Prueba de Independencia

ESTADÍSTICA APLICADA
PROF. MSC. VILMA S. ROMERO ROMERO
5TO CICLO | 17-10-2022

UNIVERSIDAD DE LIMA

Introducción

- Evaluar si dos variables cualitativas son independientes, es decir, si existe una relación entre dos variables categóricas.
  - Si las dos variables son independientes, conocer el valor de una variable no proporciona información sobre el valor de la otra variable.
- Se utiliza para determinar si los valores de una de las 2 variables cualitativas dependen de los valores de la otra variable cualitativa.
- Por ejemplo:
  - ¿Es el tipo de vuelo independiente de la clase de cabina del boleto adquirido?
  - ¿Los niveles de estrés de los atletas universitarios están relacionados con la ocurrencia de lesiones?
  - ¿La calidad de la gestión de una empresa y la reputación de la misma están relacionadas?

Estructura de los datos

Las **n** observaciones de la muestra se pueden clasificar de acuerdo a dos criterios diferentes:

Y
(Segundo criterio de clasificación)

X **(Primer criterio de clasificación)**

	Y_1	Y_2	$\dots$	Y_J	**Total**
X_1	O_{11}	O_{12}	$\dots$	O_{1J}	$O_{1.}$
X_2	O_{21}	O_{22}	$\dots$	O_{2J}	$O_{2.}$
.	.	.	$\dots$	.	.
.	.	.	$\dots$	.	.
X_I	O_{I1}	O_{I2}	$\dots$	O_{IJ}	$O_{I.}$
Total	$O_{.1}$	$O_{.2}$	$\dots$	$O_{.J}$	**n**

Donde:

O_{ij} : Número de observaciones de la celda (ij).
$O_{i.}$: Número de observaciones de la fila i.
$O_{.j}$: Número de observaciones de la columna j.

UNIVERSIDAD DE LIMA

Procedimiento general

1) Se formulan las hipótesis:

   H_0 : Las variables X e Y son independientes
   H_1 : Las variables X e Y no son independientes

   O equivalentemente:

   H_0 : No existe relación entre las variables X e Y
   H_1 : Existe relación entre las variables X e Y

2) Fijar el nivel de significación α.

3) Calcular el valor del estadístico de prueba:

$$\chi_0^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)}$$

Donde:

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{n} = \frac{\text{(Total fila i)(Total columna j)}}{\text{Total general}}$$

UNIVERSIDAD DE LIMA

Procedimiento general

4) Determinar la región crítica y regla de decisión:

$$RC = \left\langle \chi^2_{((I-1)(J-1);1-\alpha)}; \infty \right\rangle \Rightarrow \text{Se } \textbf{rechaza } H_0 \text{ si} \chi^2_0 > \chi^2_{((I-1)(J-1);1-\alpha)}$$

5) Decisión y conclusión: **Rechazar** H_0 si el valor del estadístico de prueba pertenece a la región crítica.

Nota: Esta prueba es apropiada si todas las frecuencias esperadas son al menos 5 ($E_{ij} \geq 5$ para todo i y j).

Ejemplo 1:

El propietario de un restaurante tiene el objetivo comercial de aprender más sobre los patrones de demanda de los clientes durante el periodo de fin de semana (de viernes a domingo). Se recopilaron datos de 556 clientes sobre el tipo de plato principal solicitado y el tipo de postre ordenado. Los resultados se organizan en la siguiente tabla

		Tipo de plato principal				
		Carne de ave	Carne de res	Pescado	Pasta	**Total**
Tipo de postre	Helado	49	35	17	14	115
	Fruta	51	27	19	41	138
	Torta	67	41	41	22	171
	Ninguno	47	26	26	33	132
	Total	214	129	103	110	**556**

¿Hay alguna evidencia de una asociación entre el tipo de postre y el tipo de plato principal? Usar $\alpha = 0.01$

UNIVERSIDAD DE LIMA

Solución:

1) Se formulan las hipótesis:
   H_0 : El tipo de postre y el tipo de plato principal son independientes.
   H_1 : El tipo de postre y el tipo de plato principal no son independientes.

2) $\alpha = 0.01$

3) Valor del estadístico de prueba:

$$\chi_0^2 = \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 26.6081 \sim \chi_9^2$$

- El estadístico de prueba tiene una distribución Chi-Cuadrado con $(I - 1)(J - 1) = (4 - 1)(4 - 1) = 9$ grados de libertad.

UNIVERSIDAD DE LIMA

Tabla de frecuencias esperadas

Tipo de postre	Tipo de plato principal: Carne de ave	Carne de res	Pescado	Pasta
Helado	44.26	26.68	21.30	22.75
Fruta	53.12	32.02	25.56	27.30
Torta	65.82	39.67	31.68	33.83
Ninguno	50.81	30.63	24.45	26.12

- $E_{11} = \frac{O_{1.} \times O_{.1}}{n} = \frac{115 \times 214}{556} = 44.26$
- $E_{12} = \frac{O_{1.} \times O_{.2}}{n} = \frac{138 \times 129}{556} = 26.68$

. . . .

- $E_{44} = \frac{O_{4.} \times O_{.4}}{n} = \frac{132 \times 110}{556} = 26.12$

$$\chi_0^2 = \frac{(49 - 44.26)^2}{44.26} + \frac{(35 - 26.68)^2}{26.68} + \cdots + \frac{(33 - 26.12)^2}{26.12} = 26.6081$$

UNIVERSIDAD DE LIMA

4) Región crítica y regla de decisión:

$$RC = \left\langle \chi^2_{((4-1)(4-1);1-\alpha)}; \infty \right\rangle = \left\langle \chi^2_{(9;0.99)}; \infty \right\rangle = \langle 21.666; \infty \rangle$$

Además, se rechaza H_0 si $\chi^2_0 > 21.666$.

5) **Decisión y conclusión:** Como $\chi^2_0 = 26.6081 \in RC$, entonces se rechaza la hipótesis nula. Por lo tanto, existe suficiente evidencia para concluir que el tipo de postre y el tipo de plato principal no son independientes, al 1% de significación.

Ingreso de datos en R

```r
ave <- c(49,51,67,47)
res <- c(35,27,41,26)
pescado <- c(17,19,41,26)
pasta <- c(14,41,22,33)

resultados <- matrix(c(ave,res,pescado,pasta), nrow = 4, ncol = 4,
                     dimnames = list(c("Helado","Fruta","Torta","Ningunc
                                     c("Ave","Res","Pescado","Pasta")))
resultados
```

```
##          Ave Res Pescado Pasta
## Helado    49  35      17    14
## Fruta     51  27      19    41
## Torta     67  41      41    22
## Ninguno   47  26      26    33
```

Prueba de Independencia en R

```r
chisq.test(resultados)

##
##  Pearson's Chi-squared test
##
## data:  resultados
## X-squared = 26.608, df = 9, p-value = 0.001623
```

Decisión y conclusión: Como p-value $= 0.001623 < \alpha = 0.01$, entonces se rechaza la hipótesis nula. Por lo tanto, existe suficiente evidencia para concluir que el tipo de postre y el tipo de plato principal no son independientes, al 1% de significación.

Ejemplo 2:

El archivo "`Preferencias.csv`" contiene información de una encuesta realizada por una Asociación de Chocolateros para determinar las preferencias de los tipos de chocolate de los consumidores. ¿Existe alguna evidencia de asociación entre los tipos de chocolate y el sexo de la persona consumidora de chocolate? Usar $\alpha = 0.05$

Solución:

1) Se formulan las hipótesis:
   H_0 : El tipo de chocolate y el sexo del consumidor son independientes.
   H_1 : El tipo de chocolate y el sexo del consumidor no son independientes.
2) $\alpha = 0.05$

Lectura de datos en R

```
datos <- read.csv("Preferencias.csv")

datos$Preferencia <- factor(datos$Preferencia)
datos$Sexo <- factor(datos$Sexo)
```

```
table(datos$Preferencia, datos$Sexo)

##
##          Femenino Masculino
##   Tipo 1        8        25
##   Tipo 2       39        51
##   Tipo 3       21        56
```

Prueba de Independencia en R

```
chisq.test(datos$Preferencia, datos$Sexo)

##
##  Pearson's Chi-squared test
##
## data:  datos$Preferencia and datos$Sexo
## X-squared = 6.4468, df = 2, p-value = 0.03982
```

3) **Decisión y conclusión:** Como p-value $= 0.03982 < \alpha = 0.05$, entonces se rechaza la hipótesis nula. Por lo tanto, existe suficiente evidencia para concluir que el tipo de chocolote y el sexo no son independientes, al 5% de significación.

UNIVERSIDAD DE LIMA

Recursos Adicionales I

Devore, J. (2019). *Introducción a la probabilidad y estadística para ingeniería y ciencias*. Cengage, 1 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_CENGAGE.

Johnson, R. A. (2012). *Probabilidad y estadística para ingenieros*. Pearson Educación, 8 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_PEARSON.

Kokoska, S. (2015). *Introductory Statistics*. W. H. Freeman and Company, 2 edition.

Mendenhall, W., Beaver, R. J., and Beaver, B. M. (2015). *Introducción a la probabilidad y estadística*. Cengage, 14 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_CENGAGE.

Recursos Adicionales II

Millones, R., Barreno, E., Vásquez, F., and Castillo, C. (2017). *Estadística Descriptiva y Probabilidades: Aplicaciones en la ingeniería y los negocios*. Lima: Fondo Editorial de la Universidad de Lima, 1 edition. Código Biblioteca U.Lima: 519.53 E.

Triola, M. (2018). *Estadística*. Pearson Educación, 12 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_PEARSON.

UNIVERSIDAD DE LIMA