ESTIMACION DE PROPORCIONES
¿Un problema aún no resuelto?

JAIRO ALFONSO CLAVIJO MENDEZ
Universidad del Tolima

Este documento tiene como fin hacer una síntesis tal vez incompleta del trabajo que se ha realizado a lo largo de casi 100 años para lograr una manera práctica de estimar una proporción.

Presentación del problema

Supóngase que estamos en una población finita de tamaño N conformada por elementos de dos clases: A de ellos del tipo E (éxitos) y $N-A$ del tipo F(fracasos).

La fracción $\pi = \frac{A}{N}$, usualmente desconocida, se denomina "Proporción de elementos A " o simplemente ***proporción*** cuando es claro a cuáles elementos estamos haciendo referencia.

Nuestro objetivo inmediato es estimar π mediante una muestra $\underline{\Psi}$ de n elementos seleccionados de la población bajo muestreo aleatorio simple (M.A.S).

Desarrollo

El problema, aparentemente sencillo, ha sido objeto de estudio durante casi 100 años sin que se pueda afirmar que en este momento haya una solución completa y definitiva para el mismo.

Comencemos diciendo que al aplicar M.A.S para seleccionar la muestra, la primera unidad muestral tiene probabilidad $\frac{A}{N}$ de ser seleccionada. Las subsecuentes unidades tienen probabilidades que dependen del método de selección: si el muestreo se hace ***con*** reemplazamiento, todas tienen la misma probabilidad $\frac{A}{N}$ de ser seleccionadas, pero si el muestreo se hace ***sin*** reemplazamiento, esta probabilidad va cambiando. En el primer caso, la variable X que cuenta el número de éxitos en la muestra, se ajusta a un modelo binomial de parámetros π, n. En

el segundo caso, a un modelo hipergeométrico de parámetros N, A, n. Es decir, la probabilidad de que haya x éxitos en la muestra está dada por:

$$\Pr(X = x) = \begin{cases} \binom{n}{x} \pi^x (1-\pi)^{n-x} & \text{CON reemplazamiento} \\ \dfrac{\binom{n}{x}\binom{N-A}{n-x}}{\binom{N}{n}} & \text{SIN reemplazamiento} \end{cases}$$

En ambos casos se tiene $\mathrm{E}(X) = n\pi$ pero las varianzas tienen expresiones diferentes, dadas por:

$$\mathrm{V}(X) = \begin{cases} n\pi(1-\pi) & \text{CON reemplazamiento} \\ n\pi(1-\pi)\dfrac{N-n}{N-1} & \text{SIN reemplazamiento} \end{cases}$$

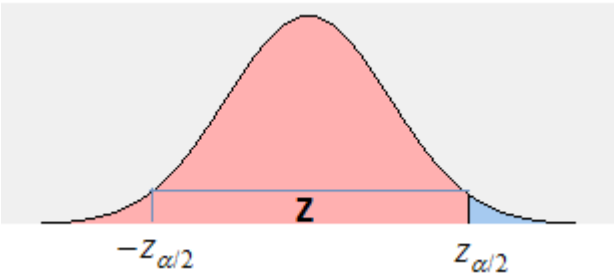
Nótese que la diferencia entre ambas varianzas está determinada por el factor $\frac{N-n}{N-1}$ que tiende a 1 cuando N tiende a infinito. Es decir, que solamente en poblaciones **infinitas o muy grandes**, podría decirse que $\mathrm{V}(X) = n\pi(1-\pi)$ a menos que la muestra se tome CON reemplazamiento. En la práctica se utiliza más el muestreo SIN reposición pero se usan las fórmulas del muestreo CON reposición, lo que ya trae consecuencias indeseables especialmente si la población es finita y pequeña o si π .es cercano a 0 o a 1.

Por lo anterior, ***si tomamos una muestra grande,*** podemos asumir que $X \sim \mathrm{B}(n,\pi)$ y -siguiendo a A. Wald- usar una aproximación normal para calcular $\Pr(X = x)$. Esto es:

$$Z = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{V}(X)}} = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} = \frac{\frac{X}{n} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim \mathrm{N}(0,1)$$

Al hacer $p = \frac{X}{n}$ se cumple $\mathrm{E}(p) = \frac{1}{n}\mathrm{E}(X) = \frac{1}{n}(n\pi) = \pi$, razón por la cual podemos utilizar $p = \frac{X}{n}$ como estimador insesgado de π

Se cumple entonces que $Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim \mathrm{N}(0,1)$ lo que nos lleva a considerar el intervalo $\left(-z_{\alpha/2}, z_{\alpha/2}\right)$ (ver figura) que cubre una probabilidad $1-\alpha$ para Z bajo la normal estándar.



Este intervalo, mediante transformaciones algebráicas simples, puede ser reescrito como:

$$\left(p - z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \;,\; p + z_{\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \right)$$

Que es la fórmula conocida y más popular para construir el intervalo de confianza para una proporción.

La fórmula anterior, además de la fuerte exigencia de una muestra y una población grandes, tiene el inconveniente de que depende de π, el parámetro desconocido que se pretende estimar. En la práctica π se reemplaza por la estimación puntual

p lograda con la muestra, lo que no está del todo mal, ya que como lo muestra Cochran, este cambio debería llevar a una expresión del tipo:

$$\left(p - t_{n-1,\,\alpha/2}\sqrt{\frac{p(1-p)}{n-1}} \;,\; p + t_{n-1,\,\alpha/2}\sqrt{\frac{p(1-p)}{n-1}} \right)$$

Pero, al ser n grande, el cuantil $t_{n-1,\,\alpha/2}$ puede considerarse bastante bien aproximado por el cuantil normal $z_{\alpha/2}$ y dividir entre $n-1$ es casi igual que dividir entre n.

El uso de la aproximación normal exige entonces, además de una población muy grande, un tamaño mínimo de muestra que debe ser conocido previamente para garantizar la validez de dicha aproximación. Infortunadamente dicho tamaño de muestra depende del valor de π que es obviamente desconocido. Cochran establece que para $\pi = 0.5$ es necesario cuando menos $n = 30$. Para $\pi = 0.2$ se necesita $n = 200$ y para $\pi = 0.05$ es necesario $n = 1400$. Sin embargo Newcombe pone en duda la validez de estos valores señalando que con frecuencia ellos son insuficientes.

En la práctica, como es sabido, los tamaños de muestra mínimos para poder usar aproximación normal con un nivel de confianza y un error máximo de estimación dados, se suelen calcular con las fórmulas:

$$n_0 = \frac{z^2_{\alpha/2}\,p(1-p)}{e^2} \qquad \text{y} \qquad n = \frac{n_0}{1 + \frac{n_0}{N}}$$

cuya aplicación no se discute en este documento, señalando únicamente que con frecuencia, son mal utilizadas.

Baste con decir, por ejemplo, que si el valor de π estuviese alrededor de 0.2, la estimación con un 95% de confianza y un error no mayor a 0.01 (que representa el 5% del valor de π), exigiría una muestra no menor a 6147 unidades en una población infinita o muy grande. Si el tamaño de la población fuese, por ejemplo, $N = 6000$, la muestra se reduciría a 3037 elementos como mínimo.

Existe un consenso generalizado de que el método de aproximación normal visto anteriormente tiene muy pobre desempeño, llegando incluso a proporcionar

estimaciones erróneas, sobre todo cuando la muestra es insuficiente, razón por la cual nos gustaría contar con otros métodos que en lo posible no dependieran del tamaño muestral. Varios de tales métodos han sido propuestos desde 1934 cuando Clopper y Pearson en un famoso artículo hicieron una propuesta basada en la distribución Beta. La idea de tales métodos es usar la verdadera y exacta probabilidad binomial y no la probabilidad aproximada con la normal. Por esta razón son conocidos como ***métodos exactos***. Son exactos no porque proporciones intervalos exactos –que también son aproximados- sino porque usan la probabilidad exacta. Estos métodos en principio no dependen del tamaño de n para la construcción del Intervalo de confianza (IC) que siempre estará bien construido aunque, como parece natural, valores muy pequeños de n producirán estimaciones poco precisas, es decir, intervalos muy amplios, con las consecuencias desagradables que se derivan de este hecho.

Supóngase pues que se tiene una muestra de tamaño n (el valor de n no ha sido calculado por un método especial y, en principio, podría ser cualquiera definido razonablemente, por ejemplo, por los costos del muestreo). Se busca un intervalo $(\pi_L\ ,\ \pi_U)$ dentro del cual se encuentre π con probabilidad $1-\alpha$ donde α es un valor pequeño arbitrario (usualmente $\alpha = 0.05$). Es claro que X toma valores enteros entre 0 y n, lo que nos dice que para x éxitos en la muestra, p tomará valores en el conjunto $\left\{0 = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \cdots, \frac{n}{n} = 1\right\}$ por lo que, al ser una variable discreta no será posible construir un intervalo con valores exactos. Sin embargo sí es posible construir un intervalo con un cubrimiento de probabilidad de $1-\alpha$ si se resuelve simultáneamente las dos ecuaciones siguientes:

$$\pi_L = Sup\left\{\pi,\ \sum_{j=x}^{n} \pi^x(1-\pi)^{n-x} \le \frac{\alpha}{2}\right\}$$

$$\pi_U = Inf\left\{\pi,\ \sum_{j=0}^{x} \pi^x(1-\pi)^{n-x} \le \frac{\alpha}{2}\right\}$$

Tanto π_L como π_U son funciones de $N,\ n,\ x$ lo que hace particularmente engorroso el cálculo de las expresiones anteriores.

Clopper y Pearson (1934) dieron una primera solución utilizando la distribución Beta y la igualdad $\sum_{j=x}^{n}\binom{n}{j}\pi^{j}(1-\pi)^{n-j} = \int_{0}^{\pi} f_B(t)dt$ con $f_B(t) =$ Probabilidad acumulada bajo dicha distribución.

Más exactamente se trata de lo siguiente:

Se busca un intervalo (π_L, π_U) para el cual se cumpla:

$\pi_L = \text{Sup}\left\{\pi / \Pr(X \le x) > \frac{\alpha}{2}\right\}$ y $\pi_U = \text{Inf}\left\{\pi / \Pr(X \ge x) > \frac{\alpha}{2}\right\}$ bajo el supuesto de que $X \sim B(n,\pi)$

Es claro que $P(X \le x) = \sum_{I=0}^{x}\binom{n}{i}\pi^{i}(1-\pi)^{n-i}$ y $P(X \ge x) = 1 - P(X \le x-1)$.

Usando las identidades $i\binom{n}{i} = n\binom{n-1}{i}$ y $(n-i)\binom{n}{i} = n\binom{n-1}{i}$ podemos calcular

$$\begin{aligned}\frac{\partial}{\partial\pi}P(X \ge x) &= \sum_{i=x}^{n}\binom{n}{i}i\pi^{i-1}(1-\pi)^{n-i} - \sum_{i=x}^{n-1}\binom{n}{i}(n-i)\pi^{i}(1-\pi)^{n-i-1}\\ &= n\sum_{i=x}^{n}\binom{n-1}{i-1}\pi^{i-1}(1-\pi)^{n-i} - \sum_{i=x}^{n-1}\binom{n-1}{i}\pi^{i}(1-\pi)^{n-i-1}\\ &= x\binom{n}{x}\pi^{x-1}(1-\pi)^{n-x} > 0 \qquad (1)\end{aligned}$$

Recordando la distribución Beta dada por:

$$X \sim \text{B}(a,b) \quad \text{ssi} \quad f^X(x;\, a,\, b) = \frac{x^{a-1}(1-x)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}du} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$$

Concluímos que la probabilidad acumulada bajo esta distribución, hasta un punto $\pi \in (0,1)$, está dada por $\text{B}_{\pi}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^{\pi} t^{a-1}(1-t)^{b-1}dt$.

Nótese que, según el teorema fundamental del cálculo, la derivada de

$$B_{\pi}(x,\ n-x+1)=\frac{\Gamma(n+1)}{\Gamma(x)\Gamma(n-x+1)}\int_0^{\pi} t^{x-1}(1-t)^{n-x}dt = x\binom{n}{x}\pi^{x-1}(1-\pi)^{n-x}$$

es $x\binom{n}{x}\pi^{x-1}(1-\pi)^{n-x}$ que no es otra cosa que la expresión (1).

En consecuencia,

$$P(X \geq x)=\sum_{i=x}^{n}\pi^{i}(1-\pi)^{n-i}=B_{\pi}(x,\ n-x+1) \qquad (2)$$

Expresión que nos proporciona una relación entre la sumatoria en la cola derecha de la binomial y la probabilidad acumulada bajo una Beta de parámetros

$$a=x,\quad b=n-x+1$$

Consideraciones similares, haciendo los cambios pertinentes, permiten establecer la siguiente identidad para el lado izquierdo de la sumatoria en la binomial:

$$P(X \leq x)=1-P(X \geq x+1)=1-B_{\pi}(x+1,\ n-x) \qquad (3)$$

Las expresiones (2) y (3) nos permiten afirmar que el intervalo de confianza para π está dado por:

$$(\pi_L\ ,\ \pi_U) = B_{\alpha/2}(x,\ n-x+1)\ ,\ B_{1-\alpha/2}(x+1,\ n-x)$$

Expresión que más frecuentemente se presenta como

$$(\pi_L\ ,\ \pi_U) = \left(B(x,\ n-x+1,\ \frac{\alpha}{2})\ ,\ B(x+1,\ n-x,\ 1-\frac{\alpha}{2}) \right) \qquad (CP)$$

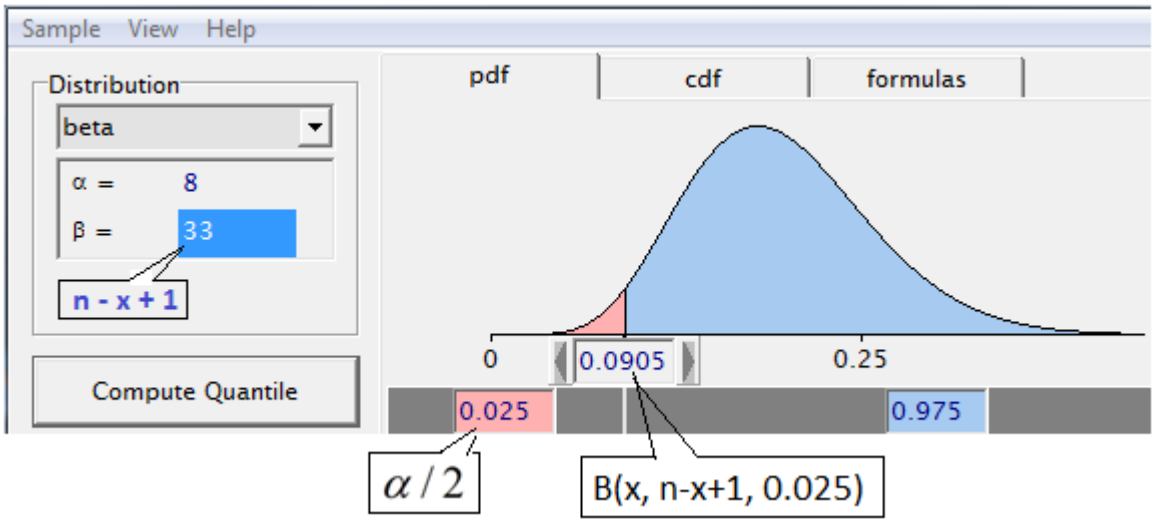
y que corresponde a la fórmula propuesta por Clopper y Pearson para construir el intervalo de confianza.

Históricamente esta fórmula ha sido de gran importancia porque marcó el inicio de una carrera investigativa alrededor del tema de la estimación de proporciones y además porque ofreció una solución al problema en dos casos extremos: cuando $x=0$ y cuando $x=1$, para los que se tienen los correspondientes intervalos: $\left(0,\ 1-\left(\frac{\alpha}{2}\right)^{\frac{1}{n}}\right)$ y $\left(\left(\frac{\alpha}{2}\right)^{\frac{1}{n}},\ 1\right)$ situación insoluble bajo el método asintótico de Wald

El intervalo CP nunca será menor que la cobertura nominal pudiendo, por ende, resultar más grande que lo deseado. A pesar de ser muy interesante y promocionado por sus inventores como insuperable, puede haber otros métodos aproximados que dan mejor cobertura. Así lo anuncia Agresti en un conocido artículo, cuyo título es muy sugestivo (ver referencia 13).

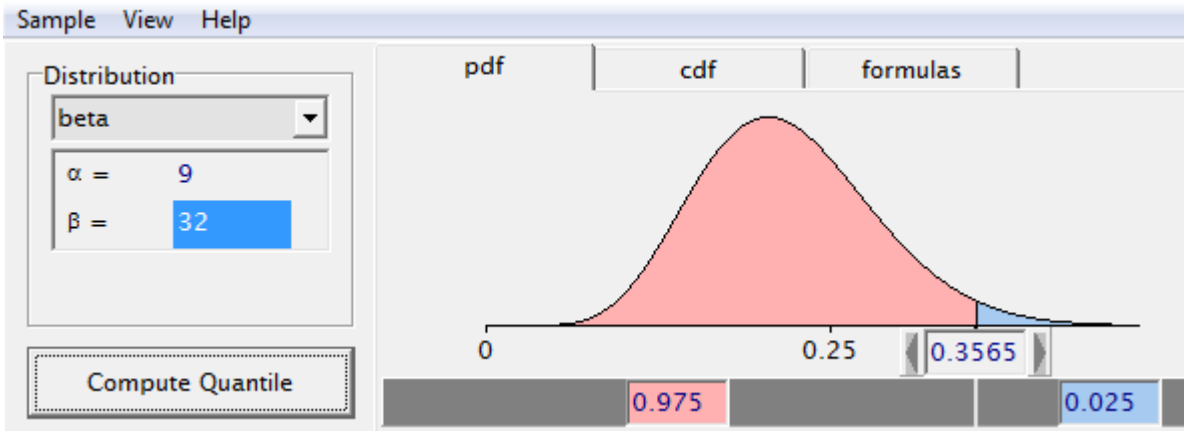
A manera de ejemplo, mostraremos cómo usar un paquete muy versátil y gratuito que puede ser descargado de Internet para construir intervalos de confianza usando la fórmula de Clopper y Pearson. Se trata de PQRS (http://members.home.nl/sytse.knypstra/PQRS/)

Vamos a construir el intervalo de confianza del 95% para la proporción sabiendo que en una muestra de tamaño 40 hubo en total 8 éxitos. En este ejemplo $\alpha/2 = 0.025$ y $1-\alpha/2 = 0.975$, por tanto:



De la figura anterior se deduce $B(x,\ n-x+1,\ \alpha/2) = B(8,\ 33,\ 0.025) = 0.0905$

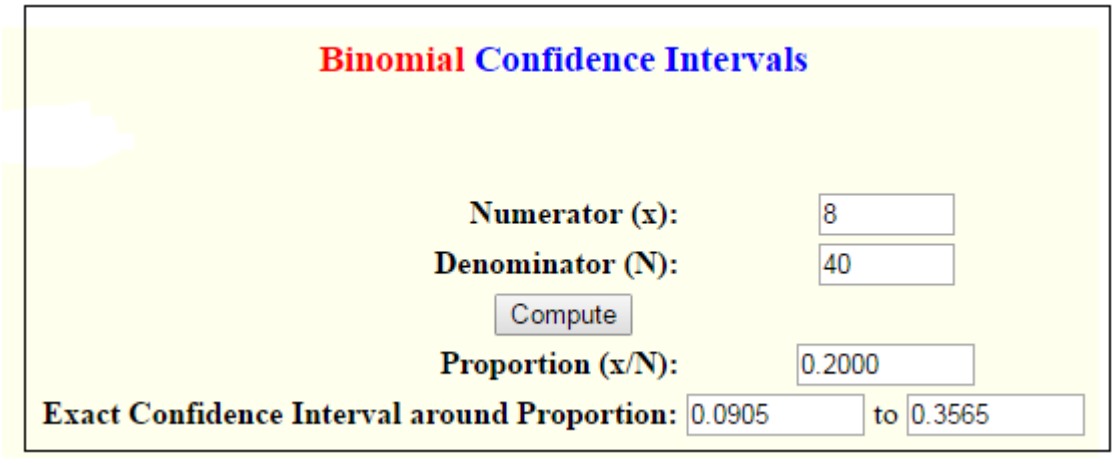
Análogamente:



De donde: $B(x+1,\ n-x,\ 1-\alpha/2) = B(9,\ 32,\ 0.975) = 0.3565$

En conclusión, el intervalo de confianza correspondiente es: **(0.0905 , 0.3565)**

La página http://statpages.info/confint.html contiene una calculadora *on line* que igualmente permite el cálculo del intervalo, como se ve en la siguiente figura:



Durante algún tiempo el método CP, propuesto por Clopper y Pearson fue considerado como la *regla de oro* para estimar proporciones, sin embargo con el paso de los años tal prestigio fue decayendo debido principalmente a que el método CP es muy conservativo en el sentido de que $1-\alpha$ no es el *inf* para la probabilidad de cobertura. Esto es, los intervalos obtenidos resultan en general más grandes que el verdadero.

En 1960 Blyth y Hutchinson publicaron un método que mejoraba la construcción de los intervalos a partir de un artículo de profundización escrito por J. Neyman en

1935. Este mismo autor junto con H.A. Still publica en 1983 otro artículo en el que utiliza la distribución F para hacer una revisión de la construcción dada por Clopper y Pearson.La razón? Tal vez era más fácil lidiar con la F que con la Beta

En 1986, usando una relación existente entre las distribuciones Beta y F transformó la fórmula de Clopper y Pearson en otra de más fácil cálculo puesto que solo depende de F, la cual está dada por:

$$\left(\frac{1}{1+\frac{n-x+1}{x}F_{2(n-x+1),\,2x,\,\alpha/2}} \;,\; \frac{\frac{x+1}{n-x}F_{2(x+1),\,2(n-x),\,\alpha/2}}{1+\frac{x+1}{n-x}F_{2(x+1),\,2(n-x),\,\alpha/2}} \right)$$

Esta fórmula ha tenido amplia difusión y ha sido usada bajo una presentación equivalente un poco más compacta, como se muestra a continuación:

$$\left(\frac{x}{x+(n-x+1)F_i} \;,\; \frac{(x+1)F_s}{n-x+(x+1)F_s} \right)$$

donde:

$$F_i = F_{2(n-x+1),\,2x,\,\alpha/2} \qquad F_s = F_{2(x+1),\,2(n-x),\,\alpha/2}$$

Han sido numerosos los investigadores que han trabajado sobre el tema, tantos que los portugueses Pires y Amado han realizado un trabajo de comparación mediante simulación de nada menos que 20 propuestas para determinar su desempeño.

En estas notas mencionaremos solamente algunos autores, los que son más reconocidos, sin profundizar en sus propuestas. Solamente nos interesa darlos a

conocer y mencionar sus fórmulas para que el lector interesado profundice en la teoría buscando el material correspondiente.

Cabe mencionar de manera especial a Goodman, Fitzpatrick , Scott, Sison y Glaz. Pero también son ampliamente reconocidos:

1. Wilson, quien propuso la fórmula siguiente:

$$\frac{p+\frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\, n^{-\frac{1}{2}}\sqrt{p(1-p)+\frac{1}{4n}z^2_{\alpha/2}}}{1+\frac{z^2_{\alpha/2}}{n}}$$

2. Agresti y Coull (**AC**) quienes proponen agregar cuatro observaciones (dos éxitos y dos fracasos) y tomar $\tilde{x} = x + \frac{1}{2}z^2_{\alpha/2}$, $\tilde{n} = n + z^2_{\alpha/2}$ y $\tilde{p} = \frac{\tilde{x}}{\tilde{n}}$ con lo cual el IC quedará de la forma: $\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$

3. Una propuesta muy interesante, hecha por Jeffrey, puede ser vista como una estimación de carácter bayesiano con distribución a priori $B\left(\frac{1}{2},\frac{1}{2}\right)$ que es no informativa. Según esta propuesta el IC tiene la forma

   $\left(L_j(x)\ ,\ U_j(x)\right)$ donde

$$L_j(x) = \begin{cases} 0 & \text{si } x = 0 \\ B^{-1}\left(x+\frac{1}{2},\ n-x+\frac{1}{2},\frac{\alpha}{2}\right) & \text{en otro caso} \end{cases}$$

$$U_j(x) = \begin{cases} 1 & \text{si } x = n \\ B^{-1}\left(x+\frac{1}{2},\ n-x+\frac{1}{2},\frac{\alpha}{2}\right) & \text{en otro caso} \end{cases}$$

4. Es conocido también el método no paramétrico que surge a través de la propuesta bootstrap de Efron (1979) pero que requiere el uso de computador y la programación de un algoritmo que tome muchas submuestras de la muestra

dada, construya la distribución empírica de las proporciones obtenidas con esas muestras y finalmente calcule los percentiles $P_{\alpha/2}$ y $P_{1-\alpha/2}$ que corresponden a los límites del intervalo de confianza.

5. Haremos una breve referencia a un método de reciente aparición propuesto por Zhou, Li y Yang (2008), denominado ***método ZL***, según el cual el CI puede construirse de acuerdo con la siguiente expresión:

$$\left(\frac{\left(\left(\frac{p}{(1-p)} \right)^{-1} - g^{-1}\left(z_{\alpha/2}\right) \right)}{1+\exp\left(\log\left(\frac{p}{1-p} \right) - np(1-p)^{-1/2} g^{-1}\left(z_{\alpha/2}\right) \right)} , \frac{\left(\left(\frac{p}{(1-p)} \right)^{-1} - g^{-1}\left(z_{1-\alpha/2}\right) \right)}{1+\exp\left(\log\left(\frac{p}{1-p} \right) - np(1-p)^{-1/2} g^{-1}\left(z_{1-\alpha/2}\right) \right)} \right)$$

Siendo $$g^{-1}(T) = \sqrt{n}\left(\frac{6}{\gamma} \right)^{-1} \left[1 - \frac{\gamma}{2}\left(n^{-\frac{1}{2}} T - \frac{1}{6} n^{-1} \gamma \right)^{3} - 1 \right] \quad \text{con} \quad \gamma = \frac{1-2p}{\sqrt{p(1-p)}}.$$

Para un nivel de confianza del 95% se tiene $z_{1-\alpha/2} = 1.96$ y $z_{\alpha/2} = -1.96$

Si $x = 0$ o $x = n$ se toma $x + 0.5$ en vez de x y $n+1$ en vez de n.

Este método se encuentra implementado en un programa Matlab que se presenta al final del documento (ver Apéndice, Programa No 1).

Los autores del método ZL hacen las siguientes recomendaciones:

a. Proscribir el método de Wald.
b. Usar el método de Wilson cuando no se conozca el posible valor de π.
c. Si se tiene alguna idea del posible valor de π y éste es cercano a 0.5 usar el método AC de Agresti Coull, pero si el valor de π es cercano a los extremos 0 o 1 usar el método ZL.

6. Finalmente: de muy reciente aparición (diciembre de 2014) en el *Journal of Statistical Theory and Applications* (Vol 13, No 4) un artículo de D. Habtzghi, C.K. Midha y A. Das, propone un método radicalmente diferente para construir los intervalos de confianza. Este método calcula los valores esperados de los intervalos mediante la búsqueda de sus límites que son modelados a través de

la variación de los niveles $1-\alpha$ y la aplicación de dos modelos logísticos especiales.

En este artículo los autores comparan los métodos de Wald, Clopper y Pearson, Wilson, Agresti-Coull y Jeffrey con el que ellos proponen, denominado *Mnew*. Los resultados muestran que el método de Wald fue el de peor cobertura, el Mnew fue el mejor y los otros cuatro fluctúan entre los dos anteriores. El método ZL no fue incluído en las comparaciones.

La tabla siguiente proporciona los IC del 95% de confianza para muestras de tamaños comprendidos entre 5 y 16, según el número X de éxitos presentes en la muestra. Una tabla más completa (hasta $n = 40$) se encuentra en el artículo original.

Binomial Proportions. The 95% confidence limits based on "*Mnew*" approach for $5 \leqslant n \leqslant 16$

	L_{new}	U_{new}	L_{new}	U_{new}	L_{new}	U_{new}	L_{new}	U_{new}	L_{new}	U_{new}	L_{new}	U_{new}
x	5		6		7		8		9		10	
0	0.000	0.360	0.000	0.314	0.000	0.279	0.000	0.250	0.000	0.228	0.000	0.208
1	0.020	0.597	0.016	0.528	0.013	0.474	0.011	0.428	0.010	0.391	0.009	0.359
2	0.100	0.774	0.078	0.696	0.065	0.631	0.055	0.575	0.048	0.528	0.043	0.487
3			0.174	0.826	0.141	0.758	0.120	0.697	0.104	0.644	0.092	0.598
4							0.202	0.798	0.173	0.743	0.152	0.693
5											0.224	0.776
x	11		12		13		14		15		16	
0	0.000	0.193	0.000	0.179	0.000	0.166	0.000	0.155	0.000	0.145	0.000	0.136
1	0.008	0.332	0.007	0.309	0.006	0.289	0.006	0.271	0.006	0.255	0.005	0.241
2	0.039	0.452	0.035	0.422	0.032	0.395	0.030	0.372	0.028	0.351	0.026	0.332
3	0.082	0.557	0.075	0.521	0.068	0.489	0.063	0.461	0.058	0.436	0.054	0.413
4	0.136	0.648	0.123	0.609	0.112	0.573	0.103	0.541	0.096	0.512	0.089	0.486
5	0.199	0.729	0.179	0.687	0.163	0.648	0.150	0.614	0.138	0.582	0.129	0.554
6			0.243	0.757	0.220	0.717	0.201	0.681	0.186	0.647	0.172	0.616
7							0.258	0.742	0.237	0.707	0.220	0.674
8											0.271	0.729

A manera de ejemplo presentaremos el IC calculado según algunos de los métodos mencionados a lo largo de este documento.

INTERVALOS DE CONFIANZA CALCULADOS POR DIFERENTES METODOS: n = 20 x = 6

METODO EMPLEADO	CARACTERISTICAS DE CALCULO	IC-Linf	IC-Lsup
Wald	p=0.3 e = 0.2008 IC = (p-e , p+e)	0.0992	0.5008
Clopper Pearson	B(6,15, 0.025) , B(7,14, 0.975) con PQRS	0.1189	0.5428
Blyth Hutchinson	$F_i = F_{30,12,0.05} = 2.466$ $F_s = F_{14,28,0.05} = 2.064$	0.1396	0.5079
Jeffrey	B(6.5, 14.5, 0.025), B(6.5, 14.5, 0.975)	0.1361	0.5172
AC Agresti Coull	x0 = 7.928 n0 = 21.96 y fórmula de Wald	0.1599	0.5615
ZL de Zhou, Li y Yang	Con el programa Matlab del apéndice A	0.1354	0.5170
Mnew de Habtzgui et al.	Tomado de la tabla parcial anexa al documento	0.134	0.516

NOTA: Los intervalos Clopper-Pearson y Jeffrey se calcularon usando PQRS un paquete gratuito para distribuciones en Internet

Como puede apreciarse en esta tabla el IC más desfasado es el correspondiente al método tradicional de Wald

Existe también un paquete implementado en R que permite, entre otras varias cosas, estimar proporciones mediante 8 métodos, entre ellos, cuatro de los que hemos mencionado en este documento. Es el paquete BINOM.

Los métodos y la sintaxis para el uso de BINOM son los siguientes:

Clopper – Pearson ("exact")
Asintótico (Wald) ("asymptotic")
Agresti – Coull ("ac")
Wilson ("wilson")
Todos ("all") (Está por defecto)

Sintaxis:

```
binom.confint(x, n, conf.level = 0.95, methods = "nombre")
```

El nombre que se encuentra entre comillas se utiliza para invocar cada procedimiento como se muestra en el ejemplo siguiente:

Ejemplo:

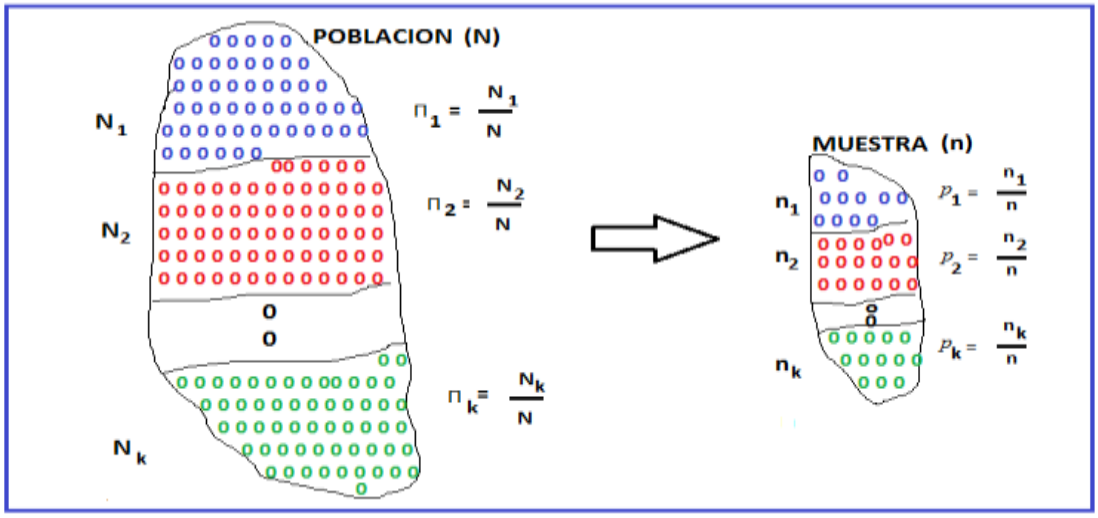
```
library("binom")
binom.confint(6, 20, conf.leve = 0.95, methods = "exact")
binom.confint(6, 20, conf.leve = 0.95, methods = "asymptotic")
binom.confint(6, 20, conf.leve = 0.95, methods = "ac")
binom.confint(6, 20, conf.leve = 0.95, methods = "wilson")
binom.confint(6, 20, conf.leve = 0.95, methods = "all"
```

method	x	n	mean	lower	upper
exact	6	20	0.3	0.1189316	0.5427892
asymptotic	6	20	0.3	0.09916346	0.5008365
agresti-coull	6	20	0.3	0.1431593	0.5212908
wilson	6	20	0.3	0.1454772	0.5189728

La opción “all” se encuentra por defecto.

CASO DE PROPORCIONES MULTINOMIALES

Terminaremos estas notas con una muy breve referencia al caso de proporciones multinomiales, es decir, a aquel tipo de proporciones que hace referencia a poblaciones cuyos elementos están clasificados en k categorías disjuntas.



Este caso presenta aún mayores dificultades teóricas y hay muchas menos referencias bibliográficas. Su estudio merece un capítulo aparte por lo que en estas notas solamente se hará un resumen muy sucinto de algunos trabajos y una rápida referencia a uno de los métodos de estimación más versátiles que fue propuesto por Quesenberry y Hurst (1964), el cual es aplicable siempre que se tengan muestras de gran tamaño.

Supóngase que se tiene una población de N elementos partida en k categorías $A_1, A_2, \cdots A_k$ con $N_1, N_2, \cdots, N_k$ elementos respectivamente. Se extrae una muestra de *n* elementos y, en general, se desea saber cuál es la probabilidad de que haya xi elementos de la categoría A_i, para $i = 1, 2, \cdots, k$. Obviamente se ha de cumplir $\sum_{i=1}^{k} N_i = N$ y $\sum_{i=1}^{k} n_i = n$.

Igual que en el caso binomial, puede suceder que la probabilidad $\pi_i = \frac{N_i}{N}$ de que el elemento seleccionado pertenezca a la categoría A_i no cambie en las sucesivas extracciones, lo que ocurre si dichas extracciones son independientes. Esto sucede únicamente cuando se hace muestreo CON reemplazamiento ya que obviamente las categorías no son infinitas. En tal caso se tiene un modelo ***multinomial*** de k categorías.

Por el contrario, si la probabilidad de que el elemento extraído pertenezca a A_i cambia con cada extracción, lo que ocurre, como es usual, si se hace muestreo SIN reemplazamiento, se tiene un modelo ***hipergeométrico k-variado***.

Si X_i es la variable aleatoria que cuenta el número de elementos de la categoría A_i que aparecen en la muestra, el cuadro siguiente resume las principales propiedades de los modelos multinomial e hipergeométrico k-variado:

MODELO MULTINOMIAL DE k CATEGORIAS	**MODELO HIPERGEOMETRICO k VARIADO:**
$f(x_1, x_1, \cdots, x_1) = \Pr(X_1 = x_1, \cdots, X_k = x_k) =$ $\frac{n!}{x_1!x_2!\ldots x_k!} p^{x_1} p^{x_2} \cdots p^{x_3}$	$f(x_1, x_1, \cdots, x_1) = \Pr(X_1 = x_1, \cdots, X_k = x_k) =$ $\binom{N}{n}^{-1} \prod_{i=1}^{k} \binom{N_i}{x_i}$
Medias: $\mathrm{E}(X_i) = np_i \quad i = 1,2,\cdots,k$	Medias: $\mathrm{E}(X_i) = n\left(\frac{N_i}{N}\right) \quad i = 1,2,\ldots,k$
Varianzas: $\mathrm{V}(X_i) = np_i(1-p_i) \quad i = 1,2,\cdots,k$	Varianzas: $\mathrm{V}(X_i) = n\left(\frac{N_i}{N}\right)\left(1-\frac{N_i}{N}\right)\left(\frac{N-n}{N-1}\right) \quad i = 1,2,\ldots,k$
Covarianzas: $\mathrm{Cov}(X_i, X_j) = -np_ip_j \quad (i \neq j)$	Covarianzas: $\mathrm{Cov}(X_i, X_j) = n\left(\frac{N_iN_j}{N^2}\right)\left(\frac{N-n}{N-1}\right) \quad (i \neq j)$

Sucede algo similar a lo que sucedía en el caso (dicotómico) de dos categorías: Si la población de donde se extrae la muestra es grande, comparada con la muestra, no hay diferencias apreciables entre las varianzas en los dos modelos lo que permite afirmar que las estimaciones puntuales correspondientes a las proporciones de cada categoría están dadas por

$$\hat{\pi_i} = p_i = \frac{x_i}{n} \qquad i = 1,2,\cdots,k$$

Sin embargo la construcción de los intervalos de confianza para cada proporción es un problema aún más complicado que en caso dicotómico. Frecuentemente se construyen estos intervalos aplicando erróneamente el caso de dos categorías: una categoría formada por los elementos de A_i y la otra categoría formada por los demás elementos de la población. Esta técnica, implementada en varios paquetes estadísticos es incorrecta debido a que construye intervalos que no son independientes entre sí ya que hay elementos comunes en todo par de intervalos.

Quesenberry y Hurst (1964), Goodman (1965) y Fitzpatrick (1987) proporcionaron fórmulas que permiten construir los IC para las proporciones.

W.May y W. Johnson proporcionaron un macro en SAS para la construcción de tales intervalos. Cfr: A SAS® macro for constructing simultaneous confidence intervals for multinomial proportions by Warren L. May and William D. Johnson.

(DOI: http://dx.doi.org/10.1016/S0169-2607(97)01809-9)

La fórmula propuesta por Quesenberry y Hurst es relativamente fácil de implementar y está dada por:

$$\frac{(w+2x_i) \mp \sqrt{w\left(w+4x_i\left(1-\frac{x_i}{n}\right)\right)}}{2(n+w)}$$

Donde $w = \chi^2_{k-1,\alpha}$.

En el apéndice se encuentra un programa en Matlab para su aplicación, realizando los cambios que sean necesarios de acuerdo con la información en cada caso.

Goodman, basado en argumentos similares a los de Bonferroni para intervalos simultáneos, propuso usar $w = \chi^2_{1,\alpha/k}$.

Existe igualmente un paquete en R, denominado **CoinMinD**, con el mismo propósito. Su uso está dado por:

QH(Frec, alfa)

Donde *Frec* es el vector de frecuencias y alfa el valor de α que define el nivel de confianza de los intervalos

EJEMPLO:

Supóngase que en un país se presentan cuatro candidatos a la presidencia de la república . Se hace una encuesta que se aplica a 2954 votantes. Se obtuvieron 546 votos a favor del candidato A, 658 a favor de B, 935 a favor de C y 815 a favor de D. Para realizar las estimaciones correspondientes a un 95% de confianza, adecuamos el programa No 2 del apéndice con la siguiente información:

```
% w es el valor Ji2 con k-1 GL al nivel 1-alfa
% Toma de información:
w = 7.81;
n = [546 658 935 815];
k = 4;
```

se obtienen los siguientes resultados:

L =

0.1848	0.1657	0.2056
0.2227	0.2021	0.2449
0.3165	0.2931	0.3409
0.2759	0.2535	0.2994

Que se interpretan así:

	Proporción	Intervalo del 95%
Candidato A	0.1848	(0.1657, 0.2056)
Candidato B	0.2227	(0.2021, 0.2449)
Candidato C	0.3165	(0.2931, 0.3409)
Candidato D	0.2759	(0.2535, 0.2994)

Como se ve, realmente hay un "empate técnico" entre los candidatos C y D, ya que sus IC se alcanzan a traslapar, lo que indica que las proporciones correspondientes a ellos dos no difieren significativamente.

Aquí debemos llamar la atención sobre el mal uso que se hace de los llamados "empates técnicos" presentados en medios de comunicación no científicos: en primer lugar tales empates son determinados según el error máximo de estimación estipulado para la muestra la que generalmente se diseña con las fórmulas de aproximación normal del caso binomial, lo cual no es correcto. En segundo lugar aplican el mismo nivel de error para comparar las estimaciones de las proporciones en cada par de categorías como si este error fuese el mismo para todos los pares lo que tampoco es cierto.

El mismo ejemplo ya propuesto se resuelve en R de la siguiente manera:

Lo primero que debe hacerse es instalar el paquete CoinMinD en R, para lo cual puede usarse el comando

Install.packages("CoinMinD")

O simplemente usar la opción de instalación de paquetes que presenta el programa.

Una vez que se tenga instalado el programa, cada vez que se quiera usar la función QH de estimación, se debe cargar la librería correspondiente, como se muestra en el código siguiente:

```
library(CoinMinD)
Frec  <- c(546, 658, 935, 815)
QH(Frec, 0.05)
```

La ejecución de este código produce:

```
Original Intervals
Lower Limit
[1] 0.1657099 0.2020948 0.2931078 0.2535218
Upper Limit
[1] 0.2056215 0.2448659 0.3409004 0.2994549
Adjusted Intervals
Lower Limit
[1] 0.1657099 0.2020948 0.2931078 0.2535218
Upper Limit
[1] 0.2056215 0.2448659 0.3409004 0.2994549
Volume
[1] 3.75e-06
```

Donde se ven los límites de los intervalos en forma original y corregidos más el volumen del hipercubo encerrado por ellos. Este volumen podría interpretarse como una estimación de la totalidad de individuos que son de interés para el estudio dentro de toda la población

APENDICE:

PROGRAMA No 1 Método ZL

El siguiente programa en Matlab permite construir el IC del 95% de confianza según el método ZL de D. Habtzgui, C.K. Midha y A. Das:

```
% Intervalo de confianza para una proporción binomial
% Método ZL de Zhou, Li y Yang. Programó J.A.Clavijo
clear
clc
% Introduzca la siguiente información:
% x es número de éxitos en la muestra
% n es el tamaño de muestra
n = input('          Ingrese el tamaño de muestra: ');
x = input('Ingrese el número de éxitos en la muestra: ');
% inicio de calculos
if(x==0)
  x=0.5;
  n = n+1;
end
if(x==n)
  x=n+0.5;
  n = n+1;
end
p = x/n;
gam = (1-2*p)/sqrt(p*(1-p));
gsup = sqrt(n)*inv(-gam/6)*((1-(gam/2)*(-1.96/sqrt(n)-(1/6)*(gam/n)))^(1/3)-1);
ginf = sqrt(n)*inv(-gam/6)*((1-(gam/2)*(1.96/sqrt(n)-(1/6)*(gam/n)))^(1/3)-1);
a = exp(log(p/(1-p))-inv(sqrt(n*p*(1-p)))*ginf);
b = exp(log(p/(1-p))-inv(sqrt(n*p*(1-p)))*gsup);
Li = a/(1+a);
Ls = b/(1+b);
% intervalo:
disp('Intervalo del 95% de confianza para la proporción:')
[Li Ls]
```

PROGRAMA No 2 – Fórmula de Qusenberry y Hurst

El siguiente programa en Matlab proporciona las estimaciones puntuales y los IC para cada proporción. Se deben introducir el valor de w y las frecuencias observadas en cada categoría.

```
% PROGRAMA MATLAB PARA CONSTRUIR IC MULTINOMIALES
% FORMULA DE QUESSENBERRY Y HURST

%  w  es el valor Ji2 con k-1 GL al nivel 1-alfa
% Toma de información:
w = 9.49;
n = [218 639 545 483 515];
k = 5;

% Inicio de calculos
N = sum(n);
d = 2*(N+w);
for i=1:k
    p(i) = n(i)/N;
    a(i) = (w+2*n(i))/d;
    b(i) = sqrt(w^2 + 4*n(i)*w*(1-n(i)/N))/d;
end

for i = 1:k
    li(i) = a(i) - b(i);
    ls(i) = a(i) + b(i);
end

L = [p' li' ls'];
L
```

REFERENCIAS:

1. Clopper and Pearson. The use of Confidence or Fiducial Limits ilustrated in the case of Binomials. Biometrika 26(1934), 404 – 413
2. J. Neyman. On the problem of Confidence Limits.. The Annals of Mathematical Statistics 6(1935)
3. Blyth C.R. 1986. Approximate Binomial Confidence Limits. Journal of the American Statistical Asociation, JASA. 81(395), 843 – 855
4. Hsiuging Wang. Exact Coeficients of Simultaneous CI for Multinomial Proportions. Journal of Multivariate Analysis. 99(2008), 896 – 911
5. J.T. Morisette and S. Khorram; Exact binomial CI for Proportions. Photogrammetric Engineering & Remot Sensing. Abril 1988
6. L.Brown, T. Cai and A. DasGupta; Interval estimation for a Binomial Proportion. Statistical Science, 2001. Vol 16, No 2. 101 – 133
7. X.H. Zhou, C.M. Li y Z. Yang; Improving Interval estimation of Binomial Proportions. Philosophical Transactions of the Royal Society. A(2008)366, 2405 – 2418
8. A.M. Pires et C. Amado; Interval estimators for a binomial Proportion: Comparison of Twenty Methods. Statistical Journal. Vol 6 No 2, June 2008. 165 – 197
9. C.R. Blyth and D.W. Hutchinson; Table of Neyman-Shortest unbiased Confidence Intervals for the Binomial Parameter. Biometrika(1960), 47 3 and 4, p. 381
10. F. Scholz. Confidence Bounds & Intervals for Parameters Relating to the Binomial, Negative Binomial, Poisson and Hypergeometric Distributions. Sl. 2008.
11. D. Habtzghi, C.K. Midha and A. Das; Modified Clopper-Pearson Confidence Interval for Binomial Proportion. Journal of Statistical Theory and Applications. Vol 13 No 4, December 2014, 296 – 310
12. E. Cepeda et al.; Intervalos de confianza e Intervalos de credibilidad para una Proporción. Revista Colombiana de Estadística. Diciembre 2008. Vol 31 No 2. 211-228
13. A. Agresti and B.A. Coull; Aproximate is better than “Exact” for Interval estimation of Binomial Proportions. American Statistician. Vol 52 No 2, may 1998. 119-126
14. C.P. Quesenberry & D.C. Hurst; Large Sample Simultaneous Confidence Intervals for Multinomial Proportions. 1964. Technometrics 6, 191-195