Tema 1: Introducción al Análisis de datos Categóricos

Introducción

Los datos categóricos aparecen cuando una variable se mide en una escala que sólo incluye a los posibles encuestados en un número limitado de grupos. Por ejemplo, una encuesta donde se recoge información sobre variables como el género, estado civil o afiliación política.

Además de distinguir una variable como categórica (*cualitativa*) o continua (*cuantitativa*), las variables también se pueden clasificar como *independientes* o *dependientes*. El término independiente se refiere a una variable que se puede manipular experimentalmente (e.j. el tipo de tratamiento que se le asigna a cada persona), pero también se aplica a menudo a una variable que se utiliza para predecir otra variable (e.j. nivel socio–económico).

El término *dependiente* se refiere en general a una variable cuyo interés primordial es el resultado o la respuesta. Por ejemplo, se pueden considerar como variables dependientes, el resultado de un tratamiento (basado en el tipo del mismo) o el nivel educativo previsto a partir de una situación socio–económica,.

Ejemplo: supongamos que se desea determinar si los colegios concertados difieren de manera sustancial de los colegios privados y públicos en base a ciertos datos demográficos como la *ubicación* (urbano, suburbano o rural), el *tipo* (pública o privada), la *situación predominante socio-económica de los estudiantes* (bajo, medio o alto) etc. Para este tipo de análisis es necesario usar técnicas de análisis de datos categóricos, porque todas las variables involucradas son categóricas.

Ejemplo: supongamos que un sociólogo quiere predecir si un estudiante se graduará en *secundaria* en base a cierta información como el número de días de asistencia, promedio de las calificaciones y los ingresos familiares. En este caso, se necesita un enfoque de análisis categórico donde la graduación (*sí* o *no*) sirve como variable dependiente en función de otras variables explicativas.

Escalas de medida

La escala de medida de una variable de respuesta categórica es fundamental para la elección del análisis estadístico apropiado.

Las variables de respuesta categórica pueden ser

- Dicotómicas
- Ordinales
- Nominales
- De recuento

Respuestas dicotómicas son aquellas que tienen dos posibles resultados que a menudo son *sí* y *no*. ¿Se desarrollará la enfermedad? ¿El votante votará por el candidato A o por el B? ¿Aprobará el examen?

Con frecuencia, las respuestas de los datos categóricos representan más de dos resultados posibles y a veces en estos resultados es posible considerar algún orden inherente. Estas variables tienen una escala de respuesta *ordinal* de medición. ¿El nuevo plan de estudios gusta a los estudiantes? ¿La muestra de agua es de dureza baja, media o alta?

En el primer caso del nuevo plan de estudios, el orden de los niveles de respuesta es claro, pero no hay ninguna pista en cuanto a las *distancias relativas* entre los niveles.

En el segundo caso de la dureza del agua, hay una distancia posible entre los niveles: *medio* podría tener el doble de la dureza de *baja* y *alta* podría tener tres veces la dureza de *baja*.

Si existen más de dos categorías posibles y no hay un orden inherente entre ellas entonces se tiene una escala de medida **nominal**. No existe una escala subyacente en esos resultados y no hay una forma aparente de ordenarlos.

Ejemplos: ¿A cuál de los cuatro candidatos votaste en las elecciones municipales de la ciudad? ¿Prefieres la playa, la montaña o la ciudad para ir de vacaciones?

Las variables categóricas a veces contienen recuentos. En lugar de considerar las categorías que presenta cada observación (sí, no) o (bajo, medio, alto), los resultados que se estudian son los mismos números o recuentos de apariciones.

Ejemplos: El tamaño de una camada, ¿fue de 1, 2, 3, 4 ó 5 animales? La oficina tiene ¿1, 2, 3 ó 4 equipos de aire acondicionado?

En la metodología clásica habitual se analiza la *media* de los recuentos, pero los supuestos que se tienen que cumplir en un modelo lineal estándar con datos continuos no se cumplen a menudo con datos discretos. En general, los recuentos no se distribuyen según una distribución normal y la varianza no suele ser homogénea.

Revisión de distribuciones de Probabilidad

Distribución binomial

Ensayos de *Bernoulli*: Habitualmente, los datos proceden de n ensayos independientes e idénticos con dos posibles resultados para cada uno: *éxito* y *fracaso* con igual probabilidad de éxito para cada prueba. Ensayos independientes significa que los resultados son variables aleatorias independientes. En particular, el resultado de una prueba no afecta al resultado de otra.

Así, se denota como π a la probabilidad de éxito para un ensayo dado e Y denota el número de éxitos de las n pruebas:

$$P(y) = \frac{n!}{y!\,(n-y)!}\pi^y(1-\pi)^{n-y}$$

para $0, 1, 2, \ldots, n$.

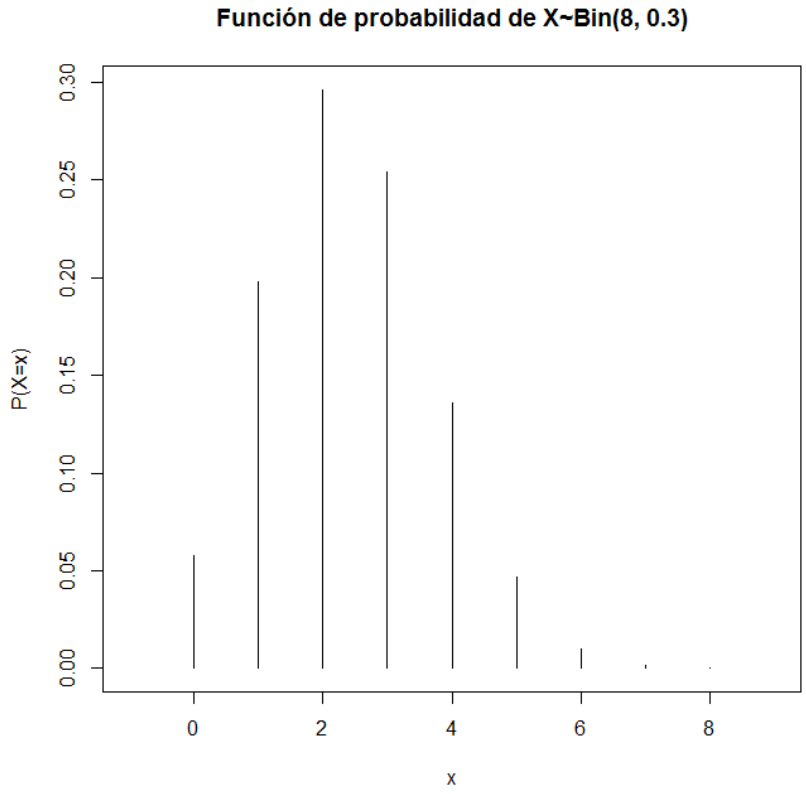
La distribución binomial para n ensayos con parámetro π tiene como media y desviación estándar:

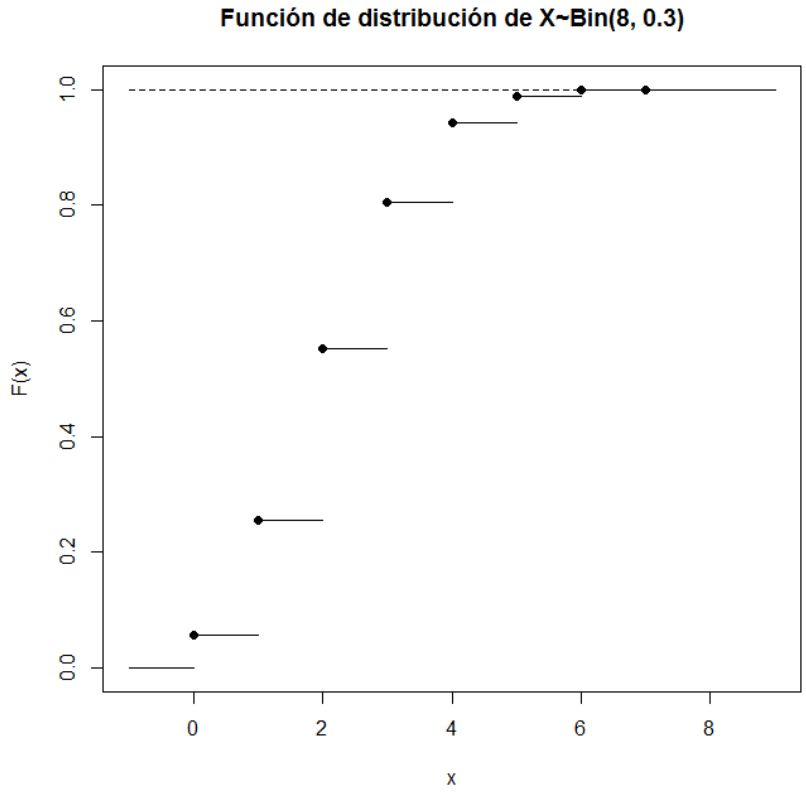
$$E(Y) = \mu = n\pi$$

$$\sigma = \sqrt{n\pi(1-\pi)}$$

Las gráficas de las funciones de probabilidad y distribución son, respectivamente,

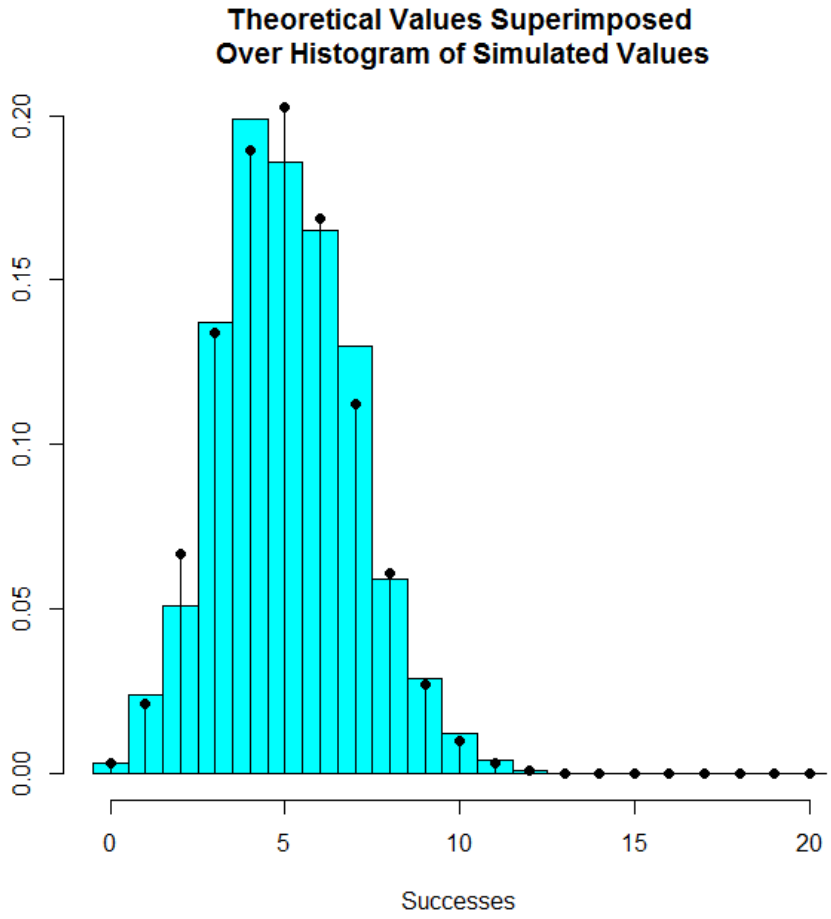
```
# Script de R
# Funcion de probabilidad de una binomial
X11()
plot(0:8, dbinom(0:8,8,0.3), type="h", xlab="x",ylab="P(X=x)",
xlim=c(-1,9))
title("Funcion de probabilidad de X~Bin(8, 0.3)")

# Funcion de distribucion de una binomial
X11()
plot(0:8, pbinom(0:8,8,0.3), type="n", xlab="x", ylab="F(x)",
xlim=c(-1,9), ylim=c(0,1))
segments(-1,0,0,0)
segments(0:8, pbinom(0:8,8,.3), 1:9, pbinom(0:8,8,.3))
lines(0:7, pbinom(0:7,8,.3), type="p", pch=16)
segments(-1,1,9,1, lty=2)
title("Funcion de distribucion de X~Bin(8, 0.3)")
```





Para generar 1000 observaciones de una distribución binomial $Bin\,(n = 5, p = 0{,}5)$:

```
# Uso la libreria PASWR
library(PASWR)
bino.gen(1000, 5, 0.5)
```



Distribución Multinomial

Algunos ensayos tienen más de dos resultados posibles. Por ejemplo, el resultado de un accidente de automóvil se puede clasificar en varias posibles categorías:

1. *Sin lesiones*,
2. *Lesiones que no requieren hospitalización*,
3. *Lesiones que requieren hospitalización*,
4. *Muerte*.

Cuando los ensayos son independientes respecto a cada categoría, la distribución de los recuentos en cada categoría sigue una distribución **multinomial**.

Supongamos que se tienen c posibles categorías. Dados n sucesos, se puede definir una variable aleatoria X_i (para $i = 1, \ldots, c$) que indica el número de veces que aparece la categoría i.

Se denota la probabilidad de obtener cada categoría i como $\{\pi_1, \pi_2, \ldots, \pi_c\}$ donde $\sum_i \pi_i = 1$.

Para n observaciones independientes, la probabilidad de que n_1 observaciones caigan en la categoría 1, n_2 caigan en la categoría 2, ..., n_c caigan en la categoría c, (donde $\sum_i n_i = n$) es igual a

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

La distribución binomial es, en realidad, un caso particular de la distribución multinomial cuando $c = 2$.

Ejemplos

Se puede generar una muestra de tamaño 10 de una multinomial $Mult(12, (0{,}1, 0{,}2, 0{,}7))$, y calcular la probabilidad conjunta del vector $(3, 7, 2)$ y del vector $(1, 2, 9)$.

```
# Distribucion multinomial con R
rmultinom(10, size=12, prob=c(0.1, 0.2, 0.7))
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    2    1    1    2    0    3    2    1     3
[2,]    1    1    2    2    2    2    1    1    2     2
[3,]   10    9    9    9    8   10    8    9    9     7
```

```
dmultinom(c(3, 7, 2), prob=c(0.1, 0.2, 0.7))
```

```
[1] 4.967424e-05
```

```
dmultinom(c(1, 2, 9), prob=c(0.1, 0.2, 0.7))
```

```
[1] 0.1065335
```

La esperanza y varianza de observar el suceso i en n ensayos es

$$\begin{aligned} E\left(X_i\right) &= np_i \\ Var\left(X_i\right) &= np_i(1 - p_i) \end{aligned}$$

La covarianza entre los sucesos i y j observados en n ensayos es

$$Cov\left(X_i, X_j\right) = -np_i p_j \qquad (i \neq j)$$

Inferencia para la distribución binomial

El método habitual en Inferencia Estadística, desde el punto de vista clásico, es la estimación por máxima verosimilitud. El estimador de máxima verosimilitud de un parámetro es el valor del parámetro para el que la probabilidad de obtener los datos observados es mayor.

Por ejemplo, si en $n = 10$ ensayos se obtienen 0 éxitos, la función de verosimilitud en este caso $L(\pi) = (1 - \pi)^{10}$ que alcanza el máximo en $\hat{\pi} = 0$. Es decir el resultado de 0 éxitos en 10 ensayos es más probable que ocurra cuando $\pi = 0$ que para cualquier otro valor.

Así, en general, si una variable aleatoria X se observa con x éxitos en n ensayos, el estimador de máxima verosimilitud (EMV) de la probabilidad de éxito p es simplemente $\hat{p} = x/n$, la proporción observada de éxitos entre n ensayos. La varianza es

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

y un intervalo de confianza al $100\times(1-\alpha)$ aproximado para p es

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

que se denomina habitualmente **intervalo de Wald**.

Una mejor alternativa es el llamado **intervalo de Wilson** (o *q-interval*) que se calcula como un subproducto del *Teorema Central del Límite*.

$$\widehat{p}\left(\frac{n}{n+z^2_{\frac{\alpha}{2}}}\right) + \frac{1}{2}\left(\frac{z^2_{\frac{\alpha}{2}}}{n+z^2_{\frac{\alpha}{2}}}\right) \pm \sqrt{\left[\frac{\widehat{p}(1-\widehat{p})}{n}\right]\left[\frac{n^2 z^2_{\frac{\alpha}{2}}}{\left(n+z^2_{\frac{\alpha}{2}}\right)^2}\right] + \frac{1}{4}\left[\frac{z^4_{\frac{\alpha}{2}}}{\left(n+z^2_{\frac{\alpha}{2}}\right)^2}\right]}$$

Esta aproximación funciona mejor que el intervalo de Wald para valores pequeños de n.

Estimación numérica de la función de verosimilitud

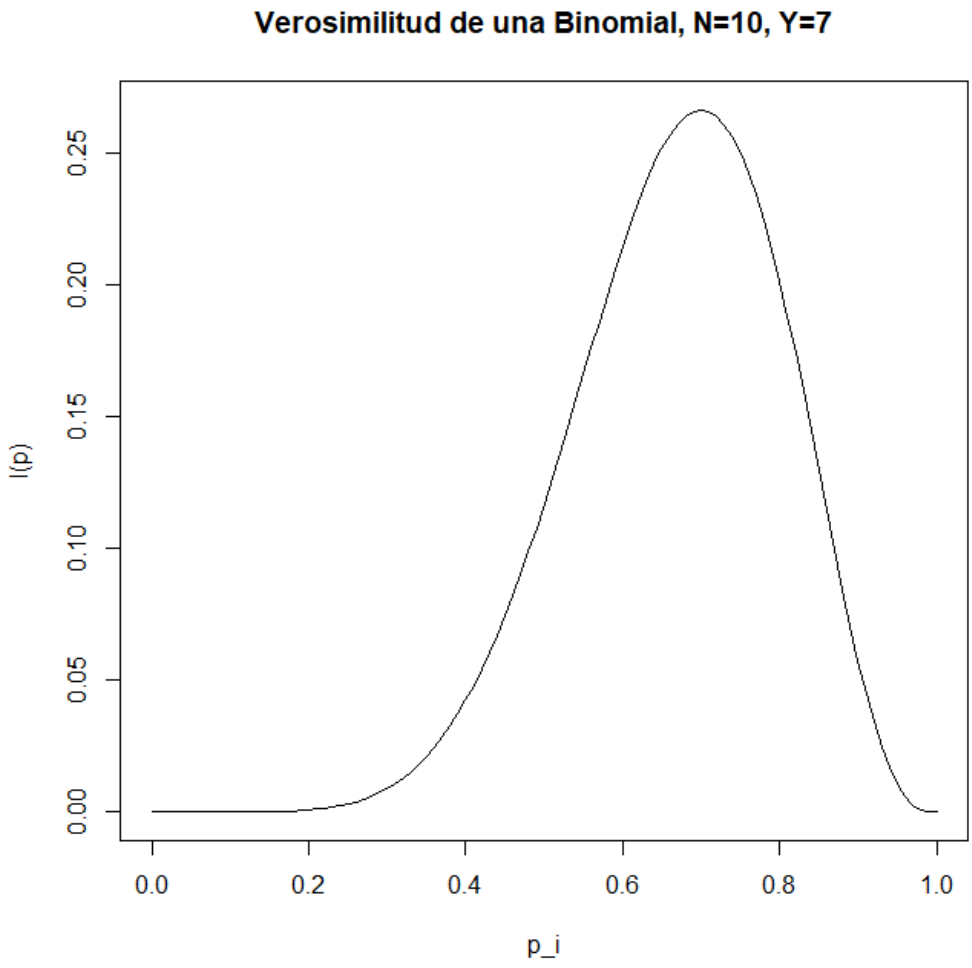
En `R` se pueden calcular de manera numérica los estimadores de máxima verosimilitud.

Por ejemplo:

```
# Funcion de verosimilitud de una binomial
# Se define la funcion de verosimilitud para una muestra de
# una binomial con N=10 Y=7 exitos
```

```r
lklhd = function(p){ dbinom(7,10,p) }

# Grafica de la funcion de verosimilitud
plot(lklhd, 0, 1, xlab="p_i", ylab="l(p)",
main="Verosimilitud de una Binomial, N=10, Y=7")

# Estimador de maxima verosimiltud
optimize(lklhd, c(0,1), maximum=TRUE)
```

```
$maximum
[1] 0.6999843

$objective
[1] 0.2668279
```



Los intervalos de confianza *exactos*, o intervalos de *Clopper-Pearson*, se basan en el cálculo mediante la función de distribución de la distribución binomial, no en aproximaciones mediante la normal. Sin embargo, no son exactos en el sentido de que la función de distribución binomial es discontinua en sí, lo que impide en la realidad el cálculo exacto de los intervalos de confianza fijados en un $(1-\alpha)$.

Estos intervalos son conservadores, es decir, suelen ser mayores que los calculados según los métodos asintóticos.

Test de hipótesis usando el método de la razón de verosimilitudes

En el método de la razón de verosimilitudes se compara la verosimilitud (*plausibilidad*) de los datos observados usando la proporción especificada bajo la hipótesis nula, respecto a la verosimilitud de los datos observados usando la estimación muestral.

La verosimilitud obtenida bajo la hipótesis nula se denota mediante L_0 y la verosimilitud obtenida usando el estimador muestral se denota como L_1.

El cociente L_0/L_1 representa la razón de verosimilitudes. Si L_1 (la verosimilitud obtenida a partir de los datos observados) es mucho mayor que L_0 (la verosimilitud bajo la hipótesis nula H_0) la razón de verosimilitudes será pequeña e indicará que los datos muestran evidencias en contra de la hipótesis nula.

El test de la razón de verosimilitudes se obtiene tomando el logaritmo (log) de la razón de verosimilitudes y multiplicándolo por –2. En concreto,

$$G^2 = -2\log\left(\frac{L_0}{L_1}\right) = -2\left[\log\left(L_0\right) - \log\left(L_1\right)\right].$$

Un valor alto de G^2 (más positivo) indica una mayor evidencia en contra de H_0. Bajo la hipótesis nula y para una muestra razonablemente grande, G^2 sigue una distribución χ^2 con los grados de libertad igual al número de parámetros libres que existen bajo la hipótesis nula. En el caso del problema de las proporciones es igual a 1.

Ejemplo

Se tiene una muestra de 2818 personas de modo que el 34 % bebe la cantidad diaria de agua recomendada en general (1.5 litros). El intervalo de confianza aproximado del 95 % para la proporción de toda la población es

$$0{,}34 \pm \sqrt{\frac{0{,}34 \cdot 0{,}66}{2818}} = 0{,}34 \pm 0{,}017 = (0{,}323, 0{,}357).$$

El intervalo de Wilson correspondiente es

$$\widehat{p}\left(\frac{n}{n+z^2_{\frac{\alpha}{s}}}\right)+\frac{1}{2}\left(\frac{z^2_{\frac{\alpha}{s}}}{n+z^2_{\frac{\alpha}{s}}}\right)\pm\sqrt{\left[\frac{\widehat{p}(1-\widehat{p})}{n}\right]\left[\frac{n^2z^2_{\frac{\alpha}{s}}}{\left(n+z^2_{\frac{\alpha}{s}}\right)^2}\right]+\frac{1}{4}\left[\frac{z^4_{\frac{\alpha}{s}}}{\left(n+z^2_{\frac{\alpha}{s}}\right)^2}\right]}=$$

$$0{,}34\cdot\left(\frac{2818}{2818+1{,}96^2}\right)+\frac{1}{2}\left(\frac{1{,}96^2}{2818+1{,}96^2}\right)\pm$$

$$\pm\sqrt{\frac{0{,}34\cdot0{,}66}{2818}\cdot\left(\frac{2818^2\cdot1{,}96^2}{(2818+1{,}96^2)^2}\right)+\frac{1}{4}\cdot\left(\frac{1{,}96^4}{(2818+1{,}96^2)^2}\right)}$$

$$=(0{,}32274,0{,}35770)$$

Ejemplo

Se puede usar `R` para calcular intervalos de confianza para el parámetro p de la binomial. Supongamos un ensayo con 50 observaciones, entre los que se encuentran 46 éxitos. Se trata de calcular los intervalos de confianza.

```
# x=46 exitos
# n=50 observaciones
prop.test(x=46, n=50, conf.level=0.95, correct=F)
```

```
        1-sample proportions test without continuity correction

data:  46 out of 50, null probability 0.5
X-squared = 35.28, df = 1, p-value = 2.855e-09
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.8116175 0.9684505
sample estimates:
   p
0.92
```

```
# Alternativa
library(Hmisc)
binconf(46, 50, method="all")
```

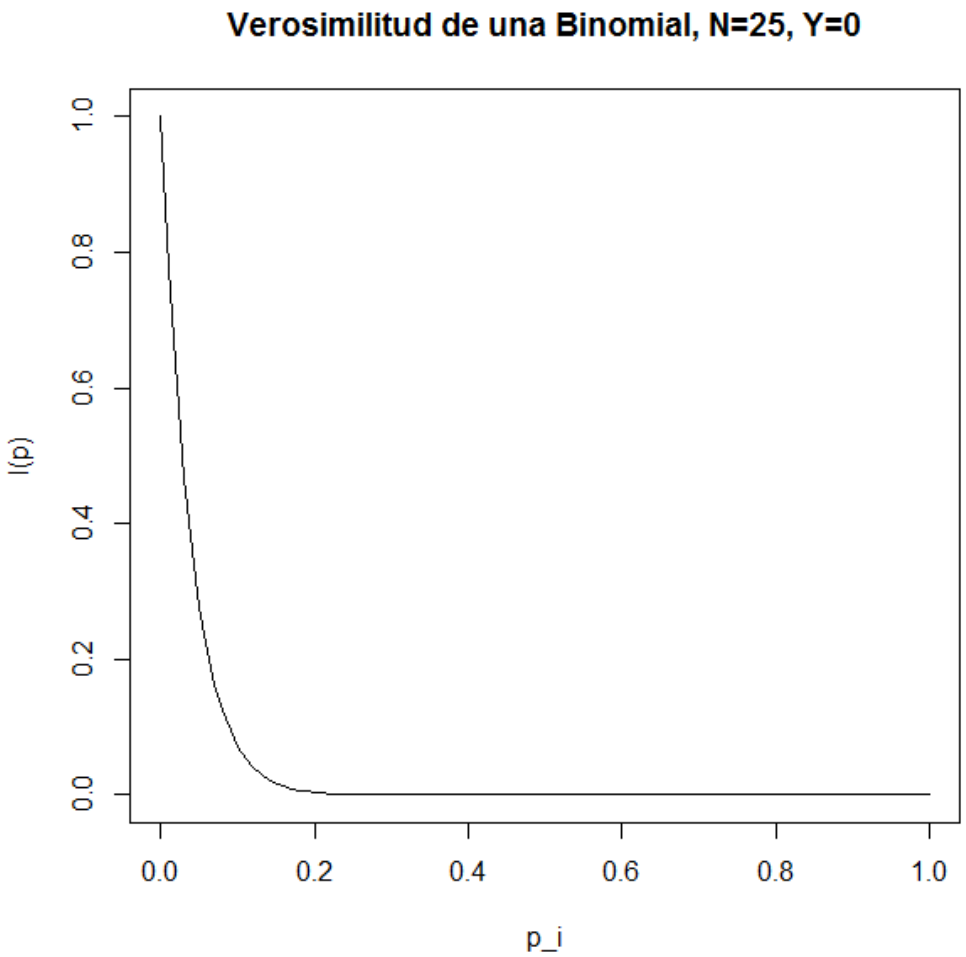
```
           PointEst     Lower     Upper
Exact          0.92 0.8076572 0.9777720
Wilson         0.92 0.8116175 0.9684505
Asymptotic     0.92 0.8448027 0.9951973
```

Ejemplo

Supongamos que se plantea un cuestionario y se toma una muestra aletoria de un grupo de estudiantes donde se pregunta si son *vegetarianos* o no. De entre $n = 25$ estudiantes, no hay ninguno que dice ser vegetariano.

Se trata de calcular los intervalos de confianza al 95 % para el parámetro p de la binomial.

```
lklhd = function(p){ dbinom(0,25,p) }

# Grafica de la funcion de verosimilitud
plot(lklhd, 0, 1, xlab="p_i", ylab="l(p)",
main="Verosimilitud de una Binomial, N=25, Y=0")
```



Se obtiene que

```
# Estimador de maxima verosimilitud mediante
optimize(lklhd, c(0,1), maximum=TRUE)
```

```
$maximum
[1] 6.610696e-05

$objective
[1] 0.9983486
```

```
res = prop.test(x=0, n=25, conf.level=0.95, correct=F)
res$conf.int
```

```
[1] 0.0000000 0.1331923
attr(,"conf.level")
[1] 0.95
```

Inferencia para la distribución multinomial

Los parámetros que se tienen que estimar son $(\pi_1, \ldots, \pi_c)$ donde $\pi_c = 1 - \sum_{i=1}^{c-1} \pi_i$.

Se obtiene que los estimadores por máxima verosimilitud son

$$\widehat{\pi_j} = \frac{n_j}{n}$$

para $j = 1, \ldots c$.

Es decir, los estimadores son simplemente las proporciones muestrales de cada categoría.

Contraste para la distribución de una multinomial

En 1900 Karl Pearson presentó una prueba de hipótesis que fue uno de los primeros métodos de inferencia que se inventaron y tuvo un gran impacto en el análisis de datos categóricos, que hasta ese momento se había centrado en la descripción de las asociaciones entre variables.

La prueba de Pearson evalúa si los parámetros de una multinomial son iguales a unos valores previos especificados.

Se considera como hipótesis nula

$H_0 : \pi_j = \pi_{j0}$

para $j = 1, \ldots, c$ donde $\sum_j \pi_{j0} = 1$.

Cuando la hipótesis nula es cierta, entonces (para $j = 1, \ldots, c$) los valores esperados de n_j o *frecuencias esperadas* son iguales a $n\pi_{j0}$.

Se define el siguiente estadístico:

$$\mathbb{X}^2 = \sum_j \frac{(n_j - n\pi_{j0})^2}{n\pi_{j0}}.$$

Intuitivamente, diferencias elevadas entre lo que se espera según H_0 y lo observado implica a su vez valores grandes de $\mathbb{X}^2$.

Para un el tamaño muestral *no pequeño*, $\mathbb{X}^2$ se distribuye como una chi cuadrado χ^2 con $c-1$ grados de libertad.

El test es muy sensible a los tamaños muestrales. Si alguna categoría tiene una frecuencia esperada baja (menor que 5) el test pierde mucha potencia ya que se basa en la aproximación asintótica a la distribución χ^2.

Ejemplo

El test de Pearson se usó en Genética para contrastar las teorías de Mendel sobre la herencia. Mendel cruzó guisantes amarillos puros con guisantes verdes puros. Su predicción era que 3/4 tenían que ser amarillos y 1/4 verdes.

En un experimento se obtuvo $n = 8023$ guisantes, de los cuales $n_1 = 6022$ fueron amarillos y $n_2 = 2001$ verdes.

Las frecuencias esperadas son, entonces,

$$\begin{aligned} H_0: \quad & \pi_A = 8023 \cdot 0{,}75 \\ & \pi_V = 8023 \cdot 0{,}25 \end{aligned}$$

De modo que

$$\mathbb{X}^2 = \frac{(6022 - 8023 \cdot 0{,}75)^2}{8023 \cdot 0{,}75} + \frac{(2001 - 8023 \cdot 0{,}25)^2}{8023 \cdot 0{,}25} = 1{,}59 \times 10^{-2}$$

```
# Con R
pchisq(0.015, 1, lower.tail=FALSE)
```

```
[1] 0.9025233
```

Es decir, comparando con una distribución χ^2 con 1 grado de libertad, se obtiene un p-valor igual a 0.90.

Alternativamente con el comando `chisq.test` se obtiene

```
chisq.test(x=c(6022, 2001), p=c(.75,.25))
```

```
        Chi-squared test for given probabilities

data:  c(6022, 2001)
X-squared = 0.015, df = 1, p-value = 0.9025
```

Pero Fisher (y otros) sospechó que Mendel tuvo *demasiada* suerte en su experimentación...

Fisher comentó textualmente:

El nivel general de acuerdo entre las expectativas de Mendel y sus resultados obtenidos muestra que está más cerca de lo esperado en el mejor caso de varios miles de repeticiones... No tengo duda de que Mendel fue engañado por un asistente de jardinería, que sabía muy bien lo que su jefe esperaba en cada ensayo.

Con `SAS` se usa el siguiente programa:

```
OPTIONS ls=70 nodate;
/* Para SAS University */
ODS rtf file='/folders/myfolders/resultado.rtf' style=minimal
    startpage=no;

DATA mendel;
INPUT guisante $ recuento;
DATALINES;
amarillos 6022
verdes 2001
;

PROC freq order=data;
weight recuento;
tables guisante / binomial (p=0.75) alpha=0.05;
exact binomial;
RUN;
ODS rtf close;
```

Se obtiene el siguiente resultado:

Procedimiento FREQ

guisante	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
amarillo	6022	75.06	6022	75.06
verdes	2001	24.94	8023	100.00

Proporción binomial	
guisante = amarillo	
Proporción (P)	0.7506
ASE	0.0048
95% Límite conf. inferior	0.7411
95% Límite conf. superior	0.7601
Límites conf. exactos	
95% Límite conf. inferior	0.7410
95% Límite conf. superior	0.7600

Test de H0: Proporción = 0.75	
ASE bajo H0	0.0048
Z	0.1225
Pr de un lado > Z	0.4513
Pr de dos lados > |Z|	0.9025
Test exacto	
Pr de un lado >= P	0.4572
Dos colas = 2 * Una cola	0.9144

Tamaño de la muestra = 8023

Ejemplo

El departamento de instrucción pública de Wisconsin usa cuatro categorías para medir las habilidades matemáticas: *advanced*, *proficient*, *basic* y *minimal*.

Se considera una muestra de 71709 estudiantes de 10º grado en 2006 y se supone que las proporciones se mantienen en los mismos niveles que en años anteriores. Los datos se recogen en la siguiente tabla:

Nivel matemáticas	**Proporción esperada**	**Frecuencia esperada**	**Frecuencia observada**
Advanced	15 %	10756.35	18644
Proficient	40 %	28683.60	32269
Basic	30 %	21512.70	10039
Minimal	15 %	10756.35	10757

Con `R` se usa el comando `chisq.test`.

```
chisq.test(x=c(18644, 32269, 10039, 10757),
p=c(0.15, 0.40, 0.30, 0.15))
```

```
        Chi-squared test for given probabilities

data:  c(18644, 32269, 10039, 10757)
X-squared = 12352, df = 3, p-value < 2.2e-16
```

Con SAS se usaría el siguiente programa:

```
OPTIONS ls=70 nodate;
/* Para SAS University */
ODS rtf file='/folders/myfolders/resultado.rtf' style=minimal
    startpage=no;

DATA Wisconsin;
INPUT niveles $ recuento;
DATALINES;
advanced 18644
proficient 32269
basic 10039
minimal 10757
;

PROC freq order=data;
weight recuento;
tables niveles / testp=(0.15 0.40 0.30 0.15);
RUN;
ODS rtf close;
```

Procedimiento FREQ

niveles	Frecuencia	Porcentaje	Porcentaje de test	Frecuencia acumulada	Porcentaje acumulado
advanced	18644	26.00	15.00	18644	26.00
proficie	32269	45.00	40.00	50913	71.00
basic	10039	14.00	30.00	60952	85.00
minimal	10757	15.00	15.00	71709	100.00

Test chi-cuadrado para proporciones especificadas	
Chi-cuadrado	12351.6415
DF	3
Pr > ChiSq	<.0001