Análisis de datos Categóricos

Introducción

Ms Carlos López de Castilla Vásquez

Universidad Nacional Agraria La Molina

2017-1

Variable cualitativa

- Una *variable cualitativa* es aquella cuya escala de medida consiste de un conjunto de categorías.
- Por ejemplo: la orientación política se mide como izquierda, centro o derecha.
- Por ejemplo: el diagnóstico de cáncer de mama se mide como normal, benigno, probablemente benigno, sospechoso o maligno.
- Por ejemplo: las enfermedades mentales pueden ser clasificadas en esquizofrenia, depresión o neurosis.

Variable respuesta y explicativa

- Muchas de las herramientas estadísticas hacen una distinción entre la variable *respuesta* (o *dependiente*) y variables *explicativas* (o *independientes*).
- Por ejemplo: los modelos de regresión describen como la media de una variable respuesta, como el precio de venta de una casa, cambia de acuerdo a los valores de variables explicativas, como el área total y la ubicación.
- En este curso el análisis se enfoca al caso en que la variable respuesta es de tipo cualitativa.

Variable de conteo y proporción

- Una *variable de conteo* es aquella que representa la frecuencia de ocurrencia de un evento.
- Por ejemplo: el número de personas que responden correctamente una encuesta, el número de autos mal estacionados en un centro comercial, etc.
- Las *variables de tipo proporción* representan la razón entre el número de éxitos y el número de eventos.
- Por ejemplo: la proporción de pacientes que responden satisfactoriamente a un antibiótico, la proporción de estudiantes de Biología que se matriculan en el curso de Estadística General, etc.

Escalas de medición

- Las variables cualitativas cuyas categorías no presentan un orden natural son llamadas *nominales*. Para una variable nominal el orden de listado de las categorías es irrelevante.
- Por ejemplo: la filiación religiosa, el tipo de transporte utilizado para ir al trabajo, género musical favorito, etc.
- Las variables cualitativas cuyas categorías presentan un orden natural son llamadas *ordinales*. Los métodos de análisis de estas variables consideran el orden de las categorías.
- Por ejemplo: el tamaño de un automóvil, el nivel socioeconómico, el grado de instrucción, etc.

Escalas de medición

- Una variable cuantitativa de *intervalo* es aquella en la que solo tienen significado las distancias numéricas entre dos valores cualesquiera. En esta escala el cero no indica la ausencia de la característica que se mide.
- Por ejemplo: la temperatura, las puntuaciones del coeficiente intelectual, las fechas de calendario, etc.
- Una variable cuantitativa de *razón* es aquella que permite calcular razones o proporciones entre dos valores cualesquiera.
- Por ejemplo: la presión sanguinea, el ingreso familiar, la edad de un paciente, etc.

Escalas de medición

- La forma en que se mide una variable determina como ésta se clasifica.
- Por ejemplo: la variable educación es nominal si se mide como pública o privada; es ordinal si se mide como el máximo grado obtenido (primaria, secundaria, superior) y es de razón si se mide como el número de años de educación.
- La escala de medición de una variable determina el método estadístico a utilizar para su análisis.
- Las variables cuantitativas se clasifican en *continuas* o *discretas* de acuerdo al número de valores que puedan tomar.

Escalas de medición

- La medición de las variables se hace de manera discreta debido a las limitaciones en los instrumentos de medición.
- En la práctica se considera que una variable continua es aquella que toma un conjunto muy grande de valores mientras que una variable discreta es aquella que toma un conjunto pequeño de valores.
- Las variables ordinales suelen tratarse como variables cualitativas usando métodos para variables nominales o asignando puntuaciones a las categorías para darle una naturaleza cuantitativa.

Distribución binomial

- Sean $y_1, y_2, \cdots, y_n$ las respuestas obtenidas en n ensayos independientes.
- Suponga que:

$$\Pr(Y_i = 1) = \pi \quad \text{y} \quad \Pr(Y_i = 0) = 1 - \pi$$

- Se usa *éxito* y *fracaso* para denotar los valores 1 y 0 respectivamente.
- Se asume que la probabilidad de éxito, π, es constante.
- Las variables aleatorias independientes $\{Y_i\}$ son llamadas *ensayos de Bernoulli*.

Distribución binomial

- El número total de éxitos $Y = \sum_{i=1}^{n} Y_i$ tiene *distribución binomial* con parámetros n y π.
- La función de probabilidad es:

$$f(y) = \frac{n!}{y!\,(n-y)!}\pi^y\,(1-\pi)^{n-y} \qquad y = 0, 1, \cdots, n$$

- Se denota como $Y \sim \mathcal{BI}(n, \pi)$. La media y varianza estan dadas por:

$$\mu = \mathrm{E}(Y) = n\pi \qquad \sigma^2 = \mathrm{Var}(Y) = n\pi(1-\pi)$$

Distribución multinomial

- Suponga que un conjunto de n ensayos independientes puede resultar en cualquiera de c categorías.
- Sea $y_{ij} = 1$ si el ensayo i resulta en la categoría j y $y_{ij} = 0$ en caso contrario.
- Se tiene que:

$$\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{ic})$$

representa un ensayo multinomial.

- Si $n_j = \sum_i y_{ij}$ el número de ensayos que caen en la categoría j entonces $(n_1, n_2, \cdots, n_c)$ tiene *distribución multinomial*.
- Se denota como $(n_1, n_2, \cdots, n_c) \sim \mathcal{M}(n, \{\pi_j\})$.

Distribución multinomial

- Sea $\pi_j = \Pr(Y_{ij} = 1)$ que denota la probabilidad que el ensayo i resulte en la categoría j.
- La función de probabilidad es:

$$f(n_1, n_2, \cdots, n_c) = \left(\frac{n!}{n_1! n_2! \cdots n_c!}\right) \prod_{i=1}^{c} \pi_i^{n_i}$$

con $\sum_j n_j = n$.

- La media, varianza y covarianza son:

$$\mathrm{E}(n_j) = n\pi_j \quad \mathrm{Var}(n_j) = n\pi_j(1-\pi_j) \quad \mathrm{Cov}(n_j, n_k) = -n\pi_j\pi_k$$

Distribución de Poisson

- Muchas veces los procesos de conteo no se obtienen a partir de un número fijo de ensayos por lo que no es posible establecer un límite superior.
- En este caso suele utilizarse la distribución de *Poisson*.
- La función de probabilidad es:

$$f(y) = \frac{e^{-\mu}\mu^{y}}{y!} \qquad y = 0, 1, \cdots$$

- Se denota por $Y \sim \mathcal{P}(\mu)$. La media y la varianza son:

$$\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$$

Sobredispersión

- En la práctica los conteos presentan mayor variablidad de la que se asume con la distribución binomial o Poisson. Este fenómeno es llamado *sobredispersión*.
- Por ejemplo: en lugar de considerar que cada persona tiene la misma probabilidad de tener un accidente fatal debería considerase que ésta depende de *factores* como la velocidad, el uso del cinturón, etc.
- Los factores anteriores son los responsables de tener mayor variación de la que se establece con las distribuciones mencionadas.
- La distribución *binomial negativa* permite que la varianza exceda el valor de la media en situaciones como la descrita.

Poisson y multinomial

- Sean c variables aleatorias independientes con distribución de Poisson tal que $\mathrm{E}(Y_i) = \mu_i$.
- La distribución condicional para $Y_1 = n_1, \cdots, Y_c = n_c$ dado $\sum Y_i = n$ es:

$$\begin{aligned} \frac{\Pr(Y_1 = n_1, \cdots, Y_c = n_c)}{\Pr(\sum Y_i = n)} &= \frac{\prod_i \left[\exp\{-\mu_i\}\,\mu_i^{n_i}/n_i!\right]}{\exp\{-\sum \mu_i\}\left(\sum \mu_i\right)^n/n!} \\ &= \left(\frac{n!}{n_1!n_2!\cdots n_c!}\right)\prod_i \pi_i^{n_i} \end{aligned}$$

donde $\pi_i = \mu_i / \sum \mu_j$, es decir la distribución es $\mathcal{M}(n, \{\pi_i\})$.

Estimadores de máxima verosimilitud

- Los estimadores de *máxima verosimilitud* tienen propiedades importantes cuando el tamaño de muestra es grande.
- Son asintóticamente consistentes, eficientes y convergen hacia la distribución normal.
- El *estimador de máxima verosimilitud* (EMV) es el valor del párametro que maximiza la *función de verosimilitud*.
- La función de verosimilitud es la *distribución de probabilidad conjunta* para la data luego de haberla observado.
- Se denota la función de verosimilitud por $L(\beta)$ y su logaritmo por $l(\beta)$.

Estimadores de máxima verosimilitud

- El estimador de máxima verosimilitud de $\boldsymbol{\beta}$ se denota por $\hat{\boldsymbol{\beta}}$ y es la solución de:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \quad \text{siempre que} \quad \left.\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} < 0$$

- La matriz de covarianza asintótica $\text{Cov}(\hat{\boldsymbol{\beta}})$ es la inversa de la *matriz de información* y los *errores estándar* son las raíces cuadradas de los elementos en su diagonal.
- El elemento (j, k) de la matriz de información es:

$$-\text{E}\left(\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right)$$

Prueba de hipótesis

- Suponga que se desea probar $H_0 : \beta = \beta_0$
- El estadístico de prueba de *Wald* es:

$$Z_W = \frac{\hat{\beta} - \beta_0}{\widehat{\mathrm{EE}}(\hat{\beta})}$$

- El *estadístico Score* es:

$$Z_S = \frac{\frac{\partial l(\beta)}{\partial \beta_0}}{\sqrt{\iota(\beta_0)}}$$

donde $\iota(\beta_0) = -\mathrm{E}\left(\frac{\partial^2 l(\beta)}{\partial \beta_0^2}\right)$.

Prueba de hipótesis

- Sea L_0 el máximo valor de la función de verosimilitud bajo las restricciones que impone H_0.
- Sea L_1 el máximo valor de la función de verosimilitud sin restricción.
- La prueba de *razón de verosimilitud* es:

$$X^2_{RV} = -2 \log \Lambda = -2 \log \left(L_0 / L_1 \right) = -2 \left(l_0 - l_1 \right)$$

y su distribución límite es chi-cuadrado.

- Los grados de libertad se obtienen como la diferencia de dimensiones del espacio paramétrico completo y el que se establece por H_0.

Intervalos de confianza

- En los tres métodos el intervalo de confianza se obtiene invirtiendo la prueba de hipótesis con respecto de β_0.
- El intervalo basado en el estadístico de Wald:

$$|Z_W| < z_{1-\alpha/2}$$

- El intervalo basado en el estadístico score:

$$Z_S^2 < \chi^2_{1,1-\alpha}$$

- El intervalo basado en la prueba de razón de verosimilitud:

$$X^2_{RV} < \chi^2_{d,1-\alpha}$$

Estimador de máxima verosimilitud

- El estimador de máxima verosimilitud:

$$\hat{\pi} = \frac{y}{n} = \frac{\sum_{i=1}^{n} y_i}{n}$$

es un estimador insesgado. Además:

$$-\mathrm{E}\left(\frac{\partial^2 l(\pi)}{\partial \pi^2}\right) = \frac{n}{\pi(1-\pi)}$$

- La varianza asintótica es:

$$\mathrm{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

Prueba de hipótesis

- Considere $H_0 : \pi = \pi_0$. El estadístico de Wald es:

$$z_W = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$$

- El estadístico score es:

$$z_S = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

- La prueba de razón de verosimilitud es:

$$X^2_{RV} = 2\left(y \log \frac{\hat{\pi}}{\pi_0} + (n-y) \log \frac{1-\hat{\pi}}{1-\pi_0}\right)$$

Intervalo de confianza

- El intervalo basado en el estadístico de Wald:

$$\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

- El intervalo basado en el estadístico score:

$$\hat{\pi}\left(\frac{n}{n+z_{1-\alpha/2}^2}\right) + \frac{1}{2}\left(\frac{z_{1-\alpha/2}^2}{n+z_{1-\alpha/2}^2}\right) \pm z_{1-\alpha/2}$$

$$\sqrt{\frac{1}{n+z_{1-\alpha/2}^2}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+z_{1-\alpha/2}^2}\right) + \left(\frac{1}{4}\right)\left(\frac{z_{1-\alpha/2}^2}{n+z_{1-\alpha/2}^2}\right)\right]}$$

Intervalo de confianza

- El intervalo basado en la prueba de razón de verosimilitud:

$$2\left(y\log\frac{\hat{\pi}}{\pi}+(n-y)\log\frac{1-\hat{\pi}}{1-\pi}\right)\leq\chi^2_{1,1-\alpha}$$

Intervalos de confianza

- Suponga que en una muestra de $n=250$ estudiantes $y=50$ respondieron ser vegetarianos. Hallar un intervalo del 95 % de confianza para π usando los métodos mencionados.

```
> library(Hmisc)
> binconf(x = 50, n = 250, method = "asymptotic")
> prop.test(x = 50, n = 250, conf.level = 0.95, correct = F)
```

Estimadores de máxima verosimilitud

- El logaritmo de la función de verosimilitud multinomial es:

$$l(\boldsymbol{\pi}) = \sum_j n_j \log \pi_j$$

- Como $\pi_c = 1 - (\pi_1 + \cdots + \pi_{c-1})$ y $\sum_j n_j = n$ entonces:

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0$$

- Los estimadores de máxima verosimilitud son $\hat{\pi}_j = \frac{n_j}{n}$ para $j = 1, \cdots, c-1$.

Estadístico de Pearson

- Se usa para establecer una prueba de hipótesis sobre los parámetros de la distribución multinomial.
- Considere $H_0 : \pi_j = \pi_{j0}$ para $j = 1, \cdots, c$ donde $\sum_j \pi_{j0} = 1$.
- Si la hipótesis nula es verdadera las *frecuencias esperadas* son:

$$\mu_j = n\pi_{j0}$$

para $j = 1, \cdots, c$.

- Pearson (1900) propuso el estadístico:

$$X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j} \sim \chi^2_{c-1}$$

Estadístico de razón de verosimilitud

- Esta prueba es una alternativa para el estadístico de Pearson.
- La razón de verosimilitud es:

$$\Lambda = \frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j (\hat{\pi}_j)^{n_j}}$$

- El estadístico de prueba es:

$$G^2 = -2 \log \Lambda = 2 \sum n_j \log(\hat{\pi}_j / \pi_{j0})$$

- Para n grande, G^2 tiene distribución chi-cuadrado con $c - 1$ grados de libertad.

Ejemplo: Teoría de Mendel

Prueba de Pearson

- Mendel cruzó plantas de arvejas de pura cepa color amarillo con plantas de pura cepa color verde y predijo que la segunda generación de semillas seria 75 % amarilla y 25 % verde.
- Un experimento produjo $n = 8023$ semillas donde $n_1 = 6022$ fueron amarillas y $n_2 = 2001$ verdes.
- ¿Los resultados anteriores contradicen la hipótesis de Mendel?

```
> n <- c(6022, 2001)
> prob <- c(0.75, 0.25)
> chisq.test(x = n, p = prob)
```