Capítulo 14 Análisis de datos categóricos

14.1 Descripción del experimento

Numerosos experimentos resultan en mediciones que son *cualitativas* o *categóricas* más que *cuantitativas*, como muchas de las mediciones estudiadas en capítulos previos. En estos experimentos una cualidad o característica es identificada por cada una de las unidades experimentales. Entonces se puede tener un conteo del número de mediciones que caen en cada categoría. Un ejemplo pudiera ser el hecho de clasificar los artículos manufacturados en tres categorías de acuerdo a su calidad o pinturas se pueden clasificar en una de k categorías de acuerdo con su estilo y periodo.

En ocasiones no es difícil aproximar de manera razonable el hecho de que este tipo de experimentos tienen características asociadas con el *experimento multinomial*. Estas se muestran a continuación.

1. EL experimento consiste en n ensayos idénticos.
2. EL resultado de cada ensayo cae en exactamente una de k categorías o celdas distintas.
3. La probabilidad de que el resultado de un solo ensayo caiga en una celda particular, la celda i, es p_i donde $i = 1, 2, \ldots, k$ y continúa igual de un ensayo a otro. Observe que

$$p_1 + p_2 + p_3 + \cdots + p_k = 1$$

4. Los ensayos son independientes.
5. Estamos interesados en $n_1, n_2, n_3, \ldots, n_k$, donde n_i para $i = 1, 2, \ldots, k$ es igual al número de ensayos para los cuales el resultado cae en la celda i. Observe que $n_1 + n_2 + n_3 + \cdots + n_k = n$.

Entonces el objetivo de este capítulo es hacer inferencias acerca de las probabilidades por celda $p_1, p_2, \ldots, p_k$. Las inferencias se expresarán en término de pruebas estadísticas de hipótesis que se refieran a valores numéricos específicos de las probabilidades por celda o a su relación mutua.

14.2 Prueba ji cuadrada

Suponga que proponemos valores para $p_1, p_2, \ldots, p_k$. Estos valore representan la probabilidad por celda. Se quiere calcular el valor esperado por cada celda, entonces si la hipótesis de que se tiene un experimento mulitnomial es cierto, los conteos n_i por celda no deben desviarse demasiado de sus valores esperados np_i para $i = 1, 2, \ldots, k$. Por lo tanto parecería intuitivamente razonable usar un estadístico de prueba que comprenda las k desviaciones, esto es,

$$n_i - E(n_i) = n_i - np_i$$

para $i = 1, 2, \ldots, k$. En 1900, Karl Pearson propuso el siguiente estadístico de prueba, que es una funición de los cuadrados de las deviaciones de las cantidades observadas respecto de sus valores esperados, ponderados por los recíprocos de sus valores esperados:

$$X^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)} = \sum_{i=1}^{k} \frac{[n_i - np_i]^2}{np_i}$$

Se puede demostrar que cuando n es grande X^2 tiene una distribución de probabilidad Ji cuadrada (χ^2) aproximada. Es es difícil demostrar para k categorías pero se puede hacer una demostración para $k = 2$ categorías.

Demostración

Si $k = 2$, entonces $n_2 = n - n_1$ y $p_1 + p_2 = 1$. Entonces,

$$\begin{aligned} X^2 &= \sum_{i=1}^{2} \frac{[n_i - E(n_i)]^2}{E(n_i)} = \frac{(n_1 - np_1)^2}{np_1} + \frac{(n_2 - np_2)^2}{np_2} \\ &= \frac{(n_1 - np_1)^2}{np_1} + \frac{[(n - n_1) - n(1 - p_1)]^2}{n(1 - p_1)} \\ &= \frac{(n_1 - np_1)^2}{np_1} + \frac{(-n_1 + np_1)^2}{n(1 - p_1)} \\ (n_1 - np_1)^2 &\left(\frac{1}{np_1} + \frac{1}{n(1 - p_1)}\right) = \frac{(n_1 - np_1)^2}{np_1(1 - p_1)} \end{aligned}$$

De la Sección 7.5 se tiene que para n granda lo siguiente,

$$\frac{n_1 - np_1}{\sqrt{np_1(1 - p_1)}}$$

tiene aproximadamente una distribución normal estándar. Como el cuadrado de una variable aleatorias normal estándar tiene una distribución χ^2, para $k = 2$ y n grande, X^2 tiene una distribución χ^2 aproximada con 1 grado de libertad (gl).

La experiencia ha demostrado que las cantidades n_i en la celda no debe ser tan pequeñas si la distribución χ^2 debe proporcionar una aproximación adecuada para la distribución X^2.

Como grandes diferencias entre las cantidades observadas y esperadas por celda contradicen la hipótesis nula, rechazaremos la hipótesis nula cuando X^2 sea grande y emplearemos una prueba estadística de cola superior. *El número apropiado de grados de libertad será igual al número de celdas, k, menos 1 grado de libertad para cada restricción lineal independiente colocada en las probabilidades por celda*.

14.3 Prueba de una hipótesis con respecto a probabilidades especificadas por celda: una prueba de bondad de ajuste

La hipótesis más sencilla respecto a probabilidades por celda es aquella que especifica valores numéricos para cada una. En este caso, estamos probando $H_0 : p_1 = p_{1,0}, p_2 = p_{2,0}, \ldots, p_k = p_{k,0}$, donde $p_{i,0}$ denota un valor especificado para p_i. La hipótesis alternativa expresa, en general, que al menos una de las siguientes igualdades no se cumple. Como la úniva restricción en las probabilidades por celda es que $\sum_{i=1}^{k} p_i = 1$, estadístico de prueba X^2 tiene aproximadamente una distribución χ^2 con $k - 1$ grados de libertad.

Ejemplo

Un grupo de ratas, una por una, bajan por una rampa que conduce a tres puertas. Deseamos probar la hipótesis de que las ratas no tienen preferencia respecto a la elección de una puerta. Entonces, la hipótesis nula apropiada es

$$H_0 : p_1 = p_2 = p_3 = \frac{1}{3}$$

donde p_i es la probabilidad de que una rata escogerá la puerta i, par $i = 1, 2, 3$. Suponga que las ratas bajaron por la rampa $n = 90$ veces y que las tres frecuencias por celda observadas fueron $n_1 = 23$, $n_2 = 36$ y $n_3 = 31$. La frecuencia esperada por celda es igual para cada celda: $E(n_i) = np_i = (90)(1/3) = 30$

El estadístico de prueba χ^2 para el ejemplo tendrá $(k-1) = 2$ grados de libertad porque la única restricción en las probabilidades por celda es que

$$p_1 + p_2 + p_3 = 1$$

Por lo tanto, si escogemos $\alpha = 0.05$, rechazaríamos la hipótesis nula cuando $\chi^2 > 5.991$. Sustituyendo en la fórmula para X^2 obtenemos

$$\begin{aligned} X^2 &= \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \\ &= \frac{(23 - 30)^2}{30} + \frac{(36 - 30)^2}{30} + \frac{(31 - 30)^2}{30} = 2.87 \end{aligned}$$

Como X^2 es menor que el valor crítico de χ^2, la hipótesis nula no es rechazada y concluimos que los datos no presentan suficiente evidencia para indicar que las ratas tienen preferencia por cualquiera de las puertas.

El estadístico χ^2 también se puede usar para probar si los datos

muestrales indican que un modelo específico para una distribución poblacional no se ajusta a los datos. Esto se llamaría una *prueba de bondad de ajuste*.

14.4 Tablas de contingencia

Cuando se quiere investigar la *dependencia* (o *contingencia*) entre dos criterios de clasificación se necesita usar lo que se conoce como Tablas de contingencia. Uno de los criterios de clasificación se ponen en los renglones y el otro en las columnas. De esta manera se $r \times c$ celdas, donde r es el número de renglones y c es el número de columnas. Las columnas serán las categorías de uno de los criterios de clasificación y los renglones serán las categorías del otro. Cabe mencionar que el número de columnas no tiene que ser igual al número de renglones.

Se construye una tabla de contingencia con el objetivo de probar H_0: clasificación de columna es independiente de clasificación de renglón. Como aun se esta tratando con la suposición de que los datos vienen de un experimento mulitnomial entonces se tiene que cada criterio tiene su probabilidad asociada. Entonces hay *probabilidades de columna* y *probabilidades de renglón*. Estos tienen la siguiente restricción,

$$p_1 + p_2 + \cdots + p_c = 1 \qquad p_1 + p_2 + \cdots + p_r = 1$$

En otras palabras la suma de las probabilidades de las columnas tiene que dar 1 y la suma de las probabilidades de los renglones tiene que dar 1. Si la hipótesis nula es cierta entonces la probabilidad por celda será igual al producto de sus respectivas probabilidades de renglón y columna. Ya que esto implica independencia de las dos clasificaciones.

Para lograr esto se necesita estimar las probabilidades de renglón y columna y con estas estimar las frecuencias por celda esperadas. El estimador de máxima probabilidad (MLE) para cualquier probabilidad de renglón o columna se encuentra como sigue.

Con n_{ij} denotemos la frecuencia observada en el renglón i y la columna j de la tabla de contingencia y con p_{ij} denotemos la probabilidad de que una observación caiga en esta celda. Si las observaciones se seleccionan de manera independiente, entonces las frecuencias por celda tienen una distribución multinomial y el MLE de p_{ij} es simplemente la frecuencia relativa observada para esa celda. Esto es,

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \qquad i = 1, 2, \ldots, r \qquad j = 1, 2, \ldots, c$$

Del mismo modo, viendo el renglón i como una sola celda, la probabilidad para el renglón i está dado por p_i y si r_i denota el número de observaciones en el renglón i,

$$\hat{p}_i = \frac{r_i}{n}$$

es el MLE de p_i. Por argumento semejante, el MLE de la probabilidad de la j-ésima columna es c_j/n, donde c_j denota el número de observaciones en la columna j.

Si la hipótesis nula es verdadera, el valor esperado y estimado de la frecuencia por cada celda, n_{ij} para una tabla de contingencia es igual al producto de sus respectivos totales de renglón y columna divididos entre el tamaño de muestra total. Esto es,

$$\widehat{E(n_{ij})} = \frac{r_i c_j}{n}$$

Entonces el estadístico de prueba tendría la siguiente forma.

$$X^2 = \sum_{i=1}^{k} \frac{\left[n_{ij} - \widehat{E(n_{ij})}\right]^2}{\widehat{E(n_{ij})}}$$

Los *grados de libertad asociados con una tabla de contingencia que tenga r renglones y c columnas siempre será igual a* $(r-1)(c-1)$.

Recuerde que el número de grados de libertad asociado con el estadístico χ^2 será igual al número de celdas menos 1 grado de libertad por cada restricción lineal independiente colocada en las probabilidades por celda. Entonces, en el caso de una tabla de contingencia de r renglones y c columnas se resta 1 grado de libertad por la restricción de que la suma de las probabilidades por celda debe ser igual a 1, esto es,

$$p_{11} + p_{12} + \cdots + p_{rc} = 1$$

Como se necesitan estimar las probabilidades de los renglones y tienen la restricción de sumar 1 entonces se pierden $r-1$ grados de libertad. De la misma manera se pierden $c-1$ por las probabilidades de las columnas y por su restricción de sumar 1. Entonces se puede ver que el número total de grados de libertad asociado con una tabla de contingencia $r \times c$ es

$$\text{gl} = rc - 1 - (r-1) - (c-1) = (r-1)(c-1)$$

14.5 Tablas $r \times c$ con totales fijos de renglón o columna

En las secciones pasadas se vio como se ajusta un modelo multinomial a una tabla de contingencia $r \times c$. Muchos experimentos se pueden hacer de esta manera sin embargo en ocasiones puede que no se quiera muestrear aleatoriamente. Por ejemplo suponga que se decidió de antemano entrevistar a un número especificado de personas de cada categoría de columna, con lo cual se fija anticipadamente los totales de columna. Entonces se tendrían c experimentos binomiales separados e independientes correspondientes a las columnas.

Se desea probar la hipótesis nula de que las c columnas tienen la misma probabilidad, o $H_0 : p_1 = p_2 = \cdots = p_c$. De acuerdo con esta hipótesis, los MLE de las frecuencias esperadas por celda son los mismos que en la sección pasada, es decir,

$$\widehat{E(n_{ij})} = \frac{r_i c_j}{n}$$

En cuanto a los grados de libertad como se están fijando los totales de columna y la suma de probabilidades en cada columna debe ser igual a uno se tiene que hay c restricciones lineales en las p_{ij}. Además se necesitan estimar $r-1$ probabilidades de renglón. Con esto se tiene que el número de grados de libertad asociados con X^2 calculado para una tabla $r \times c$ con totales fijos de columna es $gl = rc - c - (r-1) = (r-1)(c-1)$.

Es frecuente que esta prueba se denomine *prueba de homogeneidad* de las poblaciones binomiales. Si hay más de dos categorías de renglón y los totales de columna son fijos, la prueba χ^2 es una prueba de la equivalencia de las proporciones en c poblaciones multinomiales.