Transformaciones e Inferencia Estadística

Resuelva los siguientes problemas de forma individual. Para cada problema, incluir la formulación de las hipótesis; gráficas y tablas necesarias; y la interpretación de los resultados. Puede utilizar Excel, Minitab o cualquier otro software o lenguaje (Python o R) como apoyo para su solución.

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.

WORKSHEET 2

Descriptive Statistics: Salario, Costo de Capacitación, Producción Generada

Statistics

Variable	Total Count	N	N*	Mean	SE Mean	StDev	Variance	CoefVar
Salario	10	10	0	4812.5	58.0	183.5	33656.1	3.81
Costo de Capacitación	10	10	0	401.2	17.7	56.0	3140.4	13.97
Producción Generada	10	10	0	9831.6	62.5	197.8	39123.6	2.01
Satisfacción del Cliente Intern	10	10	0	7.500	0.500	1.581	2.500	21.08
Ventas Generadas	10	10	0	75449	1178	3725	13874148	4.94
Ausentismo	10	10	0	3.600	0.452	1.430	2.044	39.72

Variable	Minimum	Q1	Median	Q3	Maximum	Mode	N for Mode
Salario	4550.0	4658.3	4799.5	4959.0	5100.0	*	0
Costo de Capacitación	330.0	353.0	387.0	455.0	499.0	*	0
Producción Generada	9500.0	9672.5	9793.0	10013.3	10100.0	*	0
Satisfacción del Cliente Intern	5.000	6.000	7.500	9.000	10.000	6, 7, 8, 9	2
Ventas Generadas	69000	72500	75750	78829	80014	*	0
Ausentismo	2.000	2.000	3.500	5.000	6.000	2	3

Variable	Skewness	Kurtosis
Salario	0.19	-1.16
Costo de Capacitación	0.58	-0.84

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

La variable que tiene el coeficiente de variación más grande, con un valor de 39.72, es la de ausentismo, por lo que es la que tiene mayor variabilidad y es la menos predecible. La que tiene el coeficiente de variación más pequeño, con un valor de 2.01, es la de producción generada, por lo tanto es la que tiene menor variabilidad y es más predecible.

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 ranking de cada uno de los empleados es:

d. Empleado 9
e. Empleado 10
f. Empleado 5
g. Empleado 4
h. Empleado 7
i. Empleado 8
j. Empleado 1
k. Empleado 2
l. Empleado 6
m. Empleado 3

	Menos	Menos	Más	Más	Más	Menos
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 1	0.9848485	0.9322034	0.990198	0.7	1	0.4
Empleado 2	0.8921569	0.6613226	0.970297	0.8	0.937336	0.3333333
Empleado 3	1	0.7333333	0.9405941	0.6	0.8623491	0.5
Empleado 4	0.9576931	0.7021277	0.99	0.9	0.8873447	0.6666667
Empleado 5	0.9385314	0.8684211	0.9653465	0.7	0.9560827	1
Empleado 6	0.9225466	0.8918919	0.9584158	0.6	0.9975004	0.4
Empleado 7	0.9027778	1	0.9689109	0.8	0.9705552	0.5
Empleado 8	0.9740955	0.9428571	0.9554455	0.5	0.9810783	1
Empleado 9	0.9682911	0.7951807	1	0.9	0.9123403	1
Empleado 10	0.9259259	0.8375635	0.9950495	1	0.9248382	0.6666667
	9.4668668	8.3649013	9.7342574	7.5	9.4294249	6.4666667

0.1040311	0.1114422	0.101723	0.0933333	0.106051	0.0618557
0.0942399	0.0790592	0.0996786	0.1066667	0.0994054	0.0515464
0.1056316	0.0876679	0.0966272	0.08	0.091453	0.0773196
0.1011626	0.0839374	0.1017027	0.12	0.0941038	0.1030928
0.0991385	0.1038173	0.09917	0.0933333	0.1013935	0.1546392
0.09745	0.1066231	0.098458	0.08	0.1057859	0.0618557
0.0953618	0.1195471	0.0995362	0.1066667	0.1029284	0.0773196
0.1028952	0.1127159	0.0981529	0.0666667	0.1040443	0.1546392
0.1022821	0.0950616	0.10273	0.12	0.0967546	0.1546392
0.097807	0.1001283	0.1022214	0.1333333	0.09808	0.1030928
1	1	1	1	1	1

	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo	PROMEDIO PONDERADO	
Empleado 1	0.00624187	0.00334327	0.01627568	0.02333333	0.0424204	0.00618557	0.01630002	
Empleado 2	0.0056544	0.00237178	0.01594857	0.02666667	0.03976217	0.00515464	0.01592637	
Empleado 3	0.00633789	0.00263004	0.01546035	0.02	0.0365812	0.00773196	0.01479024	
Empleado 4	0.00606976	0.00251812	0.01627243	0.03	0.03764152	0.01030928	0.01713518	
Empleado 5	0.00594831	0.00311452	0.0158672	0.02333333	0.04055741	0.01546392	0.01738078	
Empleado 6	0.005847	0.00319869	0.01575329	0.02	0.04231437	0.00618557	0.01554982	
Empleado 7	0.00572171	0.00358641	0.01592579	0.02666667	0.04117134	0.00773196	0.01680065	
Empleado 8	0.00617371	0.00338148	0.01570446	0.01666667	0.04161774	0.01546392	0.01650133	
Empleado 9	0.00613693	0.00285185	0.0164368	0.03	0.03870184	0.01546392	0.01826522	
Empleado 10	0.00586842	0.00300385	0.01635543	0.03333333	0.03923201	0.01030928	0.01801705	
							0.01826522	MAYOR

2. **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:**

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.

	Menos	Menos	Más	Más	Más	Menos							
	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo							
Empleado 1	0.9848485	0.9322034	0.990198	0.7	1	0.4		0.1040311	0.1114422	0.101723	0.0933333	0.106051	0.0618557
Empleado 2	0.8921569	0.6613226	0.970297	0.8	0.937336	0.3333333		0.0942399	0.0790592	0.0996786	0.1066667	0.0994054	0.0515464
Empleado 3	1	0.7333333	0.9405941	0.6	0.8623491	0.5		0.1056316	0.0876679	0.0966272	0.08	0.091453	0.0773196
Empleado 4	0.9576931	0.7021277	0.99	0.9	0.8873447	0.6666667		0.1011626	0.0839374	0.1017027	0.12	0.0941038	0.1030928
Empleado 5	0.9385314	0.8684211	0.9653465	0.7	0.9560827	1		0.0991385	0.1038173	0.09917	0.0933333	0.1013935	0.1546392
Empleado 6	0.9225466	0.8918919	0.9584158	0.6	0.9975004	0.4		0.09745	0.1066231	0.098458	0.08	0.1057859	0.0618557
Empleado 7	0.9027778	1	0.9689109	0.8	0.9705552	0.5		0.0953618	0.1195471	0.0995362	0.1066667	0.1029284	0.0773196
Empleado 8	0.9740955	0.9428571	0.9554455	0.5	0.9810783	1		0.1028952	0.1127159	0.0981529	0.0666667	0.1040443	0.1546392
Empleado 9	0.9682911	0.7951807	1	0.9	0.9123403	1		0.1022821	0.0950616	0.10273	0.12	0.0967546	0.1546392
Empleado 10	0.9259259	0.8375635	0.9950495	1	0.9248382	0.6666667		0.097807	0.1001283	0.1022214	0.1333333	0.09808	0.1030928
	9.4668668	8.3649013	9.7342574	7.5	9.4294249	6.4666667		1	1	1	1	1	1

jupyter 1.3 Last Checkpoint: 2 minutes ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

[1]:
```
import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
```

[3]:
```
data_existing = {
    '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],
    'Ausentismo': [5, 6, 4, 3, 2, 5, 2, 4, 2, 3],
    'Ventas Generadas': [80014, 75000, 69000, 71000, 76500, 79814, 77658, 78502, 73000, 74000]
}

df_existing = pd.DataFrame(data_existing)

data_new = {
    '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]
}

df_new = pd.DataFrame(data_new)

X_existing = df_existing[['Salario', 'Costo de Capacitación', 'Producción Generada', 'Satisfacción del Cliente Interna'
y_existing = df_existing['Ventas Generadas']

# Aplicar el escalado MinMax
scaler = MinMaxScaler()
X_existing_scaled = scaler.fit_transform(X_existing)

# Ajustar el modelo de regresión lineal
```

jupyter 1.3 Last Checkpoint: 3 minutes ago

File Edit View Run Kernel Settings Help Trusted

Code JupyterLab Python 3 (ipykernel)

```
    'Producción Generada': [9800, 9600, 10000],
    'Satisfacción del Cliente Interna': [8, 7, 8],
    'Ausentismo': [3, 5, 4]
}

df_new = pd.DataFrame(data_new)

X_existing = df_existing[['Salario', 'Costo de Capacitación', 'Producción Generada', 'Satisfacción del Cliente Interna'
y_existing = df_existing['Ventas Generadas']

# Aplicar el escalado MinMax
scaler = MinMaxScaler()
X_existing_scaled = scaler.fit_transform(X_existing)

# Ajustar el modelo de regresión lineal
model = LinearRegression()
model.fit(X_existing_scaled, y_existing)

X_new_scaled = scaler.transform(df_new)

predicted_sales = model.predict(X_new_scaled)

df_new['Ventas Generadas Predichas'] = predicted_sales
print(df_new)
```

```
   Salario  Costo de Capacitación  Producción Generada  \
0     4700                    420                 9800
1     4900                    450                 9600
2     4850                    380                10000

   Satisfacción del Cliente Interna  Ausentismo  Ventas Generadas Predichas
0                                 8           3                71324.110588
1                                 7           5                72924.511302
2                                 8           4                78410.725525
```

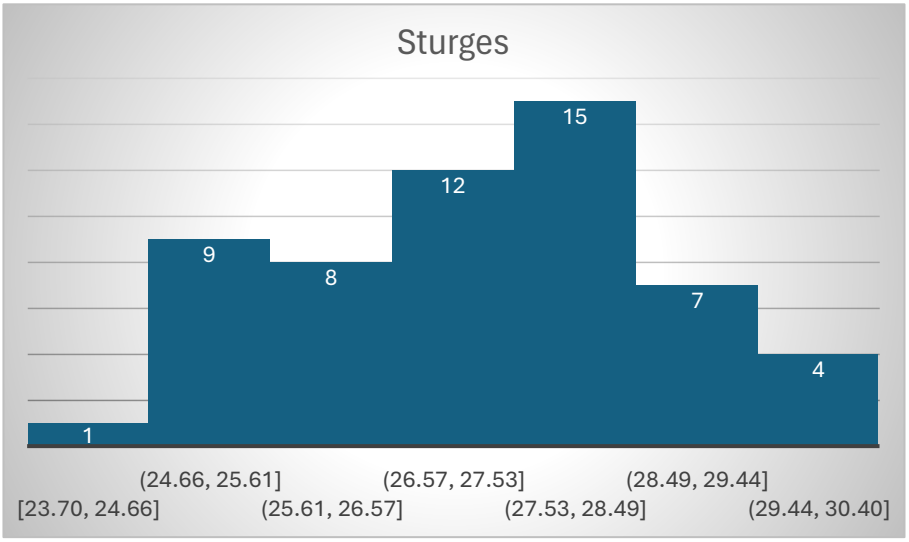
2.- En la elaboración de envases de plástico es necesario garantizar que cierto tipo de botella en posición vertical tenga una resistencia mínima de 20kg de fuerza. Para garantizar esto, se aplica fuerza a la botella hasta que ésta cede, y el equipo registra la resistencia que alcanzó la botella. Se 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												

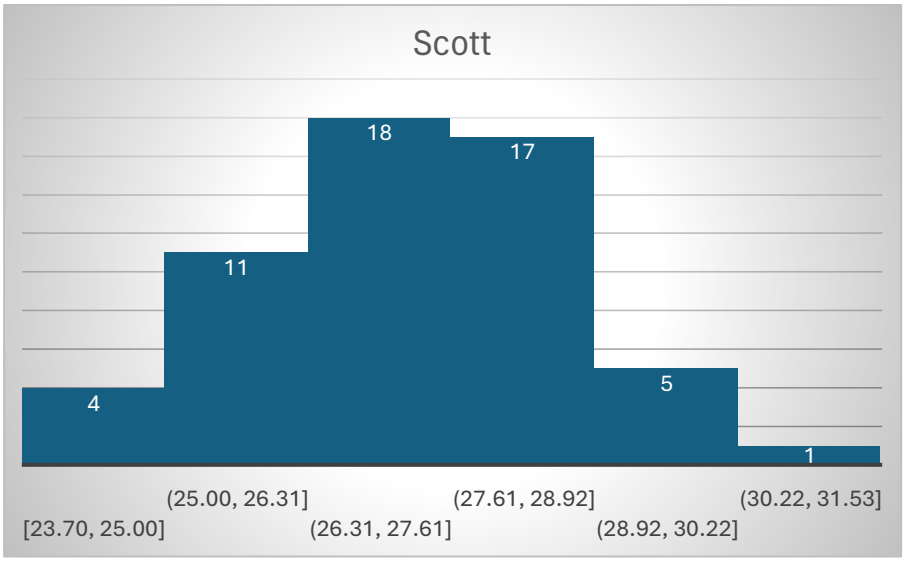
A. a) ¿Qué tipo de variable se está midiendo? ¿Discreta o continua? Explique.

Se está midiendo la variable de resistencia y es una variable continua, ya que se puede tomar cualquier valor dentro de un rango, no se limita a los valores enteros, sino que también se pueden incluir valores decimales, etc.

B. 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 distribución de los datos parece estar sesgada hacia la derecha (ligeramente), con la mayoría de los datos concentrados en los intervalos del centro. Hay un pico en el intervalo (27.53, 28.49) indicando que la mayoría de las botellas tienen esa resistencia máxima.



La distribución está hacia la derecha. El intervalo con la mayor concentración es (26.31,27.61) por lo que la mayoría de las botellas tienen esa resistencia máxima.

En los dos histogramas los datos tienen a concentrarse en los intervalos centrales, lo que significa que la mayoría de las botellas tienen una resistencia máxima alrededor de esos valores.

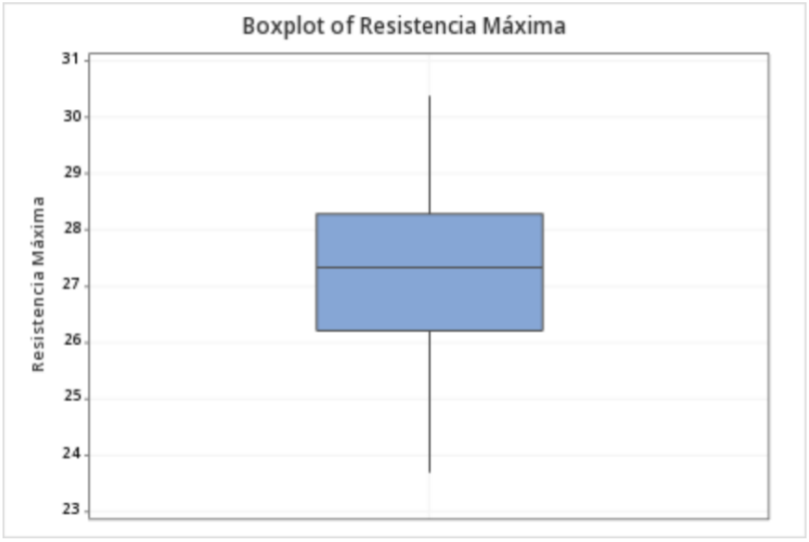
En el de Sturges, la mayor concentración en un intervalo del centro indica una menor dispersión de los datos en comparación al de Scott.

En el de Scott, los intervalos más chicos permiten observar mejor la variabilidad y dispersión, mostrando que hay más datos dispersos hacia la derecha e izquierda.

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

WORKSHEET 6

Boxplot of Resistencia Máxima



En esta gráfica de boxplot se puede ver que no contiene datos atípicos (outliers). La caja parece estar bastante centrada entre los bigotes, lo que significa que la distribución de los datos es aproximadamente simétrica, lo que indica que los valores no están sesgados hacia ningún extremo.

Se podría confirmar la consistencia de las medidas para asegurar que las mediciones son precisas, sin embargo, el que la distribución sea simétrica puede suponer que los datos son fiables.

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

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

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: μ = 25
Alternative hypothesis	H_1: μ ≠ 25

T-Value	P-Value
11.75	0.000

Se rechaza la hipotesis porque el P value es menor a 0.05.

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

One-Sample T: Resistencia Máxima

Descriptive Statistics

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

μ: population mean of Resistencia Máxima

La desviación estandar sigue siendo la misma. Sin embargo, la media con este nivel de confianza va de 26.788 a 27.704.

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.

Son independientes, ya que se evalúan por separado y las respuestas de uno no dependen de los datos o respuestas del otro y viceversa.

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.

WORKSHEET 7

Two-Sample T-Test and CI: Mujer, Hombre

Method

μ_1: population mean of Mujer
μ_2: population mean of Hombre
Difference: $\mu_1 - \mu_2$

Equal variances are not assumed for this analysis.

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

Se debe realizar una prueba de 2 sample t.

Puesto que el P value es menor a 0.05, se rechaza la hipótesis nula.

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

Correlations

	Mujer
Hombre	0.374

Hay una correlación positiva leve entre las dos variables, esto nos dice que hay una tendencia leve a que, a medida que una variable aumenta, la otra también, pero la relación no es alta o fuerte.

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

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

   Las muestras son dependientes, cada disco es evaluado con el método actual y el nuevo, lo que significa que los resultados de cada disco bajo los dos métodos están parejos. Esto crea relación entre las mediciones del mismo disco, lo que hace que las muestras no sean independientes.

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

   Se debe realizar dos prueba t para las muestras independientes.

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
-0.06	33	0.954

   Como el P value es mayor a 0.05, no se rechaza la hipótesis nula.

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

   Sí se recomienda adoptar el nuevo método porque como las medias son iguales, el método actual entonces se podrá reducir un 60% el tiempo de prueba.