Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Guadalajara

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

M1.1 Actividad Distribuciones

Samuel Padilla Esqueda | A01641383

Agosto 23

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":

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	N for Mode
Salario	4812,5	183,5	3,81	4799,5	*	0
Costo de Capacitación	401,2	56,0	13,97	387,0	*	0
Producción Generada	9831,6	197,8	2,01	9793,0	*	0
Satisfacción del Cliente Intern	7,500	1,581	21,08	7,500	6; 7; 8; 9	2
Ventas Generadas	75449	3725	4,94	75750	*	0
Ausentismo	3,600	1,430	39,72	3,500	2	3

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 la mayor **desviación estándar** y mayor **coeficiente de variación** tendrá la mayor variabilidad. Mientras que la variable con la menor desviación estándar y menor coeficiente de variación tendrá la menor variabilidad.

Esto es así debido a que la **desviación estándar** es la medida directa de la variabilidad de los datos, entre más grande sea más grande es la variación.

El **coeficiente de variación** es importante porque ayuda a comparar la variabilidad ente variables de diferentes unidades, representa variabilidad respecto a la media.

Sabiendo esto establecemos que la variable con mayor variabilidad es:

- Ausentismo

La variable con menor variabilidad es:

- Producción Generada

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.

Antes de realizar el análisis multifactor primero debemos normalizar los datos, ya que no es posible llevar a cabo en análisis si los datos se encuentran en diferentes unidades.

Para normalizar utilizamos estas fórmulas dependiendo si un valor mas alto es mejor o si un valor más bajo es mejor:

$$n_i = \frac{X_{ij}}{\max(X_j)} \qquad n_i = \frac{\min(X_j)}{X_{ij}}$$

Tras normalizar aplicamos el analisis multifactor utilizando los niveles de importancia determinados para cada variable, obtenemos el siguente ranking:

promedio ponderado	Ranking	Puntaje
1,63000197	Empleado 9	1,8265222
1,59263703	Empleado 10	1,80170523
1,47902397	Empleado 5	1,73807829
1,71351841	Empleado 4	1,71351841
1,73807829	Empleado 7	1,68006467
1,55498199	Empleado 8	1,65013291
1,68006467	Empleado 1	1,63000197
1,65013291	Empleado 2	1,59263703
1,8265222	Empleado 6	1,55498199
1,80170523	Empleado 3	1,47902397

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:

La predicción que obtenemos es la siguiente

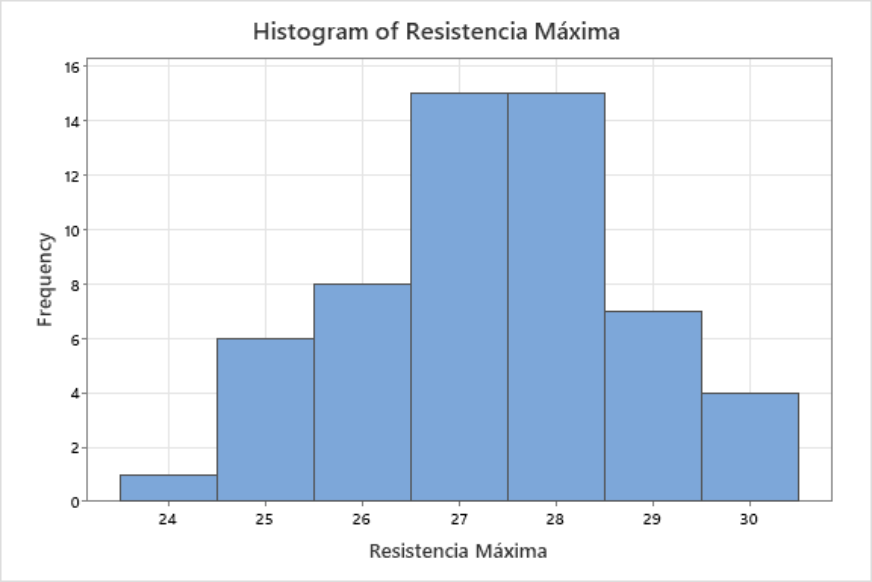
VENTAS GENERADAS PREDICCION
71178,6
72703,5

78412,0

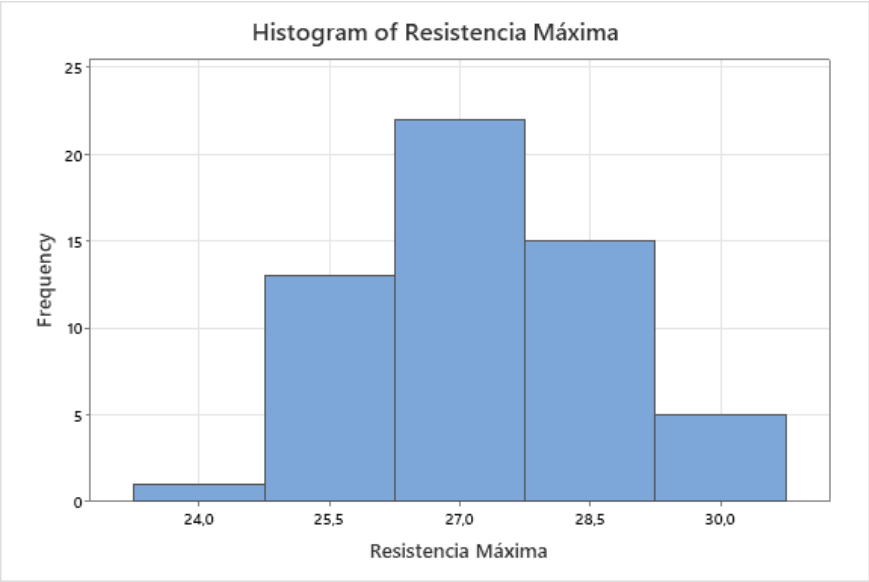
2.- En la elaboración de envases de plástico es necesario garantizar que cierto tipo de botella en posición vertical tenga una resistencia mínima de 20kg de fuerza. Para garantizar esto, se aplica fuerza a la botella hasta que ésta cede, y el equipo registra la resistencia que alcanzó la botella. Se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

a) **¿Qué tipo de variable se está midiendo? ¿Discreta o continua? Explique.**
   Es una variable continua, ya que puede tomar valores numéricos infinitos dentro de un rango determinado, nos damos cuenta de esto ya que las mediciones usan decimales.
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



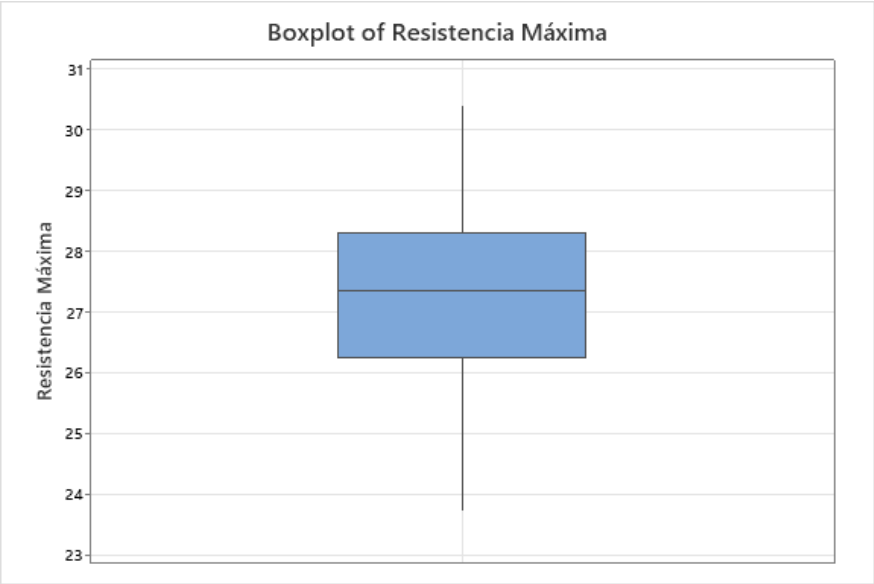
Regla de Scott



Ambos histogramas tienen la forma aproximada de campana de Gauss, esto indica que los datos siguen una distribución normal o cercana a la normal.

Esto nos indica que la mayoría de los valores se concentran cerca de la media.

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



Observando la grafica vemos que no hay puntos alejados de la caja, esto significa que no hay valores atípicos y por lo tanto no hay nada que hacer al respecto.

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

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

El intervalo de confianza del 94% indica que, con un 94% de confianza, la verdadera resistencia promedio de los envases está entre 26.879 kgf y 27.614 kgf. Sin embargo, la mejor estimación puntual de la resistencia promedio es 27.246 kg.

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.

Dado la evidencia obtenida en el análisis anterior, vemos que el supuesto era incorrecta y que la resistencia promedio es mayor a lo que se pensaba: 27.246 kg

Para determinar si el supuesto de que la resistencia promedio es de 25 kg es correcto, debes realizar una prueba de hipótesis (una prueba t).

Test

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

T-Value	P-Value
11,75	0,000

Observamos que el P-values es tan bajo que se muestra como 0, esto significa que la evidencia contra la hipótesis es extremadamente fuerte y podemos rechazarla con gran confianza. (Ya demostramos en el paso anterior que la hipótesis no era cierta).

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

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

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:

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

Ya que las respuestas obtenidas vienen de grupos de personas diferentes, y no se toman datos de las mismas personas en diferentes momentos, u otras condiciones parecidas, podemos decir que las muestras 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.

Para saber si ambos promedios son iguales debemos hacer una prueba t de 2 variables. Tenemos las siguientes hipótesis y resultados de la prueba:

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

Ya que el P value tiene un valor menor a 0.05 rechazamos la hipótesis nula, esto significa que tenemos evidencia para concluir que hay una diferencia importante entre los promedios de las temperaturas confortables entre hombres y mujeres.

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

Par saber si los datos tienen la misma variabilidad hacemos una prueba F para la igualdad de varianzas.

Tenemos las siguientes hipótesis y resultados de la prueba:

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

Ya que ambos Valores p de los modelos usados por minitab son > 0.05: No se rechaza la hipótesis nula. Esto indica que no hay suficiente evidencia para afirmar que las varianzas de las temperaturas confortables entre hombres y mujeres son diferentes.

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.

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

Las muestras son dependientes ya que están directamente relacionadas, se están utilizando los mismos discos para ambos métodos.

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

Ya que las variables son dependientes, hay que usar la prueba t para muestras pareadas.

Tenemos las siguientes hipótesis y resultados de la prueba:

Test

Null hypothesis	$H_0: \mu\_difference = 0$
Alternative hypothesis	$H_1: \mu\_difference \neq 0$

T-Value	P-Value
-0,24	0,814

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)

Tomando en cuenta los resultados, el valor p es mucho mayor que 0.05, esto indica que no hay suficiente evidencia para afirmar que existe una diferencia significativa en las medias de las mediciones entre el método actual y el nuevo método.

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

Si recomiendo el nuevo método, ya que los resultados del análisis indican que las diferencias entre las mediciones de los métodos son prácticamente insignificantes, esto quiere decir que el nuevo modelo es igual de efectivo que el modelo original, pero tiene la ventaja de reducir en 60% el tiempo de prueba.