Universidad Nacional Autónoma de México

Facultad de Estudios Superiores Acatlán

Estadística 2

Tarea unidad 1

Autor:

Jorge Miguel Alvarado Reyes - 421010301
Samuel Eduardo Mariche Wajsfeld - 421040001
Jiménez Pineda Leydi Monserrat - 421089037
Monroy Alarcon Omar Ulises - 421098277

23 de marzo de 2024

Índice

1. Problema 1

En algunas pruebas de salud en ancianos, un nuevo medicamento ha restaurado su memoria casi como la de jóvenes. Pronto se probará en pacientes con enfermedad de Alzheimer, esa fatal enfermedad del cerebro que destruye la mente. Según el Dr. Gary Lynch, de la Universidad de California en Irvine, el medicamento, llamado ampakina CX-516, acelera señales entre células cerebrales que parecen agudizar significativamente la memoria. 2En una prueba preliminar en estudiantes de poco más de 20 años y en hombres de entre 65 y 70 años de edad, los resultados fueron particularmente sorprendentes. Después de recibir dosis moderadas de este medicamento, las personas de entre 65 y 70 años de edad calificaron casi tan alto como los jóvenes. Los datos siguientes son los números de sílabas sin sentido recordadas después de 5 minutos, para 10 hombres de poco más de 20 años de edad y 10 señores de entre 65 y 70 años. Determina si las distribuciones para el número de sílabas sin sentido recordadas son iguales para estos dos grupos.

20s	3 6 4 8 7 1 1 2 7 8
65-70s	1 0 4 1 2 5 0 2 2 3

- $H_0 : F_x = F_y$ No hay diferencia entre las distribuciones de las cantidades de sílabas sin sentido recordadas por los dos grupos de edad.
- $H_1 : F_x \neq F_y$ Existe una diferencia entre las distribuciones de las cantidades de sílabas sin sentido recordadas por los dos grupos de edad.

Ordenación de los Datos

$$\{0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 6, 7, 7, 8, 8\}$$

Cálculo de Rangos

- 0: 1.5
- 1: 4.5
- 2: 8.5
- 3: 11.5
- 4: 13.5
- 5: 15.0
- 6: 16.0
- 7: 17.5
- 8: 19.5

Suma de rangos para cada grupo

$$\text{20s: } 11{,}5 + 16 + 13{,}5 + 19{,}5 + 17{,}5 + 4{,}5 + 4{,}5 + 8{,}5 + 17{,}5 + 19{,}5 = 132{,}5$$

$$\text{65-70s: } 4{,}5 + 1{,}5 + 13{,}5 + 4{,}5 + 8{,}5 + 15 + 1{,}5 + 8{,}5 + 8{,}5 + 11{,}5 = 77{,}5$$

Estadisticos U

$$u_1 = 10 \cdot 10 + \frac{10(10+1)}{2} - 132{,}5 = 22{,}5$$
$$u_2 = 10 \cdot 10 + \frac{10(10+1)}{2} - 77{,}5 = 77{,}5$$

$$min(u_1, u_2) = 22{,}5$$

Calculando la **esperanza** tenemos:

$$E(u) = \frac{10 \cdot 10}{2} = 50$$

Calculando la **varianza** tenemos:

$$Var(u) = \frac{10 \cdot 10(10+10+1)}{12} = 175$$

Calculando **z** tenemos:

$$z = \frac{u - E(u)}{\sqrt{Var(u)}} = \frac{22{,}5 - 50}{\sqrt{175}} = -2{,}0788$$

$$|z| = 2{,}0788$$

Realizando la prueba $|z| > q_Z\left(1 - \frac{\alpha}{2}\right)$ obtenemos que.

$$q_Z(0{,}975) = 1{,}959964$$

$$2{,}0788 > 1{,}959964$$

De igual manera podemos hacer la prueba calculando el p-valor correspondiente.

$$\begin{aligned} \text{p-valor} &= 2P(Z > 2{,}0788) \\ &= 2\left[1 - P(Z < 2{,}0788)\right] \\ &= 2\left[0{,}01881\right] \\ &= 0{,}03763 \end{aligned}$$

Sabemos que se rechaza H_0 (Se acepta H_1) si se cumple

$$\text{p-valor} < \alpha$$

En este caso la condicion

$$0{,}03763 < 0{,}05$$

se cumple asi que concluimos que se rechaza H_0 asi que las distribuciones son diferentes

2. Problema 2

El tratamiento de cáncer por medios químicos, llamado quimioterapia, mata células cancerosas y células normales. En algunos casos, la toxicidad del medicamento para el cáncer, es decir, su efecto sobre células normales, puede reducirse con la inyección simultánea de un segundo medicamento. Se realizó un estudio para determinar si la inyección de un medicamento en particular reducía los efectos dañinos de un tratamiento de quimioterapia en el tiempo de sobrevivencia de ratas. Dos grupos de 12 ratas seleccionados al azar se emplearon en un experimento en el que ambos grupos, llamémoslos A y B, recibieron la droga tóxica en una dosis lo suficientemente grande para causarles la muerte, pero, además, el grupo B recibió la antitoxina que iba a reducir el efecto tóxico de la quimioterapia en células normales. La prueba finalizó al término de 20 días, o sea, 480 horas. Los tiempos de sobrevivencia para los dos grupos de ratas, a las 4 horas más cercanas, se muestran en la tabla siguiente. ¿Los datos dan suficiente evidencia para indicar que las ratas que recibieron la antitoxina tienden a sobrevivir más después de la quimioterapia que las que no recibieron la antitoxina?

Sólo quimioterapia	Quimioterapia más droga
84	140
128	184
168	368
92	96
184	480
92	188
76	480
104	244
72	440
180	380
144	480
120	196

Ordenación de los Datos

$$\{72, 76, 84, 92, 92, 96, 104, 120, 128, 140, 144, 168, 180, 184, 184, 188, 196, 244, 368, 380, 440, 480, 480, 480\}$$

Cálculo de Rangos

- 72: 1
- 76: 2
- 84: 3
- 92: 4.5
- 96: 6
- 104: 7
- 120: 8
- 128: 9
- 140: 10
- 144: 11

- 168: 12
- 180: 13
- 184: 14.5
- 188: 16
- 196: 17
- 244: 18
- 368: 19
- 380: 20
- 440: 21
- 480: 23

Suma de rangos para cada grupo

Quimioterapia: $3{,}0 + 9{,}0 + 12{,}0 + 4{,}5 + 14{,}5 + 4{,}5 + 2{,}0 + 7{,}0 + 1{,}0 + 13{,}0 + 11{,}0 + 8{,}0 = 89{,}5$

Quimioterapia + droga: $10{,}0+14{,}5+19{,}0+6{,}0+23{,}0+16{,}0+23{,}0+18{,}0+21{,}0+20{,}0+23{,}0+17{,}0 = 210{,}5$

Estadisticos U

$$U_1 = 12 * 12 + \frac{12(12+1)}{2} - 89{,}9 = 132{,}5$$
$$U_2 = 12 * 12 + \frac{12(12+1)}{2} - 210{,}5 = 11{,}5$$

$$U = \text{mín}(U_1, U_2) = 11{,}5$$

Esperanza, varianza y z

$$E(u) = \frac{12 * 12}{2} = 72$$
$$Var(u) = \frac{12 * 12(12+12+1)}{12} = 300$$
$$z = \frac{11{,}5 - 72}{\sqrt{300}} = -3{,}4929$$
$$|z| = 3{,}4929$$

conclusiónes

Se rechaza la hipótesis nula H_0 con un nivel de significancia α, si el valor absoluto de Z es mayor que el cuantil crítico $q_z(1-\frac{\alpha}{2})$ de la distribución normal estándar:

$$|Z| > q_z\left(1-\frac{\alpha}{2}\right)$$

$$q_Z(0{,}975) = 1{,}959964$$

$$3{,}4929 > 1{,}959964$$

Se rechaza H0, se acepta H1. Las distribuciones son diferentes.
De igual manera podemos hacer la prueba calculando el p-valor correspondiente. Si p-valor $< \alpha$, se rechaza H_0

$$\begin{aligned} \text{p-valor} &= 2P(Z > 3{,}4929) \\ &= 2\left[1 - P(Z < 3{,}4929)\right] \\ &= 0{,}00047 \end{aligned}$$

$$0{,}00047 < 0{,}05$$

Por lo tanto se confirma que se rechaza H0 y se acepta H1

3. Problema 3

Dos chefs, A y B, calificaron 22 comidas en una escala del 1 al 10. Los datos se muestran en la tabla. ¿Los datos dan suficiente evidencia para indicar que uno de los chefs tiende a dar calificaciones más altas que el otro?

Comida	A	B	Comida	A	B
1	6	8	12	8	5
2	4	5	13	4	2
3	7	4	14	3	3
4	8	7	15	6	8
5	2	3	16	9	10
6	7	4	17	9	8
7	9	9	18	4	6
8	7	8	19	4	3
9	2	5	20	5	4
10	4	3	21	3	2
11	6	9	22	5	3

Linea A	Linea B	$\|A-B\|$	Rango	R con signo
6	8	2	13	-13
4	5	1	5.5	-5.5
7	4	3	18	18
8	7	1	5.5	5.5
2	3	1	5.5	-5.5
7	4	3	18	18
9	9	0	0	0
7	8	1	5.5	-5.5
2	5	3	18	-18
4	3	1	5.5	5.5
6	9	3	18	-18
8	5	3	18	18
4	2	2	13	13
3	3	0	0	0
6	8	2	13	-13
9	10	1	5.5	-5.5
9	8	1	5.5	5.5
4	6	2	13	-13
4	3	1	5.5	5.5
5	4	1	5.5	5.5
3	2	1	5.5	5.5
5	3	2	13	13

$$T_+ = 113$$

$$T_- = 97$$

$$T = min(T_+, T_-) = 97$$

$$E(T) = \frac{n(n+1)}{4} = \frac{22(22+1)}{4} = 126{,}5$$

$$Var(T) = \frac{n(n+1)(2n+1)}{24} = 948{,}75$$

$$z = \frac{T - E(T)}{\sqrt{Var(T)}} - 0{,}9577366819967589$$

$$\alpha = 0{,}05$$

$$q_z(1 - \frac{\alpha}{2}) = 1{,}96$$

En este caso Se rechaza H1, se acepta H0. Las distribuciones son iguales

4. Problema 4

Dos métodos para controlar el tránsito, A y B, se usaron en cada una de n = 12 cruceros durante una semana y los números de accidentes que ocurrieron durante ese tiempo se registraron. El orden de uso (cuál se emplearía para la primera semana) se seleccionó de una manera aleatoria. Se desea saber si los datos dan suficiente evidencia para indicar una diferencia en las distribuciones de porcentajes de accidentes para los métodos A y B de control de tránsito.

Crucero	A	B	Crucero	A	B
1	5	4	7	2	3
2	6	4	8	4	1
3	8	9	9	7	9
4	3	2	10	5	2
5	6	3	11	6	5
6	1	0	12	1	1

Solución

Dado que tenemos datos en pares y lo que buscamos demostar es si la distribucion de los datos diferen podemos utilizar una prueba Wilcoxon

Linea A	Linea B	$A - B$	$\lvert A - B \rvert$	Rango	R con signo
5	4	1	1	3.5	3.5
6	4	2	2	7.5	7.5
8	9	−1	1	3.5	-3.5
3	2	1	1	3.5	3.5
6	3	3	3	10	10
1	0	1	1	3.5	3.5
2	3	−1	1	3.5	-3.5
4	1	3	3	10	10
7	9	−2	2	7.5	-7.5
5	2	3	3	10	10
6	5	1	1	3.5	3.5
1	1	0	0	0	0

$$T_+ = 3{,}5 + 7{,}5 + 3{,}5 + 10 + 3{,}5 + 10 + 10 + 3{,}5 = 51{,}5$$

$$T_- = 3{,}5 + 3{,}5 + 7{,}5 = 14{,}5$$

$$T = min(T_+, T_-) = 14{,}5$$

$$E(T) = \frac{n(n+1)}{4} = \frac{12(12+1)}{4} = 39$$

$$Var(T) = \frac{n(n+1)(2n+1)}{24} = \frac{12(12+1)(2(12)+1)}{24} = 162{,}5$$

$$z = \frac{T - E(T)}{\sqrt{Var(T)}} = \frac{14{,}5 - 39}{\sqrt{162{,}5}} = -1{,}9219$$

Tomando $\alpha = 0{,}05$

$$q_z(1 - \frac{\alpha}{2}) = 1{,}96$$

Se rechaza H_0 si

$$|z| > q_z(1 - \frac{\alpha}{2})$$

$$1{,}65385 > 1{,}96$$

Por lo anto las distribuciones son iguales (No se rechaza H_0), se acepta H_0 lo que indica que la lina A y la linea B no son diferentes

5. Problema 5

Los resultados de un experimento para investigar el reconocimiento de productos, durante tres campañas publicitarias, se muestran en la siguiente tabla. Las respuestas fueron el porcentaje de 400 adultos que estaban familiarizados con el producto recién anunciado. La gráfica de probabilidad normal indicó que los datos no eran aproximadamente normales y debía usarse otro método de análisis. ¿Hay una diferencia significativa entre las tres distribuciones poblacionales de donde vinieron estas muestras?

Campaña		
1	2	3
.33	.28	.21
.29	.41	.30
.21	.34	.26
.32	.39	.33
.25	.27	.31

Solución

Como tenemos tres muestras, las purebas de Whitney y Wilcoxon quedan descartadas, usaremos una prueba de de Kruskal-Wallis

Asignacion de rangos

Campaña		
1	2	3
$,33_{11,5}$	$,28_{6}$	$,21_{1,5}$
$,29_{7}$	$,41_{15}$	$,30_{8}$
$,21_{1,5}$	$,34_{13}$	$,26_{4}$
$,32_{10}$	$,39_{14}$	$,33_{11,5}$
$,25_{3}$	$,27_{5}$	$,31_{9}$

Cuadro 1: **Número de observaciones y suma de rangos**

1	2	3
$n_1 = 5$	$n_2 = 5$	$n_3 = 5$
$R_1 = 33$	$R_2 = 53$	$R_3 = 34$

$$H = \frac{12}{15 \times 16}\left(\frac{33^2}{5} + \frac{53^2}{5} + \frac{34^2}{5}\right) - 3(15+1) = 2{,}5399$$

Se rechaza H_0 si

$$H > q_{X^2_{k-1}}(1-\alpha)$$
$$q_{X^2_2}(0{,}95) = 5{,}991$$
$$2{,}5399 > 5{,}991$$

No rechaza H_0

6. Problema 6

En un estudio reciente que involucró una muestra aleatoria de 300 accidentes automovilísticos, se clasificó la información de acuerdo con el tamaño del automóvil.

	Pequeño	Mediano	Grande
Por lo menos un muerto	42	35	20
Ningún muerto	78	65	60

Con estos datos, ¿puedes afirmar que la frecuencia de accidentes depende del tamaño del automóvil?

	Pequeño	Mediano	Grande	Suman Filas
Por lo menos un muerto	42	35	20	97
Ningún muerto	78	65	60	203
Suma Columnas	120	100	80	300

	Pequeño	Mediano	Grande
Por lo menos un muerto	38.8	32.3	26.867
Ningún muerto	81.2	67.67	54.13

$$\sum_i 0{,}26391753 + 0{,}21993127 + 1{,}33058419 + 0{,}12610837 + 0{,}10509031 + 0{,}63579639 = 2{,}68142806$$

Se rechaza H_0 si $X^2 > Q_{X^2}((r-1)(c-1))^{(1-\alpha)}$

$$2{,}68142806 > Q_{X^2}((r-1)(c-1))^{(1-\alpha)} = Q_{X^2}((1)(2))^{(0{,}95)} = 0{,}103$$

Se rechaza H_0 si P-valor $< \alpha$

$$\text{P-valor} = 1 - P\left(z < \frac{2{,}6814 - E(X_2^2)}{\sqrt{Var(X_2^2)}}\right) = 1 - P\left(z < \frac{2{,}6814 - 2}{\sqrt{4}}\right) = 1 - 0{,}34071 = 0{,}6593$$

$$0{,}6593 \not\leq 0{,}05$$

$\therefore$ Se acepta H_0, lo que significa que el tamaño del automóvil si afecta, por ende es más seguro un auto grande o mediano que uno pequeño.

7. Problema 7

Verifica si los siguientes datos provienen de una distribución normal.

1.0672029	2.3103976	0.8193199	-0.8588287	0.8003015
-0.8404432	0.2049356	-0.9665391	1.7639849	-0.7825124
-2.9712801	-0.8181979	-2.0191393	-1.6289196	2.4613544
-2.6406738	-2.6125324	-1.1968322	-0.1210923	-2.1779296
2.5898699	-2.6133718	-1.6105999	1.0137149	-2.3441204

x	n	$F_n(x)$	F(x)	$\lvert F_n(x) - F(x)\rvert$
-2.97	1	0.04	0.0015	0.0385
-2.64	2	0.08	0.0041	0.0759
-2.61	3	0.12	0.0045	0.1155
-2.61	4	0.16	0.0045	0.1555
-2.34	5	0.2	0.0096	0.1904
-2.18	6	0.24	0.0146	0.2254
-2.02	7	0.28	0.0217	0.2583
-1.63	8	0.32	0.0516	0.2684
-1.61	9	0.36	0.0537	0.3063
-1.20	10	0.4	0.1151	0.2849
-0.97	11	0.44	0.166	0.274
-0.86	12	0.48	0.1949	0.2851
-0.84	13	0.52	0.2005	0.3195
-0.82	14	0.56	0.2061	0.3539
-0.78	15	0.6	0.2177	0.3823
-0.12	16	0.64	0.4522	0.1878
0.20	17	0.68	0.5793	0.1007
0.80	18	0.72	0.7881	0.0681
0.82	19	0.76	0.7939	0.0339
1.01	20	0.8	0.8438	0.0438
1.07	21	0.84	0.8577	0.0177
1.76	22	0.88	0.9608	0.0808
2.31	23	0.92	0.9896	0.0696
2.46	24	0.96	0.9931	0.0331
2.59	25	1	0.9952	0.0048

$D_0 = \sup_x |F_n(x) - F(x)| = 0{,}382$
Valor de la tabla K-S = 0.32
Por lo que $Q_n(1-\alpha), n = 25, \alpha = 0{,}01 \rightarrow 0{,}382 \not\leq 0{,}32$
$\therefore$ Se rechaza H_0, lo que significa que los datos no provienen de una distribución normal.

8. Problema 8

Demuestra que el coeficiente de la τ de Kendall es:

$$\tau = 1 - \frac{4Q}{n(n-1)}$$

Hint: P+Q=$\frac{n(n-1)}{2}$
La fórmula para calcular τ de Kendall es:

$$\tau = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}}$$

donde:

- P = pares concordantes
- Q = pares discordantes
- T = empates en el primer conjunto
- U = empates en el segundo conjunto

Podemos considerar que no hay empates, por lo cual:

$$\tau = \frac{P - Q}{\sqrt{(P + Q)(P + Q)}} = \frac{P - Q}{P + Q}$$

Ahora como $P + Q = \frac{n(n-1)}{2}$, entonces:

$$\tau = \frac{P - Q}{P + Q} = \frac{P - Q}{\frac{n(n-1)}{2}} = \frac{2(P - Q)}{n(n-1)}$$

de la hint, sabemos que $b = \frac{n(n-1)}{2} - Q$, entonces:

$$\frac{\frac{n(n-1)}{2} - Q - Q}{\frac{n(n-1)}{2}} = \frac{n(n-1) - 4Q}{n(n-1)}$$

$$\frac{n(n-1)}{n(n-1)} - \frac{4Q}{n(n-1)} = 1 - \frac{4Q}{n(n-1)}$$

Así, llegamos a que:

$$\boxed{\tau = 1 - \frac{4Q}{n(n-1)}}$$

9. Problema 9

Investiga la prueba de Friedman y resuelve un problema práctico.

La prueba de Friedman

La prueba de Friedman es un test no paramétrico utilizado para detectar diferencias en tratamientos a lo largo de múltiples intentos de prueba. La prueba de Friedman es aplicable cuando tienes dos o más tratamientos dependientes (relacionados) y deseas comparar sus efectos. Se utiliza comúnmente en estudios donde los mismos sujetos son sometidos a diferentes tratamientos en un orden aleatorio. Se basa en el rango que ocupan las observaciones dentro de cada uno de los bloques (donde un bloque podría ser un sujeto o unidad experimental que recibe todos los tratamientos en un orden aleatorio). En esencia, ordena las observaciones dentro de cada bloque y asigna rangos, donde el tratamiento con el mejor resultado recibe el rango más alto. Luego, evalúa si las diferencias entre los rangos de los tratamientos son mayores de lo esperado por casualidad.

PASOS DE LA PRUEBA DE FRIEDMAN.

1. Rankear los datos dentro de cada bloque: Para cada bloque (sujeto), ordena las observaciones de los tratamientos de menor a mayor y asigna rangos. En caso de empates, asigna un rango promedio.

2. Calcular el estadístico de Friedman: Se calcula un valor estadístico basado en la suma de los rangos asignados a cada tratamiento a través de todos los bloques. Este estadístico se compara con una distribución de Friedman para determinar si las diferencias observadas son estadísticamente significativas.

   Para calcular el estadístico de Friedman, usamos la fórmula:

$$\chi_F^2 = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1)$$

   donde:

   - N es el número de bloques.
   - k es el número de tratamientos.
   - R_j es la suma de rangos para el tratamiento j.

3. Determinar la significancia: Usar el valor estadístico calculado y compárarlo con un valor crítico de la distribución de Friedman (o usa un valor p) para determinar si hay una diferencia estadísticamente significativa entre los tratamientos.

4. Post-hoc (si es necesario): Si la prueba indica diferencias significativas, se pueden realizar análisis post-hoc para identificar específicamente entre qué tratamientos existen las diferencias.

EJEMPLO
Supongamos que un investigador desea comparar la efectividad de tres dietas (A, B y C) en la pérdida de peso. Cinco individuos son sometidos a cada una de las dietas durante tres meses, uno después del otro en un orden aleatorio. La pérdida de peso (en kilogramos) se registra para cada dieta en cada individuo.

Participante	Dieta A	Dieta B	Dieta C
1	3	5	2
2	4	3	5
3	2	4	3
4	5	2	4
5	1	3	2

1.- Rankear los datos: Dentro de cada fila (participante), se asignan rangos a las pérdidas de peso, donde el mayor peso perdido recibe el rango más alto.
Para cada participante, comparamos las pérdidas de peso obtenidas con las tres dietas y asignamos rangos.
2.-Ordenamos las pérdidas de peso de menor a mayor para cada participante Asignamos rangos a estas pérdidas de peso, donde la mayor pérdida de peso recibe el rango más alto. En caso de empates, se asignarían rangos promedio, pero no aplican en este caso.
La asignación de rangos queda de la siguiente manera:

Participante	Dieta A (Rango)	Dieta B (Rango)	Dieta C (Rango)
1	2	3	1
2	2	3	1
3	3	1	2
4	2	1	3
5	3	1	2

Paso 2: Calcular el Estadístico de Friedman
Para calcular el estadístico de Friedman, usamos la fórmula:

$$\chi_F^2 = \frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3N(k+1)$$

donde:

- N es el número de bloques (participantes, en este caso 5).
- k es el número de tratamientos (dietas, en este caso 3).
- R_j es la suma de rangos para el tratamiento j.

El estadístico de Friedman calculado para nuestro ejemplo es aproximadamente 0.4. Este valor se utiliza para evaluar si existen diferencias significativas entre los tratamientos (en este caso, las dietas).
Paso 3: Determinar la Significancia
Para determinar si las diferencias entre las dietas son estadísticamente significativas, comparamos el valor del estadístico de Friedman calculado (0.4) con el valor crítico de la distribución de χ^2 correspondiente a los grados de libertad $k-1$ (donde k es el número de tratamientos, es decir, 3 en nuestro caso, lo que nos da 2 grados de libertad) y el nivel de significancia deseado (comúnmente 0.05).
El valor p es aproximadamente 0.819. Dado que este valor p es mayor que 0.05, concluimos que no hay evidencia suficiente para rechazar la hipótesis nula, lo que significa que no hay diferencias estadísticamente significativas en la efectividad de las tres dietas en términos de pérdida de peso.

10. Problema 10

Investiga la prueba de independencia basada en el coeficiente de Pearson y resuelve un problema práctico.

La prueba de independencia basada en el coeficiente de correlación de Pearson

La prueba de independencia basada en el coeficiente de correlación de Pearson es una técnica estadística utilizada para evaluar si existe una relación lineal entre dos variables cuantitativas. El coeficiente de correlación de Pearson, denotado como r, mide el grado y la dirección de la relación lineal entre dos variables. Su valor varía entre -1 y 1, donde:

- $r = 1$ indica una correlación positiva perfecta: a medida que una variable aumenta, la otra también lo hace en proporción directa.
- $r = -1$ indica una correlación negativa perfecta: a medida que una variable aumenta, la otra disminuye en proporción directa.
- $r = 0$ sugiere que no hay correlación lineal entre las variables.

Teoría

La correlación de Pearson se calcula usando la fórmula:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

donde:

- x_i e y_i son los valores individuales de las variables X e Y,
- $\bar{x}$ y $\bar{y}$ son las medias de X e Y, respectivamente.

Para probar la independencia utilizando el coeficiente de Pearson, se realiza una prueba de hipótesis:

- Hipótesis nula (H_0): No hay relación lineal entre las dos variables (es decir, $r = 0$).
- Hipótesis alternativa (H_1): Existe una relación lineal entre las dos variables (es decir, $r \neq 0$).

El valor de r se utiliza para calcular un valor t, que luego se compara con un valor crítico de la distribución t con $n - 2$ grados de libertad (donde n es el número de pares de datos). La fórmula para calcular el valor t es:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Pasos Necesarios

1. Calcular el coeficiente de correlación de Pearson (r) entre las dos variables usando la fórmula proporcionada.
2. Calcular el valor t usando el valor de r y el número total de pares de datos (n).
3. Determinar el valor p asociado con el valor t calculado, utilizando la distribución t con $n - 2$ grados de libertad.
4. Concluir si se rechaza o no la hipótesis nula basándose en el valor p y el nivel de significancia elegido (comúnmente, 0.05).

EJEMPLO PRACTICO
Supongamos que queremos investigar la relación entre las horas estudiadas y las calificaciones obtenidas por un grupo de 10 estudiantes.

Estudiante	Horas Estudiadas (X)	Calificación (Y)
1	1	2
2	2	3
3	3	6
4	4	8
5	5	10
6	6	12
7	7	14
8	8	16
9	9	18
10	10	20

Paso 1: Calcular el Coeficiente de Correlación de Pearson (r)
Primero, necesitamos calcular las medias ($\bar{x}$ y $\bar{y}$) de las horas estudiadas y las calificaciones, respectivamente.
Luego, aplicamos la fórmula del coeficiente de Pearson:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Donde x_i e y_i son los valores individuales de las horas estudiadas y las calificaciones, respectivamente.
Paso 2: Calcular el Valor t
Con el valor de r calculado, procedemos a calcular el valor t usando la fórmula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Donde n es el número de pares de datos (en este caso, 10).
Paso 3: Determinar el Valor p
Utilizamos el valor t calculado y el número de grados de libertad $(n-2)$ para determinar el valor p asociado, lo cual nos dirá si la correlación observada es estadísticamente significativa.
Ahora, hagamos los cálculos paso a paso. Comenzaremos calculando las medias de X e Y, y luego aplicaremos estos valores en la fórmula de Pearson para obtener r. Finalmente, calcularemos el valor t y determinaremos el valor p.
Paso 1: Calcular el Coeficiente de Correlación de Pearson (r)

- Media de las horas estudiadas ($\bar{x}$): 5.5
- Media de las calificaciones ($\bar{y}$): 10.9
- Utilizando estos valores y la fórmula de Pearson, calculamos el coeficiente de correlación de Pearson (r) como aproximadamente 0.999. Esto indica una correlación positiva muy fuerte entre las horas estudiadas y las calificaciones.

Paso 2: Calcular el Valor t

- Aplicando el valor de r en la fórmula del valor t, obtenemos un valor t de aproximadamente 60.53. Este valor se utilizará para evaluar la significancia de la correlación.

Paso 3: Determinar el Valor p

- El valor p asociado con el valor t calculado y 8 grados de libertad (10 pares de datos menos 2) es aproximadamente $6{,}17 \times 10^{-12}$. Este valor p es significativamente menor que el nivel de significancia estándar de 0.05, lo que indica que la correlación observada es estadísticamente significativa.

Estos cálculos muestran que hay una relación lineal positiva muy fuerte entre las horas estudiadas y las calificaciones obtenidas por los estudiantes, y esta relación