

Inferència Estadística

Yolanda Colom Torrens

Introducción

Durante esta actividad estaremos trabajando con herramientas de inferencia estadística las cuales nos ayudan a conseguir conclusiones y tomar decisiones a partir de estas. Gracias a los intervalos de confianza y pruebas de hipótesis estudiaremos los resultados con el fin de seleccionar las mejores resoluciones basadas en problemas matemáticos

Problema 1

En este problema debemos evaluar la asociación que puedan tener dos variables categóricas, estas pueden ser nominales u ordinales. Para este caso lo mejor es utilizar el test de chi-cuadrado el cual nos ayudara a determinar si acaso hay una relación existente entre el trastorno del sueño y la práctica de fumar regularmente.

Bajo la hipótesis nula de que no existe una relación entre A y B, aplicamos el cálculo de frecuencias esperadas con la siguiente formula: **F.E = (total fila * total columna) / tamaño de muestra**

A	B		
	No	Si	Total
Mucho	7	30	
Moderadamente	12	36	
Nada	29	36	
			150

$$X^2 = \sum [(frecuencia\ observada - frecuencia\ esperada)^2 / frecuencia\ esperada]$$

Si sustituimos los valores de la tabla en la formula, **$X^2 = 13,026$** .

Luego de obtener este valor se debe comparar con el valor crítico del test a partir de un nivel de confianza específico y un número determinado de variables independientes. Para esta situación establecemos:

- Nivel de confianza = 5%
- Grado de libertad = (N.filas - 1) * (N.columnas - 1) = 2 * 1 = 2

Al comparar estos resultados con la tabla de chi-cuadrado, podemos analizar que con un nivel de confianza de 5% y 2 grados de libertad, el valor crítico será de 5,99. Luego de esto debemos comparar los resultados de valor crítico con el valor de X^2 .

$$13,026 > 5,99$$

Es por esto que tenemos la capacidad de rechazar la hipótesis nula, gracias a que en la muestra de 150 personas se concluye que hay una relación significativa entre el habito de fumador y los problemas para dormir.

1.2

Nuevamente al tener que evaluar dos variables categóricas para responder al psicólogo si acaso existe una relación entre los síntomas depresivos (A) y la ansiedad (B), en este caso tendremos una muestra de 100 historiales clínicos, utilizaremos nuevamente el test chi-cuadrado.

		B	
		Si	No
A	Si	38	31
	No	9	22

Establecemos una hipótesis nula de que no hay relación entre las variables y aplicamos el uso de la formula. **F.Esperada = (total fila * total columna) / tamaño de muestra**

Estadístico de chi-Cuadrado:

$$X^2 = \sum [(frecuencia\ observada - frecuencia\ esperada)^2 / frecuencia\ esperada]$$

$$X^2 = 11,634.$$

Ahora, debemos comparar el valor calculado del estadístico χ^2 con el valor crítico de la distribución Chi-cuadrado. Para ello, necesitamos determinar los grados de libertad (df). En este caso, $df = (\text{número de filas} - 1) * (\text{número de columnas} - 1) = (2 - 1) * (2 - 1) = 1$.

Dado que hemos seleccionado un nivel de significancia del 2% (0.02), el valor crítico de χ^2 con 1 grado de libertad es aproximadamente 5.41.

Dado que el estadístico χ^2 calculado (6.76) es mayor que el valor crítico (5.41), rechazamos la hipótesis nula y concluimos que los síntomas depresivos (A) y ansiedad (B) son dependientes en la muestra de 100 historiales clínicos, considerando un nivel de significación de 0.02.

Problema 2

Para elaborar un informe en el que se interprete la prueba de bondad de ajuste a la normalidad de los gráficos de las variables Var1 y Var2, seguiremos los pasos utilizados en cualquier prueba estadística:

Planteamiento de la hipótesis nula.

- H_0 : La variable sigue una distribución normal.
- H_1 : La variable no sigue una distribución normal.

Var1:

1. Interpretación del estadístico de contraste (Shapiro-Wilk):

$$W = 0.93938, p\text{-valor} = 0.01274$$

2. Decisión estadística a partir del estadístico y su correspondiente p-valor, con su conclusión:

Dado que el p-valor (0.01274) es menor que el nivel de significancia comúnmente utilizado de 0.05, rechazamos la hipótesis nula (H_0) y concluimos que la variable Var1 no sigue una distribución normal.

3. Interpretación de las gráficas de la normalidad:

- Densidad suavizada: La densidad suavizada de Var1 muestra cierto alejamiento de la forma de campana típica de la distribución normal.
- Histograma de intervalos: El histograma de intervalos de Var1 no muestra una distribución simétrica alrededor de la media.
- QQ Plot: Los puntos en el QQ plot de Var1 no siguen una línea recta, lo que indica que la distribución de Var1 difiere de una distribución normal.
- Gráfico de cuantiles: El gráfico de cuantiles para Var1 muestra desviaciones en la probabilidad acumulada en comparación con las puntuaciones Z, lo que refuerza la idea de que Var1 no sigue una distribución normal.

Var2:

4. Interpretación del estadístico de contraste (Shapiro-Wilk):

$W = 0.95394$, $p\text{-valor} = 0.04965$

5. Decisión estadística a partir del estadístico y su correspondiente p-valor, con su conclusión:

Dado que el p-valor (0.04965) es ligeramente menor que el nivel de significancia de 0.05, podríamos rechazar la hipótesis nula (H_0) y concluir que la variable Var2 no sigue una distribución normal. Sin embargo, cabe destacar que el p-valor está muy cerca del límite, por lo que la evidencia en contra de la normalidad es relativamente débil. En este caso, podríamos argumentar que no hay suficiente evidencia para rechazar la hipótesis nula y, por lo tanto, no podemos afirmar con certeza que Var2 no sigue una distribución normal.

6. Interpretación de las gráficas de la normalidad:

- Densidad suavizada: La densidad suavizada de Var2 muestra una forma más cercana a la campana típica de la distribución normal en comparación con Var1, pero aún con algunas diferencias.
- Histograma de intervalos: El histograma de intervalos de Var2 presenta una distribución que se aproxima a la simetría alrededor de la media.
- QQ Plot: Los puntos en el QQ plot de Var2 siguen una línea recta en su mayoría, pero con algunas desviaciones, lo que sugiere que la distribución de Var2 puede no ser exactamente normal.
- Gráfico de cuantiles: El gráfico de cuantiles para Var2 muestra una menor desviación en la probabilidad acumulada en comparación con las puntuaciones Z que Var1, pero aún con ciertas discrepancias.

En resumen, aunque la prueba Shapiro-Wilk indica que Var2 no sigue exactamente una distribución normal, la evidencia en contra de la normalidad es débil debido a que el p-valor está muy cerca del nivel de significancia de 0.05. Por lo tanto, no podemos afirmar con certeza que Var2 no sigue una distribución normal.

Problema 3

Relación entre género y edad:

1. Planteamiento de la hipótesis nula:
 - H_0 : No existe relación entre género y edad.
 - H_1 : Existe relación entre género y edad.
2. Cálculo del estadístico de contraste:

```
Medianas para EDAD segun Género

      F      M
41.83025 41.17878

Resumen descriptivo por rangos

Rango promedio Suma Rangos
F      26.24    656
M      24.76    619

Prueba U de Mann-Whitney para EDAD por Género

Wilcoxon rank sum test

data: EDAD by Género
W = 331, p-value = 0.7196

alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
-5.406874  6.958391

sample estimates:
difference in location
0.7463992
```

3. Decisión estadística:

Se rechaza la hipótesis nula $0.71 > 0.5$

4. Conclusión:

Dado que se rechaza la hipótesis nula, se puede decir que no hay una relación entre género y edad.

Relación entre género y sueldo:

1. Planteamiento de la hipótesis nula:
 - H_0 : No existe relación entre género y sueldo.
 - H_1 : Existe relación entre género y sueldo.

2. Cálculo del estadístico de contraste:
Como el p-valor es menor que el nivel de significancia ($0.54 > 0.174$), no podemos asumir que la distribución de los datos es normal

```
> UMW # Prueba U de Mann-Whitney para Sueldo segun Género
Medianas para Sueldo segun Género
  F    M
2165.266 2145.312
Resumen descriptivo por rangos
Rango promedio Suma Rangos
F      26.64    666
M      24.36    609
Prueba U de Mann-Whitney para Sueldo por Género
Wilcoxon rank sum test
data: Sueldo by Género
```

```
W = 341, p-value = 0.5803
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
-269.3285 442.8126
sample estimates:
difference in location
85.20958
```

3. Decisión estadística:
No se puede rechazar la hipótesis nula.

Relación entre sueldo y posición en la empresa:

1. Planteamiento de la hipótesis nula:
 - H_0 : No existe relación entre sueldo y posición en la empresa.
 - H_1 : Existe relación entre sueldo y posición en la empresa.
2. Cálculo del estadístico de contraste:
3. Decisión estadística:
4. Conclusión: