Guía de estudio - Estadística descriptiva y probabilidades I

¡Hola! Te damos la bienvenida a esta nueva guía de estudio.

¿En qué consiste esta guía?

La siguiente guía de estudio tiene como objetivo practicar y ejercitar los contenidos que hemos visto en clase, además de profundizar temas adicionales que complementan los que hemos visto en la clase.

¡Bienvenidos al mundo de la estadística!

¡Vamos con todo!

Tabla de contenidos

¡Comencemos!

¿Qué es la estadística?

La estadística es una disciplina que se ocupa de recopilar, analizar e interpretar datos con el fin de comprender y describir los fenómenos que ocurren en diferentes áreas de estudio. A través de métodos y técnicas estadísticas, se busca extraer información significativa y tomar decisiones fundamentadas con base en la evidencia proporcionada por los datos.

En su recopilación, la estadística utiliza técnicas de muestreo para seleccionar una muestra representativa de una población más grande. Luego, se aplican métodos de análisis estadístico para resumir y visualizar los datos, identificar patrones y tendencias, calcular medidas de centralidad y dispersión, y realizar inferencias y predicciones sobre la población objetivo. En resumen, la estadística es una disciplina esencial para la comprensión y aplicación de los datos en diversas áreas, proporcionando herramientas fundamentales para el análisis, la toma de decisiones y la generación de conocimiento.

¿Qué tipos de análisis estadísticos existen?

El análisis estadístico se refiere al proceso de examinar y estudiar los datos recopilados mediante técnicas y métodos estadísticos. Su objetivo es comprender, resumir y sacar conclusiones significativas a partir de los datos, identificando patrones, tendencias y relaciones entre variables.

El análisis estadístico implica la aplicación de diferentes técnicas estadísticas, como medidas de centralidad (como la media, la mediana y la moda), medidas de dispersión (como la desviación estándar y el rango), pruebas de hipótesis, análisis de correlación y regresión, entre otros. Estas técnicas permiten examinar los datos desde diferentes perspectivas y obtener información valiosa sobre la población o el fenómeno estudiado.

Además, el análisis estadístico puede incluir la visualización de datos a través de gráficos y tablas, lo que facilita la interpretación y la comunicación de los resultados. En última instancia, el análisis estadístico proporciona una base sólida para la toma de decisiones informadas y la generación de conocimiento en diversas áreas de estudio.

1. **Análisis descriptivo:** Se utiliza para resumir y describir los datos, utilizando medidas de tendencia central (como la media, la mediana y la moda) y medidas de dispersión (como la desviación estándar y el rango) para caracterizar la distribución de los datos.

2. **Pruebas de hipótesis:** Se emplean para evaluar afirmaciones o hipótesis sobre los datos. Estas pruebas permiten determinar si hay evidencia suficiente para aceptar o rechazar una hipótesis planteada sobre una población.

3. **Análisis de correlación:** Se usa para examinar la relación entre dos o más variables. Permite determinar si existe una asociación estadística entre las variables y, en caso afirmativo, la fuerza y dirección de esa asociación.

4. **Regresión:** Se usa para investigar la relación entre una variable dependiente y una o más variables independientes. Permite predecir el valor de la variable dependiente en función de las variables independientes y entender cómo estas influyen en la variable de interés.

5. **Análisis de varianza (ANOVA):** Se utiliza para comparar las medias de tres o más grupos para determinar si hay diferencias estadísticamente significativas entre ellos.

6. **Análisis de series temporales:** Se aplica a datos secuenciales en el tiempo para identificar patrones, tendencias y estacionalidad. Se utiliza para predecir y modelar el comportamiento futuro de la serie temporal.

7. **Análisis factorial:** Se utiliza para identificar y describir la estructura subyacente en un conjunto de variables observadas, permitiendo reducir la complejidad y encontrar patrones subyacentes.

En este módulo nos enfocaremos en el análisis descriptivo de los datos, empleando las medidas de tendencia central y la dispersión de la información.

¿Qué tipos de datos existen en estadística?

En estadística, podemos encontrar diferentes tipos de datos, que se clasifican en cuatro categorías principales: nominal, ordinal, de intervalo y de razón. Estas categorías se basan en la naturaleza y las propiedades de los datos.

1. **Datos nominales:** Son datos que se utilizan para categorizar o clasificar elementos, sin un orden o jerarquía específica. Los valores en esta categoría no se pueden ordenar o medir de forma cuantitativa. Ejemplos de datos nominales son el género (masculino/femenino), el estado civil (soltero/casado/divorciado) o la nacionalidad.

2. **Datos ordinales:** Son datos que mantienen un orden o jerarquía, pero la distancia entre las categorías no es cuantificable. Los valores en esta categoría se clasifican en función de su posición relativa. Ejemplos de datos ordinales son las escalas de clasificación, como la calificación de satisfacción (baja/media/alta) o el nivel de acuerdo (totalmente en desacuerdo/en desacuerdo/neutral/de acuerdo/totalmente de acuerdo).
3. **Datos de intervalo:** Son datos numéricos que tienen un orden y una distancia cuantificable entre las categorías. En este caso, se puede determinar la diferencia entre dos valores, pero no hay un punto cero absoluto. Ejemplos de datos de intervalo son la temperatura en grados Celsius o Fahrenheit, donde la diferencia de 10 grados tiene el mismo significado en cualquier parte de la escala, pero no existe un valor de temperatura que indique la ausencia total de calor.

4. **Datos de razón:** Son datos numéricos que tienen un orden, una distancia cuantificable y un punto cero absoluto. En esta categoría, las diferencias y las relaciones entre las categorías tienen un significado absoluto. Ejemplos de datos de razón son la edad, el peso, la altura o el tiempo transcurrido. En estos casos, el valor de cero representa la ausencia total de la característica medida.

Estos tipos de datos tienen implicaciones en la elección de las técnicas estadísticas adecuadas para analizarlos y en la interpretación de los resultados obtenidos. Es importante identificar correctamente el tipo de datos con el que se está trabajando para realizar un análisis estadístico apropiado.

¡Reforcemos lo aprendido!

A partir de lo que has revisado respecto al análisis de datos, ¿De qué manera podemos clasificar los datos para hacerlos más comprensibles? Haz un mapa conceptual que te permita ordenar dichas clasificaciones para facilitar tu aprendizaje. Puedes hacerla en papel o con alguna herramienta en línea como Miro, Lucidchart, entre otras.

Medidas de tendencia central

Las medidas de tendencia central son estadísticas utilizadas para resumir y representar la ubicación o centralidad de un conjunto de datos. Estas medidas proporcionan información sobre el valor típico o representativo de los datos. Las tres medidas de tendencia central más comunes son la media, la mediana y la moda.

Estas medidas de tendencia central son útiles para resumir la distribución de los datos y proporcionar una visión general de su ubicación central. Sin embargo, es importante considerar el contexto y la naturaleza de los datos al interpretar y usar estas medidas. Cada medida tiene sus ventajas y limitaciones, y es recomendable emplear varias medidas de tendencia central en conjunto para obtener una imagen más completa de los datos.

Media Aritmética

La media es una medida de tendencia central que representa el valor promedio de un conjunto de datos. Se calcula sumando todos los valores en el conjunto y dividiendo esa suma por el número total de elementos. La media es ampliamente utilizada debido a su simplicidad y capacidad para resumir la distribución de los datos en un solo número. Sin embargo, es valioso tener en cuenta que la media puede verse afectada por valores extremos, lo que puede distorsionar su representatividad en casos donde existen valores atípicos.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mediana

La mediana es una medida de tendencia central que se utiliza para representar el valor central de un conjunto de datos ordenados de manera ascendente o descendente. Para calcular la mediana, se coloca el conjunto de datos en orden y se selecciona el valor que se encuentra en la posición central. Si el número de elementos en el conjunto es impar, la mediana es el valor exactamente en el medio. En cambio, si el número de elementos es par, la mediana se obtiene promediando los dos valores centrales.

De cualquier manera, la mediana tiene sentido para una gran cantidad de datos que, además, pueden considerarse como continuos, por lo que estos procedimientos específicos solo son referenciales: lo que importa es encontrar un valor que sea mayor o igual que el 50 % del conjunto, y menor o igual que el 50 % restante.

Moda

La moda es una medida de tendencia central utilizada en estadística para representar el valor o valores que aparecen con mayor frecuencia en un conjunto de datos. En otras palabras, la moda es el valor que se repite con mayor frecuencia en el conjunto. Puede haber una moda única (y el conjunto ser unimodal), o múltiples modas (conjunto multimodal).

Indicadores de posición

Percentiles

Los percentiles permiten identificar el valor donde un porcentaje en específico cae. Si tomamos el percentil 15 %, por ejemplo, este corresponde a un valor que es mayor o igual que el 15 % menor de los valores. Una manera de calcularlo es mediante el método del ranking más cercano que resume el punto de una lista ordenada de menor a mayor.

Un uso muy habitual es el peso y talla de los recién nacidos, para tener una referencia de su curva de crecimiento.

Cuartiles, quintiles y deciles

A partir de los percentiles, suelen definirse algunos de uso común. Lo más usual es separar el conjunto en 4, 5 o 10 partes iguales que definen, respectivamente, los **cuartiles**, **quintiles** o **deciles**.

De esta manera, por ejemplo, los datos que se ubican entre los percentiles 40 y 60 corresponden al **tercer quintil,** así como los que se ubican entre los percentiles 0 (es decir, el valor mínimo) y el percentil 10 corresponden al **primer decil**. Hay que tener en cuenta que estos términos suelen usarse para denominar al límite superior de cada intervalo.

Medidas de dispersión

Las medidas de dispersión son indicadores utilizados para describir la variabilidad o dispersión de un conjunto de datos. Mientras que las medidas de tendencia central, como la media o la mediana, nos dan una idea de la ubicación central de los datos, las medidas de dispersión nos indican cómo se extienden o dispersan los valores alrededor de esa ubicación central.

Algunas medidas de dispersión de uso común son:

1. **Rango:** Es la diferencia entre el valor máximo y el valor mínimo en un conjunto de datos. Proporciona una idea básica de cuán dispersos están los datos.

2. **Rango intercuartil:** corresponde a la diferencia entre los percentiles 75 y 25. Da una medida de la dispersión de los datos "centrales" del conjunto.

3. **Desviación media:** Es la media de las diferencias absolutas entre cada valor y la media. Indica la dispersión promedio de los valores respecto a la media.

4. **Varianza:** Es una medida de dispersión que representa la media de los cuadrados de las desviaciones de cada valor respecto a la media. La varianza nos indica cuánto se alejan los valores individuales de la media.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}$$

5. **Desviación estándar:** Es la raíz cuadrada de la varianza. La desviación estándar mide la dispersión de los valores alrededor de la media y se expresa en las mismas unidades que los datos originales.

6. **Coeficiente de variación:** Es una medida relativa de dispersión que se calcula dividiendo la desviación estándar por la media y multiplicando por 100. Proporciona una indicación de la dispersión relativa en relación con la media, con la ventaja de que "estandariza" la unidad de medida. Así, por ejemplo, si queremos comparar la dispersión entre las estaturas y los pesos de un grupo de personas, la desviación estándar, en cada caso, tendrá unidades diferentes, pero el coeficiente de variación sí corresponde a una medida absoluta y, por ello, comparable.

Reflexiona

- ¿Qué es la media, la moda y la mediana? Identifica cada una
- ¿Para qué sirven cada una de ellas en el análisis de datos?
- ¿Qué son las medidas de dispersión?
- ¿En qué casos ocuparías los percentiles o la desviación estándar? Da un ejemplo por cada uno.

Indicadores estadísticos con Python

Ya sabemos trabajar con DataFrames, y conocemos las ventajas que esto tiene. Veremos cómo utilizarlos para nuestro análisis estadístico. Supondremos que contamos con un DataFrame llamado df.

Indicadores del DataFrame

```
df.describe()
```

Este comando nos permitirá conocer el conteo de valores, media, desviación estándar, mínimo, percentil 25, percentil 50, percentil 75 y valor máximo de cada columna de nuestro Data Frame que contenga datos cuantitativos.

Indicadores por columna

```
media = df['nombre_columna'].mean()
mediana = df['nombre_columna'].median()
moda = df['nombre_columna'].mode()
desviacion_estandar = df['nombre_columna'].std()
minimo = df['nombre_columna'].min()
maximo = df['nombre_columna'].max()
percentil=df['nombre_columna'].quantile(0.15) #percentil 15
```

Estos comandos nos permitirán determinar individualmente los indicadores estadísticos, por columna. Para el caso de la moda, podemos poner una columna con variables cualitativas.

Medidas agrupadas

```
medida= df.groupby('columna_de_categorías')['columna_de_datos'].mean()
```

Podemos agrupar considerando las categorías de una columna, y calcular para cada una de ellas el indicador que deseemos (en este caso, el promedio)

Tablas pivote

```
tabla_pivote = pd.pivot_table(df, values=['columna_1', 'columna_2', 'columna_3],
index='columna_indice', aggfunc=['funcion1', 'funcion2'])
```

Esto nos permite generar una tabla pivote en la que, para las categorías de la **columna_indice**, se calculará la **funcion1** y la **funcion2** correspondientes a las columnas **columna_1**, **columna_2** y **columna_3**

¡Manos a la obra! - Indicadores estadísticos con python

1. Carga el dataset POKEMON.XLSX, y a partir de sus datos:
    a. Calcula el promedio, mediana y moda de las variables Altura (Height) y peso (Weight)
    b. Calcula los quintiles, rango intercuartil y varianza de las mismas variables.
    c. Calcula el promedio de peso y altura de los Pokemon, agrupados por Shapes
    d. En cada uno de los casos anteriores, ¿cuál presenta una media más representativa? ¿Cómo lo puedes determinar? Explica.

2. Carga el dataset PS4_GamesSales.csv, y a partir de sus datos:
    a. Calcula el promedio y la desviación estándar de ventas de los videojuegos según cada región
    b. ¿Qué género presenta mayor dispersión en sus ventas globales? ¿Y en sus ventas por región?
3. Carga el dataset earnings.csv, y a partir de sus datos:
    a. Calcula los indicadores correspondientes para las variables cuantitativas. ¿Qué cuidados hay que tener?
    b. ¿Qué variable presenta una mayor dispersión? Explica.

Preguntas de cierre

- ¿Entiendo los conceptos clave presentados en la clase? Pregúntate si comprendes los fundamentos de la estadística, como las medidas de tendencia central y las medidas de dispersión.
- ¿Puedo aplicar los conceptos en situaciones reales? Reflexiona sobre si te sientes cómodo aplicando los conceptos de estadística en problemas prácticos y si puedes identificar qué medidas son más apropiadas en diferentes contextos.
- ¿Puedo interpretar los resultados estadísticos correctamente? Evalúa si puedes interpretar correctamente los resultados de un análisis estadístico y si comprendes las implicaciones de esos resultados.