Capítulo 1

Tablas de Contingencia

1.1. Introducción

Las tablas de contingencia son utilizadas principalmente para resumir datos categóricos, dentro de ellas se van a encontrar las frecuencias observadas entre dos variables de interés; el término de tablas de contingencia fue introducido por Pearson en 1904.

El interés de estudio de las tablas de contingencia se centra en saber si hay asociación entre dos variables, dicho esto si se logra determinar asociación entonces lo que nos preguntaríamos ahora sería ¿qué tanta es la intensidad de asociación entre estas variables?. Esta pregunta la resolveremos con la teoria que desarrollemos en este tema.

Sean X y Y dos variables categóricas con K y P categorías respectivamente. Un sujeto puede venir clasificado en una de las $K \times P$ categorías, que es el número posible de categorías existentes. La tabla de contingencia para las variables X y Y con K renglones y P columnas:

	Y_1	Y_2	$\ldots$	Y_P
X_1	O_{11}	O_{12}	$\ldots$	O_{1P}
X_2	O_{21}	O_{22}	$\ldots$	O_{2P}
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
X_K	O_{K1}	O_{K2}	$\ldots$	O_{KP}

Donde O_{ij} expresa la frecuencia absoluta observada en las modalidades X_i y Y_j. La misma tabla puede expresarse en frecuencias relativas o proporciones en cuyo caso solo hay que dividir

cada casilla O_{ij} entre el total N, donde:

$$N = \sum_{i=1}^{K} \sum_{j=1}^{P} O_{ij}$$

Por ejemplo, considere una muestra de pacientes de un hospital, donde:

- $X \equiv$ Se toma aspirina o placebo ($K = 2$).
- $Y \equiv$ Se sufre ataque cardíaco o no ($P = 3$).

Se tienen una muestra de 22,071 pacientes, a 11,034 pacientes se les suministro placebo y a los demás se les suministro aspirina, lo que se busca con este estudio es ver si alguno de estos tratamientos tiene como consecuencia un ataque cardíaco; para los pacientes que se les dio placebo 189 sufieron ataque cardíaco, de los cuales 18 murieron; por otra parte a los que tomaron aspirina 10,933 no sufrieron ataque al corazón y 5 murieron por un ataque.

Entonces la tabla de contingencia para este caso queda de la siguiente manera:

	Ataque mortal	Ataque no mortal	No ataque
Placebo	18	171	10845
Aspirina	5	99	10933

1.2. Características de la tabla de contingecia

Para poder determinar la relación entre dos variables categóricas necesitamos conocer su densidad conjunta, asi mismo esta densidad nos ayudará a determinar las densidades marginales y condicionales.

Definición 1.2.1 (Función de densidad conjunta)**.** *Sean X y Y variables aleatorias discretas, de ciertas distribuciones $F_X(x)$ y $F_Y(y)$ respectivamente, se define la función de densidad conjunta como:*

$$f_{XY}(i, j) = \mathbb{P}(X = i, Y = j)$$

con $i = 1, \ldots, K$ y $j = 1, \ldots, P$.

En la tabla de contingencia $f_{XY}(i, j)$ puede ser estimada por O_{ij} si la tabla contiene frecuencias relativas, en caso de tener frecuencias absolutas, entonces $f_{XY}(i, j)$ se estimaria como $\frac{O_{ij}}{N}$

Definición 1.2.2 (Función de densidad marginal)**.** *Sean X y Y variables aleatorias discretas, con función de densidad conjunta $f_{XY}(x, y)$, se definen las funciones de densidad marginal como:*

Densidad marginal de X:

$$f_X(i) = m_i = \mathbb{P}(X = i) = \sum_{j=1}^{P} \mathbb{P}(X = i, Y = j)$$

Densidad marginal de Y:

$$f_Y(j) = n_j = \mathbb{P}(Y = j) = \sum_{i=1}^{K} \mathbb{P}(X = i, Y = j)$$

Las funciones marginales cumplen con:

$$\sum_{i=1}^{K} f_X(i) = \sum_{j=1}^{P} f_Y(j) = \sum_{i=1}^{K} \sum_{j=1}^{P} f_{XY}(i, j) = 1$$

En la tabla de contingencia frecuencias relativas, las desidades marginales pueden estimarse de la siguiente manera:

$$f_X(i) = \hat{m}_i = \sum_{j=1}^{P} O_{ij}$$

$$f_Y(j) = \hat{n}_j = \sum_{i=1}^{K} O_{ij}$$

Definición 1.2.3 (Función de densidad condicional)**.** *Sean X y Y variables aleatorias, con función de densidad conjunta $f_{XY}(x, y)$, se definen las funciones de densidad condicional como:*

Densidad condicional de X respecto a Y:

$$f_{X|Y}(i \mid j) = \mathbb{P}(X = i \mid Y = j) = \frac{f_{XY}(i, j)}{f_Y(j)}$$

Densidad condicional de Y respecto a X:

$$f_{Y|X}(j \mid i) = \mathbb{P}(Y = j \mid X = i) = \frac{f_{XY}(i, j)}{f_X(i)}$$

Las funciones condicionadas cumplen con:

$$\sum_{i=1}^{K} f_{X|Y}(i \mid j) = 1$$

$$\sum_{j=1}^{P} f_{Y|X}(j \mid i) = 1$$

Para efectos de este curso consideraremos a Y como variable respuesta de X, por lo que consideraremos muchas veces la densidad condicional de Y respecto a X. Entonces sabremos que **que no hay relación entre X y Y** si hay *Independencia* entre las variables.

Definición 1.2.4 (Independecia)**.** *Sean X y Y variables aleatorias, con función de densidad conjunta $f_{XY}(x, y)$, se dice que las variables son independientes si:*

$$f_{XY}(i, j) = f_X(i) f_Y(j)$$

Lo que tiene como implicación directa que:

$$f_{Y|X}(j \mid i) = \frac{f_{XY}(i, j)}{f_X(i)} = \frac{f_X(i) f_Y(j)}{f_X(i)} = f_Y(j)$$

Veamos el siguiente ejemplo. Doce individuos (muestra) se clasificaron según el sexo (hombre, mujer) y su deseo de ver o no una final de campeonato de fútbol que será televisada:

Sexo	Fútbol
Hombre	Sí
Mujer	No
Hombre	Sí
Hombre	No
Hombre	Sí
Mujer	No
Mujer	No
Mujer	Sí
Hombre	Sí
Hombre	Sí
Hombre	Sí
Mujer	No

```r
datos=read.csv('ejemplo1.csv')
datos
```

```
##    sexo futbol
## 1     h      1
## 2     m      0
## 3     h      1
## 4     h      0
## 5     h      1
## 6     m      0
## 7     m      0
## 8     m      1
## 9     h      1
## 10    h      1
## 11    h      1
## 12    m      0
```

```r
datos=data.frame(datos)
datos$sexo=factor(datos$sexo, labels=c("Hombre", "Mujer")) #Declara factor con etiquetas
datos$futbol= factor(datos$futbol, labels=c("No","Si")) #Declara factor con etiquetas
datos
```

```
##      sexo futbol
## 1  Hombre     Si
## 2   Mujer     No
## 3  Hombre     Si
## 4  Hombre     No
## 5  Hombre     Si
## 6   Mujer     No
## 7   Mujer     No
## 8   Mujer     Si
## 9  Hombre     Si
## 10 Hombre     Si
## 11 Hombre     Si
## 12  Mujer     No
```

```r
#Tabla de contingencia (con frecuencias absolutas):
ftable(datos)
```

```
##        futbol No Si
## sexo
## Hombre         1  6
## Mujer          4  1
```

```r
#Total de datos
sum(ftable(datos))
```

```
## [1] 12
```

```r
#Tabla de contingencia (con frecuencias relativas):
prop.table(table(datos))
```

```
##         futbol
## sexo             No         Si
##   Hombre 0.08333333 0.50000000
##   Mujer  0.33333333 0.08333333
```

```r
#Marginales:
td=prop.table(table(datos))
addmargins(td)
```

```
##         futbol
## sexo             No         Si        Sum
##   Hombre 0.08333333 0.50000000 0.58333333
##   Mujer  0.33333333 0.08333333 0.41666667
##   Sum    0.41666667 0.58333333 1.00000000
```

```r
#Condicional de futbol dado el sexo
prop.table(table(datos),1)
```

```
##         futbol
## sexo            No        Si
##   Hombre 0.1428571 0.8571429
##   Mujer  0.8000000 0.2000000
```

```r
#Condicional del sexo dado el futbol
prop.table(table(datos),2)
```

```
##         futbol
## sexo            No        Si
##   Hombre 0.2000000 0.8571429
##   Mujer  0.8000000 0.1428571
```

1.3. Estudio de la asociación

Sean X e Y dos variables categóricas, con i=1,. . . ,k y j=1,. . . ,p modalidades o categorías, respectivamente, presentadas en una tabla de $k \times p$. Nos interesa ver si hay relación entre las variables categóricas X y Y, por lo que nos planteamos la siguiente prueba de hipótesis:

H_0: Las variables categóricas son independientes.

vs

H_1: Las variables categóricas no son independientes

Dicho de otra forma, la hipótesis nula se puede ver en términos de la tabla de contingencia como: el evento "de la observación perteneciente al i-ésimo reglón" es independiente del evento "de la misma observación que pertenece a la j-ésima columna" $\forall\, i, j$.

La proposición anterior puede traducirse en términos probabilísticos de la siguiente forma:

Sea P_i la probabilidad de pertenecer al i-ésimo renglón y P_j la probabilidad de pertenecer a la j-ésima columna.

Esto tiene como consecuencia que la hipótesis nula se transforme de la siguiente manera:

$$H_0 : P_{ij} = P_i P_j \quad \forall\, i, j$$

$$vs$$

$$H_1 : P_{ij} \neq P_i P_j \quad p.a.\, i, j$$

1.4. Prueba χ^2

Se define la frecuencia esperada E_{ij} correspondiente a la celda con coordenadas (i, j) como el producto de la suma por renglón por la suma por columna asociada a la celda de referencia, dividida entre el total de observaciones.

$$E_{ij} = N P_i P_j = N \left(\frac{m_i}{N}\right)\left(\frac{n_j}{N}\right) = \frac{m_i n_j}{N}$$

Hagamos una observación, en el caso de las tablas de contingencia no sabemos cual es la distribución exacta de nuestros datos, por lo cual no conocemos exactamente las probabilidades $(\mathrm{P}_i P_j), entoncesprocederemosatrabajarconlosestimadores, esdecirestimaremoslaesperanzadelosestimado$

$$\hat{E}_{ij} = N \hat{P}_i \hat{P}_j = N \left(\frac{\hat{m}_i}{N}\right)\left(\frac{\hat{n}_j}{N}\right) = \frac{\hat{m}_i \hat{n}_j}{N} \approx E_{ij}$$

Bajo H_0 la estadística de prueba χ^2 se denota por T y su fórmula de cálculo es:

$$T = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

La distribución de T bajo H_0 se aproxima a una $\chi^2_{((p-1)(k-1))}$.

Especial atención merecen las tablas de contingencia de 2×2. Para este caso la fórmula para el cálculo de T se puede simplificar con la siguiente expresión.

$$T = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{m_1 m_2 n_1 n_2}$$

Sin embargo, es posible mejorar el cálculo de T con la corrección de $Yates$ para la continuidad, ya que en esencia se aproxima la distribución de una variable aleatoria continua, como lo es la χ^2 , a una estructura discreta correspondiente a las frecuencias de la tabla.

La introducción de la correcci?n para la continuidad cambia la expresión de T de la forma siguiente:

$$T = \frac{N(\mid O_{11}O_{22} - O_{12}O_{21} \mid -\frac{N}{2})^2}{m_1 m_2 n_1 n_2}$$

Luego entonces, se rechaza H_0 si T es grande, es decir:

$$T > \chi^{2^{(1-\alpha)}}_{((p-1)(k-1))}$$

Veamos un ejemplo sencillo aplicado en R de esta prueba.

```
# Prueba de Chi-cuadrado para tablas de contingencia
# Creacin de los datos de sexo
sexo <- c(rep("H",52), rep("M",48))

# Creacin de los datos zurdo y diestro
dz <- c(rep("d",43),rep("z",9),rep("d",44),rep("z",4))

# Elaboracin de la tabla sexo/diestro-zurdo
tabla <- table(sexo,dz)
tabla
```

```
##     dz
## sexo  d  z
##    H 43  9
##    M 44  4
```

```r
test <- chisq.test(tabla)
test

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tabla
## X-squared = 1.0725, df = 1, p-value = 0.3004

# Dado que p > 0.05 no hay diferencias entre hombres y mujeres
# en cuanto a ser diestro o zurdo

# Tambin se pueden introducir directamente las frecuencias y los factores
test2 <- chisq.test(sexo,dz)
test2

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sexo and dz
## X-squared = 1.0725, df = 1, p-value = 0.3004
```

1.5. Otras Estadísticas e índices complementarios

1.5.1. El Coeficiente de Contingencia

Para medir el grado de relación o dependencia de las clasificaciones de una tabla de contingencia se utiliza el coeficiente de contingencia (CC).

$$CC = \sqrt{\frac{T}{T+N}}$$

Un valor alto de CC significa un alto grado de asociación, cuyo valor máximo está determinado por el número de filas y columnas de la tabla. Está acotado entre 0 y 1. Un valor cercano a 1 refleja una fuerte asociación entre las variables. En el caso de tablas cuadradas si consideramos a R como el número de renglones su cota superior es:

$$Cota = \sqrt{\frac{R}{R+1}}$$

1.5.2. La Prueba exacta de Fisher

Una popular prueba no paramétrica aplicable a tablas de contingencia 2×2 es la prueba exacta de Fisher. la prueba se apoya en el cálculo de la probabilidad exacta de observar un conjunto particular de frecuencias en una tabla 2×2, cuando los totales marginales se consideran fijos.

Si se considera la tabla:

O_{11}	O_{12}	m_1
O_{21}	O_{22}	m_2
n_1	n_2	N

La probabilidad exacta se obtiene mediante una distribución hipergeométrica

$$P = \frac{\binom{n_1}{O_{11}}\binom{n_2}{O_{12}}}{\binom{N}{m_1}}$$

Esta forma de calcular probabilidades es generalizable a tablas $2 \times p$.

$$P = \frac{\binom{n_1}{O_{11}}\binom{n_2}{O_{12}}\cdots\binom{n_p}{O_{1p}}}{\binom{N}{m_1}}$$

Estas probabilidades se usan para calcular valor de la p asociado al test exacto de Fisher. Este valor de p indicará la probabilidad de obtener una diferencia entre los grupos mayor o igual a la observada, bajo la hipótesis nula de independencia. Si esta probabilidad es pequeña ($p < 0.05$) se deberá rechazar la hipótesis de partida y deberemos asumir que las dos variables no son independientes, sino que están asociadas. En caso contrario, se dirá que no existe evidencia estadística de asociación entre ambas variables.

Veamos un ejemplo sencillo aplicado en R de esta prueba.

```
# Test exacto de Fisher para tablas de contingencia
# Creacin de los datos de sexo
sexo <- c(rep("H",52), rep("M",48))

# Creacin de los datos zurdo y diestro
dz <- c(rep("d",43),rep("z",9),rep("d",44),rep("z",4))

# Elaboracin de la tabla sexo/diestro-zurdo
tabla <- table(sexo,dz)
tabla
```

```
##     dz
## sexo  d  z
##    H 43  9
##    M 44  4

test <- fisher.test(tabla)
test

##
##  Fisher's Exact Test for Count Data
##
## data:  tabla
## p-value = 0.2392
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.09150811 1.71527769
## sample estimates:
## odds ratio
##  0.4378606

# Dado que p > 0.05 no hay diferencias entre hombres y mujeres
# en cuanto a ser diestro o zurdo

# Tambin se pueden introducir directamente las frecuencias y los factores
test2 <- fisher.test(sexo,dz)
test2

##
##  Fisher's Exact Test for Count Data
##
## data:  sexo and dz
## p-value = 0.2392
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.09150811 1.71527769
## sample estimates:
## odds ratio
##  0.4378606
```

Por último veamos un ejemplo má de tablas de contingencia.

La siguiente tabla representa un grupo de 11047 trabajadores clasificados segń el SALARIO que perciben y el tipo de TRABAJO que estos realiza, (ya sea manual o intelectual), se quiere

ver si existe dependencia entre las variables de edad y salario dependiendo del tipo de trabajo.

Salario	*T. Manual*			*T. Intelectual*		
(*Miles*)	$18-25$	$25-35$	$35-65$	$18-25$	$25-35$	$35-65$
$20-50$	165	644	1800	170	378	332
$50-100$	168	672	1763	234	757	664
$100-150$	17	84	187	21	757	2234

```
d=read.csv('ejemplo2.csv')
d
```

```
##    Salario Edad Tipo_T N_Personas
## 1       35 21.5      1        165
## 2       75 21.5      1        168
## 3      125 21.5      1         17
## 4       35 30.0      1        644
## 5       75 30.0      1        672
## 6      125 30.0      1         84
## 7       35 50.0      1       1800
## 8       75 50.0      1       1763
## 9      125 50.0      1        187
## 10      35 21.5      2        170
## 11      75 21.5      2        234
## 12     125 21.5      2         21
## 13      35 30.0      2        378
## 14      75 30.0      2        757
## 15     125 30.0      2        757
## 16      35 50.0      2        332
## 17      75 50.0      2        664
## 18     125 50.0      2       2234
```

```
tab1=ftable(xtabs(N_Personas ~ Edad+Salario, subset= Tipo_T== 1 , data = d))
tab1
```

```
##      Salario   35   75  125
## Edad
## 21.5          165  168   17
## 30            644  672   84
## 50           1800 1763  187
```

```
chisq.test(tab1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 3.2136, df = 4, p-value = 0.5227

##La tabla anterior muestra que para el grupo de trabajadores "manual" no existe
##asociacion entre salario y edad.
##Tal como muestra el p-valor correspondiente al contraste de hipotesis nula
##de salario y edad.
##La relacion entre salario y edad presenta un nivel de significancia igual a 0,523 > 0,05.
##No se rechaza la hipotesis de independencia del salario y edad para los trabajadores
##de la categora manual.
tab2=ftable(xtabs(N_Personas ~ Edad+Salario, subset= Tipo_T== 2 , data = d))
tab2

##      Salario   35   75  125
## Edad
## 21.5          170  234   21
## 30            378  757  757
## 50            332  664 2234

chisq.test(tab2)

##
##  Pearson's Chi-squared test
##
## data:  tab2
## X-squared = 882.5, df = 4, p-value < 2.2e-16

##Por el contrario, si se aprecia fuerte relacion entre las variables sexo y salario
##para el tipo de trabajador "intelectual", tal como muestra el resultado anterior.
##Aqui la relacion es altamente significativa,
##el nivel de significancia es (2.2e-16) la cual nos permite rechazar la hipotesis nula.


##Continuemos el analisis mostramos las tablas condicionadas relativas a salario y edad,
##distinguiendo por tipo de trabajador.

##Tabla condicinada de salario a edad para el grupo de trabajadores de tipo manual.
prop.table(tab1,1)

##      Salario         35         75        125
```

```
## Edad
## 21.5         0.47142857 0.48000000 0.04857143
## 30           0.46000000 0.48000000 0.06000000
## 50           0.48000000 0.47013333 0.04986667

##Aqui podemos ver que la asociacion entre edad y salario no es significativa, pues si
##esta fuera significativa entonces al paso del tiempo el salario aumentaria o disminuiria
##pero en este caso no es asi, ya que es indiferente la edad para este tipo de trabajo
##por lo cual podemos ver que no hay relacion entre estas variables.

##Tabla condicinada de salario a edad para el grupo de trabajadores de tipo intelectual.
prop.table(tab2,1)

##      Salario         35         75        125
## Edad
## 21.5         0.40000000 0.55058824 0.04941176
## 30           0.19978858 0.40010571 0.40010571
## 50           0.10278638 0.20557276 0.69164087

##En cambio en esta tabla se puede apreciar que si hay una estrecha relacion entre
##las variables de edad y salario, pues podemos ver que a menor edad menor salario
##y a mayor edad, mejores son los ingresos en este tipo de trabajo.

##Para terminar con el analisis mostraremos otra alternativa para para ver el grado
##de relacion entre las variables de salario y edad calcularemos
##los coeficientes de contingencia de la tabla de trabajo manual e intelectual

##Calculemos su Coeficiente de Contingencia para la tabla de trabajo manual

##Extraemos nuestra estadistica T
T1=as.vector(chisq.test(tab1)$statistic)
T1

## [1] 3.213551

##Calculamos el numero total de trabajadores
N1=sum(tab1)
N1

## [1] 5500

##Calculamos CC
CC1=sqrt(T1/(T1+N1))
CC1
```

```
## [1] 0.02416487

##Como podemos observar el CC es muy pequeo, por lo que notamos que el grado
##de dependencia entre las variables es muy pequeo, por lo que para este tipo
##de trabajo no contradice la hipotesis nula, es decir podemos decir que
##la variable Salario es independiente de la Edad

##Calculemos su Coeficiente de Contingencia para la tabla de trabajo intelectual

##Extraemos nuestra estadistica T
T2=as.vector(chisq.test(tab2)$statistic)
T2

## [1] 882.5047

##Calculamos el numero total de trabajadores
N2=sum(tab2)
N2

## [1] 5547

##Calculamos CC
CC2=sqrt(T2/(T2+N2))
CC2

## [1] 0.3704842

##Como podemos observar el CC es significativo, pues es superior al 30%, entonces
##podemos decir que si hay relacion entre las variables, por lo que contradice
##la hipotesis nula, es decir para los trabajadores de tipo intelectual
##hay dependencia entre el Salario y la Edad
```