Biometría

Análisis de datos categóricos

Introducción

Muchos estudios resultan en datos que son categóricos o cualitativos antes que cuantitativos y que admiten **más de dos resultados posibles**:

- Pacientes clasificados según evolución (mejora, sin cambios, empeora)
- Individuos clasificados según estadio (larva, pupa, imago)
- Votantes clasificados según intención de voto

Estos datos tienen las características de un experimento **multinomial**

Ejemplo: grupos sanguíneos

- La distribución en Buenos Aires de los grupos sanguíneos es de un 35%, 10%, 6% y un 49% para los grupos A, B, AB y O respectivamente.
- Se desea saber si la distribución de los grupos sanguíneos en la provincia de Formosa difiere de la de Buenos Aires

El experimento multinomial

- El experimento consiste de n ensayos idénticos
- El resultado de cada repetición es una de **k categorías**
- La probabilidad de que el resultado sea una determinada categoría i se denomina π_i y permanece constante de ensayo en ensayo
- La suma de las k probabilidades, $\pi_1 + \pi_2 + .. + \pi_k = 1$
- Los ensayos son **independientes**

El experimento binomial

- Es un caso especial del experimento multinomial con k = 2
- Las 2 categorías se denominan éxito y fracaso
- π_1 y π_2 son π y 1- π
- Nosotros hacemos inferencia sobre π (y 1- π)
- En un experimento multinomial hacemos inferencia sobre todas las probabilidades, $\pi_1, \pi_2, \ldots, \pi_k$

Pruebas de bondad de ajuste

- Se mide **una única variable categórica**, por lo tanto cada elemento de la población se asigna a una y sólo una de varias categorías k
- Para cada categoría se posee un valor **preconcebido o supuesto o histórico** de π_i y usamos información muestral para determinar si dichos valores son correctos

- Para determinar saber si la distribución de los grupos sanguíneos en Formosa difiere de la de Buenos Aires se extrajo una muestra aleatoria de 200 formoseños y se les determinó el grupo sanguíneo.
- Los resultados fueron:

Grupo A	Grupo B	Grupo AB	Grupo 0
61	15	6	118

- En este caso, la población es **multinomial**: cada formoseño se clasifica según su grupo sanguíneo en una de 4 categorías (k= 4)

frecuencias observadas FO_i

¿La distribución difiere?

- Dado que se cuenta solo con una muestra y se desea inferir sobre toda la población, la pregunta se resuelve mediante una **prueba de hipótesis**
- Las hipótesis puestas a prueba son:

Ho: Las proporciones de cada grupo sanguíneo en Formosa no difieren de las de Bs As; $\pi_1=0.35, \pi_2= 0.10, \pi_3=0.06, \pi 4=0.49$

H1: Las proporciones sí difieren; al menos una π_i cambia

- ¿Cómo se resuelve?
- Se contrastan **frecuencias observadas** FO_i en la muestra con las **frecuencias que se esperaría observar** FE_i si las proporciones no cambiasen (es decir si Ho fuera verdadera)

- Se calculan las frecuencias esperadas:

$$E_i = np_i$$

	Grupo A	Grupo B	Grupo AB	Grupo 0	**TOTAL**
FO_i	61	15	6	118	**200**
P_i	0.35	0.10	0.06	0.49	**1**
FE_i	200x0.35 = **70**	200x0.10 = **20**	200x0.06 = **12**	200x0.49 = **98**	**200**

- ¿Las diferencias son **lo suficientemente grandes** como para afirmar que las proporciones de la población son diferentes a las de Buenos Aires? (α = 0.05)

Estadístico chi-cuadrado

- Para cuantificar las diferencias en un único número se utiliza el estadístico

$$\chi^2_{muestral} = \sum \frac{(FO_i - FE_i)^2}{FE_i}$$

- Cuando Ho es verdadera, las diferencias entre FO_i y FE_i serán pequeñas, pero cuando Ho es falsa, serán grandes
- Para determinar si la discrepancia entre FO y FE es lo suficientemente grande, se utiliza la distribución chi-cuadrado con cierta cantidad de grados de libertad
- Sin embargo este estadístico tiene una distribución que se **aproxima** a la chi-cuadrado

Grados de libertad

- Varían según la aplicación
- Se comienza con el número de categorías o celdas k
- Se le resta un GL por cada restricción sobre las probabilidades (siempre se perderá un GL ya que $\pi_1 + \pi_2 + \ldots + \pi_k = 1$)
- Se pierde un GL por cada parámetro que se debe estimar para calcular FE_i
- Es decir

$$GL = k - 1 - m$$

$$siendo\ k = cantidad\ de\ categorías$$

$$m = cantidad\ de\ parámetros\ estimados\ \ para\ calcular\ las\ FE$$

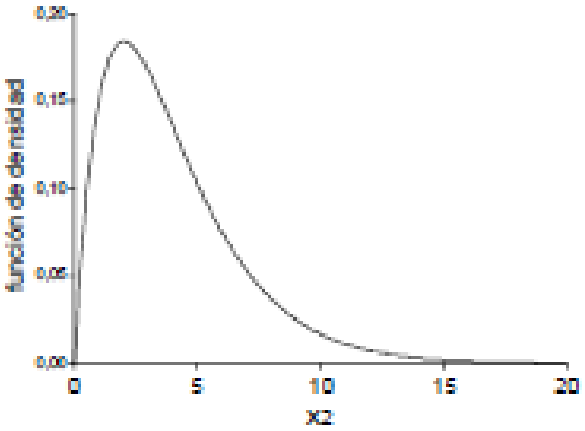
En el ejemplo:

	Grupo A	Grupo B	Grupo AB	Grupo 0	**TOTAL**
FO_i	61	15	6	118	**200**
FE_i	70	20	12	98	**200**

$$\chi^2_{muestral} = \sum \frac{(FO_i - FE_i)^2}{FE_i}$$

$\chi^2_{muestral}$ = 1,16 + 1,25 + 3,00 + 4,08 = 9,49

Conclusión:



Comentarios

- Para que las conclusiones sean válidas:
  - La muestra debe ser aleatoria y su tamaño n debe ser ≥50
  - Las observaciones deben ser independientes
  - Las FE_i deben ser > 0. Y se admite solo un 20% de casillas con FE_i < *5.* Si esto no se cumple, puede solucionarse agrupando categorías.
- La distribución del estadístico es aproximada, pero si el tamaño de la muestra es grande (FE > 10) la aproximación es muy buena

Comentarios

- A diferencia de las pruebas anteriores, la Ho indica que existe buen ajuste a un modelo o a ciertas proporciones supuestas:

Ho: el modelo es correcto, hay buen ajuste a las proporciones supuestas

H1: el modelo no es correcto, hay mal ajuste

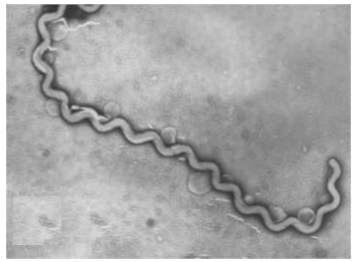
Otras aplicaciones

- Las pruebas de bondad de ajuste pueden utilizarse para determinar si una variable ajusta a una determinada distribución de probabilidades, como por ejemplo:
  - Normal
  - Binomial
  - Poisson
- En estos casos se deben estimar algunos parámetros a partir de la muestra:
  - Normal: el promedio μ y el desvío estándar σ
  - Binomial: la probabilidad de éxito π
  - Poisson: la cantidad esperada de eventos en un continuo λ

Tablas de contingencia

- El investigador mide dos variables cualitativas, de manera tal que los eventos son clasificados según dos criterios:
  - Personas clasificadas según intención de voto y nivel socioeconómico
  - Pacientes clasificados según presencia de daltonismo y género
  - Individuos clasificados según estadio (larva, pupa, imago) y según respuesta al tratamiento (sobreviven o no sobreviven)
- Los datos son resumidos en tablas de doble entrada (o de contingencia), donde en cada cruce se indican las frecuencias

Ejemplo: ¿la prevalencia de leptospirosis en población canina depende de los hábitos?



- La leptospirosis es una enfermedad infecciosa que afecta a diversos animales. El hombre puede ser huésped accidental. La principal fuente de contagio son las ratas, aunque se están estudiando otros animales domésticos, como los perros, como vectores.
- Se desea analizar la relación entre los hábitos de salida de los perros y la prevalencia de leptospirosis
- Se eligen al azar 278 perros en hogares de Florencio Varela y se clasifican según:

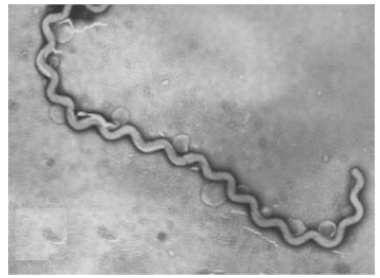
	No sale	1 salida diaria	Más de 1 salida diaria	TOTAL
Positivo	12	47	99	**158**
Negativo	21	35	64	**120**
TOTAL	**33**	**82**	**163**	**278**

	No sale	1 salida diaria	Más de 1 salida diaria	TOTAL
Positivo	12	47	99	**158**
Negativo	21	35	64	**120**
TOTAL	**33**	**82**	**163**	**278**

- La tabla de doble entrada
  posee F filas y C columnas (FxC)

- Se estudia la relación entre las dos variables: ¿es un método de clasificación **dependiente** del otro?
- O dicho de otra manera: ¿la distribución de los casos en las categorías de una variable dependen o cambian según la categoría de la otra variable que está siendo observada? Si la respuesta es no, entonces las variables son **independientes**

Prueba de independencia



- ¿la presencia de leptospirosis en perros en hogares **depende** de los hábitos de salida?

- Las hipótesis puestas a prueba son:

  Ho:

  H1:

- ¿Cómo se resuelve?
- Se contrastan **frecuencias observadas** en la muestra con las **frecuencias que se esperaría observar** si Ho fuese verdadera

	No sale	1 salida diaria	Más de 1 salida diaria	TOTAL
Positivo	12	47	99	**158**
Negativo	21	35	64	**120**
TOTAL	**33**	**82**	**163**	**278**

frecuencias observadas FO

$$P(positivo) =$$

$$P(no\ salidas) =$$

$$Si\ los\ sucesos\ son\ independientes \Rightarrow P(positivo\ y\ no\ salidas) =$$

- Se calculan las frecuencias esperadas: $E_i = np_i$

$$FE(positivo\ y\ no\ salidas) =$$

	No sale	1 salida diaria	Más de 1 salida diaria	TOTAL
Positivo	18,76	46,60	92,64	**158**
Negativo	14,24	35,40	70,36	**120**
TOTAL	**33**	**82**	**163**	**278**

frecuencias esperadas FE

Estadístico chi-cuadrado

- Para cuantificar las diferencias en un único número se utiliza el mismo estadístico que en el ejemplo anterior

$$\chi^2_{muestral} = \sum \frac{(FO_i - FE_i)^2}{FE_i}$$

- Cuando Ho es verdadera, las diferencias entre O_i y E_i serán pequeñas, pero cuando Ho es falsa, serán grandes
- Para determinar si la discrepancia entre O y E es lo suficientemente grande, se utiliza la distribución chi-cuadrado con GL = (F-1)(C-1)
- Cuando la distribución del estadístico es aproximada, si el tamaño de la muestra no es muy grande (5<FE<10) y la tabla es de 2x2 se puede aplicar la **corrección de Yates**

En el ejemplo:

FO

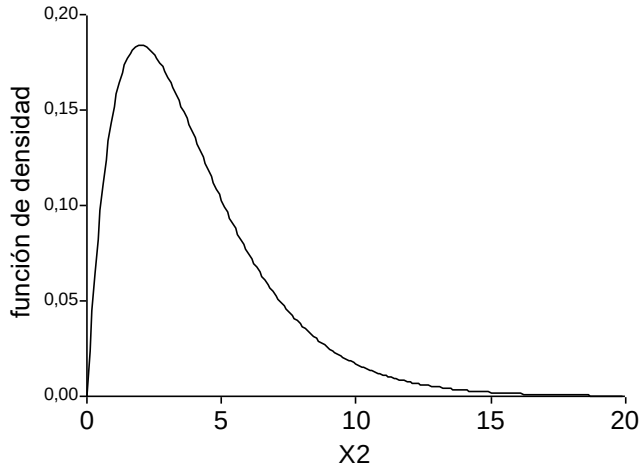
	No sale	1 salida diaria	Más de 1 salida diaria	TOTAL
Positivo	12	47	99	**158**
Negativo	21	35	64	**120**
TOTAL	**33**	**82**	**163**	**278**

FE

	No sale	1 salida diaria	Más de 1 salida diaria	TOTAL
Positivo	18,76	46,60	92,64	**158**
Negativo	14,24	35,40	70,36	**120**
TOTAL	**33**	**82**	**163**	**278**

$$\chi^2_{muestral} = \sum \frac{(FO_i - FE_i)^2}{FE_i} =$$

$$= 6{,}66$$



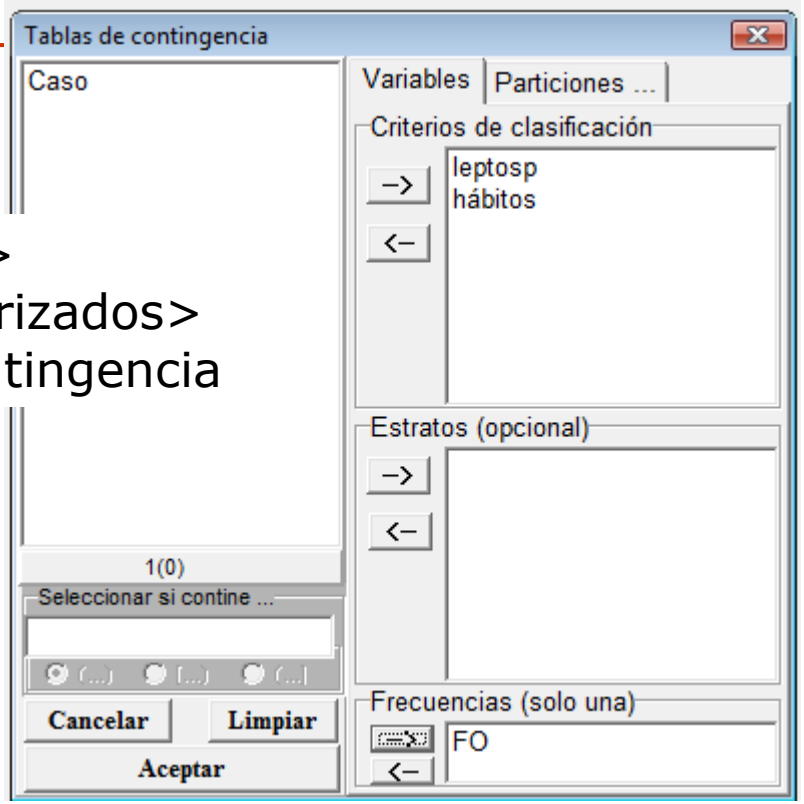
Conclusión:

En Infostat:

chi2

Caso	leptosp	hábitos	FO
1	positivo	nunca	12
2	positivo	una salida	47
3	positivo	más de una	99
4	negativo	nunca	21
5	negativo	una salida	35
6	negativo	más de una	64

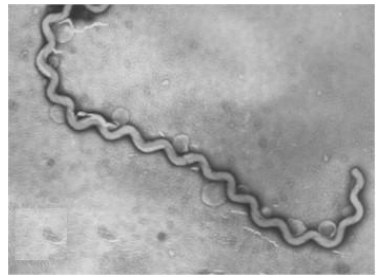
Estadísticas >
Datos categorizados>
Tablas de contingencia



Estadístico	Valor	gl	p
Chi Cuadrado Pearson	6,66	2	0,0359
Chi Cuadrado MV-G2	6,61	2	0,0367
Coef.Conting.Cramer	0,11		
Coef.Conting.Pearson	0,15		

Existe otro estadístico que se utiliza en estas pruebas, con la misma distribución de probabilidades, denominado **G** de máxima verosimilitud

Explorando los resultados cuando la prueba dio significativa



Frecuencias relativas por filas (expresadas en porcentajes)
En columnas:hábitos

leptosp	nunca	una salida	más de una	Total
positivo	7,59	29,75	62,66	100,00
negativo	17,50	29,17	53,33	100,00
Total	11,87	29,50	58,63	100,00

Frecuencias relativas por columnas (expresadas en porcentajes)
En columnas:hábitos

leptosp	nunca	una salida	más de una	Total
positivo	36,36	57,32	60,74	56,83
negativo	63,64	42,68	39,26	43,17
Total	100,00	100,00	100,00	100,00

- Se puede estimar la prevalencia general de leptospirosis canina
- Se puede estimar la prevalencia según hábitos
- Se puede estimar riesgo relativo

Comentarios

- Si existiesen diferencias, podrían explorarse los datos para concluir acerca de la naturaleza de las mismas

- Al igual que en el caso anterior, para que las conclusiones sean válidas existen ciertos supuestos que deben cumplirse:
  - La muestra debe ser aleatoria y su tamaño n debe ser ≥50
  - Las observaciones deben ser independientes
  - Las FE_i deben ser > 0. Y se admite solo un 20% de casillas con $FE_i < 5$. Si esto no se cumple, puede solucionarse agrupando categorías.

- Cuando la tabla es de 2 x 2, la prueba es equivalente a la prueba Z para diferencia de dos proporciones bilateral

Otras aplicaciones

- En las tablas de doble entrada se observan generalmente dos situaciones:
  - Que ninguno de los totales está predeterminado
  - Que uno de los totales esté fijado de antemano

	No sale	1 salida diaria	Más de 1 salida diaria	TOTAL
Positivo	12	47	99	**158**
Negativo	21	35	64	**120**
TOTAL	**33**	**82**	**163**	**278**

- En el primer caso, se trata de una prueba de independencia
  - Ho: la variable X es independiente de la variable Y
  - Es posible estimar proporciones usando los marginales

- En el segundo, de una prueba de homogeneidad
  - Ho: las i categorías (totales fijos) son homogéneas con respecto a…
    o bien la k proporciones son iguales ($\pi_1 = \pi_2 = \ldots = \pi_k$)
  - No tiene sentido estimar proporciones con respecto a los marginales, ya que no varían libremente

¿Independencia u homogeneidad?

- Para determinar si la prevalencia del parasitismo por tordo en nidos de calandrias varía según la ubicación de los nidos se seleccionaron 40 nidos de calandria en ambientes urbanos y 50 en suburbanos y se determinó la presencia de huevos de tordo
  - Ho:

- En una encuesta preelectoral, se tomó una muestra de 500 individuos que fueron clasificados según su intención de voto y su nivel educativo
  - Ho:

Prueba exacta de Fisher

- Se utiliza cuando:
  - los tamaños de muestra son pequeños (n<50) o FE<5
  - La tabla es de 2 x 2
- Se basa en la distribución hipergeométrica
- Permite establecer con exactitud el correspondiente P-valor, y no de manera aproximada como es el caso cuando se recurre a la distribución χ^2.