Inferencia Estadística

Ejercicio: Contraste no paramétrico

DSLab

noviembre, 2024

Ejercicio 1:

Pregunta: Un estudio quiere analizar si hay diferencias significativas en la distribución de la duración del sueño entre dos grupos de personas: estudiantes universitarios y trabajadores a tiempo completo. Para ello, se recolectaron dos muestras independientes de tamaño $n1 = 20$ para estudiantes y $n2 = 30$ para trabajadores. Los datos representan las horas de sueño por noche durante una semana para cada participante.

A partir de los datos recopilados, se desea determinar si hay evidencia estadística para afirmar que las distribuciones de horas de sueño difieren entre los dos grupos.

```r
# Muestras de horas de sueño para estudiantes universitarios (grupo 1)
muestras_estudiantes <- c(7.2, 7.0, 6.8, 7.5, 7.1, 7.3, 7.2, 7.4, 7.0,
                          7.6, 6.5, 6.9, 7.2, 7.3, 7.1, 6.4, 7.5, 7.0,
                          7.2, 7.3)

# Muestras de horas de sueño para trabajadores a tiempo completo (grupo 2)
muestras_trabajadores <- c(7.5, 6.3, 6.7, 6.9, 7.3, 6.8, 7.3, 6.9, 6.5,
                           6.2, 6.8, 7.0, 6.3, 7.6, 6.9, 6.5, 7.4, 6.7,
                           6.6, 7.2, 6.7, 6.9, 7.1, 6.3, 6.6, 6.1, 6.9,
                           7.1, 6.3, 6.6)
```

Solución:

1. **Formulación de Hipótesis**:
   - Hipótesis Nula (H_0): Las muestras de estudiantes y trabajadores provienen de la misma distribución de horas de sueño.
   - Hipótesis Alternativa (H_1): Las muestras de estudiantes y trabajadores provienen de distribuciones diferentes de horas de sueño.
2. **Nivel de Significación**:

   Se establece un nivel de significancia de $\alpha = 0.05$.
3. **Prueba Estadística**:

Se utilizará la Prueba de Kolmogorov-Smirnov para dos muestras.

4. **Procedimiento de Prueba**:

   Se ordenan los datos de ambas muestras y se calculan las funciones de distribución acumulativa (CDF) empíricas para cada grupo.

   Luego, se calcula la estadística de prueba D que representa la máxima diferencia vertical entre las dos CDF empíricas.

5. **Regla de Decisión**:

   Si el valor D calculado es mayor que el valor crítico de Kolmogorov-Smirnov para dos muestras con un nivel de significancia de $\alpha = 0.05$, se rechaza la hipótesis nula y se concluye que hay diferencias significativas en la distribución de horas de sueño entre los dos grupos.

6. **Interpretación de Resultados**:

   Se interpreta el valor D calculado junto con el valor crítico de la tabla de Kolmogorov-Smirnov para dos muestras. Si D es mayor que el valor crítico, se rechaza H_0 y se concluye que hay diferencias significativas en las distribuciones de horas de sueño entre estudiantes y trabajadores. De lo contrario, no hay suficiente evidencia para rechazar H_0 y se concluye que no hay diferencias significativas en las distribuciones de horas de sueño entre los dos grupos.

```
# Realizar la prueba de Kolmogorov-Smirnov para dos muestras
ks_test <- ks.test(muestras_estudiantes, muestras_trabajadores)

# Mostrar el resultado
print(ks_test)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  muestras_estudiantes and muestras_trabajadores
## D = 0.5, p-value = 0.004958
## alternative hypothesis: two-sided
```

En este caso, dado que el $valor$ es menor que un grado de significación 0.05 podemos rechazar la hipótesis nula en favor de la alternativa.

Ejercicio 2:

Pregunta: Un estudio realizado en una comunidad recopiló información sobre el género (hombre, mujer) y las preferencias de deportes (fútbol, baloncesto, tenis) de un grupo de personas. Los datos se muestran en la siguiente tabla de contingencia:

	Fútbol	Baloncesto	Tenis
Hombre	30	20	10
Mujer	20	25	15

Se desea determinar si hay una asociación significativa entre el género y las preferencias de deportes en la población utilizando la prueba de chi-cuadrado de independencia.

Solución:

- Hipótesis Nula (H_0): No hay asociación entre el género y las preferencias de deportes en la población.
- Hipótesis Alternativa (H_1): Hay una asociación entre el género y las preferencias de deportes en la población.

Para resolver el ejercicio anterior, primero necesitamos calcular la estadística de prueba chi-cuadrado y luego compararla con el valor crítico de chi-cuadrado para determinar si se rechaza o no la hipótesis nula.

Entendido, resolveremos el ejercicio calculando la estadística de prueba chi-cuadrado sin usar R. Aquí están los pasos a seguir:

1. Calcular las frecuencias esperadas bajo la hipótesis nula de independencia.
2. Calcular la estadística de prueba chi-cuadrado utilizando la fórmula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde O_{ij} son las frecuencias observadas y E_{ij} son las frecuencias esperadas.

Tomando los datos del ejercicio anterior, primero calcularemos las frecuencias esperadas:

$$\text{Total} = 30 + 20 + 10 + 20 + 25 + 15 = 120$$

$$\text{Frecuencia Total Hombres} = 30 + 20 + 10 = 60$$
$$\text{Frecuencia Total Mujeres} = 20 + 25 + 15 = 60$$

$$\text{Probabilidad de ser Hombre} = \frac{\text{Frecuencia Total Hombres}}{\text{Total}} = \frac{60}{120} = 0.5$$
$$\text{Probabilidad de ser Mujer} = \frac{\text{Frecuencia Total Mujeres}}{\text{Total}} = \frac{60}{120} = 0.5$$

$$\text{Frecuencia Total Fútbol} = 30 + 20 = 50$$
$$\text{Frecuencia Total Baloncesto} = 10 + 25 = 35$$
$$\text{Frecuencia Total Tenis} = 10 + 15 = 25$$
$$\text{Frecuencia Esperada Hombres Fútbol} = (\text{Probabilidad de ser Hombre}) \times (\text{Frecuencia Total Fútbol})$$
$$\text{Frecuencia Esperada Hombres Fútbol} = 0.5 \times 50 = 25$$
$$\text{Frecuencia Esperada Hombres Baloncesto} = (\text{Probabilidad de ser Hombre}) \times (\text{Frecuencia Total Baloncesto})$$
$$\text{Frecuencia Esperada Hombres Baloncesto} = 0.5 \times 35 = 17.5$$
$$\text{Frecuencia Esperada Hombres Tenis} = (\text{Probabilidad de ser Hombre}) \times (\text{Frecuencia Total Tenis})$$
$$\text{Frecuencia Esperada Hombres Tenis} = 0.5 \times 25 = 12.5$$
$$\text{Frecuencia Esperada Mujeres Fútbol} = (\text{Probabilidad de ser Mujer}) \times (\text{Frecuencia Total Fútbol}$$
$$\text{Frecuencia Esperada Mujeres Fútbol} = 0.5 \times 50 = 25$$
$$\text{Frecuencia Esperada Mujeres Baloncesto} = (\text{Probabilidad de ser Mujer}) \times (\text{Frecuencia Total Baloncesto})$$

$$\text{Frecuencia Esperada Mujeres Baloncesto} = 0.5 \times 35 = 17.5$$

$$\text{Frecuencia Esperada Mujeres Tenis} = (\text{Probabilidad de ser Mujer}) \times (\text{Frecuencia Total Tenis})$$

$$\text{Frecuencia Esperada Mujeres Tenis} = 0.5 \times 25 = 12.5$$

Una vez que tenemos las frecuencias esperadas, podemos calcular la estadística de prueba chi-cuadrado usando la fórmula dada. Luego, comparamos el valor obtenido con el valor crítico de la distribución chi-cuadrado para determinar si se rechaza o no la hipótesis nula de independencia.

Por tanto, utilizando los datos proporcionados:

$$O_{11} = 30, \quad O_{12} = 20, \quad O_{13} = 10, \quad O_{21} = 20, \quad O_{22} = 25, \quad O_{23} = 15$$

$$E_{11} = 25, \quad E_{12} = 17.5, \quad E_{13} = 12.5, \quad E_{21} = 25, \quad E_{22} = 17.5, \quad E_{23} = 12.5$$

Aplicamos la fórmula:

$$\chi^2 = \frac{(30-25)^2}{25} + ... + \frac{(15-12.5)^2}{12.5} \approx 6.571$$

Por lo tanto, la estadística de prueba chi-cuadrado es aproximadamente $\chi^2 \approx 6.571$.

Para determinar si el valor calculado de la estadística de prueba chi-cuadrado es significativo, necesitamos compararlo con el valor crítico de la distribución chi-cuadrado. El valor crítico depende del nivel de significancia (α) y los grados de libertad del contraste.

En este caso, para una prueba de independencia con un nivel de significancia de $\alpha = 0.05$ y un total de $df = (r-1)(c-1)$ grados de libertad, donde r es el número de filas y c es el número de columnas en la tabla de contingencia. Para la tabla de contingencia dada, tenemos $r = 2$ y $c = 3$, por lo tanto, $df = (2-1)(3-1) = 2$.

Podemos consultar una tabla de distribución chi-cuadrado o usar una función en R para encontrar el valor crítico correspondiente para $\alpha = 0.05$ y $df = 2$.

El valor crítico de chi-cuadrado para $\alpha = 0.05$ y $df = 2$ es aproximadamente 5.99.

Dado que el valor calculado de la estadística de prueba chi-cuadrado ($\chi^2 \approx 6.571$) es mayor que el valor crítico de referencia ($\chi^2 \approx 5.99$), rechazamos la hipótesis nula. Podemos concluir que hay una asociación significativa entre el género y las preferencias de deportes en la población.

Ejercicio 3:

Prueba:

Se llevó a cabo un estudio para comparar los tiempos de respuesta entre dos grupos de personas expuestas a diferentes condiciones de entrenamiento. Para ello, se registraron los tiempos de respuesta (en milisegundos) de 15 personas en el Grupo A y 18 personas en el Grupo B. Los datos obtenidos se muestran a continuación:

Grupo A: 45.1, 38.0, 52.4, 40.1, 47.5, 41.2, 43.5, 39.7, 36.1, 44.2, 50.5, 48.1, 42.4, 46.1, 49.5

Grupo B: 55.3, 60.4, 58.9, 62.0, 54.1, 57.8, 59.3, 63.6, 56.4, 61.7, 58, 64.3, 53.7, 66.1, 67.1, 65.2, 68.3, 50.1

Determine si hay una diferencia significativa en los tiempos de respuesta entre los dos grupos. Para determinar si hay una diferencia significativa en los tiempos de respuesta entre los dos grupos, puedes utilizar la prueba de Mann-Whitney U. Esta prueba no paramétrica se utiliza para comparar las distribuciones de dos muestras independientes y determinar si una muestra tiene valores significativamente mayores o menores que la otra.

Solucion:

Para determinar si hay una diferencia significativa en los tiempos de respuesta entre los dos grupos, podemos aplicar la prueba de Mann-Whitney U. A continuación, se detallan los pasos para realizar esta prueba.

Paso 1: Formular las hipótesis

- H_0: No hay diferencia significativa en los tiempos de respuesta entre los dos grupos (las distribuciones son iguales).
- H_1: Hay una diferencia significativa en los tiempos de respuesta entre los dos grupos (las distribuciones no son iguales).

Paso 2: Combinar y ordenar las observaciones

Combinamos todas las observaciones y las ordenamos en orden ascendente:

Datos	Grupo
36.1	A
38.0	A
39.7	A
40.1	A
41.2	A
42.4	A
43.5	A
44.2	A
45.1	A
46.1	A
47.5	A
48.1	A
49.5	A
50.1	B
50.5	A
52.4	A
53.7	B
54.1	B
55.3	B
56.4	B
57.8	B
58.0	B
58.9	B
59.3	B
60.4	B
61.7	B
62.0	B

Datos	Grupo
63.6	B
64.3	B
65.2	B
66.1	B
67.1	B
68.3	B

Paso 3: Asignar rangos a las observaciones

Asignamos rangos a las observaciones, donde el rango más bajo es 1 y el rango más alto es el número total de observaciones (33). En caso de empate, se asigna el promedio de los rangos para los valores empatados.

Datos	Grupo	Rango
36.1	A	1
38.0	A	2
39.7	A	3
40.1	A	4
41.2	A	5
42.4	A	6
43.5	A	7
44.2	A	8
45.1	A	9
46.1	A	10
47.5	A	11
48.1	A	12
49.5	A	13
50.1	B	14
50.5	A	15
52.4	A	16
53.7	B	17
54.1	B	18
55.3	B	19
56.4	B	20
57.8	B	21
58.0	B	22
58.9	B	23
59.3	B	24
60.4	B	25
61.7	B	26
62.0	B	27
63.6	B	28
64.3	B	29
65.2	B	30
66.1	B	31
67.1	B	32

Datos	Grupo	Rango
68.3	B	33

Paso 4: Sumar los rangos de cada grupo

- Suma de rangos del Grupo A (R_A): $1+2+3+4+5+6+7+8+9+10+11+12+13+15+16 = 122$
- Suma de rangos del Grupo B (R_B): $14 + 17 + 18 + 19 + 20 + 21 + 22 + 23 + 24 + 25 + 26 + 27 + 28 + 29 + 30 + 31 + 32 + 33 = 439$

Paso 5: Calcular el estadístico U de Mann-Whitney

$$U_A = n_A \times n_B + \frac{n_A(n_A + 1)}{2} - R_A$$

$$U_B = n_A \times n_B + \frac{n_B(n_B + 1)}{2} - R_B$$

Donde $n_A = 15$ y $n_B = 18$.

$$U_A = 15 \times 18 + \frac{15 \times 16}{2} - 122 = 268$$

$$U_B = 15 \times 18 + \frac{18 \times 19}{2} - 439 = 2$$

El valor de U es el mínimo de U_A y U_B:

$$U = \min(268, 2) = 2$$

Paso 6: Determinar el valor crítico

Para un nivel de significancia del 5% (α = 0.05) y tamaños de muestra $n_A = 15$ y $n_B = 18$, el valor crítico de U se puede buscar en una tabla de la distribución de Mann-Whitney U. Para estos tamaños de muestra, el valor crítico de U es aproximadamente 84.

Paso 7: Tomar una decisión

Como $U = 2$ es menor que el valor crítico 84, rechazamos la hipótesis nula H_0.

Conclusión

Hay una diferencia significativa en los tiempos de respuesta entre los dos grupos. Esto sugiere que las diferentes condiciones de entrenamiento tienen un impacto significativo en los tiempos de respuesta.

Ejercicio 4

Pregunta: Supongamos que estamos trabajando con un conjunto de datos de un modelo de clasificación que predice si los clientes de una tienda en línea harán una compra (Sí o No) basado en varias características. Queremos investigar si existe una asociación entre dos variables categóricas en nuestro conjunto de datos: "Genero" (Hombre, Mujer) y "Compra" (Sí, No).

Hemos entrenado un modelo de clasificación y obtenido las siguientes predicciones y datos observados para una muestra de 200 clientes:

Género	Compra: Sí	Compra: No
Hombre	50	30
Mujer	70	50

Solución:

Paso 1: Formular las Hipótesis

- **Hipótesis Nula (H_0)**: No hay asociación entre el género y la compra (son independientes).
- **Hipótesis Alternativa (H_1)**: Existe una asociación entre el género y la compra (no son independientes).

Paso 2: Crear la Tabla de Contingencia La tabla de contingencia muestra el número de observaciones para cada combinación de las categorías:

Género	Compra: Sí	Compra: No	Total
Hombre	50	30	80
Mujer	70	50	120
Total	120	80	200

Paso 3: Calcular los Valores Esperados Para cada celda de la tabla de contingencia, calculamos el valor esperado bajo la hipótesis nula de independencia:

$$E_{ij} = \frac{(F_i \times F_j)}{N}$$

donde E_{ij} es el valor esperado para la celda en la fila i y la columna j, F_i es el total de la fila i, F_j es el total de la columna j, y N es el total de observaciones.

Género	Compra: Sí	Compra: No
Hombre	$\frac{80 \times 120}{200} = 48$	$\frac{80 \times 80}{200} = 32$
Mujer	$\frac{120 \times 120}{200} = 72$	$\frac{120 \times 80}{200} = 48$

Paso 4: Calcular el Estadístico Chi-Cuadrado El estadístico chi-cuadrado se calcula usando la fórmula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde O_{ij} es el valor observado y E_{ij} es el valor esperado.

$$\chi^2 = \frac{(50-48)^2}{48} + \frac{(30-32)^2}{32} + \frac{(70-72)^2}{72} + \frac{(50-48)^2}{48} \approx 0.3472$$

Paso 5: Determinar el Valor Crítico y Tomar una Decisión Para un nivel de significancia del 5% (α = 0.05) y con grados de libertad:

$$df = (n_{filas} - 1) \times (n_{columnas} - 1) = (2-1) \times (2-1) = 1$$

El valor crítico de chi-cuadrado para 1 grado de libertad al nivel de significancia 0.05 es aproximadamente 3.841.

Conclusión Dado que el estadístico chi-cuadrado calculado (0.3472) es menor que el valor crítico (3.841), no rechazamos la hipótesis nula H_0.

Por lo tanto, no hay evidencia suficiente para concluir que existe una asociación significativa entre el género y la compra en esta muestra de datos.

Ejercicio 5

Pregunta: Supongamos que hemos desarrollado un modelo de clasificación para predecir el riesgo crediticio de los clientes de un banco, clasificándolos como “Alto Riesgo” o “Bajo Riesgo”. Queremos evaluar si existe una asociación significativa entre las predicciones del modelo y los valores reales observados en un conjunto de prueba.

Se toma una muestra de 300 clientes, y se obtiene la siguiente tabla de contingencia que muestra la distribución de las predicciones del modelo y los valores reales observados:

	Real: Alto Riesgo	Real: Bajo Riesgo	Total
Predicción: Alto Riesgo	80	30	110
Predicción: Bajo Riesgo	50	140	190
Total	130	170	300

Solución:

Paso 1: Formular las Hipótesis

- **Hipótesis Nula (H_0)**: No hay asociación entre las predicciones del modelo y los valores reales observados (son independientes).
- **Hipótesis Alternativa (H_1)**: Existe una asociación entre las predicciones del modelo y los valores reales observados (no son independientes).

Paso 2: Crear la Tabla de Contingencia La tabla de contingencia ya está proporcionada:

	Real: Alto Riesgo	Real: Bajo Riesgo	Total
Predicción: Alto Riesgo	80	30	110
Predicción: Bajo Riesgo	50	140	190
Total	130	170	300

Paso 3: Calcular los Valores Esperados Para cada celda de la tabla de contingencia, calculamos el valor esperado bajo la hipótesis nula de independencia:

$$E_{ij} = \frac{(F_i \times F_j)}{N}$$

donde E_{ij} es el valor esperado para la celda en la fila i y la columna j, F_i es el total de la fila i, F_j es el total de la columna j, y N es el total de observaciones.

	Real: Alto Riesgo	Real: Bajo Riesgo
Predicción: Alto Riesgo	$\frac{110 \times 130}{300} = 47.67$	$\frac{110 \times 170}{300} = 62.33$
Predicción: Bajo Riesgo	$\frac{190 \times 130}{300} = 82.33$	$\frac{190 \times 170}{300} = 107.67$

Paso 4: Calcular el Estadístico Chi-Cuadrado El estadístico chi-cuadrado se calcula usando la fórmula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde O_{ij} es el valor observado y E_{ij} es el valor esperado.

$$\chi^2 = \frac{(80 - 47.67)^2}{47.67} + \frac{(30 - 62.33)^2}{62.33} + \frac{(50 - 82.33)^2}{82.33} + \frac{(140 - 107.67)^2}{107.67} = 61.13$$

Paso 5: Determinar el Valor Crítico y Tomar una Decisión Para un nivel de significancia del 5% (α = 0.05) y con grados de libertad:

$$df = (n_{filas} - 1) \times (n_{columnas} - 1) = (2 - 1) \times (2 - 1) = 1$$

El valor crítico de chi-cuadrado para 1 grado de libertad al nivel de significancia 0.05 es aproximadamente 3.841.

Conclusión Dado que el estadístico chi-cuadrado calculado (61.13) es mucho mayor que el valor crítico (3.841), rechazamos la hipótesis nula H_0.

Por lo tanto, hay evidencia suficiente para concluir que existe una asociación significativa entre las predicciones del modelo y los valores reales observados. Esto indica que las predicciones del modelo están asociadas con los valores reales, lo que sugiere que el modelo tiene un buen desempeño en la clasificación del riesgo crediticio de los clientes.

Ejercicio 6

Pregunta: Un investigador desea evaluar si un nuevo método de enseñanza mejora las calificaciones de los estudiantes. Para ello, selecciona a 10 estudiantes y mide sus calificaciones en una prueba antes y después de aplicar el nuevo método durante un semestre. Los datos obtenidos son los siguientes:

Estudiante	Calificación Antes (X)	Calificación Después (Y)
1	65	75
2	70	78
3	60	65
4	72	80
5	68	74
6	75	85
7	63	68
8	70	77
9	67	73
10	66	69

Para determinar si hay una diferencia significativa en las calificaciones antes y después del nuevo método de enseñanza, se realizará un contraste de hipótesis utilizando la prueba de Wilcoxon para muestras pareadas.

Solución:

Paso 1: Formular las Hipótesis

- **Hipótesis Nula (H_0)**: No hay diferencia en las calificaciones antes y después del nuevo método de enseñanza.
- **Hipótesis Alternativa (H_1)**: Hay una diferencia en las calificaciones antes y después del nuevo método de enseñanza.

Paso 2: Calcular las Diferencias y Ordenarlas por Rangos Calculamos la diferencia $D = Y - X$ y luego ordenamos las diferencias por rangos (ignorando los signos).

Est.	Antes (X)	Después (Y)	D = Y - X	Rango
1	65	75	10	7.5
2	70	78	8	6
3	60	65	5	4
4	72	80	8	6
5	68	74	6	5
6	75	85	10	7.5
7	63	68	5	4
8	70	77	7	3
9	67	73	6	5
10	66	69	3	2

Paso 3: Calcular la Suma de Rangos Sumamos los rangos de las diferencias positivas (mejoras):

$$W_+ = 7.5 + 6 + 4 + 6 + 5 + 7.5 + 4 + 3 + 5 + 2 = 50$$

Sumamos los rangos de las diferencias negativas (empeoramientos). En este caso, no hay diferencias negativas.

Paso 4: Determinar el Valor Crítico y Comparar Para $n = 10$ y $\alpha = 0.05$, el valor crítico para la prueba de Wilcoxon pareada (consultando una tabla de Wilcoxon) es aproximadamente 8 para una prueba bilateral.

Comparando el valor de W_+ con el valor crítico:

$$W_+ = 50 > 8$$

Paso 5: Conclusión Dado que W_+ es mayor que el valor crítico, rechazamos la hipótesis nula H_0. Esto indica que hay una diferencia significativa en las calificaciones antes y después de aplicar el nuevo método de enseñanza.

Ejercicio 7

Pregunta: Un nutricionista quiere evaluar el efecto de una nueva dieta en la presión arterial de los pacientes. Se mide la presión arterial sistólica de 12 pacientes antes y después de seguir la dieta durante un mes. Los datos son los siguientes:

Paciente	Antes (X)	Después (Y)
1	135	130
2	140	135
3	150	145
4	145	142
5	160	155
6	155	150
7	148	150
8	160	158
9	142	140
10	138	136
11	150	149
12	145	146

Determinar si hay una diferencia significativa en la presión arterial sistólica antes y después de seguir la dieta utilizando la prueba del signo para observaciones por pares.

Solución:

1. **Hipótesis**:
   - H_0: No hay diferencia en la presión arterial sistólica antes y después de la dieta.
   - H_1: Hay una diferencia en la presión arterial sistólica antes y después de la dieta.
2. **Diferencias**:

Paciente	Antes (X)	Después (Y)	Diferencia (D = X - Y)	Signo
1	135	130	5	+
2	140	135	5	+
3	150	145	5	+
4	145	142	3	+
5	160	155	5	+
6	155	150	5	+
7	148	150	-2	-
8	160	158	2	+
9	142	140	2	+
10	138	136	2	+
11	150	149	1	+
12	145	146	-1	-

3. **Contar los Signos**:
   - $S_+ = 10$
   - $S_- = 2$
4. **Estadístico de Prueba**: El estadístico de la prueba del signo es el menor entre S_+ y S_-, que es 2 en este caso.
5. **Valor Crítico**: Para $n = 12$ (todos los pares se consideran) y $\alpha = 0.05$, el valor crítico de la prueba del signo (consultando una tabla binomial o tabla de prueba del signo) es aproximadamente 1.
6. **Decisión**:
   - Como el estadístico de la prueba (2) es mayor que el valor crítico (1), no rechazamos la hipótesis nula H_0.

No hay suficiente evidencia para concluir que hay una diferencia significativa en la presión arterial sistólica antes y después de la dieta.

Ejercicio 8

Pregunta: Un investigador desea determinar si existe una relación entre el nivel de educación y la preferencia por el tipo de música. Se realiza una encuesta a 200 personas y se obtienen los siguientes datos:

	Rock	Pop	Clásica	Total
Secundaria	30	40	10	80
Universitaria	20	50	30	100
Posgrado	10	5	5	20
Total	60	95	45	200

Determinar si hay una relación significativa entre el nivel de educación y la preferencia por el tipo de música utilizando la prueba chi-cuadrado de independencia.

Solución:

Hipótesis: - H_0: No hay relación entre el nivel de educación y la preferencia por el tipo de música (independencia). - H_1: Hay relación entre el nivel de educación y la preferencia por el tipo de música (dependencia).

Frecuencias Esperadas: - La frecuencia esperada para cada celda se calcula usando la fórmula: $E_{ij} = \frac{(\text{Total Fila i})(\text{Total Columna j})}{\text{Total General}}$.

	Rock	Pop	Clásica	Total
Secundaria	$\frac{80\times60}{200} = 24$	$\frac{80\times95}{200} = 38$	$\frac{80\times45}{200} = 18$	80
Universitaria	$\frac{100\times60}{200} = 30$	$\frac{100\times95}{200} = 47.5$	$\frac{100\times45}{200} = 22.5$	100
Posgrado	$\frac{20\times60}{200} = 6$	$\frac{20\times95}{200} = 9.5$	$\frac{20\times45}{200} = 4.5$	20
Total	60	95	45	200

Cálculo del Estadístico de Prueba Chi-cuadrado: - La fórmula del estadístico chi-cuadrado es: $\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, donde O_{ij} son las frecuencias observadas y E_{ij} son las frecuencias esperadas.

$$\chi^2 = \frac{(30-24)^2}{24} + ... + \frac{(5-4.5)^2}{4.5} \approx 15.99$$

Valor Crítico:

- Los grados de libertad df para la prueba chi-cuadrado de independencia son $(r-1)(c-1)$, donde r es el número de filas y c es el número de columnas. Aquí, $df = (3-1)(3-1) = 4$.
- Para $\alpha = 0.05$ y $df = 4$, el valor crítico es aproximadamente 9.488.

Decisión:

- Como $\chi^2 = 15.99$ es mayor que el valor crítico de 9.488, rechazamos la hipótesis nula H_0. Hay evidencia suficiente para concluir que existe una relación significativa entre el nivel de educación y la preferencia por el tipo de música.

Ejercicio 9

Pregunta: Una compañía de supermercados está interesada en la preferencia del consumidor con respecto a dos marcas (A, y B) de refresco que compiten entre sí. Se seleccionan, de modo aleatorio, 15 personas y se les pide que clasifiquen las bebidas mediante una escala del 1 (poca aceptación) al 5 (mucha aceptación). El orden en la selección de la bebida fue aleatorio. Se obtiene la siguiente información:

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	3	2	5	2	4	5	2	1	3	1	4	3	3	5	2
B	1	1	4	3	4	1	2	1	2	1	2	4	3	2	2

Mediante el uso de la prueba del signo. ¿Se tiene alguna razón para creer que existe una diferencia en la preferencia para estas dos marcas de refrescos? Supón $\alpha = 0.1$

Solución:

1. **Formulación de las Hipótesis**:
   - H_0: No hay diferencia en la preferencia entre las marcas A y B.
   - H_1: Hay una diferencia en la preferencia entre las marcas A y B.

2. **Diferencias entre las Calificaciones**:
   - Restamos las calificaciones de la marca B de las calificaciones de la marca A: $D = A - B$.

$$D = [2, 1, 1, -1, 0, 4, 0, 0, 1, 0, 2, -1, 0, 3, 0]$$

3. **Signos de las Diferencias**:
   - Ignoramos las diferencias de 0 y consideramos solo las diferencias positivas y negativas.

$$D = [2, 1, 1, -1, 4, 1, 2, -1, 3]$$

$$\text{Número de diferencias positivas} = 6$$

$$\text{Número de diferencias negativas} = 2$$

4. **Aplicación de la Prueba del Signo**:
   - Sea $n = 8$ (el número de diferencias no nulas).
   - El número de diferencias positivas (S^+) sigue una distribución binomial $B(n, 0.5)$.
   - Calculamos el valor p usando la función de masa de probabilidad binomial.

$$P(X \geq 6) = P(X = 6) + P(X = 7) + P(X = 8)$$

   Usando la fórmula de la binomial:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

   Donde $p = 0.5$.

$$P(X = 6) = \binom{8}{6} (0.5)^6 (0.5)^2 = \frac{8!}{6!2!} (0.5^8) = 28 \cdot 0.0039 = 0.1094$$

$$P(X = 7) = \binom{8}{7} (0.5)^7 (0.5)^1 = \frac{8!}{7!1!} (0.5^8) = 8 \cdot 0.0039 = 0.0313$$

$$P(X = 8) = \binom{8}{8} (0.5)^8 (0.5)^0 = \frac{8!}{8!0!} (0.5^8) = 1 \cdot 0.0039 = 0.0039$$

$$P(X \geq 6) = 0.1094 + 0.0313 + 0.0039 = 0.1446$$

5. **Decisión**:
   - Comparar el valor $p = 0.1446$ con el nivel de significancia $\alpha = 0.1$.
   - Como $p > 0.1$, no rechazamos la hipótesis nula H_0.

No hay evidencia suficiente para concluir que hay una diferencia significativa en la preferencia entre las marcas A y B con un nivel de significancia del 10%.

Ejercicio 10

Pregunta: Para los datos del ejercicio anterior, emplea la prueba de rangos de signos de Wilcoxon ¿Se obtienen las mismas conclusiones?

Solución: Para utilizar la prueba de rangos de signos de Wilcoxon, primero necesitamos calcular los rangos de las diferencias absolutas y luego sumar los rangos de las diferencias positivas. Luego, comparamos esta suma con la tabla de valores críticos para determinar si hay una diferencia significativa entre los grupos.

1. **Formulación de las Hipótesis**:
   - H_0: No hay diferencia en la preferencia entre las marcas A y B.
   - H_1: Hay una diferencia en la preferencia entre las marcas A y B.
2. **Diferencias entre las Calificaciones**:
   - Restamos las calificaciones de la marca B de las calificaciones de la marca A: $D = A - B$.

$$D = [2, 1, 1, -1, 0, 4, 0, 0, 1, 0, 2, -1, 0, 3, 0]$$

3. **Diferencias Absolutas y Rangos**:
   - Calculamos las diferencias absolutas $|D|$ y sus rangos $R(|D|)$.

$$|D| = [2, 1, 1, 1, 0, 4, 0, 0, 1, 0, 2, 1, 0, 3, 0]$$
$$R(|D|) = [10, 6, 6, 6, 2, 15, 2, 2, 6, 2, 10, 6, 2, 14, 2]$$

4. **Suma de Rangos de Diferencias Positivas**:
   - Sumamos los rangos de las diferencias positivas (R^+).

$$R^+ = 10 + 15 + 10 + 14 = 49$$

5. **Determinación del Valor Crítico**:
   - Con n_+ siendo el número de diferencias positivas y n_- siendo el número de diferencias negativas, calculamos T^+, el estadístico de prueba.

$$T^+ = \text{min}(R^+, n_+(n_+ + 1)/2)$$

   En nuestro caso, $n_+ = 6$.

$$T^+ = \text{min}(49, 6 \times (6 + 1)/2) = \text{min}(49, 21) = 21$$

6. **Decisión**:
   - Comparamos T^+ con el valor crítico de la tabla de Wilcoxon para $n = 15$ y $\alpha = 0.1$.
   - Si $T^+ < T_\alpha$, rechazamos H_0.
7. **Resultados**:
   - De la tabla de Wilcoxon para $n = 15$ y $\alpha = 0.1$, obtenemos $T_\alpha = 15$.
   - Como $T^+ = 21 > T_\alpha = 15$, rechazamos H_0.

Hay suficiente evidencia para concluir que existe una diferencia significativa en la preferencia entre las marcas A y B con un nivel de significancia del 10%.