Transformaciones e Inferencia Estadística

Ejercicio 1:

1.a) Calcular e interpretar estadísticas descriptivas de los datos: media, mediana, moda, desviación estándar,coeficiente de variación

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Variance	CoefVar	Minimum	Q1
Salario	10	0	4812.5	58.0	183.5	33656.1	3.81	4550.0	4658.3
Costo de Capacitación	10	0	401.2	17.7	56.0	3140.4	13.97	330.0	353.0
Producción Generada	10	0	9831.6	62.5	197.8	39123.6	2.01	9500.0	9672.5
Satisfacción del Cliente Intern	10	0	7.500	0.500	1.581	2.500	21.08	5.000	6.000
Ventas Generadas	10	0	75449	1178	3725	13874148	4.94	69000	72500
Ausentismo	10	0	3.600	0.452	1.430	2.044	39.72	2.000	2.000

Variable	Median	Q3	Maximum	Mode	N for Mode
Salario	4799.5	4959.0	5100.0	*	0
Costo de Capacitación	387.0	455.0	499.0	*	0
Producción Generada	9793.0	10013.3	10100.0	*	0
Satisfacción del Cliente Intern	7.500	9.000	10.000	6, 7, 8, 9	2
Ventas Generadas	75750	78829	80014	*	0
Ausentismo	3.500	5.000	6.000	2	3

1.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
Menor variabilidad: Producción generada

A pesar de que la StDev y Variance nos dan una idea de la variabilidad, estos valores no son completamente compatibles para comparar con otras unidades, ventas en dólares con ausencias no son equivalentes. Por lo que a pesar de que son buenos indicadores para cada variable como tal, para comparar entre variables usé el coeficiente de varianza.

2) 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.

	Menos	Menos	Más	Más	Más	Menos	
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo	
min	4550	330	9500	5	69000	2	
max	5100	499	10100	10	80014	6	
Empleado 1	98.5%	93.2%	99.0%	70.0%	100.0%	40.0%	
Empleado 2	89.2%	66.1%	97.0%	80.0%	93.7%	33.3%	
Empleado 3	100.0%	73.3%	94.1%	60.0%	86.2%	50.0%	
Empleado 4	95.8%	70.2%	99.0%	90.0%	88.7%	66.7%	
Empleado 5	93.9%	86.8%	96.5%	70.0%	95.6%	100.0%	
Empleado 6	92.3%	89.2%	95.8%	60.0%	99.8%	40.0%	
Empleado 7	90.3%	100.0%	96.9%	80.0%	97.1%	50.0%	
Empleado 8	97.4%	94.3%	95.5%	50.0%	98.1%	100.0%	
Empleado 9	96.8%	79.5%	100.0%	90.0%	91.2%	100.0%	
Empleado 10	92.6%	83.8%	99.5%	100.0%	92.5%	66.7%	
							Total
Empleado 1	5.9%	2.8%	15.8%	17.5%	40.0%	4.0%	86.0%
Empleado 2	5.4%	2.0%	15.5%	20.0%	37.5%	3.3%	83.7%
Empleado 3	6.0%	2.2%	15.0%	15.0%	34.5%	5.0%	77.7%
Empleado 4	5.7%	2.1%	15.8%	22.5%	35.5%	6.7%	88.4%
Empleado 5	5.6%	2.6%	15.4%	17.5%	38.2%	10.0%	89.4%
Empleado 6	5.5%	2.7%	15.3%	15.0%	39.9%	4.0%	82.4%
Empleado 7	5.4%	3.0%	15.5%	20.0%	38.8%	5.0%	87.7%
Empleado 8	5.8%	2.8%	15.3%	12.5%	39.2%	10.0%	85.7%
Empleado 9	5.8%	2.4%	16.0%	22.5%	36.5%	10.0%	93.2%
Empleado 10	5.6%	2.5%	15.9%	25.0%	37.0%	6.7%	92.6%

c) 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:

Usando Excel:

- Pendiente -

Python con SciKitLearn:

```python
import scipy.stats as stats # type: ignore
import numpy as np # type: ignore
import pandas as pd # type: ignore
from sklearn.preprocessing import MinMaxScaler # type: ignore
from sklearn.linear_model import LinearRegression # type: ignore
from sklearn.model_selection import train_test_split # type: ignore

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


df = pd.DataFrame(empleados)

X = df.drop(columns=["Ventas Generadas"])
y = df["Ventas Generadas"]

scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=0)

model = LinearRegression()
model.fit(X_train, y_train)

predictions = model.predict(X_test)
print("Coefficients:", model.coef_)
print("Intercept:", model.intercept_)
print("Predictions Empleados:", predictions)

empleados_nuevos = {
    "Salario": [4700, 4900,4850],
    "Costo de Capacitación": [420, 450,380],
    "Producción Generada": [9800, 9600,10000],
    "Satisfacción del Cliente Interna": [8, 7,8],
    "Ausentismo": [3, 5, 4]
}

new_df = pd.DataFrame(empleados_nuevos)

new_X_scaled = scaler.transform(new_df)
new_predictions = model.predict(new_X_scaled)

print("Nuevos Empleados:", new_predictions)
```

```
Nuevos Empleados: [67178.42835048 67568.55102981 80945.08766012]
```

Ejercicio 2:

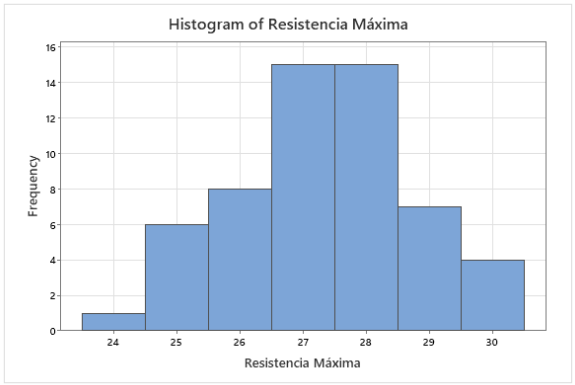
a) ¿Qué tipo de variable se está midiendo? ¿Discreta o continua? Explique.

"Una *variable continua* puede tomar un valor fijo dentro de un intervalo determinado" por lo que al tener valores entre el rango de 23.7 y 30.4 es cuantitativa y continua

b)

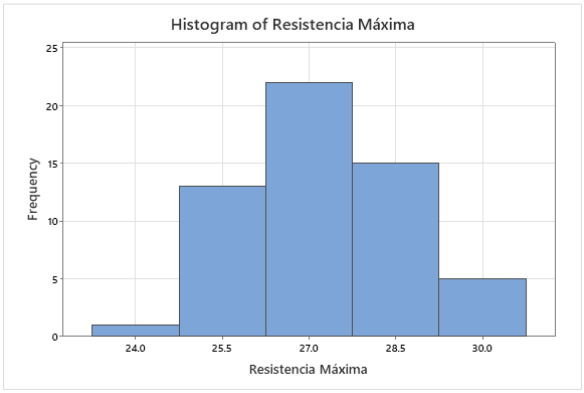
EJERCICIO 2

Histogram of Resistencia Máxima Sturges



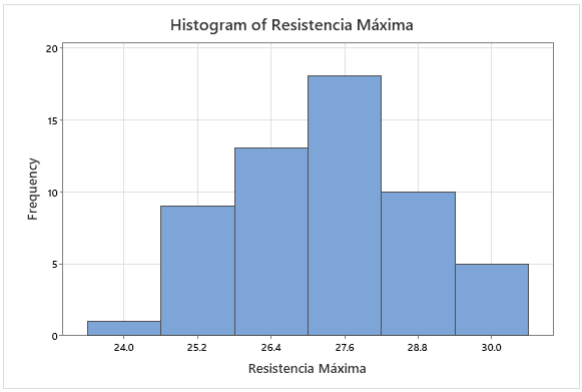
EJERCICIO 2

Histogram of Resistencia Máxima Scott



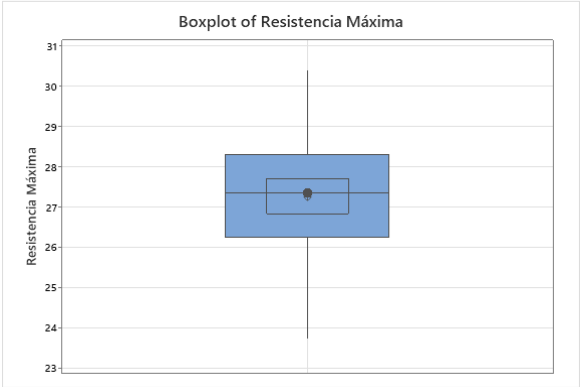
EJERCICIO 2

Histogram of Resistencia Máxima Diacosis



EJERCICIO 2

Boxplot of Resistencia Máxima



La media y mediana se encuentran prácticamente al centro de la caja, de la misma manera el minimo y maximo se encuentran alejados una distancia similar del centro. Con un máximo de 30.4 y un mínimo de 23.7, y asemejando una distribución normal como se puede ver en los histogramas.

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

Usando Minitab

EJERCICIO 2

One-Sample T: Resistencia Máxima

Descriptive Statistics

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

μ: population mean of Resistencia Máxima

Python

```python
import scipy.stats as stats # type: ignore
import numpy as np # type: ignore

mean = 27.246
sigma = 1.43
n=56
confidence_level = 0.94

z_critical = stats.norm.ppf((1+confidence_level)/2)

margin_of_error = z_critical * (sigma / np.sqrt(n))

lower_bound = mean - margin_of_error
upper_bound = mean + margin_of_error

print(f"Confidence Interval: ({lower_bound}, {upper_bound})")
```

```
Confidence Interval: (26.88659578662075, 27.60540421337925)
```

Promedio del intervalo: 27.15

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? Planteé las hipótesis correspondientes y concluya adecuadamente.

Test

Null hypothesis	H_0: $\mu = 25$
Alternative hypothesis	H_1: $\mu \neq 25$

T-Value	P-Value
11.75	0.000

Al realizar una prueba de 1 sample t con una confianza de 94%, un valor de T de 11.7 (alto) y un valor de P de 0, podemos concluir que el promedio no es de 25kg.

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

```python
import scipy.stats as stats # type: ignore
import numpy as np # type: ignore

resistencia_maxima = [
    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
]

n = len(resistencia_maxima)
s = np.std(resistencia_maxima, ddof=1)

#print(f"StDev: {s}") 1.430 - Correct!

confidence_level = 0.98

chi2_lower = stats.chi2.ppf((1 - confidence_level)/2,df=n-1)
chi2_upper = stats.chi2.ppf((1 + confidence_level)/2,df=n-1)

lower_val = (n-1)*s**2/chi2_upper
upper_val = (n-1)*s**2/chi2_lower

lower_std = np.sqrt(lower_val)
upper_std = np.sqrt(upper_val)

print(f"Confidence Interval StDev: {lower_std}, {upper_std}")
```

```
Confidence Interval StDev: 1.1694265785507387, 1.830936710513158
```

Ejercicio 3:

a) ¿Las muestras son dependientes o independientes? Explique. Como se evaluaron 2 grupos diferentes, sin una relación clara entre ambos, ni algún tipo de emparejamiento aparente, diría que son independientes.

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

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	IQR
Mujer	10	0	77.400	0.653	2.066	73.000	76.500	78.000	79.000	80.000	2.500
Hombre	10	0	74.500	0.500	1.581	72.000	73.000	74.500	76.000	77.000	3.000

Al calcular los promedios, junto con el conjunto de datos que ofrece minitab como estadísticas básicas, podemos ver que los hombres cuentan con un mínimo, máximo, mediana y promedio más alto que las mujeres.

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

Variance	CoefVar
4.267	2.67
2.500	2.12

Descriptive Statistics

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

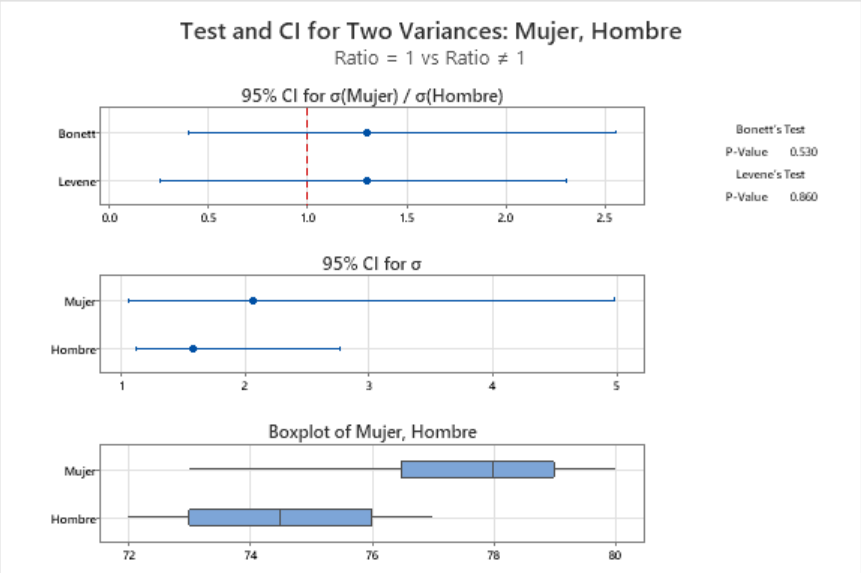
Ratio of Standard Deviations

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



La diferencia de 0.55 entre los coeficientes de varianza nos da una idea de la variabilidad. Tras realizar la prueba de Bonett y Levene, y obtener valores p de 0.530 y 0.860 (mucho mayores a 0.05) no pudimos determinar que las varianzas fueran significativamente diferentes.

Ejercicio 4:

a) ¿Las muestras son dependientes o independientes? Explique.
Dependientes porque hay un vínculo claro, y correlación directa/emparejamiento entre el método actual y el método nuevo.

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

Una prueba t pareada, para comparar ambas columnas, y su diferencia.

Paired T-Test and CI: Método Actual, Método Nuevo

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 μ_difference
-0.00222	0.03949	0.00931	(-0.02186, 0.01742)

μ_difference: population mean of (Método Actual - Método Nuevo)

Test

Null hypothesis	H_0: μ_difference = 0
Alternative hypothesis	H_1: μ_difference ≠ 0

T-Value	P-Value
-0.24	0.814

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

Con una diferencia de mínima en el promedio y en la desviación estándar, así como un valor p de 0.814 (mucho mayor a 0.05), podemos decir que el nuevo modelo es igual de efectivo que el actual.