Asociación e independencia de variable categóricas

9/5/2020

Contents

Distribución gamma y distribución exponencial

Aunque la distribución normal se puede utilizar para resolver muchos problemas de ingeniería y ciencias, aún hay numerosas situaciones que requieren diferentes tipos de funciones de densidad. La distribución gamma deriva su nombre de la función gamma, que se define como:

Sea $\Gamma : (0, \infty) \to \mathbf{R}$, donde:

$$\Gamma(\alpha) = \int_{x=0}^{\infty} x^{\alpha-1} e^{-x} dx, \, para \, \alpha > 0$$

Se integra por partes. $u = x^{\alpha-1}$, $dv = e^{-x}$ y se obtiene:

$$\Gamma(\alpha) = -e^{-x} x^{\alpha-1}|_0^{\infty} \int_{x=0}^{\infty} (\alpha - 1) x^{\alpha-2} e^{-x} dx$$

Para $\alpha > 1$, lo que nos lleva a

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

Al aplicar reiteradamente la integración, se llega a:

$$\Gamma(\alpha) = (\alpha - 1)(\alpha - 2)\Gamma(\alpha - 2) = (\alpha - 1)(\alpha - 2)(\alpha - 3)\Gamma(\alpha - 3)$$

, y así sucesivamente. Entonces, si $\alpha = n$, donde n es un entero positivo,

$$\Gamma(n) = (n - 1)(n - 2)...\Gamma(1)$$

Sin embargo, por definición $\Gamma(1) = \int_{x=0}^{\infty} e^{-x} dx = 1$, de donde

$$\Gamma(n) = (n - 1)!$$

Otras propiedades de $\Gamma(\alpha)$ son:
* $\Gamma(n + 1) = n!$ si n es un entero positivo
* $\Gamma(n + 1) = n\Gamma(n), \quad n > 0$
* $\Gamma(1/2) = \sqrt{\pi}$

La variable aleatoria continua X tiene una distribución gamma, con parámetros α y β, si su función de densidad está dada por:

$$f(x:\alpha,\beta)=\begin{cases}\frac{1}{\beta^{\alpha}\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} & para \quad x>0;\alpha,\beta>0;\\ \\ 0 & en\,otro\,caso\end{cases}$$

La distribución gamma especial para la que $\alpha = 1$ se llama distribución exponencial.

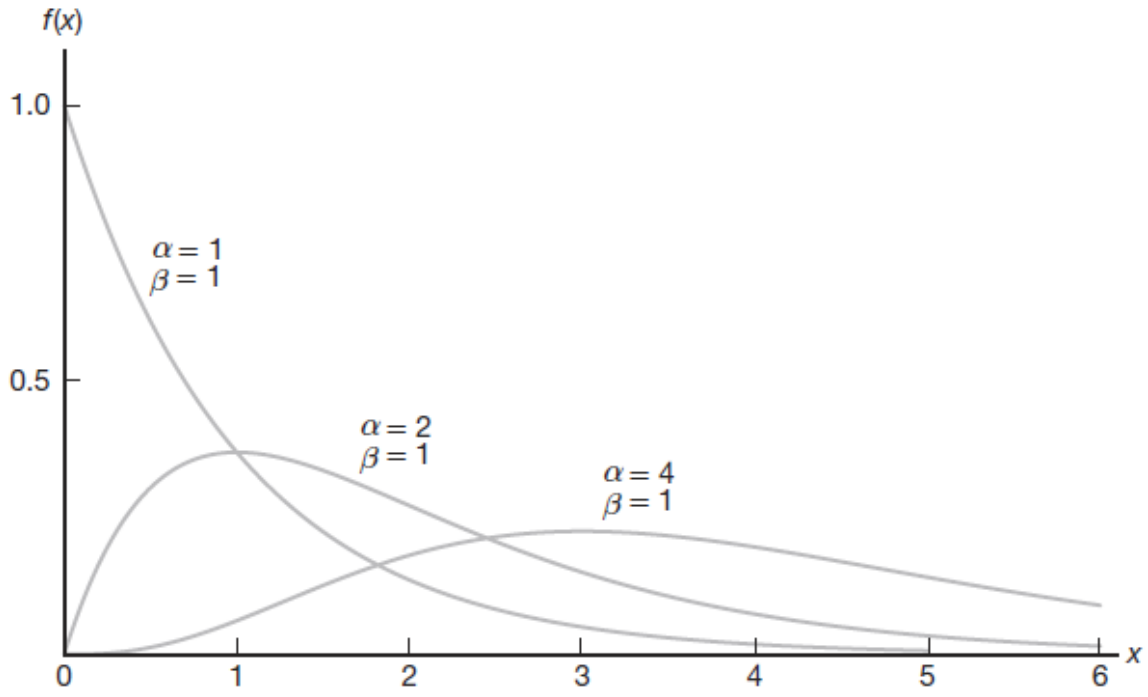


Figure 1: Distribuciones gamma de algunos valores de α y β. La distribución gamma para $\alpha = 1$ se llama distribución exponencial

La variable aleatoria continua X tiene distribución exponencial, con parámetro β, si su función de densidad está dada por:

$$f(x:\beta)=\begin{cases}\frac{1}{\beta}e^{-x/\beta} & para \quad x>0;\\ \\ 0 & en\,cualquier\,otro\,caso\end{cases}$$

donde $\beta > 0$

La media y la varianza de la distribución gamma son:

$$\mu=\alpha\beta \quad y \quad \sigma^2=\alpha\beta^2$$

La media y la varianza de la distribución exponencial son:

$$\mu=\beta \quad y \quad \sigma^2=\beta^2$$

Las aplicaciones más importantes de la distribución exponencial son situaciones donde se aplica el proceso de Poisson, que permite utilizar la distribución discreta llamada distribución de Poisson. La distribución de Poisson se utiliza para calcular la probabilidad de números específicos de eventos durante un período o espacio particulares. En muchas aplicaciones la variable aleatoria es el tiempo o la cantidad de espacio. La relación entre la distribución exponencial, también llamada exponencial negativa, y el proceso de Poisson es muy simple.

La distribución de Poisson tiene un parámetro λ,que se interpreta como el número medio de eventos por unidad de tiempo. Considere ahora la variable aleatoria descrita por el tiempo que se requiere para que ocurra el primer evento. Si se utiliza la distribución de Poisson, la probabilidad de que no ocurra algún evento, en el periodo hasta el tiempo t, es dada por:

$$p(0;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}$$

Si X es el tiempo para el primer evento de Poisson, eentonces la probabilidad de que la duración del tiempo exceda a x es la misma probabilidad de que no ocurra algún evento de Poisson en x. Esto está dado por $e^{-\lambda t}$. Por lo tanto,

$$P(X > x) = e^{-\lambda x}$$

. De esta forma, la distribución acumulada para X está dada por:

$$P(0 \leq X \leq x) = 1 - e^{-\lambda x}$$

Si se deriva la función de distribución acumulativa anterior con el fi n de obtener la función de densidad $f(x) = \lambda e^{-\lambda x}$, que es la función exponencial con $\lambda = 1/\beta$.

Ejemplo de aplicación de distribución exponencial.

Suponga que un sistema contiene cierto tipo de componente cuyo tiempo de operación antes de fallar, en años, está dado por T. La variable aleatoria T se modela bien mediante la distribución exponencial con tiempo medio de operación antes de fallar $\beta = 5$. Si se instalan 5 de estos componentes en diferentes sistemas, ¿cuál es la probabilidad de que al final de 8 años al menos dos aún funcionen?

La probabilidad de que un componente determinado siga funcionando después de 8 años es dada por:

$$P(T > 8) = \frac{1}{5}\int_8^\infty e^{-t/5}dt = e^{-8/5} \approx 0.2$$

Si X es el número de componentes que todavía funcionan después de 8 años. Entonces, utilizando la distribución binomial:

$$P(X \geq 2) = \sum_{x=2}^{5} b(x;5,0.2) = 1 - \sum_{x=0}^{1} b(x;5,0.2) = 1 - 0.7373 = 0.2627$$

```
x<- 2:5
sum(dbinom(x,5,0.2))
```

```
## [1] 0.26272
```

Suponga que las llamadas telefónicas que llegan a un conmutador particular siguen un proceso de Poisson con un promedio de 5 llamadas entrantes por minuto. ¿Cuál es la probabilidad de que transcurra hasta un minuto en el momento en que han entrado 2 llamadas al conmutador?

Se aplica el proceso de Poisson, con un lapso de tiempo hasta que ocurren 2 eventos de Poisson que sigue una distribución gamma con $\beta = 1/5$ y $\alpha = 2$.

$$P(X \leq 1) = \int_0^1 1/\beta^2 x e^{-x/\beta}dx = 25\int_0^1 x e^{-5x}dx = 1 - e^{-5}(1+5) \approx 0.92$$

```
x<-2
lambda<- 5
print(dist<-1-dpois(x,lambda))
```

```
## [1] 0.9157757
```

La distribución gamma trata con el tiempo hasta la ocurrencia de α eventos de Poisson. También funciona aunque no exista una estructura de Poisson clara. Esto es particularmente aplicable para problemas de tiempo de supervivencia en aplicaciones de ingeniería y biomédicas.

Ejemplo En un estudio biomédico con ratas se utiliza una investigación de respuesta a la dosis para determinar el efecto de la dosis de un tóxico en su tiempo de sobrevivencia. Para cierta dosis del tóxico, el estudio determina que el tiempo de sobrevivencia de las ratas, en semanas, tiene una distribución gamma con $\alpha = 5$ y $\beta = 10$. ¿Cuál es la probabilidad de que una rata no sobreviva más de 60 semanas?

$$P(X \leq 60) = \frac{1}{\beta^5} \int_0^6 \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(5)} dx$$

```
print(1-pgamma(60, 5, rate = 0.1, lower.tail = F))
```

```
## [1] 0.7149435
```

Distribución de Ji cuadrada

Un caso especial de la distribución gamma se obtiene al permitir que $\alpha = \nu/2$ y $\beta = 2$, donde ν es un entero positivo. Este resultado se conoce como distribución chi cuadrada. La distribución tiene un solo parámetro, ν, denominado grados de libertad.

$$f(x:\nu) = \begin{cases} \frac{1}{2^2\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} & para \qquad x > 0 \\ \\ 0 & en\ otro\ caso \end{cases}$$

donde ν es positiva.

La media y la varianza de la distribución χ^2 son

$$\mu = \nu \quad y \quad \sigma^2 = 2\nu$$

Aplicación de la prueba chi cuadrada

La prueba chi-cuadrada, también llamada Ji cuadrado (χ^2), se encuentra dentro de las pruebas pertenecientes a la estadística descriptiva aplicada al estudio de dos variables.

Nota: La estadística descriptiva se centra en extraer información sobre la muestra, la estadística inferencial extrae información sobre la población.

La distribución chi cuadrada no es simétrica y depende de un número específico de grados de libertad (gl). Los gl adecuados para el estadístico $\chi 2$ dependen de la aplicación para la que se utilice.

Grados de libertad

Los grados de libertad (gl) de un estadístico calculado con base en n datos, se refiere al número de cantidades independientes que se necesitan en su cálculo, menos el número de restricciones que relacionan a las observaciones y el estadístico.

Prueba de probabilidades de valores esperados

Los valores esperados para cada categoría se calculan usando probabilidades hipotéticas, $E_i = np_i$, se usan para calcular el valor observado del estadístico de prueba χ^2. Para un experimento multinomial formado por k categorías o celdas, el estadístico de prueba tiene una distribución χ^2 aproximada, con $gl = (k-1)$.

Ejemplo
Un investigador diseña en el que una rata es atraída al final de una rampa que se divide, llevando a tres

puertas de colores diferentes. Se hace que una rata baje 90 veces y se observan las siguientes elecciones. ¿La rata tiene preferencia por alguna de las tres puertas?

Cantidad Observada (O_i)

Puerta verde	Puerta roja	Puerta azul
20	39	31

$H_0 : p_1 = p_2 = p_3 = 1/3$ H_a =Al menos una p_i es diferente de 1/3

p_i es la probabilidad de que la rata elija la puerta $i, para \quad i = 1, 2\, y\, 3$. Los valores esperados son los mismos para cada una de las categorías $np_i = 90(1/3) = 30$

El estadístico de prueba $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = (20-30)^2/30 + (39-30)^2/30 + (31-30)^2/30 = 6.067$

En este caso se tienen $n - 1 = 2$ grados de libertad pues la única restricción lineal en las probabilidades es que su suma es igual a 1.

El valor observado para $\chi^2 = 6.607$ está entre 5.99 7 7.39 en la tabl, el valor de p está entre 0.025 y 0.05. EL valor que se toma como referencia en investigación es, generalmente, $p < 0.05$.

Esto significa que la hipótesis nula de no preferencia por alguna puerta se rechaza. La rata prefiere alguna de las puertas.

Puerta azul $31/90 = 0.344$
Puerta verde $20/90 = 0.222$
Puerta roja $39/90 = 0.433$

La rata tiene preferencia por la puerta roja, pero no se puede declarar una relación **causal** entre el color y la preferencia.

Prueba de independencia

La prueba chi-cuadrado es una de las más conocidas y utilizadas para analizar variables nominales o cualitativas, es decir, para determinar la existencia o no de independencia entre dos variables. Que dos variables sean independientes significa que no tienen relación, y que por lo tanto una no depende de la otra, ni viceversa.

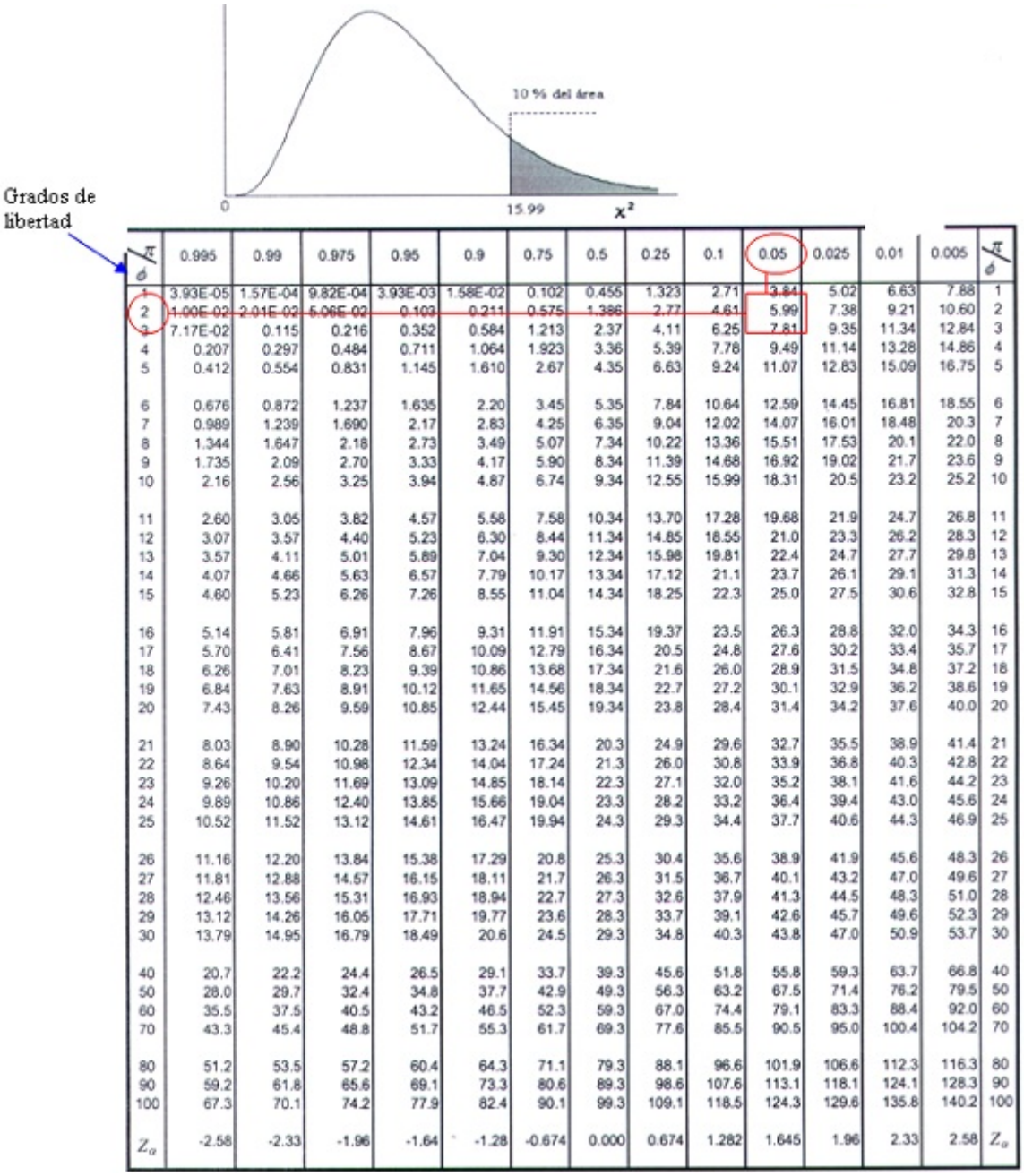
Ejemplo Suponga que se desea determinar si las opiniones de los votantes residentes del estado de Illinois respecto a una nueva reforma fiscal son independientes de sus niveles de ingreso. Los sujetos de una muestra aleatoria de 1000 votantes registrados del estado de Illinois se clasifican de acuerdo con su posición en las categorías de ingreso bajo, medio o alto, y si están a favor o no de la nueva reforma fiscal. Las frecuencias observadas se presentan en la siguiente tabla, que se conoce como tabla de contingencia.

Reforma Fiscal	Nivel bajo	Nivel medio	Nivel Alto
A favor	182	213	203
En contra	154	138	110
Total	336	351	313

Esta es una tabla de 2 x 3

Una tabla de contingencia con r renglones y c columnas se denomina tabla r $\times$ c ("r $\times$ c" se lee "r por c"). Los totales de renglones y columnas en la tabla se denominan frecuencias marginales.

La decisión de aceptar o rechazar la hipótesis nula, H_0, de que la opinión de un votante respecto a la nueva reforma fiscal es independiente de su nivel de ingreso, se basa en qué tan bien se ajusten las frecuencias



α / ϕ	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	α / ϕ
1	3.93E-05	1.57E-04	9.82E-04	3.93E-03	1.58E-02	0.102	0.455	1.323	2.71	3.84	5.02	6.63	7.88	1
2	1.00E-02	2.01E-02	5.06E-02	0.103	0.211	0.575	1.386	2.77	4.61	5.99	7.38	9.21	10.60	2
3	7.17E-02	0.115	0.216	0.352	0.584	1.213	2.37	4.11	6.25	7.81	9.35	11.34	12.84	3
4	0.207	0.297	0.484	0.711	1.064	1.923	3.36	5.39	7.78	9.49	11.14	13.28	14.86	4
5	0.412	0.554	0.831	1.145	1.610	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	5
6	0.676	0.872	1.237	1.635	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	6
7	0.989	1.239	1.690	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.3	7
8	1.344	1.647	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.1	22.0	8
9	1.735	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.7	23.6	9
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.5	23.2	25.2	10
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.9	24.7	26.8	11
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.0	23.3	26.2	28.3	12
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.4	24.7	27.7	29.8	13
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.1	23.7	26.1	29.1	31.3	14
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.3	25.0	27.5	30.6	32.8	15
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.5	26.3	28.8	32.0	34.3	16
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.5	24.8	27.6	30.2	33.4	35.7	17
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.6	26.0	28.9	31.5	34.8	37.2	18
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.7	27.2	30.1	32.9	36.2	38.6	19
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.8	28.4	31.4	34.2	37.6	40.0	20
21	8.03	8.90	10.28	11.59	13.24	16.34	20.3	24.9	29.6	32.7	35.5	38.9	41.4	21
22	8.64	9.54	10.98	12.34	14.04	17.24	21.3	26.0	30.8	33.9	36.8	40.3	42.8	22
23	9.26	10.20	11.69	13.09	14.85	18.14	22.3	27.1	32.0	35.2	38.1	41.6	44.2	23
24	9.89	10.86	12.40	13.85	15.66	19.04	23.3	28.2	33.2	36.4	39.4	43.0	45.6	24
25	10.52	11.52	13.12	14.61	16.47	19.94	24.3	29.3	34.4	37.7	40.6	44.3	46.9	25
26	11.16	12.20	13.84	15.38	17.29	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3	26
27	11.81	12.88	14.57	16.15	18.11	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6	27
28	12.46	13.56	15.31	16.93	18.94	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0	28
29	13.12	14.26	16.05	17.71	19.77	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3	29
30	13.79	14.95	16.79	18.49	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7	30
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8	40
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5	50
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0	60
70	43.3	45.4	48.8	51.7	55.3	61.7	69.3	77.6	85.5	90.5	95.0	100.4	104.2	70
80	51.2	53.5	57.2	60.4	64.3	71.1	79.3	88.1	96.6	101.9	106.6	112.3	116.3	80
90	59.2	61.8	65.6	69.1	73.3	80.6	89.3	98.6	107.6	113.1	118.1	124.1	128.3	90
100	67.3	70.1	74.2	77.9	82.4	90.1	99.3	109.1	118.5	124.3	129.6	135.8	140.2	100
Z_α	-2.58	-2.33	-1.96	-1.64	-1.28	-0.674	0.000	0.674	1.282	1.645	1.96	2.33	2.58	Z_α

Figure 2: Tabla de valores críticos de χ^2

observadas en cada una de las 6 celdas de la tabla de contingencia y en las frecuencias esperadas si supusiéramos que H_0 es verdadera. Para encontrar estas frecuencias esperadas se definen los siguientes eventos:

L: Una persona seleccionada está en el nivel de ingresos bajo.
M: Una persona seleccionada está en el nivel de ingresos medio.
H: Una persona seleccionada está en el nivel de ingresos alto.
F: Una persona seleccionada está a favor de la nueva reforma fiscal.
A: Una persona seleccionada está en contra de la nueva reforma fiscal.

Con base en las frecuencias marginales:

$$P(L) = 336/1000$$
$$P(M) = 351/1000$$
$$P(H) = 313/1000$$
$$P(F) = 598/1000$$
$$P(A) = 402/1000$$

Si H_0 es verdadera y las dos variables son independientes:

$$P(L \cap F) = P(L)P(F) = (336/1000)(598/1000)$$
$$P(L \cap A) = P(L)P(A) = (336/1000)(402/1000)$$
$$P(M \cap F) = P(M)P(F) = (351/1000)(598/1000)$$
$$P(M \cap A) = P(M)P(A) = (351/1000)(402/1000)$$
$$P(H \cap F) = P(H)P(F) = (313/1000)(598/1000)$$
$$P(H \cap A) = P(H)P(A) = (313/1000)(402/1000)$$

Las frecuencias esperadas se obtienen multiplicando la probabilidad de cada celda por el número total de observaciones. De esta manera, se estima que el número esperado de votantes de bajo ingreso en la muestra que favorecen la reforma fiscal cuando H_0 es verdadera:

$$(336/1000)(598/1000)(1000) = 200.9$$

Esto es,

$$frecuenciaesperada = \frac{(total \quad columna)(total \quad renglón)}{gran \quad total}$$

Actividad:

completa la tabla con la frecuencia esperada para cada "celda":

Reforma Fiscal	Nivel bajo	Nivel medio	Nivel Alto	Total
A favor	182 (200.9)	213()	203()	598
En contra	154()	138()	110()	402
Total	336	351	313	1000

El número de grados de libertad asociados con la prueba chi cuadrada que aquí se usa es igual al número de frecuencias de celdas que se pueden llenar libremente cuando se proporcionan los totales marginales y el gran total, y en este caso es 2.

Una fórmula sencilla que proporciona el número correcto de grados de libertad es $\nu = (r-1)(n-1) =$

$(2-1)(3-1) = 2$. Calcule

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

donde la sumatoria se extiende a todas las celdas rc en la tabla de contingencia r $\times$ c

Si $\chi^2 > \chi_\alpha$ con n grados de libertad, rechace la hipótesis nula de independencia al nivel de significancia alfa, en otro caso acéptela.

$$\chi^2 = 7.85$$

con $P \approx 0.02$ *Actividad: demuestre este resultado*

En la tabla de valores para χ^2 se encuentra que $\chi^2_{0.05}$ para 2 gl.

Rechazamos la hipótesis nula y concluimos que la opinión de un votante respecto a la reforma fiscal y su nivel de ingresos NO son independientes. Los valores χ^2 calculados dependen de las frecuencias de las celdas y, en consecuencia, son discretos.

La distribución chi cuadrada continua parece aproximarse muy bien a la distribución de muestreo discreta de χ^2, siempre y cuando el número de grados de libertad sea mayor que 1.

Corrección de Yates

En una tabla de contingencia de 2 $\times$ 2, donde sólo se tiene 1 grado de libertad, se aplica una corrección llamada corrección de Yates para continuidad. La fórmula corregida entonces se convierte en

$$\chi^2(corregida) = \sum_i \frac{(|o_i - e_i| - 0.5)^2}{e - i}$$

Si las frecuencias de las celdas esperadas son grandes, los resultados corregidos y sin corrección son casi iguales. Cuando las frecuencias esperadas están entre 5 y 10, se debe aplicar la corrección de Yates.

Prueba de homogeneidad

Otro tipo de problema para el que se aplica el método anterior es aquel en el cual los totales de renglón y de columna están predeterminados.

Ejemplo
Suponga que decidimos de antemano seleccionar 200 demócratas, 150 republicanos y 150 independientes entre los votantes del estado de Carolina del Norte y registrar si están a favor de una iniciativa de ley para el aborto, si están en contra o si están indecisos. |Ley para el aborto |Demócrata |Republicano |Independiente| Total| | —— |——| ——|——:|——:| |A favor|En contra|Indeciso|Total 82|93|25|214 70|62|18|222 62|67|21|64 200|150|150|500 Ahora bien, en vez de hacer una prueba de independencia, probamos la hipótesis de que las proporciones de población dentro de cada renglón son iguales. Es decir, probamos la hipótesis de que las proporciones de demócratas, republicanos e independientes que están a favor de la ley para el aborto son iguales; las proporciones de cada afiliación política contra la ley son iguales y las proporciones de cada afiliación política que están indecisos son iguales.

Se busca determinar si las tres categorías de votantes son homogéneas en lo que se refiere a sus opiniones acerca de la iniciativa de ley para el aborto. A esta prueba se le conoce como prueba de homogeneidad.

Al suponer homogeneidad de nuevo calculamos las frecuencias esperadas de las celdas multiplicando los totales de renglón y de columna correspondientes y después dividiendo entre el gran total. Luego continuamos el análisis utilizando el mismo estadístico chi cuadrada como antes.

1. H_0: Para cada opinión las proporciones de demócratas, republicanos e independientes son iguales.
2. H_1: Para al menos una opinión las proporciones de demócratas, republicanos e independientes no son iguales.
3. $\alpha = 0.05$.
4. Región crítica: $\chi^2 > 9.488$ con v = 4 grados de libertad (de la tabla de χ^2).

5. Calcular las 4 frecuencias de las celdas usando la fórmula de las frecuencias de las celdas esperadas $frecuenciaesperada = \frac{(total \quad columna)(total \quad renglón)}{gran \quad total}$. Todas las demás frecuencias se obtienen mediante sustracción.

$\chi^2 = 1.53$ *Actividad: demuestre este resultado* Conclusión: No rechazar H_0. No hay suficiente evidencia para concluir que la proporción de demócratas, republicanos e independientes difiere para cada opinión expresada.

Actividades

1. Individual: Resolver los ejercicios de práctica en la plataforma Connect "Aplicación chi cuadrada".
2. En grupos de tres alumnos: Para la base de datos BD_polimorf.xlsx, que está en Moodle,determina la independencia de las variables HTA, ALCOHOL y TABQ con respecto a POL1,POL2 y POL3. En tres ejercicios separados.