

ANÁLISIS DE DATOS PARA LA TOMA DE DECISIONES

CLASE 6

CHRISTIAN ARAYA

INSTITUTO DE ESTADÍSTICA
PUCV

2020



PRUEBA DE INDEPENDENCIA (DATOS CATEGÓRICOS)

Definición

- La prueba de bondad de ajuste de Chi Cuadrado, tiene una versión para estudiar la hipótesis de independencia entre dos variables categóricas.
- Por ejemplo: se desea investigar si en una organización, existe alguna relación entre el género y nivel de ingreso de sus trabajadores, siendo esta variable expresada en términos de niveles (bajo, medio, alto).
- Se ha recogido una muestra aleatoria de 1000 personas, cuya información fue resumida en la siguiente tabla:

	Nivel de Ingresos		
Género	Bajo	Medio	Alto
Masculino	182	213	203
Femenino	154	138	110

Ejemplo

- La información presentada de este modo es una tabla de doble entrada, tabla cruzada, de contingencia o 2x2.

	Nivel de Ingresos		
Género	Bajo	Medio	Alto
Masculino	182	213	203
Femenino	154	138	110

Estadístico:

- $$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$

χ^2 PARA DATOS CATEGÓRICOS

	Nivel de Ingresos		
Género	Bajo	Medio	Alto
Masculino	182	213	203
Femenino	154	138	110

Regla:

- $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$
- La regla indica: se rechaza la hipótesis nula para valores grandes del estadístico, esto es, superiores al valor crítico: χ^2_{α} con $(\text{filas}-1) * (\text{columnas}-1)$ grados de libertad.

Advertencias

- El estadístico sobre el cual decidimos la prueba tiene una distribución aproximada a Chi Cuadrado.
- Esta distribución (continua) aproxima "bien" la distribución de muestreo del estadístico.
- En una tabla de contingencia 2x2, con 1 grado de libertad, se debe aplicar corrección de Yates.

$$\chi^2 = \sum_i \frac{(|o_i - e_i| - 0,5)^2}{e_i}$$

- Frecuencias "grandes": resultados con y sin corrección son similares.
- Frecuencias inferiores a 10 (superiores a 5): se recomienda aplicar corrección.
- Frecuencias de 5 o menos: test exacto de Fisher Irwin.

PRUEBA PARA LA PROPORCIÓN

PROPORCIÓN

- Supondremos una muestra aleatoria X_1, X_2, \dots, X_n proveniente de una distribución $Ber(1, p)$.
- Estaremos interesados, en estos casos, en saber si con la información disponible en una muestra, podemos concluir sobre la proporción de una característica determinada en la población.
- La hipótesis nula será, por ejemplo, de la forma: $H_0 : p = p_0$ (en el test a 2 colas).
- Cuando la muestra es grande, emplearemos la aproximación de la distribución Binomial a la Normal para establecer que, bajo H_0 :

$$p_{obs} = E = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \sim N(0, 1)$$



- ¿Qué hacer cuando la muestra es pequeña?... se debe trabajar con la distribución Binomial.
- ¿Qué valor es "grande"? ... la aproximación funciona bien para tamaño de muestra 30 o más.

Ejemplo 1

- Luego de un período de 3 años, los pacientes que han recibido una vacuna presentan la enfermedad objetivo en una proporción de 0.4.
- Suponga que se eligen 53 pacientes al azar (datos contenidos en la base, hoja *Prop1*) y se ha registrado con 1 si durante el seguimiento durante 3 años, estos presentan dicha enfermedad cuando se les aplica una vacuna alternativa, desarrollada por otro laboratorio.
- ¿Es posible, con la información contenida en la muestra, aseverar que la vacuna alternativa es mejor? Decida el test al nivel 0,05.

!

- Supuestos para el test con la distribución Binomial:
- 1) la característica de interés es binaria.
- 2) observaciones independientes.
- 3) número de pruebas es fijo.

Ejemplo 2

- Se sabe que una línea de manufactura, en un día, produce 25 % de unidades defectuosas de un artículo.
- Se desea comparar esta información con el desempeño de una segunda línea. Suponga que se eligen 20 unidades al azar en ella.
- Se obtienen 8 unidades defectuosas. ¿Se cuenta con respaldo para aseverar que la segunda línea tiene peor desempeño que la primera?

ERROR TIPO I: α

	H_0 Verdadera	H_0 Falsa
No Rechazar H_0	Correcto	Error Tipo II
Rechazar H_0	Error Tipo I	Correcto

- **Error tipo I:** rechazar H_0 en favor de H_1 cuando en realidad es verdadera.
- Note que en el procedimiento típico, para decidir la prueba, inicialmente fijamos el valor α o nivel de significancia.
- Note que trabajar con $\mathbb{P}(\text{error tipo I})$ asume que H_0 es cierta.

ERROR TIPO II: β

	H_0 Verdadera	H_0 Falsa
No Rechazar H_0	Correcto	Error Tipo II
Rechazar H_0	Error Tipo I	Correcto

- **Error tipo II:** no rechazar H_0 cuando en realidad es falsa.
- Para estimar el valor de β , se asume el rechazo de la hipótesis nula, en favor de la alternativa, para un valor específico del parámetro en este caso (cuando estamos en H_1).
- Es decir, para un determinado caso, se estima en qué probabilidad de cometer un error de este tipo se está incurriendo.
- Aplicaremos este concepto al ejemplo Ejemplo 2 de la clase pasada.

Comentarios generales

- En general los errores tipo I y tipo II presentan la siguiente relación: a tamaño de muestra constante, por lo general una disminución en la probabilidad de cometer uno, da como resultado un incremento en la probabilidad de cometer el otro. Por este motivo se deben tener ambos en cuenta, pese a que en general se trabaja sólo con α .
- Por este motivo, es cierto que ajustando el nivel α puede cambiar el veredicto del test de hipótesis. No obstante, no se debe descuidar β y, por consiguiente, la potencia del test.
- Un aumento del tamaño de la muestra tiene como efecto la disminución de α y β .
- Si la hipótesis nula es falsa, β se maximiza cuando el valor verdadero de un parámetro se aproxima al valor hipotético.

POTENCIA DE UN TEST

Definición

- La **potencia** de un test es la probabilidad de rechazar H_0 dado que una alternativa específica es verdadera.
- Es decir, se calcula como: $1 - \beta$

[illegible]

- $1 - \beta = 0,0508$
- Es decir, si μ es verdaderamente 0,998, la prueba rechazará de forma adecuada H_0 sólo un 5 % de las veces.

POTENCIA DE UN TEST

[illegible]

- $1 - \beta = 0,0508$
- Es decir, si μ es verdaderamente 0,998, la prueba rechazará de forma adecuada H_0 sólo un 5 % de las veces.
- Es decir, la prueba no sería buena si es importante que el investigador pueda distinguir realmente entre una media poblacional de 0,995 y 0,998.
- Para aumentar la potencia de la prueba, sería necesario aumentar α o bien, aumentar el tamaño muestral.