Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Guadalajara

Inteligencia artificial avanzada para la ciencia de datos I

M1.3 Transformaciones e Inferencia Estadística

Paola Félix Torres | A00227869

23/08/2024

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.

Variable	N*	Mean	StDev	Variance	Median	Mode	N for Mode	CV
Salario	0	4812.5	183.5	33656.1	4799.5	*	0	3.812987
Costo de Capacitación	0	401.2	56.0	3140.4	387.0	*	0	13.958126
Producción Generada	0	9831.6	197.8	39123.6	9793.0	*	0	2.0118801
Satisfacción del Cliente Intern	0	7.500	1.581	2.500	7.500	6, 7, 8, 9	2	21.08
Ventas Generadas	0	75449	3725	13874148	75750	*	0	4.9371098
Ausentismo	0	3.600	1.430	2.044	3.500	2	3	39.722222

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

La variable con mayor variabilidad es el ausentismo, con un un coeficiente de variación de 39.72%. Mientras que la variable con menor variabilidad es la de producción generada con un coeficiente de variación de 2.01%. Las estadísticas más relevantes para evaluar la variabilidad son la desviación estándar y la media. La desviación estándar mide la dispersión de los datos, mientras que la media proporciona el valor promedio. Juntas, permiten calcular el coeficiente de variación, que estandariza la desviación estándar en relación con la media, permitiendo así obtener una medida de la variabilidad relativa entre las variables.

b) 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.

Basado en los puntajes globales ponderados, el ranking de los empleados de mejor a peor desempeño sería:

1. Empleado 9
2. Empleado 10
3. Empleado 5
4. Empleado 4
5. Empleado 7
6. Empleado 8
7. Empleado 1
8. Empleado 2
9. Empleado 6
10. Empleado 3

	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo	PROMEDIO	RANKING
Empleado 1	0.006241866	0.003343267	0.016275682	0.023333333	0.042420403	0.006185567	0.01630002	7
Empleado 2	0.005654396	0.002371777	0.015948574	0.026666667	0.039762169	0.005154639	0.01592637	8
Empleado 3	0.006337894	0.002630037	0.015460352	0.02	0.036581196	0.007731959	0.01479024	10
Empleado 4	0.006069758	0.002518121	0.016272428	0.03	0.03764152	0.010309278	0.017135184	4
Empleado 5	0.005948312	0.003114518	0.015867204	0.023333333	0.040557413	0.015463918	0.017380783	3
Empleado 6	0.005847003	0.003198694	0.015753285	0.02	0.04231437	0.006185567	0.01554982	9
Empleado 7	0.00572171	0.003586414	0.01592579	0.026666667	0.04117134	0.007731959	0.016800647	5
Empleado 8	0.006173714	0.003381476	0.015704463	0.016666667	0.041617737	0.015463918	0.016501329	6
Empleado 9	0.006136927	0.002851847	0.016436796	0.03	0.038701845	0.015463918	0.018265222	1
Empleado 10	0.005868421	0.003003849	0.016355425	0.033333333	0.039232007	0.010309278	0.018017052	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.

Empleados Nuevos	Salario	Costo de Capacitación	Producción Generada	Satisfacción del Cliente Interna	Ventas Generadas	Ausentismo
Empleado 11	4700	420	9800	8	71182.34101	3
Empleado 12	4900	450	9600	7	72638.35164	5
Empleado 13	4850	380	10000	8	78245.11363	4

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												

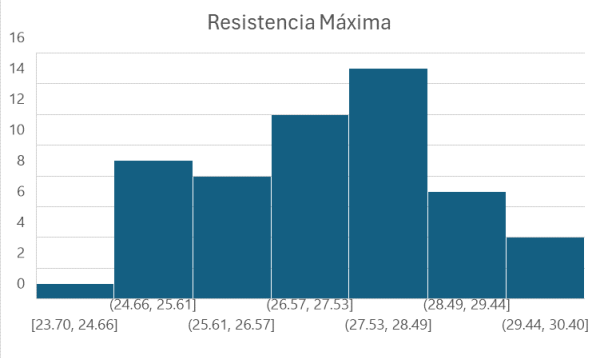
a) ¿Qué tipo de variable se está midiendo? ¿Discreta o continua? Explique.

Se está midiendo una variable continua ya que la resistencia es una medida que mayormente tiene valores decimales y puede ser medida con alta precisión.

b) Haga un análisis exploratorio de estos datos.

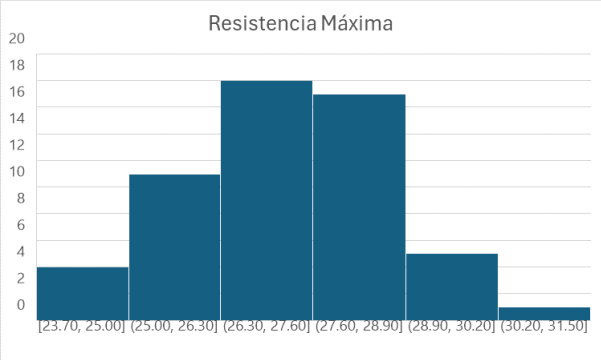
a. Realice un histograma con al menos 2 reglas para definir el número de clases (No utilizar regla empírica). Describa la forma y analice el comportamiento de los datos.

Regla de Sturges (k=7)



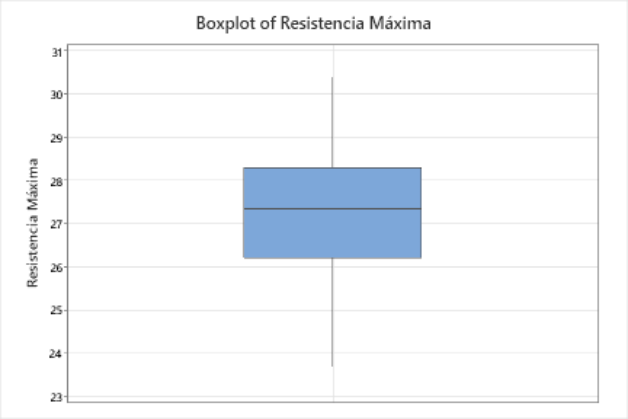
El histograma realizado con la regla de Sturges muestra una distribución asimétrica hacia la derecha. Existe una concentración de datos en el centro, lo que significa que hay un valor común, sin embargo, la asimetría a la derecha indica que hay algunos valores más altos.

Regla de Scott (h=1.30)



El histograma realizado con la regla de Scott muestra una distribución mucho más simétrica que la anterior, con la mayoría de los valores concentrados en el centro.

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



El diagrama de caja y bigotes muestra la mediana muy cerca del rango intercuartílico, lo que significa que los datos están distribuidos de manera uniforme alrededor del centro. En este caso, no se observan datos atípicos, pero si los hubiera, sería importante investigar si esos datos son precisos o si se deben a errores en la medición o recolección. Si se confirma que son errores, deberían eliminarse, ya que afectarían de gran manera la precisión del análisis de datos.

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

La resistencia promedio de los envases es 27.246 kg, asimismo estamos un 94% seguros que la media se encuentra entre 26.879 kg y 27.614 kg.

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.

Tal suposición sobre que la resistencia promedio es de 25 kg es incorrecta, esto lo podemos comprobar realizando una prueba t planteando como hipótesis nula que la resistencia promedio es igual a 25 kg, esto nos dio como resultado un p-value de 0.000, por lo tanto se rechaza la hipótesis nula y podemos afirmar que la resistencia promedio es significativamente diferente de 25 kg.

Test

Null hypothesis	H_0: μ = 25
Alternative hypothesis	H_1: μ ≠ 25

T-Value	P-Value
11.75	0.000

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

Con una confianza de 98%, se estima que la desviación estándar poblacional se encuentra entre 1.18 kg y 1.83 kg según el método de Bonett, y entre 1.17 kg y 1.83 kg según el método Chi-Cuadrado.

Method

σ: standard deviation of Resistencia Máxima
The Bonett method is valid for any continuous distribution.
The chi-square method is valid only for the normal distribution.

Descriptive Statistics

N	StDev	Variance	98% CI for σ using Bonett	98% CI for σ using Chi-Square
56	1.43	2.05	(1.18, 1.80)	(1.17, 1.83)

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.

Aunque las muestras se tomaron de hombres y mujeres sin relación entre ellos, la correlación de 0.374 sugiere cierta dependencia. Sin embargo, debido al tamaño pequeño de la muestra, se puede asumir que las muestras son **independientes** para este análisis.

Correlations

	Mujer
Hombre	0.374

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.

Los datos muestran que la temperatura promedio más confortable es diferente entre hombres y mujeres, mientras que el promedio en mujeres es de 77.4, el promedio de hombres es de 74.5, para confirmar esta hipótesis se debe de realizar una prueba t para dos muestras, dado que la hipótesis nula será que la temperatura promedio entre hombres y mujeres es igual, una vez realizada la prueba obtenemos que el p-value es de 0.003, por lo que se rechaza la hipótesis nula y concluimos que en efecto, **la temperatura promedio más confortable es significativamente diferente para hombres y mujeres.**

Descriptive Statistics

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

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

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

A simple vista los datos **NO poseen la misma variabilidad** ya que se calculó que la varianza en mujeres es de 4.2, mientras que la de los hombres fue de 2.5, ambos datos muy distintos. La prueba estadística para comprobar la hipótesis de que los datos no poseen la misma variabilidad es una prueba de igualdad de varianzas, una vez realizada la prueba se obtuvo como resultado que los valores p de las pruebas de Bonett y Levene son altos (0.530 y 0.860, respectivamente), indicando que no hay evidencia suficiente para rechazar la hipótesis nula de igualdad de varianzas, a pesar de que las varianzas parezcan muy distintas.

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

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.

Se podría asumir que las muestras son dependientes, ya que se están comparando dos conjuntos de mediciones provenientes del mismo grupo de discos. La correlación obtenida fue de 0.940, lo que indica una correlación muy fuerte y sugiere que los resultados son **dependientes**.

Correlations

	Método Actual
Método Nuevo	0.940

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

Dado que los datos son dependientes, se debe realizar una prueba t para muestras pareadas. La hipótesis planteada es que la diferencia promedio entre las mediciones es igual a 0. Al poner a prueba esta hipótesis, el p-value obtenido fue de 0.814, lo que significa que no se rechaza la hipótesis nula. Por lo tanto, el nuevo método es igual de efectivo que el método actual en términos de precisión.

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.

Sí recomendaría adoptar el nuevo método, ya que los resultados del método actual y el nuevo método son similares y esta similaridad permitiría reducir el tiempo de prueba en un 60%.