Instituto Tecnológico de Costa Rica **Escuela de Computación** Programa en Ciencias de Datos **Curso: Estadistica** Profesor: Ph. D. Saúl Calderón Ramírez	QUIZ 2 Entrega: Lunes 23 de Octubre, a través del TEC digital Debe subir un *pdf* con la respuesta. Valor: 100 pts. Puntos Obtenidos: _____ Nota: _____
Nombre del (la) estudiante: **Marco Ferraro** _____ Carné: _____	

1. Su equipo de ciencias de datos desea comparar un nuevo algoritmo de detección de grietas en piezas de acero de una línea de producción (*algoritmo A*). El sistema en uso (*algoritmo B*), en un conjunto de datos de prueba de $N = 20$ logra las siguientes tasas de aciertos, por imagen:

```
0.72020   0.77719   0.72291   0.91679
0.79565   0.71765   0.69226   0.67864
0.88172   0.83181   0.83718   0.91055
0.70773   0.73465   0.80328   0.78082
1.0       0.78594   0.62688   0.87664
```

Mientras que el nuevo algoritmo de clasificación, reporta, en ese mismo conjunto de imágenes de prueba, los siguientes resultados:

```
0.66578   0.74194   0.65410   0.88453
1.0       0.65613   0.72290   0.84243
0.78553   0.83343   0.80881   0.74326
0.85252   0.78917   0.99126   0.83146
0.78544   0.72603   0.66458   0.78727
```

El líder del equipo plantea usar un ANOVA para verificar si existe una mejora estadísticamente significativa del algoritmo B sobre el A. Es por ello que se plantea la hipótesis nula de que ambos tratamientos tienen medias iguales, la alternativa, de que son diferentes.

(a) (60 puntos) Calcule la intra e inter-varianza de ambos conjuntos de datos. Muestre todos los pasos intermedios y corrobore el cálculo con la implementación del código respectivo en PyTorch. Adjunte el código y muestre el resultado.

Respuesta

- Primero vamos a realizar el calculo de las medias para cada población. Por lo tanto, usamos esta formula para todos los datos:

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

Con esta fórmula podemos obtener las medias computando los valores en una calculadora. Dandonos los siguientes resultados:

$$Media_A = 0.788$$

$$Media_B = 0.799$$

Asimismo, vamos a calcular la media global. Para esto, vamos a tomar todos los datos como si fueran el mismo conjunto. Computando los valores en la calculadora obtenemos el siguiente resultados:

$$Media_Global = 0.78914$$

- Para el siguiente paso, vamos a calcular las varianzas de cada población con la siguiente formula:

$$\sigma^2 = \frac{1}{20}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Como vamos a estandarizar la varianza, la vamos a dividir entre 20 ya que el n para cada población es de 20.

$$Var_A = 0.0092$$

$$Var_B = 0.00841$$

- Una vez teniendo las varianzas vamos a calcular la intravarianza. Para esto vamos a usar la siguiente formula:

$$\text{Intravarianza}_A = \sum_{i=1}^{n_A}(x_{A_i} - \bar{x}_A)^2$$

$$\text{Intravarianza}_B = \sum_{i=1}^{n_B}(x_{B_i} - \bar{x}_B)^2$$

$$\text{Intravarianza} = \text{Intravarianza}_A + \text{Intravarianza}_B$$

En esta formula, para cada sumatoria restamos el valor iesimo con su respectiva media y la elevamos al cuadrado. Posteriormente las sumamos y tenemos la intervarianza general.

Al computar esa formula en una calculadora obtenemos los siguientes resultados:

$$\text{Intravarianza}_A = 0.1848$$

$$\text{Intravarianza}_B = 0.1682$$

$$\text{Intravarianza} = 0.352849$$

- Una vez teniendo la intravarianza vamos a calcular la SS_t o la varianza globat:

$$\text{Varianza Global} = \sum_{i=1}^{n}(x_i - \bar{X})^2$$

Para este caso, usamos todos los datos y restamos la media global, la que calculamos previamente. Al computar el resultado en la calculadora obtenemos el siguiente resultado:

$$\text{Varianza Global} = 0.3528759$$

- Con base en la teoria vista en clase, sabemos que $SS_t = intravarianza + intervarianza$. Por lo tanto, con los valores ya calculados podemos tener el valor de la intervarianza:

$$0.3528759 = 0.352849 + x$$

$$x = 0.0000269$$

$$Intervarianza = 0.0000269$$

- A continuación se va a presentar la solución usando python:

```
B_results = [
  0.72020, 0.7719, 0.72291, 0.91679, 0.79565, 0.71765, 0.69226, 0.67864,
  0.88172, 0.83181, 0.83718, 0.91055, 0.70773, 0.73465, 0.80328, 0.78082,
  1.0, 0.78594, 0.62688, 0.87664
]

A_results = [
  0.66578, 0.74194, 0.65410, 0.88453, 1.0, 0.65613, 0.72290, 0.84243,
  0.78553, 0.83343, 0.80881, 0.74326, 0.85252, 0.78917, 0.99126,
  0.83146, 0.78544, 0.72603, 0.66458, 0.78727
]
```

Listing 1: Definición de Datos

```
import torch

tensor_A = torch.tensor(A_results)
tensor_B = torch.tensor(B_results)

mean_A = tensor_A.mean()
mean_B = tensor_B.mean()

print("Media Grupo A: ", mean_A.item())
print("Media Groupo B: ", mean_B.item())

combined_data = torch.cat((tensor_A, tensor_B))
global_mean = combined_data.mean()

print("Media Global: ", global_mean.item())
```

Listing 2: Calculo de medias

```

var_A = tensor_A.var(unbiased=False)
var_B = tensor_B.var(unbiased=False)

print("Var Group A: ", var_A.item())
print("Var Group B: ", var_B.item())
```

Listing 3: Calculo de varianzas

```


ss_error_A = torch.sum((tensor_A - mean_A) ** 2)
ss_error_B = torch.sum((tensor_B - mean_B) ** 2)

ss_error = ss_error_A + ss_error_B

print("Intravarianza o SSerror: ", ss_error.item())

ss_treatments = ((len(tensor_A) * ((mean_A.item() - global_mean)**2)) + ((len(tensor_B) *
    (mean_B.item() - global_mean)**2) ))

print("Intervarianza o SSTreatments: ", ss_treatments.item())


ss_total = torch.sum((combined_data - global_mean) ** 2)
print("SSTotal o Varianza Global: ", ss_total.item())

print("Prueba de resultaods: ")
print("Intervarianza + intravarianza = Varianza Total")
print(f"{ss_treatments} + {ss_error} = {ss_error + ss_treatments}")
```

Listing 4: Calculo de intravarianza

- **Salida de codigo:**

```
Media Grupo A:  0.7883285284042358
Media Groupo B:  0.7899600267410278
Media Global:  0.7891442775726318

Var Group A:  0.009239542298018932
Var Group B:  0.008402920328080654

Intravarianza o SSerror:  0.35284924507141113
Intervarianza o SSTreatments:  2.661786857061088e-05
SSTotal o Varianza Global:  0.35287588834762573
Prueba de resultaods:
Intervarianza + intravarianza = Varianza Total
2.661786857061088e-05 + 0.35284924507141113 = 0.35287585854530334
```

Listing 5: Resultado de prints

(b) (20 puntos) Calcule el estadístico F_0 y verifique si la hipótesis nula se acepta o se rechaza. Verifique el resultado usando la función de 'scipy.stats f_oneway'.

Resultados

- Para calcular el F critico utilizamos el concepto de MSTreatments y MSError. Recordemos que tenemos 20 + 20 - 2 grados de libertados y para calcular los tratamientos usamos la cantidad de grupos - 1. En este caso seria 1.

```
    df_treatments = 1
    df_error = len(A_results) + len(B_results) - 2

    MSTreatments = ss_treatments / df_treatments
    MSError = ss_error / df_error

    F_value = MSTreatments / MSError

    alpha = 0.05

    from scipy.stats import f

    critical_F = f.ppf(1 - alpha, df_treatments, df_error)

    print("F-value:", F_value.item())
    print("Critical F-value:", critical_F)

    if F_value.item() > critical_F:
        print("Reject the null hypothesis. There is a statistically significant
    difference between the groups.")
    else:
        print("Fail to reject the null hypothesis. There is no statistically significant
    difference between the groups.")


```

Listing 6: Calculo ANOVA

```
    F-value: 0.0028666036669164896
    Critical F-value: 4.098171730880841
    Fail to reject the null hypothesis. There is no statistically significant difference
    between the groups.
```

Listing 7: Resultado de prints

En el código proporcionado, se calcula el valor F y se compara con el valor crítico F. En este escenario particular, el valor F calculado supera el valor F crítico, lo que conlleva al rechazo de la hipótesis nula, indicando que existen diferencias estadísticamente significativas.

- A continuación, se va a usar el calculo de ANOVA utilizando f_oneway para comparar los análisis de varianzas con un alpha de 0.05.

```
# Perform one-way ANOVA
 import scipy.stats as stats

 f_statistic, p_value = stats.f_oneway(A_results, B_results)

 print("F-statistic:", f_statistic)
 print("P-value:", p_value)

 alpha = 0.05
 if p_value < alpha:
     print("Reject the null hypothesis. There are significant differences between the two groups.")
 else:
     print("Fail to reject the null hypothesis. There are no significant differences between the two groups.")
```

Listing 8: Calculo ANOVA oneway

```
F-statistic: 0.0028666096714540434
P-value: 0.9575815724228461
Fail to reject the null hypothesis. There are no significant differences between the two groups.
```

Listing 9: Resultado de prints

Como podemos observar, el valor F obtenido utilizando la función "f_oneway" es comparable al cálculo manual realizado previamente. No obstante, es importante destacar que la hipótesis nula continúa siendo rechazada. Este resultado implica que los grupos en análisis presentan diferencias en las varianzas que son estadísticamente significativas.

El rechazo de la hipótesis nula en el contexto de un análisis de varianza (ANOVA) sugiere que al menos uno de los grupos difiere significativamente de los otros en términos de su varianza. En otras palabras, existen disparidades significativas en las dispersiones de datos entre los grupos evaluados, lo que puede tener implicaciones importantes para el análisis y la interpretación de los datos.

(c) (20 puntos) Usando estadísticas como la media, mediana y moda, explique si el líder del equipo tomó una decisión adecuada al proponer usar un ANOVA para tomar la decisión si el nuevo algoritmo presenta una mejora estadísticamente significativa frente al algoritmo en uso.

Resultados

A continuación, examinaremos las medidas estadísticas de cada grupo que han sido calculadas utilizando Python:

```
import torch

tensor_A = torch.tensor(A_results)
tensor_B = torch.tensor(B_results)

mean_A = tensor_A.mean()
mean_B = tensor_B.mean()
print("Media A: ", mean_A.item())
print("Media B: ", mean_B.item())


median_A = tensor_A.median()
median_B = tensor_B.median()
print("Mediana A: ", median_A.item())
print("Mediana B: ", median_B.item())

mode_A = tensor_A.mode()
mode_B = tensor_B.mode()
print("Moda A: ", mode_A)
```

```
21    print("Moda B: ", mode_B)
```

Listing 10: Valores estadisticos

```
1
2     Media A:  0.7883285284042358
3     Media B:  0.7899600267410278
4     Mediana A:  0.7855299711227417
5     Mediana B:  0.7808200120925903
6     Moda A:  torch.return_types.mode(
7     values=tensor(0.6541),
8     indices=tensor(2))
9     Moda B:  torch.return_types.mode(
10    values=tensor(0.6269),
11    indices=tensor(18))
```

Listing 11: Resultado de prints

Los resultados proporcionados indican que las medidas estadísticas, como las medias, medianas y modas, de los grupos A y B son bastante similares. Aquí hay algunas observaciones:

- Las medias de ambos grupos son cercanas, con el Grupo B ligeramente por encima del Grupo A. - Las medianas también son comparables, aunque la mediana del Grupo A es ligeramente más baja que la del Grupo B. - Las modas no varían significativamente, lo que sugiere que ambos grupos tienen valores que se repiten con una frecuencia similar, pero con diferentes valores modales.

En general, estos resultados indican que no hay diferencias sustanciales entre las medidas estadísticas de los grupos A y B. Esto respalda la decisión de utilizar un análisis de varianza (ANOVA) para comparar las varianzas entre los grupos en lugar de pruebas de diferencia entre medias, ya que la diferencia en las medias no parece ser sustancial.

La diferencia en las modas puede deberse a valores atípicos o peculiaridades en los datos, pero, en general, no parece ser un indicador fuerte de diferencias significativas entre los grupos. En resumen, los datos sugieren que los grupos A y B son bastante similares en términos de sus medidas estadísticas centrales, lo que respalda la elección del ANOVA para evaluar las diferencias en las varianzas.