3.6. Pruebas para diferencia de proporciones

Hasta ahora hemos estudiado pruebas de hipótesis referidos a medidas de localización generalmente representadas por medias poblacionales. A continuación veremos metodologías estadísticas apropiadas para comparar proporciones a partir de muestras seleccionadas de dos poblaciones.

3.6.1. Muestras independientes

Supongamos que hemos definido una población cuyos elementos sólo pueden clasificarse en una de dos categorías mutuamente excluyentes.

Por ejemplo, una planta puede ser clasificada en infectada o no infectada, un paciente puede ser clasificado como un caso o un control según alguna patología específica, un individuo puede ser clasificado en empleado o desempleado, etc..

Pensemos, además, que el objetivo de cierta investigación consiste en establecer hipótesis referidas a la proporción o porcentaje de elementos que pertenecen a una de las dos categorías consideradas.

Para llevar a cabo esta experiencia, seguramente se extraerán muestras de una o dos poblaciones y los individuos seleccionados que pertenezcan a una categoría de interés para el investigador se tomarán como un 'éxito' (esquema binomial estudiado en el Módulo VI).

El interés del investigador se centrará en comparar la proporción o porcentaje de éxitos en cada una de las muestras consideradas.

Este tipo de cuestiones se presentan en experimentos controlados cuando la variable de respuesta está expresada como una proporción o porcentaje.

También en estudios observacionales tales como estudios de caso control, estudios de cohorte y en investigaciones sociales cuando se extraen muestras aleatorias de dos poblaciones.

Comparación de dos Proporciones Utilizando la Aproximación Normal

Supongamos un ejemplo sencillo donde se desea estudiar la relación existente entre el hábito de fumar y la mortalidad.

De un registro de jubilados pertenecientes a una determinada Caja se extrajo una muestra aleatoria de 1.469 personas con edades entre 60 y 64 años a quienes se realizó una entrevista con un cuestionario previamente determinado.

Una de las preguntas estaba referida a si la persona fumaba o no.

El equipo de investigación efectuó un seguimiento de las personas entrevistadas durante 6 años, registrando su fallecimiento cuando éste se producía.

Una vez concluidos los años de seguimiento, los resultados obtenidos fueron resumidos en la siguiente tabla de distribución de frecuencias:

	No fumadores	Fumadores	Total
Muerto	117	54	171
Vivo	950	348	1298
Total	1067	402	1469

Podemos considerar que estamos en presencia de un estudio observacional donde la recolección de la información se ha realizado tomando dos muestras: una correspondiente a personas fumadoras y otra a personas no fumadoras.

Evidentemente, en este caso todas las frecuencias esperadas, exceden a 5.

La hipótesis de trabajo es, evidentemente, que la proporción de muertos entre las personas fumadoras es mayor que la que se produce entre las personas no fumadoras.

Prueba de Hipótesis

En el ejemplo, la hipótesis estadística planteada es la siguiente:

H_0) $P_1 = P_2$ ó $P_1 - P_2 = 0$

donde:

P_1 = proporción de muertos entre los no fumadores y P_2 = proporción de muertos entre los fumadores.

La hipótesis alternativa considerada es:

H_1) $P_1 < P_2$ ó $P_1 - P_2 < 0$

Vamos a considerar a continuación cuál es el estadístico adecuado para probar la hipótesis planteada.

El parámetro que interviene en la hipótesis es una diferencia de proporciones poblacionales y por lo tanto, es lógico pensar que su mejor estimador puntual será la diferencia de proporciones muestrales: $p_1 - p_2$.

¿Cuál será la distribución de dicho estimador?

Ya hemos visto en el Módulo VIII que, si la muestra es suficientemente grande, la variable aleatoria proporción muestral p se distribuye normalmente con media P y varianza $P(1 - P)/n$.

En el ejemplo, debido a los tamaños de las muestras seleccionadas, podemos considerar que tanto p_1 como p_2 se distribuyen normalmente y que, en consecuencia, la variable aleatoria diferencia de proporciones muestrales $p_1 - p_2$ también se distribuirá normalmente.

Nos falta aún determinar la media y la varianza de esta variable aleatoria: diferencia de proporciones muestrales.

$E(p_1 - p_2) = E(p_1) - E(p_2)$

Recordando lo aprendido anteriormente:

$E(p_1) = P_1$ y $E(p_2) = P_2$

Entonces

$E(p_1 - p_2) = P_1 - P_2$

Luego, la esperanza de la variable diferencia de proporciones muestrales es igual a la diferencia de las proporciones poblacionales.

Hemos señalado también que, cuando trabajamos con dos muestras independientes, la varianza de una diferencia de variables aleatorias es igual a la suma de sus varianzas.

También vimos anteriormente que

$$V_{(p)} = \frac{P(1 - P)}{n}$$

Entonces,

$$V_{(p_1)} = \frac{P_1(1 - P_1)}{n_1}$$

Y

$$V_{(p_2)} = \frac{P_2(1 - P_2)}{n_2}$$

Si calculamos ahora la varianza de la variable aleatoria diferencia de proporciones muestrales, tendremos:

$$V_{(p_1 - p_2)} = \frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}$$

Una vez determinada la distribución de probabilidad y sus correspondientes parámetros, podemos plantear el estadístico que utilizaremos para verificar la hipótesis.

Si la variable original se distribuye N(μ, σ), la variable estandarizada se distribuirá N(0, 1).

El estadístico será, entonces, la siguiente variable z estandarizada:

$$z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1(1 - P_1)}{n_1} + \frac{P_2(1 - P_2)}{n_2}}}$$

Analicemos un poco más el estadístico.

Si se cumple la hipótesis nula, $(P_1 - P_2) = 0$.

Sin embargo, en el denominador del estadístico aparecen los valores poblacionales P_1 y P_2 que, cuando las muestras son suficientemente grandes, pueden ser reemplazados por sus correspondientes estimadores muestrales p_1 y p_2.

Reglas de decisión

A continuación estudiaremos las reglas de decisión que aplicaremos en esta prueba.

Tal como está planteada la hipótesis alternativa, la zona de rechazo de la hipótesis nula se encuentra a la izquierda de la distribución normal.

Las reglas de decisión serán:

Si $z < z'$ se rechaza H_0 y

Si $z > z'$ no se rechaza H_0

Calcularemos ahora el estadístico con los datos proporcionados por el ejemplo.

$$p_1 = \frac{117}{1067} = 0{,}1097$$

$$p_2 = \frac{54}{402} = 0{,}1343$$

$$z = \frac{(0{,}1097 - 0{,}1343)}{\sqrt{\frac{0{,}1097 \times 0{,}8903}{1067} + \frac{0{,}1343 \times 0{,}8657}{402}}} = \frac{0{,}0246}{0{,}0195} = -1{,}26$$

Si fijamos un nivel de significación $\alpha = 0.05$, buscando convenientemente en la tabla de la distribución normal, encontramos el valor crítico $z' = -1.645$.

Como:

$-1.26 > -1.645$

no existe una evidencia muestral suficiente para rechazar la hipótesis nula.

Podemos concluir que la proporción de muertos en el grupo de ancianos fumadores no difiere significativamente de la proporción de muertos entre los no fumadores.

En el ejemplo planteado hemos trabajado con tamaños de muestras suficientemente grandes por lo cual fue sostenible el supuesto de que las proporciones muestrales se distribuyen normalmente.

Comparación de Dos Proporciones Utilizando la Distribución Ji- Cuadrado con un Grado de Libertad

También ilustraremos este tema por medio de un ejemplo.

El objetivo de una investigación médica consistió en estudiar si el ejercicio de la profesión de odontólogo era un factor de riesgo para contraer la hepatitis B.

Para implementar la experiencia se extrajo una muestra aleatoria de odontólogos (casos) y otra de individuos que ejercían otras profesiones pero con edades y niveles socioeconómicos similares (controles).

La hipótesis estadística planteada fue la siguiente:

$H_0)$ $P_1 \leq P_2$ ó $P_1 - P_2 \leq 0$

$H_1)$ $P_1 > P_2$ ó $P_1 - P_2 > 0$

siendo:

P_1 = proporción de odontólogos que presentan virus de hepatitis B

P_2 = proporción de otros profesionales que presentan virus de hepatitis B.

El único supuesto que establecemos en esta prueba de hipótesis es que las muestras fueron extraídas aleatoriamente y que son independientes.

A todos los individuos que se incluyeron en el ensayo se les extrajo una muestra de sangre y se determinó la presencia o no del virus de la hepatitis B.

Los resultados obtenidos fueron resumidos en la siguiente tabla:

	Casos	Controles	Total
Virus +	25	10	35
Virus -	4	28	32
Total	29	38	67

En este caso las frecuencias esperadas se calculan del siguiente modo:

$$n_1=\frac{C_1}{N}=29\times\frac{35}{67}=15{,}1 \qquad n_1=\frac{C_2}{N}=29\times\frac{32}{67}=13{,}9$$

$$n_2=\frac{C_1}{N}=38\times\frac{35}{67}=19{,}9 \qquad n_2=\frac{C_2}{N}=38\times\frac{32}{67}=18{,}1$$

Como todos exceden a 5 podemos aplicar indistintamente cualquiera de las dos aproximaciones, ya sea mediante la distribución normal o mediante la distribución χ^2 para mostrar su procedimiento.

En este caso el estadístico que se utiliza para probar la hipótesis nula es el siguiente:

$$T=\frac{N(n_{11}.\,n_{22}-n_{12}.\,n_{21})^2}{n_1.\,n_2.\,c_1.\,c_2}$$

siendo:

n_{11} = cantidad de casos (odontólogos) con presencia de virus n_{12} = cantidad de controles (otras profesiones) con presencia de virus

n_{21} = cantidad de casos con ausencia de virus

n_{22} = cantidad de controles con ausencia de virus

n_1 = total de casos

n_2 = total de controles

c_1 = total de individuos con presencia de virus

c_2 = total de individuos con ausencia de virus

N = total de individuos.

El estadístico T tiene distribución χ^2 con 1 grado de libertad.

$$T\sim X_1^2$$

Habiendo aclarado el significado de cada símbolo, calcularemos a continuación el valor del estadístico T.

$$T = \frac{67 \times (25 \times 28 - 4 \times 10)^2}{29 \times 38 \times 35 \times 32} = \frac{67 \times (700 - 40)^2}{1234240} = 23{,}65$$

De acuerdo a la hipótesis alternativa planteada, la zona de rechazo de la misma estará a la derecha de la distribución χ^2.

Regla de decisión

Si $T > \chi^2$, se rechazará H_0
Si $T < \chi^2$, no se rechazará H_0
Si fijamos un nivel de significación $\alpha = 0.05$, el correspondiente valor de $\chi^2 = 3.84$.
Como $T > \chi^2$, $(23.65 > 3.84)$ se rechaza H_0.

En consecuencia podemos concluir que la proporción de infectados con el virus de hepatitis B en el grupo de odontólogos es significativamente mayor a la del grupo de otras profesiones.

Calcularemos a continuación estas proporciones.

P_1 = proporción de personas con presencia de virus dentro del grupo de odontólogos = 25|29 = 0.8621.

P_2 = proporción de personas con presencia de virus dentro del grupo de otras profesiones = 10|38 = 0.2632.

Con estos resultados a la vista, se ratifica que efectivamente la proporción de personas infectadas de hepatitis B es significativamente mayor en el grupo de odontólogos. Se puede decir que ejercer esta profesión constituye un factor de riesgo para contraer la enfermedad considerada.

Relación de Eficiencia Entre Métodos Paramétricos y No Paramétricos

Cuando la prueba t de Student se utiliza con datos que no puedan considerarse como provenientes de una distribución normal, puede ocurrir:

a) Que los niveles de significación (α) no sean correctos y, en consecuencia, la probabilidad de rechazar la hipótesis nula cuando es cierta no es, por ejemplo, 0.05.

b) La potencia de la prueba para encontrar un resultado significativo cuando la hipótesis nula es falsa, se ve alterada.

En cambio, las pruebas no paramétricos basados en rangos ven afectado su nivel de significación cuando se observan varios ceros o empates.

En muestras grandes, supuestamente normales, las pruebas basadas en rangos tienen una eficiencia del 95% comparadas con las pruebas paramétricas.

En muestras pequeñas provenientes de poblaciones normales, las pruebas de rangos son ligeramente más eficientes que la prueba t de Student.

En muestras grandes extraídas de poblaciones no normales, la eficiencia de las pruebas no paramétricos basados en rangos es del 86%. Cuando las muestras son, a su vez pequeñas, la eficiencia de las pruebas de rangos respecto a la prueba t es superior en más del 100%.