MÁSTER DE INTELIGENCIA ARTIFICIAL Herramientas de estadística

Imagen que contiene Texto

Descripción generada automáticamente

Un análisis estadístico:

*Contrastes de hipótesis, intervalos de confianza, y un modelo de regresión lineal simple.*

Profesor: Walter A. Ortiz

Alumno: Mikel Etxezarraga Hernández

Tabla de contenido

[MEMORIA 3](#_Toc130430361)

[Introducción 3](#_Toc130430362)

[Datos 3](#_Toc130430363)

[Ejercicio 1 5](#_Toc130430364)

[Apartado 1 5](#_Toc130430365)

[Apartado 2 7](#_Toc130430366)

[Apartado 3 8](#_Toc130430367)

[Apartado 4 9](#_Toc130430368)

[Ejercicio 2 10](#_Toc130430369)

[Apartado 1: Recta de regresión de C2 (Y) sobre C1 (x) 10](#_Toc130430370)

[Apartado 2: Recta de regresión de C1 (Y) sobre C2 (x) 12](#_Toc130430371)

[Apartado 3: coeficiente de correlación 13](#_Toc130430372)

[ANEXOS 14](#_Toc130430373)

[Figura 1 14](#_Toc130430374)

[Figura 2 14](#_Toc130430375)

[Figura 5 14](#_Toc130430376)

[Figura 6 15](#_Toc130430377)

[Figura 7 15](#_Toc130430378)

# MEMORIA

## Introducción

El análisis estadístico es una herramienta esencial en la investigación científica, que permite evaluar la validez de las hipótesis y conclusiones a través del uso de técnicas y métodos estadísticos.

En el presente trabajo, se utilizarán tres técnicas estadísticas importantes para analizar un conjunto de datos:

En el primer ejercicio se trabajarán con contrastes de hipótesis e intervalos de confianza. Estas técnicas son útiles para evaluar si una afirmación sobre una población es verdadera o no. Los contrastes de hipótesis se utilizan para establecer dos hipótesis: la hipótesis nula y la hipótesis alternativa, y se utilizan datos de la muestra para decidir si se acepta o se rechaza la hipótesis nula. Los intervalos de confianza se utilizan para estimar el valor real de una población a partir de una muestra de datos, permitiendo calcular un rango de valores en el cual se espera que se encuentre el verdadero valor del parámetro poblacional con cierta probabilidad de acierto.

En el segundo ejercicio, se utilizará un modelo de regresión lineal simple, que es una técnica estadística que nos permite modelar la relación entre una variable dependiente y una variable independiente. En este caso, se trata de un modelo que asume que la relación entre las variables puede ser descrita mediante una línea recta. El modelo permite estimar la pendiente y la intersección de la recta que mejor se ajusta a los datos, lo que nos permite hacer predicciones y análisis sobre los datos.

En conjunto, estas tres técnicas pueden ser utilizadas para responder preguntas importantes acerca de un conjunto de datos, como, por ejemplo, si una cierta variable tiene una relación significativa con otra variable, o si hay diferencias significativas entre diferentes grupos de datos.

## Datos

Además de los datos indicados en la siguiente imagen, la media y varianzas poblacionales para las muestras 1 y 2 respectivamente son:

muestra 1:

muestra 2:

A continuación, se representan en un gráfico de dispersión creado en un jupyter notebook, y en formato tabla las dos muestras, en forma de DataFrame:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente**Figura1: datos muestrales su “scatter plot”**

Además de los datos mencionados se requiere de un nivel de significación del 10% (Alpha = 0,1), lo que significa que el nivel de confianza (1-alpha) será del 90 %.

Aunque en el ejercicio no se especifica, se trataran las variables de cada columna como numéricas y continuas. Mediante la librería pandas de Python se pueden sacar unas estadísticas básicas fácilmente:

Tabla

Descripción generada automáticamente

Se puede observar como para ambas columnas la media (mean) y la mediana (%50) son muy similares por lo que las muestras parecen ser simétricas como en una distribución normal, es decir, no parecen estar sesgadas ni positiva ni negativamente. Cabe mencionar, que los datos muestrales divergen de los poblacionales indicados anteriormente.

**Figura2: estadística básica de las muestras**

## Ejercicio 1

El presente ejercicio trata sobre contrastes de hipótesis e intervalos de confianza.

### Apartado 1

“Contrastar test de hipótesis (comparación de medias) con varianza poblacional conocida. (Los parámetros se indican con anterioridad al escoger las columnas. 𝜇,𝜎2) “

En el presente apartado se comparará la media de cada muestra con la de su respectiva población, las medias muestrales serán , donde el estimador muestral que se utilizará será la fórmula de la media:

Como se ha mencionado anteriormente las medias muestrales son inferiores a las medias poblacionales, de modo que se contrastarán dos hipótesis nulas con el nivel de confianza indicado; en la primera hipótesis alternativa se contrastará que la media muestral es diferente a la poblacional y en la segunda se contrastará que la hipótesis alternativa de la media muestral es inferior a la poblacional.

#### CASO 1: media muestral diferente a media poblacional; columna 1, análisis bilateral

En la siguiente figura se muestra la representación de un análisis bilateral

Icono

Descripción generada automáticamente

**Figura3: rango de aceptación de una hipótesis; contraste bilateral**

#### CASO 2: media muestral inferior a media poblacional; columna 2, análisis unilateral

En la siguiente figura se muestra una representación gráfica de un análisis unilateral:

Icono

Descripción generada automáticamente

**Figura4: rango de aceptación de una hipótesis; contraste unilateral**

### Apartado 2

“Contrastar test de hipótesis para comparación de dos medias (Ejemplo: datos de columna 1 y datos columna 2)”.

#### Caso 1: las poblaciones NO son iguales

población 1: , y población 2:

Se acepta la hipótesis nula, la diferencia entre las dos muestras es igual a 5

#### Caso 2: las poblaciones son iguales

En el presente apartado se considerará que ambas muestras pertenecen a la misma población para ello se usaran la media de los promedios poblacionales y la media de la desviaciones estándares poblacionales, donde:

Por lo tanto, la media de la muestra quedaría en 62,5 y la desviación en 123

Población equivalente:

En este caso acepta la hipótesis de que la diferencia entre las medias poblacionales de ambas muestras son 0.

### Apartado 3

“Usando el apartado (2.), encontrar un intervalo de confianza al 90% para las medias poblacionales”

En el presente apartado se encontrará el intervalo de confianza para las medias de las dos muestras, el intervalo de confianza para la media es el siguiente:

Para la muestra C1 el intervalo es el siguiente:

Para la muestra C2 el intervalo es el siguiente:

Por lo tanto, según lo calculado en el apartado 2, queda evidenciado de nuevo que ambas muestras pertenecen a la misma población.

### Apartado 4

“Encontrar un intervalo de confianza al 90% para la desviación típica poblacional asumiendo que esta es desconocida”.

Para el caso de un intervalo de confianza para una varianza poblacional siendo su media conocida

Para la muestra C1 sería:

Y pasando estos resultados a desviación típica:

Para la muestra C2, sería:

Y pasando estos resultados a desviación típica:

## Ejercicio 2

Redactar un problema donde se relacionen como variables las dos columnas asignadas anteriormente realizando la estimación de un modelo de regresión lineal simple.

### Apartado 1: Recta de regresión de C2 (Y) sobre C1 (x)

La ecuación para una recta de regresión lineal simple es la siguiente:

Siendo, en este caso el promedio de C2 y el promedio de C1. Para poner los puntos sobre las i-es se ha programado la siguiente tabla en juypter notebook con la librería pandas:

Tabla

Descripción generada automáticamente

**Figura5: Dataframe sobre los cálculos para la regresión lineal entre C1, y C2**

Se dispondrá del código de los cálculos realizados en el apartado de anexos, para los cuales se ha basado en los siguiente principios estadísticos:

Para este apartado me queda que:

Donde:

Entonces me queda que para este apartado a y b son:

Y la recta de regresión queda:

Graficando la ecuación sobre los datos:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

**Figura6: recta de regresión de C2 sobre C1, sobre el gráfico de dispersión**

El modelo, presenta una pendiente ligeramente positiva, lo cual indica que existe una correlación positiva entre las dos variables, sin embargo, se puede observar como el modelo de regresión se ajusta bastante mal a los datos, siendo el grado de variabilidad de estos, alto.

### Apartado 2: Recta de regresión de C1 (Y) sobre C2 (x)

En este apartado los cálculos son similares, pero intercambiando la variable dependiente por la variable independiente

Donde:

Entonces me queda que para este apartado a y b son:

Y la recta de regresión queda:

Graficando la ecuación sobre los datos:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

**Figura7: recta de regresión de C1 sobre C2, sobre el gráfico de dispersión**

La línea de regresión del presente modelo es muy similar a la anterior, de nuevo la pendiente de la recta es positiva, lo que indica que hay cierta correlación positiva entre los datos, es decir, si C1 sube C2 también se espera que suba, pero, al igual que en el anterior apartado la aproximación de la misma recta es pésima, presentando un grado de variabilidad muy alto.

Cabe mencionar que en esta línea de regresión tanto el valor interceptado a como la pendiente de la recta b son ligeramente mayores que en el apartado anterior, sin embargo, eso no indica mejor correlación, o mejor modelo.

### Apartado 3: coeficiente de correlación

A continuación, se calcula el coeficiente de correlación para los datos:

El coeficiente de correlación es de 0,35 lo cual indica una relación directa entre las dos variables como se indicaba en los apartados anteriores, pero con una asociación moderadamente baja.

# ANEXOS

En los presentes anexos se incluye el código ipython utilizado para crear las figuras de la memoria.

## Figura 1

df.plot(kind= 'scatter',

x = 'C1',

y = 'C2',

color = 'red')

plt.title('Nube de dispersion de los datos')

plt.xlabel('C1')

plt.ylabel('C2')

plt.show()

## Figura 2

df.describe()

## Figura 5

df['C1-C1\_mean'] = (df['C1'] - df['C1'].mean())\*\*2

df['C2-C2\_mean'] = (df['C2'] - df['C2'].mean())\*\*2

df['(C1-C1\_mean)(C2-C2\_mean)'] = (df['C1'] - df['C1'].mean())\*(df['C2'] - df['C2'].mean())

df

## Figura 6

plt.figure(figsize = (8,5))

plt.scatter(df['C1'], df['C2'])

plt.plot(df['C1'],

b\_c2\_sobre\_c1 \* df['C1'] + a\_c2\_sobre\_c1,

color = 'red',

)

plt.annotate(F'y = {b\_c2\_sobre\_c1:.2f} \* x + {a\_c2\_sobre\_c1:.2f}', xy = (80, 80))

plt.title('recta de regresion de C2 sobre C1')

plt.ylabel('c2')

plt.xlabel('c1')

plt.show()

## Figura 7

plt.figure(figsize = (8,5))

plt.scatter(df['C2'], df['C1'])

plt.plot(df['C2'],

b\_c1\_sobre\_c2 \* df['C2'] + a\_c1\_sobre\_c2,

color = 'red')

plt.annotate(F'y = {b\_c1\_sobre\_c2:.2f} \* x + {a\_c1\_sobre\_c2:.2f}', xy = (80, 80))

plt.title('recta de regresion de C1 sobre C2')

plt.ylabel('C1')

plt.xlabel('C2')

plt.show()