Tema 8
Pruebas de Chi-cuadrado

Contenido
Concordancia. Independencia.
Bondad de ajuste. Uso de las tablas
de contingencia.

Recordemos un poco…

Antes de comenzar a ver de que se trata este tema y las pruebas que realizaremos, recordemos algunas definiciones y veamos ejemplos.

Variable Categórica: Característica para la cual la escala de medida consiste en un conjunto de categorías (Los datos se presentan como frecuencias de observaciones que ocurren en la misma categoría).

La tabla siguiente les presenta datos estadístico sobre la población de 15 años o mas, según nivel de educación y género (ambas variables se expresan en categorías)

Ejemplo de datos categóricos

DISTRIBUCIÓN DE LA POBLACIÓN DE 15 AÑOS O MÁS SEGÚN NIVEL DE EDUCACIÓN DE Y GÉNERO. AÑO 2001 TABLA 7.3

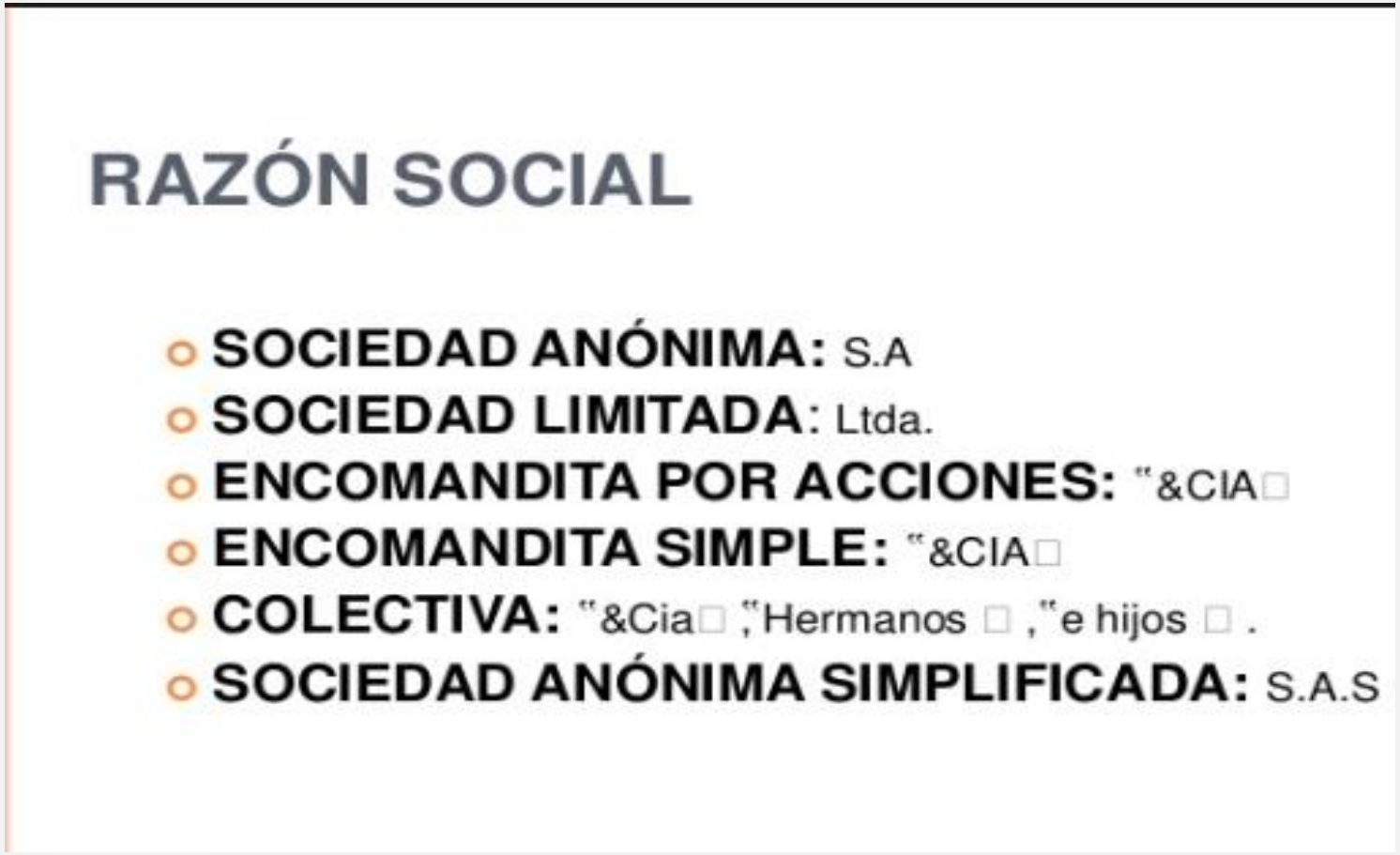
Nivel de educación	Total	Total	Género	
			Varón	Mujer
		26.012.435	12.456.479	13.555.956
	Sin instrucción (1)	**3,7%**	3,5%	3,9%
	Primario incompleto	**14,2%**	14,3%	14,1%
	Primario completo	**48,9%**	51,5%	46,5%
	Secundario completo	**24,5%**	23,7%	25,2%
	Terciario completo	**8,7%**	7,0%	10,3%

(1) incluye nunca asistió, jardín e inicial.

Fuente: *INDEC. Dirección Nacional de Estadísticas Sociales y de Población. Dirección de Estadísticas Sectoriales en base a procesamientos especiales del Censo Nacional de Población, Hogares y Viviendas 2001*

Otro ejemplo de datos categóricos

En este ejemplo la variable es razón social y las categorías son sociedad anónima, sociedad limitada, etc.



Mas ejemplo de datos categóricos: tipo de material

Según los autores, las variables categóricas se puede clasificar en dicotómicas o multiestados

Presencia – Ausencia

precoz - tardío

Normal - anormal

Dicotómicos

(pequeña, mediana, grande)

Multiestados

PRUEBAS DE CHI CUADRADO

Son comparaciones o contrastes asociados con el estadístico Chi-cuadrado; se denomina así porque se asocia a las distribuciones de probabilidades de Chi cuadrado.

Estas pruebas se aplican a variables de tipo cualitativo, trabajando con las frecuencias. Consisten en tomar una muestra y observar si hay diferencia significativa entre "frecuencias observadas" y "frecuencias esperadas".

Se llaman frecuencias esperadas a las que son especificadas o conocidas, se dice que se basan en la ley teórica del modelo propuesto.

EL ESTADÍSTICO X^2 y SU DISTRIBUCIÓN

Entonces la idea es comparar los resultados empíricos obtenidos de un experimento o de un muestreo con datos teóricos que se esperan bajo determinado modelo o ley.

Expresión matemática del estadístico.

$$\chi^2 = \sum_{i=1}^{k} \frac{(obs_i - esp_i)^2}{esp_i}$$

La variable o estadístico, tiene a una distribución de probabilidad Chi cuadrado con k-1 grados de libertad.

PRUEBAS DE CHI CUADRADO

Estas pruebas tiene varios usos y se las puede conocer como pruebas de:

- CONCORDANCIA
- INDEPENDENCIA
- BONDAD DE AJUSTE

Pruebas de Concordancia

Supongamos que se desea probar si un dado está "cargado", es decir que el resultado no es "homogéneo".
Pensemos entonces, ¿cómo sería la probabilidad de que salga cualquier cara si fuera homogéneo?. Esta probabilidad y tendría valor 1/6.
Se realiza un experimento que consiste en arrojar 60 veces el dado, si no estuviera cargado, cada una de las caras deberá aparecer 10 veces.

Este problema en términos estadísticos se plantea a través de la decisión si las frecuencias observadas son compatibles con las esperadas.

Es decir, las pruebas son comparaciones realizadas para decidir si los datos de una muestra se presentan conforme a una ley preestablecida.

Recordemos los pasos para la Prueba de Hipótesis

Formulación de Hipótesis.

Especificación del Nivel de Significación(alfa).

Selección del estadístico de prueba.

Establecimiento del criterio de decisión.

Cálculos.

Toma de decisiones.

Las hipótesis estadísticas son

H_0: Oi=Ei
Los valores observados son iguales a los esperados

H_1: Oi≠Ei
Los valores observados son distintos a los esperados

$$\chi^2 = \sum_{i=1}^{k} \frac{(obs_i - esp_i)^2}{esp_i}$$

Ejemplo

	1	2	3	4	5	6
Observados	15	7	4	11	6	17
Esperados	10	10	10	10	10	10

$$X^2 = \frac{(15-10)^2}{10} + \frac{(7-10)^2}{10} + \cdots + \frac{(17-10)^2}{10} = 13{,}6$$

Este valor (13,6) tiene una distribución X^2 con 5 grados de libertad, porque K es 6. El P-valor=0,0344 < 0.05, por lo tanto rechazo H_0 y concluyo que el dado está cargado.

Ejemplo:

El dueño de una pizzería, desea conocer si el consumo de cerveza de los jóvenes entre 18 y 35 años, es el mismo ya sea hombre o mujer. Para ello dispone de registros de consumo de cerveza; Sobre un total de 960 consumidores, 514 eran hombres y 466 mujeres.

Se espera que esta proporción sea la misma. ¿Los datos nos permiten confirmar tal relación?

Con la prueba de concordancia, nos preguntamos si los datos permiten afirmar que los valores observados concuerdan con los valores esperado, según la presunción de igualdad.

El Estadístico de Prueba

$$\chi^2 = \sum_{i=1}^{k} \frac{(obs_i - esp_i)^2}{esp_i}$$

Este valor tiene a una distribución Chi cuadrado con v grados de libertad (v=k-1)

Concordancia:

Consumidores	Observado	Esperado	O-E	$\frac{(O-E)^2}{E}$
Hombres	514	480	34	2,41
Mujeres	446	480	-34	2,41
Total	960	960	0	4,82

Se debe calcular el P-Valor de (4,82) en una distribución de Chi cuadrado con 1 grado de libertad y compararlo con alfa. P-valor=0.02813 < 0.05, por lo tanto Rechazo H0 y concluyo que el consumo de cerveza difiere según el género.

PRUEBAS DE INDEPENDENCIA

A modo de ejemplo las pruebas surgen ante determinadas preguntas

Están relacionados los hábitos de lectura con el sexo del lector?

¿Están relacionadas las calificaciones obtenidas con el número de faltas?

¿Es independiente la opinión sobre la política exterior de la política partidista?

¿Es independiente el sexo de una persona de su preferencia en colores?

¿Está relacionado el sexo con tener una educación universitaria?

Las pruebas de Independencia

Permiten probar la hipótesis de que dos características o variables observadas en los mismos individuos son independientes.

Tablas de contingencia

Son tablas de frecuencias para 2 variables y sirven para organizar los datos y poder analizarlos .

Ejemplo de pruebas de independencia

En cierta Universidad, los ingresantes fueron clasificados en tres grupos principales, de acuerdo con la orientación de sus estudios secundarios en: escuelas de comercio, normal e industrial.

Se desea saber si la elección de sus carreras universitarias guarda alguna relación con el estudio secundario. Para ello, se tomaron datos de tres grandes facultades, tales como: económicas, humanidades e ingeniería.

Tabla de contingencia

La tabla contiene la cantidad de alumnos agrupados de acuerdo a la Facultad y a su estudio Secundario.

	Comercial	Normal	Industrial	Tota Marginal Fila
Economía	524	325	200	1049
Humanidades	435	256	254	945
Ingeniería	256	320	430	1006
Total Marginal Columna	1215	901	884	3000

Es una tabla de contingencia 3x3, ya que cada variable tiene 3 categorías.
Se pueden ver los totales marginales, los cuales se agregan para el cálculo de las frecuencias esperadas.

La hipótesis nula plantea que la elección de la carrera del ingresante es independiente de su orientación secundaria, la proporción de estudiantes inscriptos en las diferentes carreras es proporcional al total en relación con las escuelas de origen.

Frecuencias Observadas (Oij)

	Comer	Normal	Industrial	Total
Economía	524	325	200	1049
Humanidades	435	256	254	945
Ingeniería	256	320	430	1006
Total	1215	901	884	3000

Frecuencias Esperadas (Eij)

	Comer	Normal	Industrial	Total
Economía	424,9	315,1	309,1	1049
Humanidades	382,7	283,8	278,5	945
Ingeniería	407,4	302,1	296,4	1006
Total	1215	901	884	3000

Las frecuencias esperadas se calculan realizando el producto de los totales marginales de las correspondientes filas y columnas dividido el total general. Ejemplo 1049*1215/3000 =424,9

Veamos en un Experimento:

100 animales fueron tratados con un antibiótico y después de un tiempo se observó la presencia de síntomas de una enfermedad, se encontraron 88 animales saludables y 12 enfermos.
A otro grupo de 200 animales no se les suministró antibiótico y al cabo del mismo tiempo fueron examinados y se encontraron 143 saludables y 57 enfermos.

Tratamiento	***Saludable***	***Enfermo***	***Total***
c/antibiótico	88	12	100
No tratado	143	57	200
Total	231	69	300

TABLA DE CONTINGENCIA 2x2

Prueba de Independencia:

Hipótesis Nula: El tratamiento con antibiótico y la incidencia de la enfermedad son independientes.

Hipótesis Alternativa: El tratamiento con antibiótico y la incidencia de la enfermedad no son independientes.

$$\chi^2 = \sum_{i=1}^{k} \frac{(obs_i - esp_i)^2}{esp_i}$$

f=nro. filas; c=nro. columnas

i=1….k K= fxc

tiene a una distribución Chi cuadrado con v grados de libertad v=(f-1)x(c-1)

Valores observados

Tratamiento	*Saludable*	*Enfermo*	*Total*
c/antibiótico	88	12	100
No tratado	143	57	200
Total	231	69	300

Valores esperados bajo la hipótesis nula

Tratamiento	**Saludable**	**Enfermo**
c/antibiótico	(100x231)/300 77	(100x69)/300 23
No tratado	(200x231)/300 154	(200x69)/300 46

$$\chi^2 = \frac{(88-77)^2}{77} + \frac{(12-23)^2}{23} + \frac{(143-154)^2}{154} + \frac{(57-46)^2}{46} = 9.34$$

El P-valor de 9.34 en una *distribución de* X^2 con 1 grado de libertad es 0.0024, por lo tanto rechazamos H0 y concluimos que el antibiótico produce efecto.

Ejemplo

En una escuela primaria se tomó una muestra de 500 niños provenientes de cuatro grupos socioeconómicos diferentes con el objeto de evaluar si la presencia de cierto defecto en la pronunciación se distribuía de manera homogénea en los cuatro grupos. Los resultados son los siguientes:

	Grupo Socioeconómico				
	Superior	Medio-Superior	Medio-Inferior	Inferior	Total
Con defecto	8	24	32	27	91
Sin defecto	42	121	138	108	409
Total	50	145	170	135	500

TABLA DE CONTINGENCIA 2x4

Si la distribución fuera homogénea, se diría que la pronunciación es independiente de la condición socioeconómica.

Pruebas de Bondad de Ajuste

Se utilizan para verificar si datos empíricos siguen una distribución ajustada a un modelo teórico.

Una de las preguntas que se quieren responder es por ejemplo ¿ la variable altura tiene distribución normal?

Ejemplo

Se presenta un ejemplo donde se quiere saber si el peso de grupos de 100 semillas tiene una distribución normal. Para evaluar eso utilizaremos la prueba de Bondad de ajuste con el estadístico de Chi cuadrado.

Se cuenta con el peso de 400 bolsas con 100 semillas fue la siguiente:

La distribución de frecuencias del peso promedio en grs. de 400 grupos de 100 semillas fue la siguiente:

Peso (grs.)	Frecuencia
hasta 6,50	14
6,51-7,00	30
7,01-7,50	85
7,51-8,00	121
8,01-8,50	75
8,51-9,00	65
mas de 9,01	10

¿Se ajusta esta distribución a la distribución normal estándar?

Para averiguarlo debemos calcular las frecuencias relativas teóricas que corresponderían a la distribución normal estándar

$\bar{x} = 7{,}81$ $S_x = 0{,}68$

El procedimiento consiste en estandarizar cada extremo superior de los intervalos de la tabla de frecuencia, y buscar la probabilidad de cada intervalo en la distribución a la distribución normal estándar?

$$Z = \frac{X - \overline{X}}{S_x}$$

$$Z = \frac{6{,}5 - 7{,}81}{0{,}68} = -1{,}93$$

P(x<-1,93) = 0,027 $\Rightarrow$ hasta x = 6,5 la probabilidad (f(x)) = 0,027

$$Z = \frac{7{,}0 - 7{,}81}{0{,}68} = -1{,}19$$

P(-1.93<x<-1,19) = 0.09 $\Rightarrow$ de 6,51 a 7,0 la probabilidad es = 0,09

$$Z = \frac{7{,}5 - 7{,}81}{0{,}68} = -0{,}45$$

P(-1.19<x<-0,45) = 0,209 $\Rightarrow$ de 7,01 a 7,5 la probabilidad es =0,209

Con las probabilidades (frecuencias relativas) de cada intervalo estamos en condiciones de calcular las frecuencias absolutas teóricas (que serian los valores esperados que necesitamos para la prueba). Para obtener las frecuencias absolutas debemos multiplicar las frecuencias relativas de cada intervalo por el número total de observaciones, que en este caso es 400.

Tabla para el cálculo de Valores Esperados

Intervalos de clase	Frecuencias relativas	Frecuencias Teóricas
hasta 6,50	0,027	0,027*400= 10,8
6,51-7,00	0,090	36,0
7,01-7,50	0,209	83,6
7,51-8,00	0,284	113,6
8,01-8,50	0,234	93,6
8,51-9,00	0,112	44,8
mas de 9,01	0,044	17,6

Tabla para el cálculo de Chi Cuadrado

Intervalos de clase	Observados	Esperados	O-E	$(O-E)^2/E$
hasta 6,50	14	10,8	3,2	0,948
6,51-7,00	30	36,0	-6,0	1,000
7,01-7,50	85	83,6	1,4	0,023
7,51-8,00	121	113,6	7,4	0,482
8,01-8,50	75	93,6	-18,6	3,696
8,51-9,00	65	44,8	20,0	9,108
mas de 9,01	10	17,6	-7,6	3,282
Total	400	400,0		18,539

Regla de decisión

X^2= 18,539 P-Valor (18.539)=0.002

Conclusión: se rechaza la H0, es decir la distribución no se ajusta a la normal.

Nota: para pruebas de bondad de ajuste, los grados de libertad se calculan restando a n-1, un grado de libertad por cada parámetro de la distribución a la que se ajusta, en este caso la normal tiene 2 parámetros μ y σ, por lo tanto al número de categorías menos 1 (7-1) se le resta 2.

Resumiendo

- Las pruebas de Chi-cuadrado analizan la concordancia entre los valores observados en una muestra o población y los esperados según alguna ley o condición preestablecida.

- Son aplicables a principalmente a variables cualitativas y en menor medida a variables cuantitativas (con pérdida de información).

- Permiten probar hipótesis referidas a frecuencias con que se presentan los diferentes valores de las variables en estudio.

FIN