

# ACTIVIDAD 2. Transformaciones e Inferencia Estadística

Instituto Tecnológico y de Estudios Superiores de Mty

Grace Aviance Silva Aróstegui A01285158

Concentración TC3006C:

Fecha, 23 de agosto del 2024.

Inteligencia Artificial Avanzada para Ciencia de Datos

Módulo: Estadística Prof: Ramiro Zermeño Díaz

Campus Guadalajara, Zapopan.

## Problema 1

Una pequeña empresa de manufactura estableció un sistema de incentivos para sus empleados basado en diferentes variables tanto de desempeño como de costo para la empresa. La empresa desea conocer cuál sería el ranking de los empleados tomando en cuenta todas las variables. A continuación, se presenta una tabla con los resultados obtenidos por cada empleado en cada uno de los rubros y si “más es mejor” o “menos es mejor”:

	Menos	Menos	Más	Más	Más	Menos
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 1	4620	354	10001	7	80014	5
Empleado 2	5100	499	9800	8	75000	6
Empleado 3	4550	450	9500	6	69000	4
Empleado 4	4751	470	9999	9	71000	3
Empleado 5	4848	380	9750	7	76500	2
Empleado 6	4932	370	9680	6	79814	5
Empleado 7	5040	330	9786	8	77658	4
Empleado 8	4671	350	9650	5	78500	2
Empleado 9	4699	415	10100	9	73000	2
Empleado 10	4914	394	10050	10	74000	3

Previamente, y con apoyo de la junta directiva, se aplicó la metodología AHP para definir los pesos de cada una de las variables y se obtuvieron los siguientes porcentajes:

	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Importancia	6%	3%	16%	25%	40%	10%

- (a) Haga un análisis exploratorio de estos datos:
- a) Calcular e interpretar estadísticas descriptivas de los datos: media, mediana, moda, desviación estándar, coeficiente de variación.

### Statistics

Variable	Mean	StDev	CoefVar	Median	Mode
Salario	4812.5	183.5	3.81	4799.5	*
Costo de Capacitación	401.2	56.0	13.97	387.0	*
Producción Generada	9831.6	197.8	2.01	9793.0	*
Satisfacción del Cliente Intern	7.500	1.581	21.08	7.500	6, 7, 8, 9
Ventas Generadas	75449	3725	4.94	75750	*
Ausentismo	3.600	1.430	39.72	3.500	2

b) ¿Cuál de las variables tiene mayor variabilidad? ¿Cuál tiene menor variabilidad? Explique, ¿cuáles estadísticas son relevantes para ello? y ¿por qué?

→ Mayor variabilidad: 'Ausentismo' y 'Satisfacción del cliente Interna'

→ Menor variabilidad: 'Produccion generada' y 'Salario'

→ Lo podemos apreciar por su coeficiente de variación. Es el único que no tiene unidades, es decir, es un porcentaje. Es la desviación estándar entre el promedio y multiplicado por 100. Entonces determina la dispersión del conjunto de datos respecto a su media.

(b) Suponga que se quiere utilizar los datos proporcionados y una regresión lineal para predecir cuáles serían las ventas generadas por 3 empleados nuevos con los siguientes valores:

Empleados Nuevos	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 11	4700	420	9800	8	?	3
Empleado 12	4900	450	9600	7	?	5
Empleado 13	4850	380	10000	8	?	4

Tip 1: Utilizar la transformación MinMax Scaler para las variables predictoras antes de realizar la regresión.

Tip 2: Transformar los datos de los nuevos empleados con los mismos parámetros de las variables originales para después meterlos en la ecuación de regresión.

→ Empleado 11: 71178.6497, Empleado 12: 72703.5438, Empleado 13: 78412.0961

```
# REGRESIÓN

# Escalamos las variables del dataset inicial
x = df[['Salario', 'Costo de Capacitación', 'Producción Generada', 'Satisfacción del Cliente Interna', 'Ausentismo']]
y = df['Ventas Generadas']
scaler = MinMaxScaler()
x_scaled = scaler.fit_transform(x)

# Escalamos las variables del dataset con los nuevos empleados
x_new = df_NewEmpleados[['Salario', 'Costo de Capacitación', 'Producción Generada', 'Satisfacción del Cliente Interna', 'Ausentismo']]
x_new_scaled = scaler.transform(x_new)

# Modelo de regresión
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(x_scaled, y)

# Predicción de Ventas Generadas para los Nuevos empleados
y_prediction = model.predict(x_new_scaled)
y_prediction

✓ 0.0s Python

array([71178.64979568, 72703.54387486, 78412.09611517])
```

- (c) Utilizando la Técnica de Análisis Multifactor, obtener cuál debería ser el ranking de cada uno de los empleados para poder definir el reparto de los incentivos.

→ El orden de los empleados de mejor a peor es: 9, 10, 5, 4, 7, 8, 1, 2, 6, 3.

*Solución:*

1 - Primeramente identificamos por cada columna cual es el valor más pequeño o grande (dependiendo de que es mejor para cada columna como se indica arriba de cada una de ellas con "Menos"ó "Más")

	Menos	Menos	Más	Más	Más	Menos
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 1	4620	354	10001	7	80014	5
Empleado 2	5100	499	9800	8	75000	6
Empleado 3	4550	450	9500	6	69000	4
Empleado 4	4751	470	9999	9	71000	3
Empleado 5	4848	380	9750	7	76500	2
Empleado 6	4932	370	9680	6	79814	5
Empleado 7	5040	330	9786	8	77658	4
Empleado 8	4671	350	9650	5	78500	2
Empleado 9	4699	415	10100	9	73000	2
Empleado 10	4914	394	10050	10	74000	3

2 - Vamos a dividir cada columna por el valor elegido como menor o mayor.

Si es valor mínimo, será divisor.

Si el valor es el máximo, será dividiendo.

	Menos	Menos	Más	Más	Más	Menos
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 1	4550/4620	330/354	10001/10100	7/10	80014/80014	2/5
Empleado 2	4550/5100	330/499	9800/10100	8/10	75000/80014	2/6
Empleado 3	4550/4550	330/450	9500/10100	6/10	69000/80014	2/4
Empleado 4	4550/4751	330/470	9999/10100	9/10	71000/80014	2/3
Empleado 5	4550/4848	330/380	9750/10100	7/10	76500/80014	2/2
Empleado 6	4550/4932	330/370	9680/10100	6/10	79814/80014	2/5
Empleado 7	4550/5040	330/330	9786/10100	8/10	77658/80014	2/4
Empleado 8	4550/4671	330/350	9650/10100	5/10	78500/80014	2/2
Empleado 9	4550/4699	330/415	10100/10100	9/10	73000/80014	2/2
Empleado 10	4550/4914	330/394	10050/10100	10/10	74000/80014	2/3

3 - Desplegamos los resultados de las divisiones. Además sumamos cada columna.

	Menos	Menos	Más	Más	Más	Menos
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 1	0.984848485	0.93220339	0.99019802	0.7	1	0.4
Empleado 2	8/9	0.661322645	0.97029703	0.8	0.937335966	0.333333333
Empleado 3	1	0.733333333	0.940594059	0.6	0.862349089	0.5
Empleado 4	0.957693117	0.70212766	0.99	0.9	0.887344715	0.666666667
Empleado 5	0.938531353	0.868421053	0.965346535	0.7	0.956082686	1
Empleado 6	0.922546634	0.891891892	0.958415842	0.6	0.997500437	0.4
Empleado 7	0.902777778	1	0.968910891	0.8	0.970555153	0.5
Empleado 8	0.974095483	0.942857143	0.955445545	0.5	0.981078311	1
Empleado 9	0.968291126	0.795180723	1	0.9	0.91234034	1
Empleado 10	0.925925926	0.837563452	0.995049505	1	0.924838153	0.666666667
Suma	9.466866764	8.36490129	9.734257426	7.5	9.429424851	6.466666667

4 - Los valores que anteriormente obtuvimos de la suma de cada columna lo utilizaremos para dividir cada valor de su respectiva columna. Observamos que si se hace correctamente, la suma de cada columna será 1

	Menos	Menos	Más	Más	Más	Menos
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 1	0.1040	0.1114	0.1017	0.1	0.1061	0.0619
Empleado 2	0.0942	0.0791	0.0997	0.1	0.0994	0.0515
Empleado 3	0.1056	0.0877	0.0966	0.1	0.0915	0.0773
Empleado 4	0.1012	0.0839	0.1017	0.1	0.0941	0.1031
Empleado 5	0.0991	0.1038	0.0992	0.1	0.1014	0.1546
Empleado 6	0.0975	0.1066	0.0985	0.1	0.1058	0.0619
Empleado 7	0.0954	0.1195	0.0995	0.1	0.1029	0.0773
Empleado 8	0.1029	0.1127	0.0982	0.1	0.1040	0.1546
Empleado 9	0.1023	0.0951	0.1027	0.1	0.0968	0.1546
Empleado 10	0.0978	0.1001	0.1022	0.1	0.0981	0.1031
Suma	1.0000	1.0000	1.0000	1.0	1.0000	1.0000

5 - Multiplicaremos cada columna por el respectivo peso de importancia que definieron en la junta directiva

Importancia	0.06	0.03	0.16	0.25	0.4	0.1
	Menos	Menos	Más	Más	Más	Menos
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 1	0.0062	0.0033	0.0163	0.023	0.0424	0.0062
Empleado 2	0.0057	0.0024	0.0159	0.027	0.0398	0.0052
Empleado 3	0.0063	0.0026	0.0155	0.020	0.0366	0.0077
Empleado 4	0.0061	0.0025	0.0163	0.030	0.0376	0.0103
Empleado 5	0.0059	0.0031	0.0159	0.023	0.0406	0.0155
Empleado 6	0.0058	0.0032	0.0158	0.020	0.0423	0.0062
Empleado 7	0.0057	0.0036	0.0159	0.027	0.0412	0.0077
Empleado 8	0.0062	0.0034	0.0157	0.017	0.0416	0.0155
Empleado 9	0.0061	0.0029	0.0164	0.030	0.0387	0.0155
Empleado 10	0.0059	0.0030	0.0164	0.033	0.0392	0.0103
Suma	0.0600	0.0300	0.1600	0.2500	0.4000	0.1000

6 - Resuelta las multiplicaciones, creamos la columna "Promedio Ponderado", en el que se suma cada renglón. Esta columna la filtramos para ordenarla de mayor a menor y así obtener el ranking de empleados

	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo	Promedio Ponderado
Empleado 9	0.0061	0.0029	0.0164	0.030	0.0387	0.0155	0.1096
Empleado 10	0.0059	0.0030	0.0164	0.033	0.0392	0.0103	0.1081
Empleado 5	0.0059	0.0031	0.0159	0.023	0.0406	0.0155	0.1043
Empleado 4	0.0061	0.0025	0.0163	0.030	0.0376	0.0103	0.1028
Empleado 7	0.0057	0.0036	0.0159	0.027	0.0412	0.0077	0.1008
Empleado 8	0.0062	0.0034	0.0157	0.017	0.0416	0.0155	0.0990
Empleado 1	0.0062	0.0033	0.0163	0.023	0.0424	0.0062	0.0978
Empleado 2	0.0057	0.0024	0.0159	0.027	0.0398	0.0052	0.0956
Empleado 6	0.0058	0.0032	0.0158	0.020	0.0423	0.0062	0.0933
Empleado 3	0.0063	0.0026	0.0155	0.020	0.0366	0.0077	0.0887

## Problema 2

En la elaboración de envases de plástico es necesario garantizar que cierto tipo de botella en posición vertical tenga una resistencia mínima de 20kg de fuerza. Para garantizar esto, se aplica fuerza a la botella hasta que ésta cede, y el equipo registra la resistencia que alcanzó la botella. Se obtuvieron los siguientes datos de la resistencia máxima alcanzada de cada botella mediante pruebas destructivas:

28.3	26.8	26.6	26.5	28.1	24.8	27.4	26.2	29.4	28.6	24.9	25.2	30.4	27.7	27.0	26.1	28.1
26.9	28.0	27.6	25.6	29.5	27.6	27.3	26.2	27.7	27.2	25.9	26.5	28.3	26.5	29.1	23.7	29.7
26.8	29.5	28.4	26.3	28.1	28.7	27.0	25.5	26.9	27.2	27.6	25.5	28.3	27.4	28.8	25.0	25.3
27.7	25.2	28.6	27.9	28.7												

(a) ¿Qué tipo de variable se está midiendo? ¿Discreta o continua? Explique.

→ Son variables Discretas ya que son listables. Si fueran continuas sería imposible listarlas todas.

(b) Haga un análisis exploratorio de estos datos.

a) Realice un histograma con al menos 2 reglas para definir el número de clases (No utilizar regla empírica). Describa la forma y analice el comportamiento de los datos.

→ La regla de Sturges es una fórmula utilizada para determinar el número óptimo de bins (intervalos) que se deben usar en un histograma. La fórmula es:

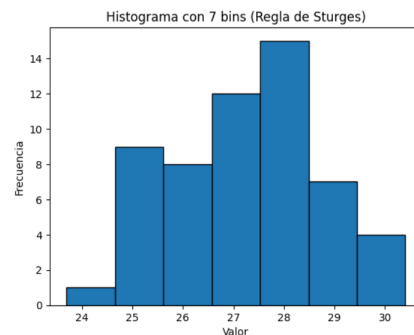
$$k = 1 + \log_2(n)$$

```
df2 = pd.read_csv('DF Problema2.csv')

# Aplicar la regla de Sturges
n = len(df2)
k = int(np.ceil(1 + np.log2(n))) # número de bins según la regla de Sturges

# Crear el histograma
plt.hist(df2, bins=k, edgecolor='black')
plt.title(f'Histograma con {k} bins (Regla de Sturges)')
plt.xlabel('Valor')
plt.ylabel('Frecuencia')
plt.show()
```

✓ 0.2s



→ La regla de Scott determina el ancho óptimo de cada bin en función de la desviación estándar de los datos y el tamaño de la muestra. La fórmula para calcular el ancho del bin según la regla de Scott es:

$$h = 3.49\sigma n^{-1/3}$$

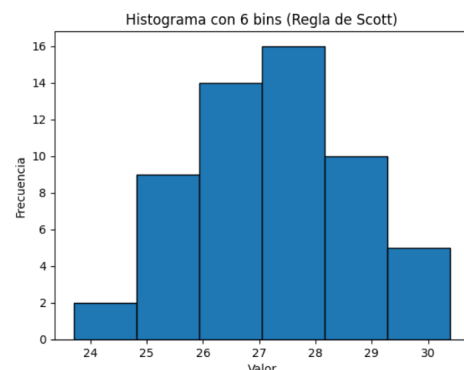
```
# Calcular la desviación estándar y el tamaño de la muestra
n = len(df2)
sigma = np.std(df2)

# Aplicar la regla de Scott
bin_width = 3.49 * sigma / (n ** (1/3))

# Calcular el número de bins
range_min, range_max = np.min(df2), np.max(df2)
num_bins = int(np.ceil((range_max - range_min) / bin_width))

# Crear el histograma
plt.hist(df2, bins=num_bins, edgecolor='black')
plt.title(f'Histograma con {num_bins} bins (Regla de Scott)')
plt.xlabel('Valor')
plt.ylabel('Frecuencia')
plt.show()
```

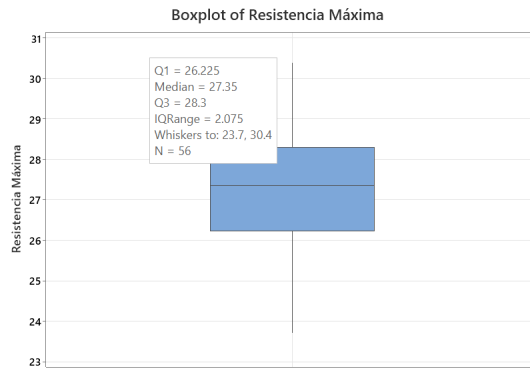
✓ 0.1s



Observamos que el histograma de Sturges no es de distribución normal y es asimétrico hacia la izquierda. Por otra parte en el de la regla de Scott con un intervalo menos se ve mucho mejor la distribución, más centrada pero no perfectamente normal.

- b) Realice un diagrama de caja y bigotes. Analice el comportamiento de los datos. ¿Existen datos atípicos? ¿Qué se debería hacer al respecto?

→ En el diagrama no se muestran datos atípicos. Por lo tanto no hay algo que se pueda hacer al respecto con ellos. Respecto al comportamiento de los datos se aprecia que tienen una distribución balanceada y tiene a ser simétrica, mas no perfectamente.



- (c) Estime, con una confianza de 94 %, ¿cuál sería la resistencia promedio de los envases?

→ El intervalo de confianza de la resistencia promedio es (26.879 , 27.614) y promedio 27.246

#### Descriptive Statistics

N	Mean	StDev	SE Mean	94% CI for $\mu$
56	27.246	1.430	0.191	(26.879, 27.614)

$\mu$ : population mean of Resistencia Máxima

- (d) Antes del estudio se suponía que la resistencia promedio era de 25kg. Dada la evidencia de los datos, ¿tal supuesto es correcto? ¿Qué tipo de prueba estadística se debe realizar? Plantee las hipótesis correspondientes y concluya adecuadamente.

→ El supuesto inicial es incorrecto ya que está fuera del intervalo de la resistencia promedio . El tipo de prueba estadística para corroborarlo fue la *prueba t* porque no conocemos la varianza de la población.

si se cumple el requisito minnimo

- (e) Con los datos anteriores estime, con una confianza del 98 %, ¿cuál es la desviación estándar poblacional (del proceso)?

→ El intervalo de confianza de la desviación estándar es (26.788 , 27.704)

#### Descriptive Statistics

N	Mean	StDev	SE Mean	98% CI for $\mu$
56	27.246	1.430	0.191	(26.788, 27.704)

$\mu$ : population mean of Resistencia Máxima

## Problema 3

En un laboratorio bajo condiciones controladas, se evaluó, para 10 hombres y 10 mujeres, la temperatura que cada persona encontró más confortable. Los resultados en grados Fahrenheit fueron los siguientes:

Mujer	75	77	78	79	77	73	78	79	78	80
Hombre	74	72	77	76	76	73	75	73	74	75

- (a) ¿Las muestras son dependientes o independientes? Explique.

→ Independientes, ya que la temperatura de uno no le afecta en absoluto al otro.

- (b) ¿La temperatura promedio más confortable es igual para hombre que para mujeres? ¿Qué tipo de prueba estadística se debe realizar? Plantee las hipótesis correspondientes y concluya adecuadamente.

→ El tipo de prueba estadística que debemos realizar es la *prueba t de 2 medias*.

Planteamiento de hipótesis:

$\mu_1$ : population mean of Mujer

$$H_0 : \mu_1 = \mu_2$$

$\mu_2$ :: population mean of Hombre

$$H_a : \mu_1 \neq \mu_2$$

### Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Mujer	10	77.40	2.07	0.65
Hombre	10	74.50	1.58	0.50

### Estimation for Difference

Difference	95% CI for Difference
2.900	(1.156, 4.644)

### Test

Null hypothesis  $H_0: \mu_1 - \mu_2 = 0$   
Alternative hypothesis  $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
3.53	16	0.003

∴ Dado que el valor  $P < \alpha$ , rechazamos  $H_0$ , es decir, no es igual la temperatura promedio de hombre y mujer.

- (c) Los datos poseen la misma variabilidad? ¿Qué tipo de prueba estadística se debe realizar? Plantee las hipótesis correspondientes y concluya adecuadamente.

→ El tipo de prueba estadística que se realizaría en este caso es la *prueba f de desviación estándar*.

Planteamiento de hipótesis:

$\sigma_1$ : standard deviation of Mujer

$$H_0 : \sigma_1 / \sigma_2 = 1$$

$\sigma_2$ :: standard deviation of Hombre

$$H_a : \sigma_1 / \sigma_2 \neq 1$$

Ratio :  $\sigma_1 / \sigma_2$

### Descriptive Statistics

Variable	N	StDev	Variance	95% CI for $\sigma$
Mujer	10	2.066	4.267	(1.064, 4.986)
Hombre	10	1.581	2.500	(1.120, 2.776)

### Ratio of Standard Deviations

Estimated Ratio	95% CI for Ratio using Bonett	95% CI for Ratio using Levene
1.30639	(0.401, 2.560)	(0.264, 2.308)

### Test

Null hypothesis  $H_0: \sigma_1 / \sigma_2 = 1$   
Alternative hypothesis  $H_1: \sigma_1 / \sigma_2 \neq 1$   
Significance level  $\alpha = 0.05$

Method	Statistic	DF1	DF2	P-Value
Bonett	0.39	1		0.530
Levene	0.03	1	18	0.860

∴ El radio es de 1.3 por lo que se puede concluir que prácticamente puede decirse que es 1. Por otra parte, los p value de ambos métodos son mayores a  $\alpha=0.05$ . Siendo así que no hay información suficiente para rechazar  $H_0$ , entonces se considera que ambas varianzas son iguales.

## Problema 4

La prueba actual de un solo disco se tarda 2 minutos. Se supone un nuevo método de prueba que consiste en medir solamente los radios 24 y 57, donde casi es seguro que estará el valor mínimo buscado. Si el método nuevo resulta igual de efectivo que el método actual se podrá reducir en 60 % el tiempo de prueba. Se plantea un experimento donde se mide la densidad mínima de metal en 18 discos usando tanto el método actual como el método nuevo. Los resultados están ordenados horizontalmente por disco. Así 1.88 y 1.87 es el resultado para el primer disco con ambos métodos.

Método Actual	1.88	1.84	1.83	1.90	2.19	1.89	2.27	2.03	1.96	1.98	2.00	1.92	1.83	1.94	1.94	1.95	1.93	2.01
Método Nuevo	1.87	1.90	1.85	1.88	2.18	1.87	2.23	1.97	2.00	1.98	1.99	1.89	1.78	1.92	2.02	2.00	1.95	2.05

(a) ¿Las muestras son dependientes o independientes? Explique.

→ Dependientes, porque se utilizan los mismos discos al mismo tiempo. De haber sido 36 discos los de prueba entonces sería independiente.

(b) ¿Qué tipo de prueba estadística se debe realizar? Plantee las hipótesis correspondientes y concluya adecuadamente.

→ Como prueba estadística de realizaría una *prueba t pareada*.

Planteamiento de hipótesis:

$\mu$  : population mean of (Método Actual - Método Nuevo)

$$H_0 : \mu = 0$$

$$H_a : \mu \neq 0$$

### Descriptive Statistics

Sample	N	Mean	StDev	SE Mean
Método Actual	18	1.9606	0.1150	0.0271
Método Nuevo	18	1.9628	0.1124	0.0265

### Estimation for Paired Difference

Mean	StDev	SE Mean	95% CI for $\mu$ difference
-0.00222	0.03949	0.00931	(-0.02186, 0.01742)

$\mu$  difference: population mean of (Método Actual - Método Nuevo)

### Test

Null hypothesis	$H_0: \mu_{\text{difference}} = 0$
Alternative hypothesis	$H_a: \mu_{\text{difference}} \neq 0$
<b>T-Value</b>	<b>P-Value</b>
-0.24	0.814

∴ Dado que el valor  $P > \alpha$ , rechazamos  $H_a$ , es decir, en promedio son iguales ambos métodos.

(c) ¿Recomienda la adopción del nuevo método? Argumente su respuesta.

→ No hay diferencia significativa entre el método actual y el nuevo. Cualquiera esta bien.