Revista Colombiana de Estadística
Volumen 26 N° 1. Págs. 61 *a* 75. Junio de 2003

Intervalos de confianza para la comparación de dos proporciones

Juan Carlos Correa[*]
Esperanza Sierra[**]

Resumen

La construcción de intervalos de confianza para la estimación de $\pi_1 - \pi_2$, la diferencia entre dos proporciones, es un problema importante en el trabajo estadístico aplicado. Revisamos diferentes procedimientos de construcción y mediante un estudio de simulación los analizamos. Proponemos un índice para comparar los distintos métodos, analizando tanto los niveles de confianza reales, como las longitudes de los intervalos.

***Palabras Clave**: Estimación, Distribución Binomial, Intervalo de confianza, Nivel de confianza real, Probabilidad de cobertura.*

Abstract

The construction of confidence intervals to estimate $\pi_1 - \pi_2$, the difference of two proportions, is an importante problem in applied work. We review different methods to construct them and we analyze their performance by a simulation study. Also we propose a index that allows us to compare these procedures, analyzing simultaneously the real confidence level and the length of them.

***Keywords**: Estimation, Binomial distribution, Confidence interval, Real confidence level, Coverage probability.*

[*]Profesor asociado, Escuela de Estadística, Universidad Nacional de Colombia - Medellín. E-mail: jccorrea@perseus.unalmed.edu.co
[**]Profesora asociada, Escuela de Estadística, Universidad Nacional de Colombia - Medellín. E-mail: esierra@perseus.unalmed.edu.co

1. Introducción

En la aplicación estadística, para el análisis de resultados, cada vez se prefiere más el uso de intervalos de confianza que las pruebas de hipótesis, debido a que el intervalo de confianza aporta información tanto de la magnitud, como de la precisión de las estimaciones, pudiéndose interpretar el intervalo en términos del margen de error de la estimación puntual. Esto hace los intervalos muy atractivos a la hora de presentar resultados, mientras que el valor-p en las pruebas de hipótesis es una elaboración probabílistica de interpretación más compleja.

La construcción de intervalos de confianza para la estimación de la diferencia de proporciones, $\pi_1 - \pi_2$, es un problema que se presenta frecuentemente en el trabajo estadístico aplicado; un caso típico, muy frecuente en ensayos clínicos, es la comparación de dos tratamientos.

Los intervalos de confianza que presentan los textos básicos de estadística, construidos con base en la aproximación normal a la binomial, tienen un desempeño pobre —pueden resultar intervalos que no tienen sentido o intervalos con probabilidad de cobertura por debajo del nivel de confianza nominal— especialmente cuando las muestras no son muy grandes. Hicimos una revisión de varios intervalos y los comparamos mediante un estudio de simulación. Cada uno de estos intervalos tiene ciertas ventajas y desventajas. Nos interesamos en analizar el comportamiento del nivel de confianza real o porcentaje de intervalos simulados que contienen el verdadero valor de $\pi_1 - \pi_2$ y en compararlo con el nivel de confianza nominal usado, 0.95. El nivel de confianza real es una estimación de la probabilidad de cobertura $P\left(L_{inf} \leq \pi_1 - \pi_2 \leq L_{sup}\right)$ que esperamos sea del 95%. Aquí L_{inf} y L_{sup} son las variables aleatorias que indican el límite inferior y superior, respectivamente, del correspondiente intervalo de confianza. También comparamos los promedios de las longitudes de los intervalos calculados con los distintos métodos. Buscamos métodos que dieran intervalos con niveles de confianza real iguales o mayores que el nominal, pero cuyas longitudes fueran pequeñas.

Chan y Zhang (1999) comparan intervalos de confianza exactos basados en cuatro estadísticos de prueba distintos, donde los límites inferior y superior se obtienen invirtiendo el procedimiento de prueba para hipótesis unilaterales sobre el valor de la diferencia. Peskun (1993) también compara cuatro intervalos, todos basados en la aproximación normal a la binomial. La diferencia ente ellos, es la forma como se hace la corrección por continuidad. Santner y Snell estudian intervalos de confianza tanto para la diferencia de proporciones como para la razón en caso de muestras pequeñas. Lloyd (1990) y Wild &

Seber (1993) estudian el caso de intervalos de confianza para dos proporciones correlacionadas. Agresti & Caffo(2000), basados en el artículo de Agresti & Coull(1998), muestran que el intervalo de Wald mejora notablemente al adicionar seudoobservaciones a las muestras: dos éxitos y dos fracasos a cada una. Pan ajusta el método de Agresti & Caffo usando la distribución t. Newcombe (1998) compara once métodos de construcción de intervalos para diferencia entre proporciones independientes. En todos estos trabajos se analiza separadamente el nivel de confianza y la longitud de los intervalos.

Notación: Sean $x_{i_1}, x_{i_2}, \cdots, x_{i_{n_i}}$, para $i = 1, 2$, muestras aletorias i, independientes, de tamaño n_i de distribuciones Bernoulli con parámetros π_i. El estimador de máxima verosimilitud para π_i está dado por

$$\hat{\pi}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i_j} = \frac{X_i}{n_i}$$

con X_i el número de éxitos en la muestra i.

2. Métodos a comparar

Analizaremos los siguientes métodos que se pueden explicar en los cursos básicos de estadística y que computacionalmente no presentan gran complejidad, así que se pueden implementar fácilmente.

2.1. Método de Wald

Este método, que presentan la mayoría de los textos básicos de Estadística, se basa en la distribución asintótica normal de la diferencia entre las proporciones muestrales, $\hat{\pi}_1 - \hat{\pi}_2$. Teóricamente este método da un intervalo con un nivel de confianza aproximado de $(1 - \alpha)100\%$, la aproximación es mejor en tanto n_1 y n_2 sean "grandes".

Los extremos del intervalo son:

$$\text{Extremo inferior} = \hat{\pi}_1 - \hat{\pi}_2 - z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

$$\text{Extremo superior } = \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}},$$

donde $z_{\alpha/2}$ es el percentil $100(1 - \alpha/2)$ de la normal estándar.

2.2. Método de Wald con corrección por continuidad

Como la aproximación normal a la binomial es de una distribución discreta por una continua es pertinente hacer la corrección por continuidad.

Los extremos de este intervalo son:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm \left(z_{\alpha/2} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} + (1/n_1 + 1/n_2)/2 \right)$$

2.3. Método de Agresti y Caffo

Sea $\hat{\pi} = X/n$ la proporción de éxitos en una muestra de tamaño n, con un nivel de significancia aproximado de α, la región de aceptación para la prueba de la hipótesis,

$$H_0 : \pi = \pi_0 \quad \text{contra} \quad H_1 : \pi \neq \pi_0$$

es el conjunto de valores π_0, tales que:

$$\frac{|\hat{\pi} - \pi_0|}{\sqrt{\pi_0(1-\pi_0)/n}} \leq z_{\alpha/2}$$

El intervalo que resulta al despejar en esta desigualdad π_0 se conoce como el intervalo de confianza de *"score"*, Agresti & Coull (1998) notaron que el punto medio de este intervalo es un promedio ponderado de $\pi = \hat{\pi}$ y $\pi = 1/2$ que es igual a una proporción muestral donde el número de éxitos es

$$X + \frac{z^2_{\alpha/2}}{2}$$

y el tamaño de la muestra es $n + z^2_{\alpha/2}$. Cuando $\alpha = 0.05$, $z^2_{\alpha/2} \approx 4$ por esto se plantea agregar cuatro seudoobservaciones: dos éxitos y dos fracasos y aplicar el intervalo de *"score"*. Con base en en estos elementos Agresti & Caffo (2000) construyen el siguiente intervalo, que llaman *intervalo ajustado*, para $\pi_1 - \pi_2$:

$$\left(\tilde{\pi}_1 - \tilde{\pi}_2 - z_{\alpha/2}\sqrt{V(\tilde{\pi}_1, \tilde{n}_1) + V(\tilde{\pi}_2, \tilde{n}_2)} \ , \ \tilde{\pi}_1 - \tilde{\pi}_2 + z_{\alpha/2}\sqrt{V(\tilde{\pi}_1, \tilde{n}_1) + V(\tilde{\pi}_2, \tilde{n}_2)} \right)$$

donde, $\tilde{\pi}_1 = \frac{X_1+1}{\tilde{n}_1}$, $\tilde{\pi}_2 = \frac{X_2+1}{\tilde{n}_2}$, $\tilde{n}_j = n_j + 2$ y donde, para $j = 1, 2$,

$$V(\tilde{\pi}_j, \tilde{n}_j) = \frac{1}{\tilde{n}_j} \left[\tilde{\pi}_j(1-\tilde{\pi}_j)\frac{n_j}{\tilde{n}_j} + \frac{1}{2}\,\frac{1}{2}\,\frac{2}{\tilde{n}_j} \right]$$

2.4. Método de Newcombe

Es una mezcla del método de Wald y el intervalo *"score"*. Para cada una de las dos muestras se hallan los límites l_j y u_j del intervalo "*score*" resolviendo para π_j con $j = 1, 2$, la ecuación

$$|\hat{\pi}_j - \pi_j| = z_{\alpha/2}\sqrt{\pi_j(1-\pi_j)/n_j}$$

El intervalo propuesto por Newcombe (1998) es

$$\left(\hat{\pi}_1 - \hat{\pi}_2 - z_{\alpha/2}\sqrt{\frac{l_1(1-l_1)}{n_1} + \frac{u_2(1-u_2)}{n_2}}\ ,\ \hat{\pi}_1 - \hat{\pi}_2 + z_{\alpha/2}\sqrt{\frac{u_1(1-u_1)}{n_1} + \frac{l_2(1-l_2)}{n_2}}\right)$$

2.5. Método de Pan

Basado en el método de Agresti y Caffo, que emplea una estimación de la varianza de los $\tilde{\pi}_j$, Pan propone un intervalo aproximado que usa la distribución t en lugar de la normal.

La fórmula para este intervalo es:

$$\left(\tilde{\pi}_1 - \tilde{\pi}_2 - t_{d,\alpha/2}\sqrt{V(\tilde{\pi}_1,\tilde{n}_1) + V(\tilde{\pi}_2,\tilde{n}_2)}\ ,\ \tilde{\pi}_1 - \tilde{\pi}_2 + t_{d,\alpha/2}\sqrt{V(\tilde{\pi}_1,\tilde{n}_1) + V(\tilde{\pi}_2,\tilde{n}_2)}\right)$$

donde d son los grados de libertad

$$d \approx \frac{2\left[V(\tilde{\pi}_1,\tilde{n}_1) + V(\tilde{\pi}_2,\tilde{n}_2)\right]}{\Omega(\tilde{\pi}_1,\tilde{n}_1) + \Omega(\tilde{\pi}_2,\tilde{n}_2)}$$

y $\Omega(\tilde{\pi}_j,\tilde{n}_j)$ es la varianza del estimador de $Var[\tilde{\pi}_j,\tilde{n}_j]$

$$\begin{aligned}\Omega(\tilde{\pi}_j,\tilde{n}_j) = \frac{\tilde{\pi}_j - \tilde{\pi}_j^2}{n^3} + \Bigg[&\tilde{\pi}_j + (6\tilde{n}_j - 7)\tilde{\pi}_j^2 \\ &+ 4(\tilde{n}_j - 1)(\tilde{n}_j - 3)\tilde{\pi}_j^2 - 2(\tilde{n}_j - 1)\frac{(2\tilde{n}_j - 3)\tilde{\pi}_j^3}{\tilde{n}_j^5} \\ &- \frac{2\tilde{\pi}_j + (2\tilde{n}_j - 3)\tilde{\pi}_j^2 - 2(\tilde{n}_j - 1)\tilde{\pi}_j^3}{\tilde{n}_j^4}\Bigg]\end{aligned}$$

3. Criterio para evaluar los intervalos de confianza

Hay dos conceptos importantes que se deben considerar al evaluar los intervalos de confianza: la precisión, indicada por la longitud del intervalo y la probabilidad de cobertura $P\left(L_{inf} \leq \pi_1 - \pi_2 \leq L_{sup}\right)$. Estos dos criterios no los podemos analizar por separado porque de poco nos sirve un intervalo con probabilidad de cobertura alta si su longitud es muy grande o un intervalo con una longitud muy pequeña pero con probabilidad de cobertura muy baja. Idealmente queremos que los intervalos sean cortos y tengan probabilidad de cobertura muy cercana al nivel de confianza nominal. Buscamos que los procedimientos que usemos para construir los intervalos de confianza nos den intervalos tales que:

1. Sus longitudes sean pequeñas, pero diferentes de cero.

2. La probabilidad de cobertura no sea inferior al nivel de confianza nominal.

Un buen método debería dar intervalos con longitudes pequeñas y probabilidad de cobertura cercana al nivel de confianza nominal; pero no necesariamente un método que produzca intervalos cortos tiene una probabilidad de cobertura cercano al nivel nominal. Para comparar mediante simulación los diferentes métodos, calcularemos para cada método el valor promedio de las longitudes de esos intervalos y el nivel de confianza real: la proporción de intervalos simulados que cubre el verdadero valor de $\pi_1 - \pi_2$, el nivel de confianza real es una estimación de la probabilidad de cobertura. Para trabajar conjuntamente con la longitud promedio del intervalo y el nivel de confianza real construimos el siguiente índice:

$$I = \frac{2 - LPI}{2} \times \frac{NR}{NN}$$

donde LPI es la longitud promedio del intervalo, NR es el nivel de confianza real, y NN es el nivel de confianza nominal. Este índice es útil para este caso, ya que la longitud de un intervalo estará siempre entre cero y dos. Idealmente la fracción NR/NN debe estar muy cercana a uno, pero si la longitud del intervalo es muy grande entonces el índice castigará el método. Por lo tanto, entre mayor sea el índice tanto mejor el método.

4. Resultados de Simulación

Para comparar los distintos intervalos se realizó una simulación en Lenguaje R versión 1.31. Se generaron 1000 muestras de tamaños $10, 20, 40, 60, 80, 100$ de distribuciones binomiales con $\pi = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ y 0.9. Para cada una de las combinaciones de $n's$ y $\pi's$ y cada método se calcularon: los 1000 intervalos de confianza del 95%, para $\pi_1 - \pi_2$, los niveles de confianza reales y los promedios de las 1000 longitudes correspondientes. Para la presentación de resultados en este documento se eligieron algunas combinaciones $(n_1, \pi_1; n_2, \pi_2)$. Las Tablas 1 a 6 muestran los índices, junto con los niveles de confianza reales y los promedios de las longitudes de los intervalos, para cada uno de los cinco métodos analizados y para diferentes valores de n_1, n_2, π_1 y π_2.

4.1. Tablas

Las siguientes son las convenciones que se usarán en las tablas:

- NR: Nivel de confianza real
- LPI: Longitud promedio del intervalo
- I: Indice, $I = \frac{2-LPI}{2} \times \frac{NR}{NN}$, donde NN es el nivel de confianza nominal que en este caso es 0.95

Tabla 1: $n_1 = n_2 = 20$ **y** $\pi_1 = 0.2$

	Wald			**Wald corregido**			**Newcombe**		
π_2	LPI	NR	I	LPI	NR	I	LPI	NR	I
0.10	0.42	0.94	0.78	0.42	0.97	0.80	0.46	0.96	0.78
0.20	0.48	0.95	0.76	0.48	0.98	0.79	0.49	0.96	0.77
0.30	0.52	0.94	0.74	0.52	0.97	0.76	0.51	0.95	0.75
0.40	0.54	0.94	0.72	0.54	0.98	0.75	0.52	0.95	0.74
0.50	0.55	0.94	0.72	0.55	0.97	0.74	0.52	0.95	0.74
0.60	0.54	0.92	0.71	0.54	0.97	0.75	0.51	0.95	0.75
0.70	0.52	0.94	0.73	0.52	0.96	0.75	0.49	0.94	0.74
0.80	0.48	0.90	0.72	0.48	0.97	0.77	0.46	0.95	0.77
0.90	0.42	0.93	0.77	0.42	0.95	0.79	0.43	0.91	0.76

Tabla 1: $n_1 = n_2 = 20$ **y** $\pi_1 = 0.2$ (Continuación)

	Agresti Caffo			**Wei Pan**		
π_2	LPI	NR	I	LPI	NR	I
0.10	0.46	0.97	0.79	0.46	0.97	0.79
0.20	0.49	0.97	0.77	0.50	0.97	0.77
0.30	0.52	0.96	0.75	0.52	0.96	0.75
0.40	0.54	0.96	0.74	0.54	0.96	0.74
0.50	0.54	0.96	0.73	0.54	0.96	0.73
0.60	0.53	0.96	0.74	0.53	0.96	0.74
0.70	0.52	0.96	0.75	0.52	0.96	0.75
0.80	0.49	0.95	0.76	0.49	0.95	0.76
0.90	0.46	0.97	0.78	0.46	0.97	0.78

La Tabla 1 compara los métodos en muestras balanceadas con tamaños muestrales pequeños, dejando fija π_1 y variando π_2.

En esta comparación el método de Wald tiene el peor desempeño, presenta para todos los valores de π_2, los índices más bajos y solo en un caso, el nivel de confianza real alcanza el nominal.

Obsérvese el buen desempeño del Wald corregido: muestra los índices mayores y niveles de confianza reales muy por encima de los nominales. Los promedios de las longitudes son iguales que los que se obtienen con el método de Wald.

Los intervalos de Agresti & Caffo y los de Pan, superan el nivel nominal. Sus índices muy similares, menores que los de Wald corregido y casi siempre mayores que los de Newcombe.

Los intervalos de Newcombe aunque tienen índices mayores que los de Wald, no siempre su nivel de confianza real alcanzan el nivel nominal del 0.95.

La Tabla 2 compara los métodos en muestras balanceadas pero con tamaños muestrales grandes, dejando fija π_1 y variando π_2.

Los tamaños muestrales grandes mejoran el desempeño de todos los métodos: las longitudes de los intervalos son menores y los niveles de confianza reales son mayores, presentando índices mejores que en el caso analizado en la Tabla 1.

La comparación que se hace en la tabla 2 muestra nuevamente el mejor desempeño para Wald corregido y el menor desempeño para el intervalo de Wald; los intervalos de Newcombe, de Agresti y Caffo y los de Pan se desempeñan muy similarmente sin alcanzar el buen desempeño de los de Wald corregido.

Tabla 2: $n_1 = n_2 = 50$ **y** $\pi_1 = 0.2$

	Wald			**Wald corregido**			**Newcombe**		
π_2	LPI	NR	I	LPI	NR	I	LPI	NR	I
0.10	0.27	0.95	0.86	0.27	0.97	0.88	0.28	0.96	0.87
0.20	0.31	0.94	0.84	0.31	0.97	0.86	0.31	0.95	0.85
0.30	0.33	0.95	0.84	0.33	0.97	0.85	0.33	0.95	0.84
0.40	0.35	0.93	0.81	0.35	0.96	0.83	0.34	0.94	0.82
0.50	0.35	0.95	0.82	0.35	0.96	0.84	0.34	0.94	0.82
0.60	0.35	0.94	0.82	0.35	0.96	0.84	0.34	0.96	0.84
0.70	0.33	0.95	0.84	0.33	0.97	0.85	0.33	0.95	0.84
0.80	0.31	0.93	0.82	0.31	0.96	0.85	0.31	0.94	0.84
0.90	0.27	0.93	0.85	0.27	0.96	0.88	0.27	0.95	0.87

Tabla 2: $n_1 = n_2 = 50$ **y** $\pi_1 = 0.2$ (continuación)

	Agresti Caffo			**Wei Pan**		
π_2	LPI	NR	I	LPI	NR	I
0.10	0.28	0.96	0.87	0.28	0.96	0.87
0.20	0.31	0.96	0.85	0.31	0.96	0.85
0.30	0.33	0.96	0.84	0.33	0.96	0.84
0.40	0.35	0.95	0.82	0.35	0.95	0.82
0.50	0.35	0.94	0.82	0.35	0.94	0.82
0.60	0.35	0.95	0.83	0.35	0.95	0.83
0.70	0.33	0.96	0.84	0.33	0.96	0.84
0.80	0.31	0.94	0.84	0.31	0.94	0.84
0.90	0.28	0.96	0.87	0.28	0.96	0.87

Obsérvese cómo, a pesar de que el tamaño de muestra es "grande", los intervalos construidos con el método de Wald todavía no alcanzan el nivel nominal.

La Tabla 3 compara los métodos para muestras desbalanceadas con tamaños muestrales pequeños, dejando fija π_1 y variando π_2.

Como se puede ver, los índices para el caso que se muestra en esta tabla son menores que los de las tablas 1 y 2, esto se explica por los tamaños muestrales pequeños.

Los intervalos de Wald corregido y los de Newcombe tienen el mejor comportamiento, aunque para los intervalos construidos con el método de Newcombe el nivel de confianza real no llega, para todos los valores de π_2, al 95%.

Los intervalos de confianza de Agresti & Caffo y de Pan tienen índices iguales y alcanzan niveles de confianza reales muy superiores, pero con prome-

dio de longitudes más grandes, produciendo índices menores que los de Wald corregido y que los de Newcombe.

Tabla 3: $\pi_1 = 0.2$**,** $n_1 = 20$ **y** $n_2 = 10$

	Wald			**Wald corregido**			**Newcombe**		
π_2	LPI	NR	I	LPI	NR	I	LPI	NR	I
0.10	0.52	0.88	0.69	0.52	0.95	0.74	0.55	0.98	0.74
0.20	0.57	0.90	0.68	0.57	0.97	0.74	0.58	0.96	0.72
0.30	0.61	0.92	0.68	0.61	0.97	0.71	0.60	0.95	0.70
0.40	0.62	0.92	0.67	0.62	0.96	0.70	0.61	0.94	0.69
0.50	0.63	0.92	0.66	0.63	0.96	0.69	0.60	0.96	0.71
0.60	0.62	0.92	0.67	0.62	0.96	0.69	0.60	0.96	0.71
0.70	0.61	0.91	0.67	0.61	0.96	0.71	0.58	0.94	0.70
0.80	0.57	0.89	0.67	0.57	0.95	0.71	0.55	0.93	0.71
0.90	0.52	0.88	0.68	0.52	0.95	0.74	0.52	0.95	0.74

Tabla 3: $\pi_1 = 0.2$**,** $n_1 = 20$ **y** $n_2 = 10$ (continuación)

	Agresti Caffo			**Wei Pan**		
π_2	LPI	NR	I	LPI	NR	I
0.10	0.57	0.99	0.74	0.57	0.99	0.74
0.20	0.60	0.97	0.71	0.60	0.97	0.71
0.30	0.62	0.97	0.70	0.62	0.97	0.70
0.40	0.64	0.96	0.69	0.64	0.96	0.69
0.50	0.64	0.97	0.70	0.64	0.97	0.70
0.60	0.64	0.96	0.69	0.64	0.96	0.69
0.70	0.62	0.96	0.69	0.63	0.96	0.69
0.80	0.60	0.96	0.71	0.60	0.97	0.71
0.90	0.57	0.96	0.72	0.57	0.96	0.72

Igual que en los casos presentados en las tablas anteriores, para el caso presentado en la tabla 3 concluimos que el método de Wald produce los intervalos con menor índice entre los métodos que estamos comparando.

La Tabla 4 compara los métodos para muestras desbalanceadas, una con tamaño muestral pequeño y la otra con tamaño muestral grande, dejando fija π_1 y variando π_2.

En este caso es el método de Newcombe el que da intervalos con mayores índices y niveles de confianza reales por encima del nominal.

Los índices de los intervalos de Agresti Caffo son iguales a los de Pan, pero superiores a los de Wald y Wald corregido.

Tabla 4: $n_1 = 10$**,** $n_2 = 50$ **y** $\pi_1 = 0.2$

	Wald			**Wald corregido**			**Newcombe**		
π_2	LPI	NR	I	LPI	NR	I	LPI	NR	I
0.10	0.48	0.88	0.70	0.48	0.89	0.71	0.49	0.97	0.77
0.20	0.51	0.88	0.69	0.51	0.92	0.72	0.51	0.94	0.74
0.30	0.52	0.87	0.68	0.52	0.94	0.73	0.52	0.96	0.75
0.40	0.53	0.88	0.68	0.53	0.94	0.73	0.52	0.95	0.74
0.50	0.54	0.89	0.69	0.54	0.94	0.73	0.52	0.95	0.74
0.60	0.52	0.87	0.68	0.52	0.93	0.72	0.52	0.96	0.75
0.70	0.52	0.89	0.69	0.52	0.94	0.73	0.51	0.96	0.75
0.80	0.51	0.91	0.71	0.51	0.94	0.74	0.49	0.96	0.76
0.90	0.48	0.87	0.70	0.48	0.89	0.71	0.47	0.95	0.77

Tabla 4: $n_1 = 10$**,** $n_2 = 50$ **y** $\pi_1 = 0.2$ (continuación)

	Agresti Caffo			**Wei Pan**		
π_2	LPI	NR	I	LPI	NR	I
0.10	0.52	0.98	0.77	0.52	0.98	0.77
0.20	0.54	0.96	0.74	0.54	0.96	0.74
0.30	0.55	0.97	0.74	0.55	0.97	0.74
0.40	0.56	0.96	0.73	0.56	0.96	0.73
0.50	0.56	0.97	0.73	0.56	0.97	0.73
0.60	0.55	0.97	0.74	0.55	0.97	0.74
0.70	0.55	0.97	0.74	0.55	0.97	0.74
0.80	0.54	0.98	0.75	0.54	0.98	0.75
0.90	0.52	0.96	0.75	0.52	0.96	0.75

El método de Wald corregido presenta niveles de confianza reales por debajo de los nominaes, no tiene en este caso tan buen comportamiento como en los casos que se analizaron en las tablas anteriores.

También, para los casos presentados en esta tabla, es el método de Wald el de más pobre desempeño.

Note que de acuerdo a la definición de los correspondientes intervalos, siempre el promedio de las longitudes de los intervalos construídos con los métodos Wald y Wald corregido es el mismo.

La Tabla 5 compara los métodos para muestras desbalanceadas con tamaños muestrales grandes, dejando fija π_1 y variando π_2.

Con muestras grandes, los índices mejoran para todos los métodos. Sin embargo no en todos los casos el nivel de confianza real alcanza el nivel nominal.

Los intervalos construidos con el método de Wald corregido tienen los mayores índices y los niveles de confianza reales alcanzan ó superan el nivel nominal.

Tabla 5: $n_1 = 50$, $n_2 = 100$, **y** $\pi_1 = 0.2$

	Wald			**Wald corregido**			**Newcombe**		
π_2	LPI	NR	I	LPI	NR	I	LPI	NR	I
0.10	0.25	0.94	0.86	0.25	0.96	0.88	0.25	0.95	0.88
0.20	0.27	0.93	0.85	0.27	0.95	0.87	0.27	0.94	0.86
0.30	0.28	0.95	0.86	0.28	0.97	0.87	0.28	0.95	0.86
0.40	0.29	0.96	0.86	0.29	0.98	0.88	0.29	0.96	0.86
0.50	0.29	0.95	0.85	0.29	0.97	0.87	0.29	0.95	0.86
0.60	0.29	0.94	0.84	0.29	0.96	0.87	0.29	0.94	0.85
0.70	0.28	0.94	0.85	0.28	0.95	0.86	0.28	0.94	0.85
0.80	0.27	0.95	0.86	0.27	0.97	0.88	0.27	0.95	0.87
0.90	0.25	0.93	0.86	0.25	0.95	0.88	0.25	0.94	0.87

Tabla 5: $n_1 = 50$, $n_2 = 100$, **y** $\pi_1 = 0.2$ (continuación)

	Agresti Caffo			**Wei Pan**		
π_2	LPI	NR	I	LPI	NR	I
0.10	0.25	0.95	0.87	0.25	0.95	0.87
0.20	0.27	0.94	0.86	0.27	0.94	0.86
0.30	0.28	0.95	0.86	0.28	0.95	0.86
0.40	0.29	0.96	0.87	0.29	0.96	0.87
0.50	0.29	0.95	0.86	0.29	0.95	0.86
0.60	0.29	0.94	0.85	0.29	0.94	0.85
0.70	0.28	0.94	0.85	0.28	0.94	0.85
0.80	0.27	0.95	0.87	0.27	0.95	0.87
0.90	0.25	0.94	0.87	0.25	0.94	0.87

Observe que el método de Wald presenta los menores índices y además la mayoría de las veces el nivel de confianza real no llega al nominal.

Para los otros intervalos el índice, en todos los casos, es igual o ligeramente inferior al que se obtiene en los intervalos por Wald corregido, pero no siempre se alcanza el nivel del 0.95.

La Tabla 6 compara los métodos para muestras balanceadas variando los tamaños muestrales, dejando fijas π_1 y π_2.

Como en todos los métodos la base es la distribución asintótica normal a la binomial, es de esperarse que a medida que aumenta el tamaño muestral aumenta el índice.

Tabla 6: $n_1 = n_2$, $\pi_1 = 0.1$ **y** $\pi_2 = 0.15$

	Wald			**Wald corregido**			**Newcombe**		
$n_1 = n_2$	LPI	NR	I	LPI	NR	I	LPI	NR	I
10.00	0.51	0.85	0.67	0.51	0.99	0.78	0.63	0.98	0.71
20.00	0.39	0.92	0.78	0.39	0.98	0.83	0.44	0.96	0.79
30.00	0.32	0.95	0.84	0.32	0.98	0.87	0.35	0.97	0.84
50.00	0.25	0.95	0.87	0.25	0.97	0.90	0.27	0.96	0.88
70.00	0.22	0.95	0.89	0.22	0.97	0.91	0.22	0.95	0.89
100.00	0.18	0.94	0.90	0.18	0.97	0.93	0.19	0.95	0.90

Tabla 6: $n_1 = n_2$, $\pi_1 = 0.1$ **y** $\pi_2 = 0.15$ (continuación)

	Agresti Caffo			**Wei Pan**		
$n_1 = n_2$	LPI	NR	I	LPI	NR	I
10.00	0.64	0.99	0.71	0.64	0.99	0.71
20.00	0.44	0.97	0.80	0.44	0.97	0.80
30.00	0.35	0.98	0.85	0.35	0.98	0.85
50.00	0.27	0.97	0.88	0.27	0.97	0.88
70.00	0.22	0.96	0.90	0.22	0.96	0.90
100.00	0.19	0.96	0.92	0.19	0.96	0.92

El método de Wald corregido muestra los índices más altos y el de Wald muestra los más bajos. Los otros métodos que hemos analizado, muestran índices muy similares entre ellos pero inferiores a los de Wald corregido.

Tabla 7: Medidas de Resumen de los Indices

	Wald	Wald corregido	Newcombe	Agresti y Caffo	Pan
Min	0.61	0.63	0.65	0.64	0.63
1st Qu	0.68	0.72	0.74	0.72	0.72
Median	0.72	0.77	0.77	0.76	0.76
Mean	0.75	0.79	0.79	0.78	0.78
3 st Qu	0.86	0.88	0.86	0.86	0.86
Max	0.91	0.94	0.92	0.93	0.93

La Tabla 7 presenta, para cada método, las medidas descriptivas para los índices. Como se observa, los índices de Wald corregido tienen el promedio, la mediana, el tercer cuartil y el máximo mayor o igual que en los otros métodos.

5. Conclusiones y recomendaciones

Como se observa en las tablas de la sección 4, el método más deficiente es el de Wald. Desafortunadamente este método es el más difundido. Tampoco los paquetes estadísticos de uso más frecuente como el SAS, el SPSS o el Minitab traen rutinas para calcular intervalos de confianza para la diferencia de proporciones y entonces el investigador usa el intervalo de Wald cuya única virtud es la de ser el más simple.

De los resultados recomendamos usar el método de Wald corregido que tiene el mejor desempeño de los métodos analizados, no es muy complicado de calcular y teóricamente no es complejo.

Queda por hacer un estudio comparativo de otros métodos, como los Bayesianos o los basados en versosimilitud. Los métodos que aquí se presentaron pueden explicarse fácilmente en un curso básico de estadística.

Bibliografía

Agresti, A. & Caffo, B. (2000), 'Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures', *The American Statistician* **54**(4), 280–288.

Agresti, A. & Coull, B. (1998), 'Approximate is better than "exact" for interval estimation of binomial proportion', *Journal of the American Statistical Association* **52**(2), 119–126.

Brown, L. D., Cai, T. T. & DasGupta, A. (2001), 'Interval estimation for binomial proportion', *Statistical Science* **16**(2), 101–133.

Chan, I. S. F. & Zhang, Z. (1999), 'Test-based exact confidence intervals for the difference of two binomial proportions', *Biometrics* **55**(4), 1202–1209.

Gart, J. J. & Nam, J. (1990), 'Approximate interval estimation of difference in binomial parameters: Correction for skewness and extension to multiple tables', *Biometrics* **46**, 637–643.

Hauck, W. W. & Anderson, S. (1986), 'A comparison of large-sample confidence interval methods for the difference of two binomial probabilities', *The American Statistician* **40**(4), 318–322.

Henderson, M. & C., M. M. (2001), 'Exploring the confidence interval for binomial parameter in the a first course in statistical computing', *The American Statistician* **55**(4), 337–344.

Johnson, N. L. & Kotz, S. (1969), *Discrete Distributions*, John Wiley and Sons.

Lloyd, C. J. (1990), 'Confidence intervals from the difference between two correlated proportions', *Journal of the American Statistical Association* **85**(412), 1154–1158.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974), *Introduction to the Theory of Statistics*, 3rd edn, McGraw-Hill, Kogasakua, Ltd: Tokyo.

Newcombe, R. G. (1998*a*), 'Improved confidence intervals for the difference between binomial proportions based on paired data', *Statistics of Medicine* **17**, 2635–2650.

Newcombe, R. G. (1998*b*), 'Interval estimation for the diference between independent proportions: Comparison of eleven methods', *Statistics of Medicine* **17**, 873–890.

Newcombe, R. G. (1998*c*), 'Two-sided confidence interval for the single proportion: Comparison of seven methods', *Statistics of Medicine* **17**, 857–872.

Pan, W. (n.d.), *Approximate Confidence Intervals for One Proportion y Difference of Two Proportions*. Email: Weip@biostat.umn.edu.

Peskun, P. H. (1993), 'A new confidence interval method based on the normal approximation for the difference of two binomial probabilities', *Journal of the American Statistical Association* **88**(422), 656–661.

Santner, T. J. & Snell, M. K. (1980).

Snedecor, G. W. & Cochran, W. G. (n.d.).

Wendell, J. P. & Schmee, J. (2001), 'Likelihood confidence intervals for proportions in finite populations', *The American Statistician* **55**(1), 55–61.

Wild, C. J. & Seber, G. A. F. (1993), 'Comparing two proportions from the same survey', *The American Statistician* **47**(3), 178–181.