Análisis de Datos Categóricos

Leticia Gracia Medrano

29 de enero de 2020

Modelo Poisson

La distribución Poisson ayuda a modelar conteos como: número de accidentes en un tramo carretero, personas que llegan a formarse en una fila. La distribución está dada por

$$P(y) = \frac{\exp^{-\mu}\,\mu^{y}}{y!}$$

Por ejemplo: Si se tiene que ocurren un promedio de 2 accidentes semanales en cierto tramo carretero entonces la probabilidad de tener 0 accidentes en una semana dada es:

$$P(0) = \frac{\exp^{-2}\,2^{0}}{0!} = \exp^{-2} = 0.135$$

En esta distribución $E(Y) = Var(Y) = \mu$, aquí entonces si la media aumenta, también aumenta la varianza.

En la práctica ocurre frecuentemente que los conteos tienen mayor varianza que la esperada, esto se conoce como *sobredispersión*.

En ocasiones suponer la distribución Poisson resulta muy simplista, pero en otras resulta muy útil.

Ejemplos

Distribución de las llamadas telefónicas diarias

La demanda diaria de servicios de urgencia en un hospital

Los arribos de los automóviles a una caseta de cobro

El número de accidentes en un cruce

El número de errores por hoja en un periódico

El número de células enfermas por cm2 en una biopsia

Modelo Binomial

En el ejemplo anterior el número de accidentes es aleatorio. Pero podría planterase algo como que se clasifican los accidentes hasta que ocurren N, con el propósito de estimar la proporción de estos que resultan fatales, entonces el total de accidentes es fijo. Ahora el número de accidentes fatales ya no es Poisson porque tiene un tope máximo de N.

Si se tienen que el número de accidentes fatales en t semanas tiene una media de $2t$, y la tasa para accidentes no-fatales es de $8t$. Cuando se junta un total de N accidentes ocurre que el número de accidentes fatales se distribuye como binomial con parámetros N y $\pi = \frac{2t}{2t+8t} = .2$, la probabilidad de cualquier accidente resulte fatal.

La función de distribución binomial recordemos está dada por:

$$P(y) = \frac{N!}{y!(N-y)!}\pi^y(1-\pi)^{N-y} \text{ con } y = 0, 1, 2, \ldots, N$$

Para el caso en que $N = 10$ y $\pi = .2$ la probabilidad de que haya $y = 0$ accidentes es

$$P(0) = \frac{10!}{0!(10)!}.2^0(.8)^{10} = (.8)^{10} = .107;$$

Para esta distribución $E(Y) = N\pi$ y $Var(Y) = N\pi(1-\pi)$, entonces la varianza siempre es menor que la media.

Cuando los resultados pueden ser más que dos, se tiene una distribución *multinomial*.

Modelo Multinomial

En este caso, el experimento tiene c posibles resultados, sus probabilidades las denotamos por $\pi_1, \pi_2, \ldots, \pi_c$, donde $\sum \pi_j = 1$.

Para n observaciones independientes, la probabilidad de que n_1 caigan en la categoría 1, n_2 caigan en la categoría 2,... y n_c caigan en la categoría c, donde $\sum n_j = n$ es

$$P(n_1, n_2, \ldots, n_c) = \left(\frac{n!}{n_1! n_2! \ldots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

La distribución multinomial es multivariada. La marginal para cualquier categoría es binomial. Para la categoría j el conteo n_j tiene media $n\pi_j$ y desviación estándar $\sqrt{n\pi_j(1-\pi_j)}$

Inferencia sobre π

En el modelo binomial el parámetro es π que generalmente es desconocido y a través de una muestra trataremos de estimarlo.

Para una $n = 10$ y una $y = 0$ la función de verosimilitud es:

$$P(0) = (10!/0!10!)\pi^0(1-\pi)^{10} = (1-\pi)^{10}$$

$$\ell(\pi) = (1-\pi)^{10}$$

que alcanza el máximo cuando $\pi = 0$, entonces el resultado $y = 0$ ocurre con una mayor probabilidad cuando $\pi = 0$.

OJO esta función depende de los valores que tome π, para cada valor de y se tiene una función de verosimilitud distinta, así para $y = 6$

$$\ell(\pi) = (10!/6!4!)\pi^6(1-\pi)^4 = 210 * \pi^6(1-\pi)^4$$

cuyo máximo lo alcanza en $\pi = .6$, en este caso $y = 6$ ocurre con una mayor probabilidad cuando $\pi = .6$

El estimador máximo verosímil para π es $\widehat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$ donde x_i es 1 o 0 según se observe éxito o fracaso.

Entonces $\widehat{p}$ es un promedio por lo que si se tiene una muestra grande se puede utilizar el Teorema Central del Límite.

Para una n grande $y = \bar{x}$ se distribuye **aproximadamente** como una normal con media $E(\widehat{p}) = \sum E(x_i)/n = n\pi/n = \pi$ y varianza $var(\widehat{p}) = \frac{\sum var(x_i)}{n^2} = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$.

Para probar la hipótesis nula $H_o : \pi = \pi_o$ se puede usar la estadística

$$z = \frac{\widehat{p} - \pi_o}{\sqrt{\pi_o(1-\pi_o)/n}}$$

.

Y se puede construir un intervalo de $100(1-\alpha)$ de confianza para π con

$$\widehat{p} \pm z_{\alpha/2} * SE, \text{donde } SE = \sqrt{\widehat{p}(1-\widehat{p})}/n$$

donde $z_{\alpha/2}$ es el percentil que deja a la derecha una cola tamaño $\alpha/2$.

Hay que tener cuidado, esta aproximación es buena cuando π está cerca de 0.5 o cuando la n es muy grande. Si no se tienen esas condiciones el nivel de confianza disminuiría. Y es muy mala cuando π se acerca a uno o al cero.

¿Qué más se podría hacer??

Como existe una correspondencia entre pruebas de hipótesis e intervalos de confianza, que dice que el intervalo de confianza es aquella región donde la hipótesis nula No se rechaza, o sea la región no crítica. Entonces deben hallarse la región donde los valores de π_o cumplan:

$$\frac{|\widehat{p} - \pi_o|}{\sqrt{(\pi_o(1-\pi_o)/n}} \leq 1.96 = z_{.05/2}$$

Elevando al cuadrado ambos lados de la ecuación anterior, resulta una ecuación cuadrática en π_o, que se resuelve con la fórmula conocida como del chicharronero.

Para el caso de $\widehat{p} = 0.90$ y $n = 10$ usando esto se encuentran las raíces $\pi_{o1} = .596$ y $\pi_{o2} = .982$, que darían un intervalo (.596,.982), mientras que usando la ecuación 'tradicional'se tendría un intervalo al 90 % de confianza (0.714, **1.086)!!!.**

Otra aproximación es la de Agresti-Coull que es muy fácil, se suman 2 a los éxitos y se suman 2 a los fracasos y entonces ... usando el ejemplo anterior

$\widehat{p} = (9+2)/(10+2+2) = .786$ y $SE = (.786)(.214)/14 = .110$ con lo que se obtiene un intervalo de (.57,1.0).

Este método funciona bien aún con muestras pequeñas.

Tablas de Contingencia

Cuando se tienen dos criterios o variables de clasificación de las observaciones, al hacer el "cruce" de éstas se genera una tabla de frecuencias como sigue:

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1j} & \cdots & n_{1J} \\ n_{21} & n_{22} & \cdots & n_{2j} & \cdots & n_{2J} \\ \vdots & \vdots & & \vdots & & \vdots \\ n_{i1} & n_{i2} & \cdots & n_{ij} & \cdots & n_{iJ} \\ \vdots & \vdots & & \vdots & & \vdots \\ n_{I1} & n_{I2} & \cdots & n_{Ij} & \cdots & n_{IJ} \end{bmatrix}$$

Donde la variable renglón tiene I categorías y la variable columna tiene J categorías.

Totales por columna son: $\sum_{i=1}^{I} n_{ij} = n_{.j}$

Totales por renglón: $\sum_{j=1}^{J} n_{ij} = n_{i.}$

Total general: $\sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij} = n$

frecuencias relativas: $r_i = n_{i.}/n$ y $c_j = n_{.j}/n$;

$i = 1, \cdots, I$ y $j = 1, \cdots, J$

Prueba de Independencia

La pregunta que surge es si las dos variables son **independientes**, es decir si los datos se acomodan en la tabla de manera proporcional al total de los renglones y el total de las columnas. Si los datos no se acomodan de manera proporcional diremos que ciertas categorías de las variables están **asociadas**. Es claro que las proporciones no resultan exactas sino que hay variaciones aleatorias, pero si las diferencias son muy grandes con esas proporciones "esperadas", se dirá que las variables no son independientes.

Siendo $p_{i,j}$ la probabilidad de que una observación pertenezca a la celda i, j, entonces la frecuencia esperada $F_{i,j}$ de observaciones despues de sacar una muestra tamaño n es $F_{ij} = n \times p_{i,j}$.

Ahora si $p_{i.}$ es la probabilidad de pertenecer al renglón i y $p_{.j}$ es la probabilidad de pertenecer a la columna j, cuando las variables son independientes se tiene $p_{ij} = p_{i.} \times p_{.j}$

Entonces las frecuencias esperadas cuando las variables son independientes son:

$$F_{ij} = n \times p_{i.} \times p_{.j}$$

Estas probabilidades no se conocen, pero pueden ser estimadas con

$$\widehat{p}_{i.} = \frac{n_{i.}}{n}$$

y

$$\widehat{p}_{.j} = \frac{n_{.j}}{n}$$

y las frecuencias esperadas se estiman con

$$E_{ij} = n\widehat{p}_{i.}\widehat{p}_{.j} = n\frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}$$

Hay ocasiones que los totales por renglón son fijos, esto por el diseño del muestreo, si la variables respuesta Y es binaria se tiene un modelo binomial, si tiene más categorías se tiene un esquema multinomial. Y en ese caso nos fijamos en las distribuciones condicionales para cada nivel dela variable X. Aquí la independencia entre X y Y puede expresarse también como que las condicionales de Y para cada nivel de la variable X son las mismas.

Otra situación es cuando n es fija y clasificamos a los individuos al «cruzar» las dos variables respuesta, en ese caso se tiene una distribución multinomial con $I \times J$ categorías.

La χ^2

Para analizar si las variables son independientes se puede utilizar la estadística χ^2, dada por

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \frac{n_{i.} n_{.j}}{n})^2}{\frac{n_{i.} n_{.j}}{n}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Cuando esta estadística toma valores "grandes" se rechaza la hipótesis nula de independencia.

La distribución asintótica de esta χ^2 puede hallarse suponiendo que las frecuencias observadas siguen una distribución multinio-mial y que las frecuencias esperadas no son muy pequeñas y corresponde a una distribución **Ji cuadrada** con $(I-1)*(J-1)$ grados de libertad.