Algoritmo que Proporciona Secuencias de Enteros Ordenadas para Intervalos de Confianza en Muestreo con Reemplazo

Briggitte Jhosselyn Vilca Chambilla
Renzo Robiño Tito Chura
Lenin Alfonso Suarez Ccama
Jhon Milfer Tintaya Sinca
Cinthia Yaneth Fonseca Lizarraga

Universidad Nacional del Altipano Puno
Facultad de Ingeniería Estadística e Informática

Noviembre 2024

Contenido

Contexto del Problema

Muestreo con Reemplazo en la Práctica

- Extracción de individuos **sin eliminación** de la población
- Distribución binomial: $f_B(y; x, m) = \binom{m}{y} \left(\frac{x}{m}\right)^y \left(1 - \frac{x}{m}\right)^{m-y}$
- Aplicaciones en física, biología, medicina, ingeniería

Problema Fundamental

- **Discreción** de la distribución de probabilidad
- **No existe fórmula matemática** óptima para IC discretos
- Métodos tradicionales requieren tamaños de muestra grandes

Limitaciones de Métodos Tradicionales

Método	**Ventajas**	**Limitaciones**
Wald	Simple de calcular	Pobre cobertura para muestras pequeñas
Clopper-Pearson	Conservador	Muy amplio, sesgado
Wilson	Mejor para proporciones	No óptimo para distribuciones discretas

Ejemplo con $m = 10$, $\alpha = 0{,}05$

Para $x = 3$: Wald da IC=[1,5] con cobertura 92.44 %, pero podría ser [0,5] (95.27 %) o [1,6] (96.12 %)

Fundamentos de Intervalos de Confianza

Definición Formal

Para un parámetro θ, un IC al $(1-\alpha) \times 100\,\%$ es:

$$IC_{1-\alpha}(\theta) = \left[\hat{\theta} - z_{\alpha/2} \cdot SE(\hat{\theta}),\ \hat{\theta} + z_{\alpha/2} \cdot SE(\hat{\theta})\right]$$

Interpretación Correcta vs Incorrecta

- **Correcto:** "Si repitiéramos el experimento muchas veces, el $(1-\alpha) \times 100\,\%$ de los ICs contienen el parámetro verdadero"
- **Incorrecto:** .El parámetro tiene probabilidad $(1-\alpha)$ de estar en el intervalo"

IC para Proporciones - Métodos Clásicos

IC de Wald (Aproximación Normal)

$$CI_W(x; m, \alpha) = x \pm z_\alpha \sqrt{\frac{x(m-x)}{m}}$$

- Basado en aproximación normal
- Problemas con muestras pequeñas o proporciones extremas

IC de Wilson (Más Robusto)

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

Enfoque Innovador del Paper

Propuesta Central

- Reemplazar **fórmulas matemáticas** por **algoritmo computacional**
- Generar **secuencias ordenadas** de enteros: $N_0, N_1, \ldots, N_m$
- Construir IC **simétricos**: $[N_x, m - N_{m-x}]$

Ventajas Clave

- ✓Óptimo para distribuciones discretas
- ✓No requiere aproximaciones
- ✓Aplicable a cualquier n
- ✓Cobertura cercana al nivel nominal
- ✓Simetría garantizada
- ✓Implementación simple

Algoritmo BalancedCI - Pseudocódigo

Require: x, m, α {x éxitos en m ensayos, nivel α}

for $y \leftarrow 1, m$ **do**
    $r_y \leftarrow f_{RB}(y; x, m)$ {Probabilidad binomial}
end for
$\beta \leftarrow 1 - \alpha$
$i \leftarrow x$; $j \leftarrow x$; $q \leftarrow r_x$
while verdadero **do**
    if $j \leq i$ **then**
    end if
    if $r_i = r_j$ **and** $|\beta - q| > |\beta - q + r_i + r_j|$ **then**
        $q \leftarrow q - r_i - r_j$; $i \leftarrow i + 1$; $j \leftarrow j - 1$
    end if
    if $r_i < r_j$ **and** $|\beta - q| > |\beta - q + r_i|$ **then**
        $q \leftarrow q - r_i$; $i \leftarrow i + 1$
    end if
    if $r_i > r_j$ **and** $|\beta - q| > |\beta - q + r_j|$ **then**
        $q \leftarrow q - r_j$; $j \leftarrow j - 1$
    end if
end while

Ensure: i, j, q {Límites del IC y cobertura real}

Características del Algoritmo

Propiedades Matemáticas

- **Monotonicidad**: $N_0 \leq N_1 \leq \cdots \leq N_m$
- **Simetría**: N_x y N_{m-x} son simétricos
- **Balance**: Error real cercano al error impuesto

Complejidad Computacional

- Complejidad: $O(\sqrt{m})$
- Más rápido para x cercano a 0 o m
- Más lento para x cerca de $m/2$ ($\approx \sqrt{m}$ evaluaciones)

Ejemplo Práctico: $m = 30$, $\alpha = 0{,}05$

Secuencia Generada

0, 0, 0, 1, 1, 2, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 23, 24, 26, 28, 30

Cálculo de IC para $x = 26$

$$
\begin{aligned}
N_x &= N_{26} = 23 \\
N_{m-x} &= N_4 = 1 \\
IC &= [N_x, m - N_{m-x}] = [23, 30 - 1] = [23, 29] \\
\text{Cobertura real} &= 94{,}8\,\%
\end{aligned}
$$

Comparación de Coberturas

Método	$m = 30$	$m = 45$	$m = 100$	$m = 900$
BalancedCI (Propuesto)	**5.10 %**	**5.00 %**	**4.96 %**	**5.00 %**
Clopper-Pearson	4.10 %	4.20 %	4.50 %	4.90 %
Wald	Variable	Variable	5.00 %	5.00 %
Jeffrey	4.80 %	4.90 %	4.95 %	4.99 %

Cuadro: No cobertura promedio ($\alpha = 5\,\%$ nominal) - **Menor es mejor**

- **BalancedCI** mantiene cobertura más cercana al nivel nominal
- Métodos tradicionales tienden a ser conservadores
- Wald tiene cobertura irregular en muestras pequeñas

Aplicación en Caso Real - Estudio Médico

Datos del Estudio [Viana, 2025]

- 20 pacientes con regeneración dérmica
- Detección bacteriana: 8/20 positivos
- Desarrollo de infección: 4/20 casos
- Pseudomonas: 3/20 casos

Análisis con BalancedCI ($m = 20$, $\alpha = 0{,}05$)

Variable	IC 95 %	Cobertura Real	Significativo
Detección bacteriana	[4, 12]	96.4 %	Sí
Desarrollo infección	[1, 7]	95.7 %	Sí
Pseudomonas	[1, 6]	93.9 %	Sí

Visualización de Resultados

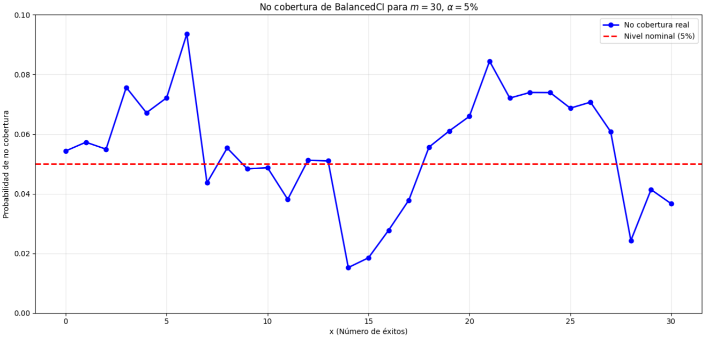

Figura: No cobertura de BalancedCI para $m = 30$, $\alpha = 5\,\%$

- Comportamiento **oscilatorio** alrededor del nivel nominal
- **Simetría** garantizada por el algoritmo
- **Convergencia** al nivel nominal conforme m aumenta
- Línea roja: nivel de significancia nominal (5 %)
- Puntos azules: no cobertura real para cada x

Posicionamiento Respecto a Enfoques Existentes

Clasificación de Métodos

Fórmulas aproximadas: Wald, Agresti-Coull, Wilson

Conservadores: Clopper-Pearson, Blyth-Still

Bayesianos: Jeffrey

Algorítmicos: BalancedCI (Propuesto)

Ventaja Competitiva

Balance óptimo entre cobertura nominal y real

No demasiado conservador • No demasiado liberal • Siempre cercano al objetivo

Ventajas Prácticas del BalancedCI

Para Investigadores

- ICs más precisos en estudios pequeños
- Fácil implementación computacional
- Resultados reproducibles
- Aplicable a múltiples disciplinas

Para Aplicaciones

- Control de calidad industrial
- Estudios clínicos y médicos
- Investigación biológica
- Encuestas y sondeos

Implementación

Disponible en `R`, `Python`, `MATLAB` mediante el algoritmo publicado

Contribuciones Principales

Avances Metodológicos

- **Primer algoritmo** específico para IC en muestreo con reemplazo
- **Solución exacta** para el problema de discreción
- **Implementación eficiente** con $O(\sqrt{m})$ complejidad
- **Validación empírica** con casos reales

Impacto Práctico

- Mejora la precisión en estudios con muestras pequeñas
- Proporciona alternativa a métodos aproximados
- Facilita análisis estadísticos más confiables

Limitaciones y Trabajo Futuro

Limitaciones Actuales

- Requiere implementación computacional
- No extensible directamente a distribuciones multinomiales
- Depende de la generación de secuencias para cada (m, α)

Direcciones de Investigación Futura

- Extensión a distribuciones multinomiales
- Optimización de performance computacional
- Integración en paquetes estadísticos comerciales
- Aplicación a problemas de hipótesis múltiples

Resumen y Recomendaciones

Mensaje Principal

Para datos binomiales discretos, los algoritmos pueden superar a las fórmulas matemáticas tradicionales

Recomendación Práctica

Usar BalancedCI cuando:

- Tamaños de muestra pequeños o moderados
- Proporciones cercanas a 0 o 1
- Se requiere máxima precisión en la cobertura
- Los métodos tradicionales muestran problemas

Referencias Bibliográficas

- Jäntschi, L. (2025). *Algorithm Providing Ordered Integer Sequences for Sampling with Replacement Confidence Intervals*. Algorithms, 18, 459.
- Wald, A. (1939). *Contributions to the theory of statistical estimation and testing hypotheses*. Annals of Mathematical Statistics, 10, 299-326.
- Clopper, C.J.; Pearson, E.S. (1934). *The use of confidence or fiducial limits illustrated in the case of the binomial*. Biometrika, 26, 404-413.
- Wilson, E.B. (1927). *Probable inference, the law of succession, and statistical inference*. Journal of the American Statistical Association, 22, 209-212.
- Agresti, A.; Coull, B.A. (1998). *Approximate is better than .exact"for interval estimation of binomial proportions*. The American Statistician, 52, 119-126.
- Viana, E.H. (2025). *Bacterial Fluorescence Signals Are Associated with Dermal Regeneration Template Infections in Burn Patients*. Journal of Burn Care Research, 46, S149.

¡Gracias por su atención!

Preguntas y Discusión

Grupo 3
Briggitte Jhosselyn Vilca Chambilla
Renzo Robiño Tito Chura
Lenin Alfonso Suarez Ccama
Jhon Milfer Tintaya Sinca
Cinthia Yaneth Fonseca Lizarraga