde la economía y la biología hasta la ingeniería y las ciencias sociales, siempre que se cumplan las condiciones necesarias (independencia y tamaño de muestra grande).

Ejercicio 9

Pregunta: Compara y contrasta la estadística paramétrica y no paramétrica, dando ejemplos de cuándo se utilizaría cada una.

Solución: La **estadística paramétrica** y la **estadística no paramétrica** son dos enfoques de análisis estadístico que tienen diferentes supuestos y aplicaciones. La **estadística paramétrica** se utiliza cuando los datos se ajustan a ciertos supuestos y se conoce la distribución subyacente, lo que permite hacer inferencias más precisas si estos supuestos

se cumplen. Por otro lado, la **estadística no paramétrica** es más flexible y se puede utilizar en una variedad más amplia de situaciones, especialmente cuando los datos no cumplen con los supuestos de los métodos paramétricos.

- **Estadística Paramétrica**:
  - **Supuestos**: Asume que los datos de la muestra provienen de una población que sigue una distribución de probabilidad conocida, generalmente la distribución normal. También asume homogeneidad de varianzas y la independencia de las observaciones.
  - **Uso**: Se utiliza cuando se conoce la forma de la distribución subyacente de los datos o cuando se tiene una muestra grande que, por el Teorema Central del Límite, tiende a una distribución normal.
  - **Ejemplos de Herramientas**: T-test, ANOVA, regresión lineal.
  - **Ejemplo de Uso**: Si queremos comparar las alturas promedio de dos grupos de personas y sabemos que las alturas siguen una distribución normal, podríamos usar un T-test paramétrico.
- **Estadística No Paramétrica**:
  - **Supuestos**: No hace suposiciones sobre la forma de la distribución de la población. Es útil cuando no se cumplen los suposiciones de normalidad o cuando se trata con muestras pequeñas.
  - **Uso**: Se aplica en situaciones donde no se conoce la distribución de los datos o cuando los datos son ordinales o nominales.
  - **Ejemplos de Herramientas**: Test de Wilcoxon, Test de Kruskal-Wallis, Test de Chi-cuadrado.
  - **Ejemplo de Uso**: Si queremos comparar las medianas de los tiempos de respuesta de dos grupos en una prueba y los datos son claramente no normales o son rangos en lugar de medidas, podríamos usar un test de Wilcoxon no paramétrico.

Ejercicio 10

Pregunta: Discute las diferencias entre los enfoques frecuentista y bayesiano en la inferencia estadística y da un ejemplo de aplicación para cada uno.

Solución: La inferencia estadística se basa en métodos que nos permiten hacer conclusiones sobre una población a partir de datos muestrales. Existen dos enfoques principales en la inferencia estadística: el enfoque frecuentista y el enfoque bayesiano. A continuación, se presentan las diferencias clave entre estos enfoques y ejemplos de aplicación para cada uno.

Enfoque Frecuentista

Características Principales:

1. **Interpretación de Probabilidad**: La probabilidad se interpreta como la frecuencia relativa de eventos en el largo plazo. Es decir, si un experimento se repite infinitas veces, la probabilidad de un evento es la proporción de veces que ocurre.
2. **Estimación de Parámetros**: Se basa en el concepto de estimación puntual y de intervalos de confianza. Los parámetros poblacionales se consideran fijos pero desconocidos, y los datos son aleatorios.
3. **Pruebas de Hipótesis**: Utiliza pruebas de hipótesis y valores p para decidir si rechazar la hipótesis nula. Las decisiones se basan en la frecuencia de observación de datos extremos bajo la suposición de que la hipótesis nula es verdadera.
4. **No usa Información Priori**: Los análisis frecuentistas no incorporan información previa sobre los parámetros; sólo se basan en los datos actuales.

Ejemplo de Aplicación Frecuentista:

Imaginemos que una empresa desea saber si un nuevo medicamento es efectivo para reducir la presión arterial. Realizan un ensayo clínico donde:

- **Hipótesis Nula (H0)**: El medicamento no tiene efecto en la presión arterial (la media de la reducción de la presión arterial es 0).
- **Hipótesis Alternativa (H1)**: El medicamento reduce la presión arterial (la media de la reducción de la presión arterial es mayor que 0).

El análisis frecuentista implicaría:

1. Recoger datos muestrales de pacientes.
2. Calcular la media y la desviación estándar de la reducción en la presión arterial.
3. Realizar una prueba t para comparar la media muestral con 0.
4. Calcular el valor p para determinar la significancia estadística.
5. Rechazar o no la hipótesis nula en función del valor p y el nivel de significancia establecido (por ejemplo, 0.05).

Enfoque Bayesiano

Características Principales:

1. **Interpretación de Probabilidad**: La probabilidad se interpreta como un grado de creencia o confianza sobre la ocurrencia de un evento, dado el conocimiento disponible.
2. **Estimación de Parámetros**: Los parámetros poblacionales se tratan como variables aleatorias con distribuciones de probabilidad. Utiliza la distribución a priori (información previa) y los datos observados para obtener la distribución a posteriori.
3. **Actualización de Conocimientos**: Aplica el Teorema de Bayes para actualizar la probabilidad a medida que se dispone de nueva información.

4. **Incorporación de Información Priori**: Utiliza información previa sobre los parámetros en forma de distribuciones a priori, que se combinan con la información de los datos para obtener las distribuciones a posteriori.

Ejemplo de Aplicación Bayesiana:

Supongamos que un médico quiere estimar la probabilidad de que un paciente tenga una enfermedad dada, basándose en un resultado positivo de una prueba diagnóstica y en el conocimiento previo sobre la prevalencia de la enfermedad y la precisión de la prueba.

1. **Información a Priori**: El médico tiene una estimación previa (a priori) de la prevalencia de la enfermedad (por ejemplo, 1% de la población tiene la enfermedad).
2. **Datos Observados**: La sensibilidad (probabilidad de un resultado positivo dado que el paciente tiene la enfermedad) es 90% y la especificidad (probabilidad de un resultado negativo dado que el paciente no tiene la enfermedad) es 95%.
3. **Aplicación del Teorema de Bayes**:
   - $P(E|+)$: Probabilidad de tener la enfermedad dado un resultado positivo.
   - $P(+|E)$: Sensibilidad.
   - $P(+|\neg E)$: Probabilidad de un falso positivo (1 - especificidad).
   - $P(E)$: Prevalencia de la enfermedad.
   - $P(\neg E)$: Probabilidad de no tener la enfermedad (1 - prevalencia).

$$P(E|+) = \frac{P(+|E) \cdot P(E)}{P(+|E) \cdot P(E) + P(+|\neg E) \cdot P(\neg E)}$$

Sustituyendo los valores:

$$P(E|+) = \frac{0.90 \cdot 0.01}{0.90 \cdot 0.01 + 0.05 \cdot 0.99} = \frac{0.009}{0.009 + 0.0495} = \frac{0.009}{0.0585} \approx 0.154$$

El resultado indica que, dado un resultado positivo de la prueba, la probabilidad a posteriori de tener la enfermedad es aproximadamente 15.4%.

Conclusión

Frecuentista: - No utiliza información previa. - Se basa en la frecuencia de los eventos. - Adecuado para análisis donde no se dispone de información previa o se quiere evitar la subjetividad.

Bayesiano: - Utiliza información previa (a priori). - Actualiza las probabilidades a medida que se dispone de nueva información. - Adecuado para situaciones donde la información previa es relevante y valiosa.

Ambos enfoques son útiles y válidos, y la elección entre ellos depende del contexto del problema, la disponibilidad de información previa y las preferencias del investigador.