# Tarea 0 Seminario ML

Guzmán Flores David
Sánchez Hernández Emanuel
Cuautencos Mora Jazmin Daniela
Flores Loredo Francisco

## Complejidad Computacional

Demuestre lo siguiente:

1. Si $f_1$ es $O(g_1)$ y $f_2$ es $O(g_2)$ entonces $(f_1 + f_2)$ es $O(max|g_1|, g_2|)$.

   **Dem.**

   Como $f_1$ es $O(g_1)$ entonces $\exists x_0, c_0 > 0$ talque $\forall x \geq x_0$ entonces $0 \leq |f_1(x)| \leq c_0|g_1(x)|$ y $f_2$ es $O(g_2)$ entonces $\exists x_1, c_1 > 0$ talque $\forall x \geq x_1$ entonces $0 \leq |f_2(x)| \leq c_1|g_2(x)|$

   Por demostrar que $(f_1 + f_2)$ es $O(max\{|g_1|, |g_2|\})$. Sea $x_3 = max\{x_0, x_1\}$ y $\frac{n}{2} = max\{c_0, c_1\}$ con $n > 0$

   Tenemos que $\forall x \geq x_3$ $0 \leq |f_1(x)+f_2(x)| \leq |f_1(x)|+|f_2(x)| \leq c_0|g_1(x)|+c_1|g_2(x)|$ la anterior desigualdad por propiedad de valor absoluto y porque $f_1$ es $O(g_1)$ y $f_2$ es $O(g_2)$

   Además $c_0|g_1(x)| + c_1|g_2(x)| \leq \frac{n}{2}|g_1(x)| + \frac{n}{2}|g_2(x)| \leq \frac{n}{2}max\{|g_1(x)|, |g_2(x)|\} + \frac{n}{2}max\{|g_1(x)|, |g_2(x)|\} = nmax\{|g_1(x)|, |g_2(x)|\}$

   Así tenemos que existen constantes positivas $x_3 = max\{x_0, x_1\}$ y $\frac{n}{2} = max\{c_0, c_1\}$ tales que $0 \leq |f_1(x) + f_2(x)| \leq nmax\{|g_1(x)|, |g_2(x)|\}$

   Por tanto $(f_1 + f_2)$ es $O(max\{|g_1|, |g_2|\})$

2. Si $f_1$ es $O(g_1)$ y $f_2$ es $O(g_2)$ entonces $(f_1 f_2)$ es $O(g_1 g_2)$.

   **Dem.**

   Como $f_1$ es $O(g_1)$ entonces $\exists x_0, c_0 > 0$ talque $\forall x \geq x_0$ entonces $0 \leq |f_1(x)| \leq c_0|g_1(x)|$ y $f_2$ es $O(g_2)$ entonces $\exists x_1, c_1 > 0$ talque $\forall x \geq x_1$ entonces $0 \leq |f_2(x)| \leq c_1|g_2(x)|$.
   Para $x_2 = max\{x_0, x_1\}$, $C = c_0 c_1$. Entonces $x > x_2$
   $\Rightarrow |f_1(x)f_2(x)| = |f_1(x)||f_2(x)| \leq c_0|g_1(x)|c_1|g_2(x)| = c_0c_1|g_1(x)||g_2(x)| = C|g_1(x)g_2(x)|$

3. ¿Cuál es la complejidad del siguiente algoritmo (O)?

```
for(i=0; i < N; i++){

 for(j=0; j< M; j++){

 print(i,j);
 }
}
```

   Sabemos que el tiempo de ejecución de un ciclo anidado es el tiempo de ejecución de la proposición multiplicado por el producto de los tamaños de todos los ciclos. En este caso tomanos como 1 la unidad de tiempo de ejecución de $print(i, j)$. Por tanto la complejidad de $print(i, j)$ es $O(1)$ y la complejidad del algoritmo completo es $O(nm)$, en el caso que n=m, entonces el algoritmo tiene complejidad $= O(n^2)$

# Estadistica Descriptiva

1. Demuestre que $-1 \leq r_{XY} \leq 1$ donde

$$r_{XX} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

   **Dem.**

   Sean X y Y variables aleatorias, sabemos que :

   $Cov(X, Y) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
   $Var(X) = \sum_{i=1}^{n}(x_i - \bar{x})^2$ y $Var(Y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$
   Así podemos reescribir como:
   $r_{XX} = \frac{Cov(X,Y)}{\sqrt{Var(x)Var(Y)}}$

Por la desigualdad de Cauchy Schwartz sabemos que $|Cov(X,Y)|^2 \leq Var(X)Var(Y)$

Entonces $|Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)}$

Dividiendo ambos terminos de la inecuación por $\sqrt{Var(X)Var(Y)}$ obtenemos que:

$$|\frac{Cov(X,Y)}{\sqrt{Var(x)Var(Y)}}| \leq \frac{\sqrt{Var(x)Var(Y)}}{\sqrt{Var(x)Var(Y)}} = 1.$$

Por lo tanto $-1 \leq r_{XX} \leq 1$

2. Demuestre que $|r_{XY}| = 1$ si y solo sí existe una relación lineal entre las variables X y Y, es decir , $y_i = \alpha + \beta x_i$ con $\beta \neq 0$

**Dem.**

$\Rightarrow$ Si X y Y son v.a tales que $Y_i = \alpha + \beta X_i$ con $\alpha \neq 0$ y $\beta$ constantes, entonces:

$$\rho(X,Y) = \frac{Cov(X,\alpha X+\beta)}{\sqrt{Var(X)Var(\alpha X+\beta)}} = \frac{a}{|a|}$$

Por lo tanto $\rho(X,Y) = 1$ cuando $a > 0$ y $\rho(X,Y) = -1$ cuando $a < 0$.

$\Leftarrow$

Supongamos que $|\rho(X,Y)| = 1$. Sean $U = \frac{X-\mu_X}{\sigma_X}$ y $V = \frac{Y-\mu_Y}{\sigma_Y}$.
Entonces $E(U) = E(V) = 0$ y $Var(U) = Var(V) = 1$, por lo tanto:

$\rho(U,V) = E(UV)$. Con esto tenemos que $|\rho(U,V)| = |\rho(X,Y)| = 1$. Así que si $\rho(U,V) = 1$ entonces:

$Var(U-V) = E[(U-V)^2] - E^2(U-V) = E[(U-V)^2] = 2[1-E(UV)] = 0$

Con esto tenemos que con probabilidad 1, la v.a U-V es constante ç", con probabilidad 1 U-V =c , esta constante es cero pues $E(U-V) = 0$, por lo tanto

$$\frac{X-\mu_x}{\sigma_x^2} = \frac{Y-\mu_Y}{\sigma_Y^2} \tag{2}$$

$\Rightarrow Y = \mu_Y + \frac{\sigma_Y}{\sigma_X}(X-\mu_X)$, esta es la relación lineal entre X y Y.

Si $\rho(U,V) = -1$ entonces:

$$\begin{aligned} Var(U+V) &= E[(U+V)^2] - E^2(U+V) \\ &= E[(U+V)^2] \\ &= 2[1+E(UV)] \\ &= 0 \qquad\qquad . \end{aligned}$$

De manera análoga obtenemos que:

$$\frac{X-\mu_X}{\sigma_X^2} = \frac{Y-\mu_Y}{\sigma_Y^2} \tag{3}$$

entonces $Y = \mu_Y + \frac{\sigma_Y}{\sigma_X}(X-\mu_X)$, esta es la relación lineal que buscabamos.

Con todo esto obtenemos que:

$$Y = (\rho(X,Y)\frac{\sigma_Y}{\sigma_X})X + (\mu_Y - \rho(X,Y)\mu_X\frac{\sigma_Y}{\sigma_X}) \tag{4}$$

donde $\alpha = \rho(X,Y)\frac{\sigma_Y}{\sigma_X}$ y $\beta = \mu_Y - \rho(X,Y)\mu_X\frac{\sigma_Y}{\sigma_X}$

3. Las calificaciones de 50 alumnos en Estadistica han sido las siguientes:

```
calif=[0, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5,↩
    5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6,↩
    6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, ↩
   10]
```

*a*) Calcular las medidas de tendencia central y de dispersión, ademas encuentre el Percentil 0.25 y 0.75 y elabore un gráfico de caja y brazos e histograma.

*b*) ¿Los distribución de estos datos es Sesgada a la Izquierda?

**Solucion**

```
import numpy as np # importando numpy
from scipy import stats # importando scipy.stats
import pandas as pd #
import matplotlib.pyplot as plt
```

Calculamos la medidas de tendencia central y dispersion

```
#Media aritmetica
np.mean(calif)
5.48
```

El promedio del grupo en Estadistica fue de 5.48

```
# mediana
np.median(calif)
6.0
```

entonces 6 es el valor central en el conjunto de datos ordenados

```
# Desviacion tipica
np.std(calif)
1.9923855048659633
```

Nuestra desviacion tipica es pequeña por lo que nos quiere decir que las calificaciones no estan muy aejadas al promedio.

```
# varianza
np.var(calif)
3.9696
```

Calculamos la moda

```
# moda
stats.mode(calif)
ModeResult(mode=array([6]), count=array([12]))
```

Notamos que la calificacion que mas se repitio fue de 6, y fueron 12 alumnos. Calculamos el percentil 0.25 y 0.75

```
np.percentile(calif, 25) # percentil al 25
4.25
```

Esto nos quiere decir que el 25 % de los alumnos del grupo obtuvo 4.25 o menos

```
np.percentile(calif, 75) # percentil 75
```

De la misma manera que el anterior, el 75 % de los alumnos del grupo de Estadistica obtuvo 7.0 o menos.

```
fig1, ax1 = plt.subplots()
ax1.set_title('Caja de brazos')
ax1.boxplot(calif)
```

Realizamos un diagrama de brazos Podemos notar en la grafica que la media esta

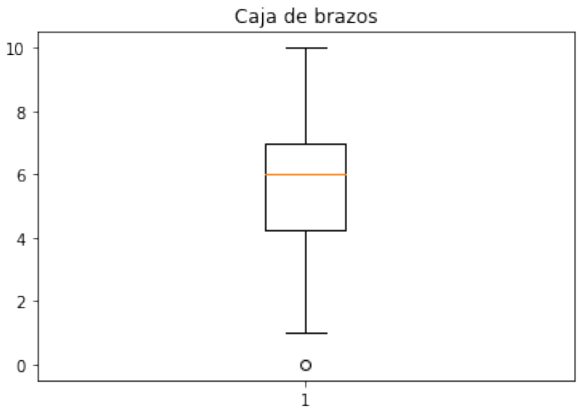


Figura 1: Caja de brazos

en 6, tambien el percentil al 75 % es 7 y al 25 % esta entre 4 y 5, tambien qu el calificacion maxima fue de 10 y la minima de 0.

```
# Histograma
plt.title('Histograma calificaciones')
plt.hist(calif, bins = 60, alpha=1, edgecolor = 'black'↩
    ,  linewidth=1)
plt.grid(True)
plt.show()
plt.clf()
```

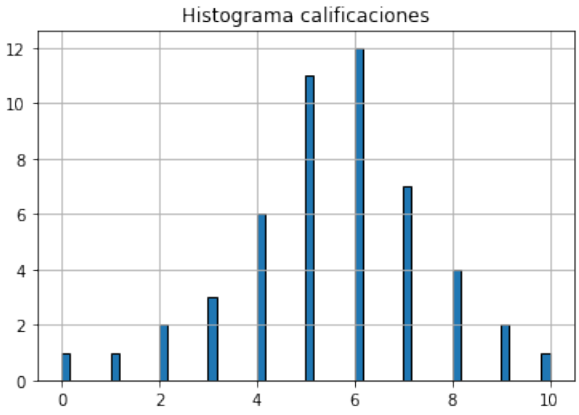


Figura 2: Histograma

¿La distribución de estos datos es Sesgada a la Izquierda? A simple vista se nota que los datos se concentran en el medio de la distribucion, de hecho se ve que siguen una distribucion normal, pero un poco mas del lado derecho, ya que las

calificaciones de 5 y 6 fueron las que mas hubo. Sacamos el coeficiente de asimetria para comprobarlo

```
    print("skew : ",skew(calif))
skew :  -0.342246184028854
```

Por lo que podemos ver la disitribucicon es sesgada a la derecha ya que el coeficiente de asimetria es de -0.03822461 .

4. Los 40 alumnos de una clase han obtenido las siguientes puntuaciones, sobre 50, en un examen de Física.

```
    califfisica=[48, 47, 44, 42, 41, 39, 39, 38, 38, 38, ↩
       37, 36, 36, 35, 35, 34, 34, 34, 33, 32, 32, 31, 29,↩
        28, 28, 27, 26, 25, 24, 23, 22, 20, 17, 15, 15, ↩
       13, 13, 11, 7, 3]
```

*a*) Calcular las medidas de tendencia central y de dispersión, ademas encuentre el Cuartil 1 y 3 así como los Deciles 1 y 9. Elabore un gráfico de caja y brazos, así como el gráfico de distribución empírica.
Primero cargamos las bibliotecas correspondientes

**Solucion**

```
        import numpy as np # importando numpy
        from scipy import stats # importando scipy.↩
           stats
        import pandas as pd #
        import matplotlib.pyplot as plt
        from scipy import stats
        import  matplotlib.pyplot  as plt
        import  numpy as np # importando  numpy
        from statistics import mode
        import matplotlib.pyplot as plt
        import pandas as pd
        import scipy.stats as ss
        import seaborn as sns
```

Calculamos las medidas de tendencia central

```
        #Media aritmetica de variable califfisica
        np.mean(califfisica)
        29.225
```

Tenemos entonces que 29 es la calificación representativa de la materia de fisica para una puntuacion máxima de 50, por lo que a los alumnos no les fue muy bien en el examen de física.

Calculamos la mediana de los datos.

```
# Mediana de la varibale califfisica
np.median(califfisica)
32.0
```

Tenemos que el valor de la variable de posicion central de las calificaciones ordenadas del examen de física es 32.

Obtenemos la moda de las calificaciones de la siguiente manera.

```
# Moda de la variable califfisica
mode(califfisica)
38
```

La moda es el valor de la calificacion que mas frecuencia tuvo, es decir que varios alumnos aprobaron el examen con una puntuacion de 38.

Procederemos a calcular las medidas de dispersión

Calculamos la varianza de los datos

```
#Varianza de califfisica
np.var(califfisica)
119.12437500000001
```

Podemos decir que dado que la varianza es muy grande entonces los datos estan muy dispersos respecto de la media de 29 puntos.

Y ahora obtenemos la desviaci'on estandar

```
#Calculamos la desviacion estandar ###
np.std(califfisica)
10.91441134463971
```

Con esta medida de dispersión corroboramos que la mayor parte de las calificaciones estan extendidas sobre un rango de valores mas amplio que su media.

Procedemos a calcular el primer y tercer cuartil

```
#Calculamos el cuartil 1 y 3
np.quantile(califfisica,[0.25,  0.75])
array([22.75, 37.25]
```

Tenemos entonces que la calificacion media entre el número más pequeño y la mediana del conjunto de calificacaiones es 22.75 o que el 25 % de los datos es menor a esa calificación y por otro lado el tercer cuartil que es la calificacion 37.25 es el valor medio entre la mediana y el valor más alto del conjunto calificaciones o bien que el 75 % de los datos es menor que ese valor de 37.25.

Por otra parte los deciles 1 y 9 son:

```
#Calculamos el decil 1 y 9
np.quantile(califfisica,[0.1,0.9])
array([13.  , 41.1])
```

Es decir que el 10 % de las calificaciones fue menor a 13 mientras que el noveno decil corresponde a la calificación de 41.1 entonces el 90 % de los alumnos de física tuvo una calificacion menor o igual a 41 por lo que muchos alumnos estuvieron lejos de alcanzar una calificación de diez en le examen.

Ahora graficamos el boxplot o diagrama de caja y brazos.

```
#Elaboramos un grafico de caja y brazos de ↩
   los datos
Grafb= plt.boxplot(califfisica)
plt.title('Caja de brazos')
plt.rcParams.update({'font.size': 15})
```

Podemos ver que la calificación máxima esta muy cerca de 10 y que el minímo esta muy cerca de cero, mientras que la media esta alrededor de 6, el percentil al 75 % esta alrededor de 7 y el percentil al 25 % casi es una calificación de 4, y además como la mayoria de los datos están por debajo de la media entonces la mayor parte del grupo reprobo el examén.

Ahora veamos la distribución empirica de los datos

```
fig, ax = plt.subplots()
sns.distplot(califfisica, bins=20, color="b↩
   ", ax = ax)
plt.show()
```

Como podemos ver la distribucion esta segada a la derecha, presenta valores

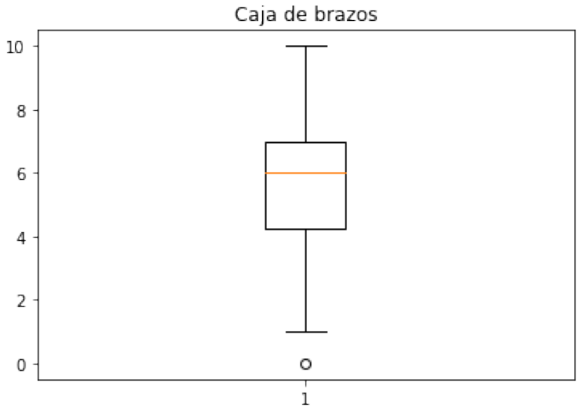


Figura 3: Caja de brazos

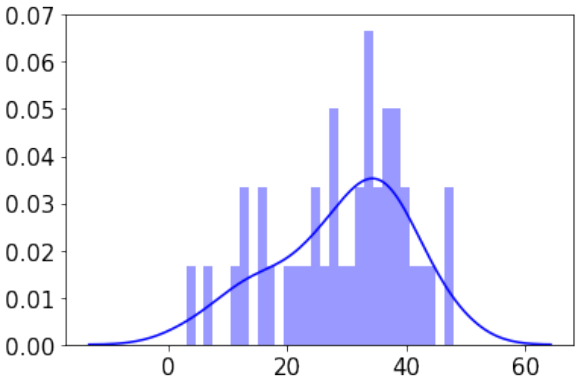


Figura 4: Caja de brazos

atipicos muy cerca de cero y tiene una función de distribución empirica que se asemeja a una normal.

*b*) ¿Los datos tiene una distribución platicurtica?

```
ss.kurtosis(califfisica)
-0.4482490835838018
```

No, porqué la curtosis no es igual a cero