Revista Colombiana de Estadística
Volumen **23** (2000) No. 1, páginas 65–70

CRITERIOS SOBRE EL USO DE LA DISTRIBUCIÓN NORMAL PARA APROXIMAR LA DISTRIBUCIÓN BINOMIAL

JORGE ORTIZ PINILLA[1]
AMPARO CASTRO
TITO NEIRA
PEDRO TORRES
JAVIER CASTAÑEDA[2]

Resumen. Las dos reglas empíricas más conocidas para aceptar la aproximación normal de la distribución binomial carecen de regularidad en el control del margen de error cometido al utilizar la aproximación normal. Se propone un criterio y algunas fórmulas para controlarlo cerca de algunos valores escogidos para el error máximo.

Abstract. The two most popular "rules of thumb" for the approximation of the binomial distribution by the normal distribution lack of regularity in the control of the maximal error. A criterion and some formulas for selected values of maximal errors are proposed to solve this lack of regularity.

Palabras Claves Distribución binomial, Aproximación normal,

1. Introducción

Tradicionalmente se ha considerado de manera empírica que los parámetros n y p deben satisfacer $np \geq 5$ y $n(1-p) \geq 5$ como condición para que la aproximación normal pueda ser utilizada en el marco de aplicación del teorema del límite central en el caso de la distribución binomial. Una segunda condición

(1) Profesor Asociado, Departamento de Estadística, Universidad Nacional de Colombia; e-mail: jortiz@matematicas.unal.edu.co
(2) Estudiantes Carrera de Estadística, Universidad Nacional de Colombia.

conocida establece más estrictamente que $np(1-p) \geq 9$, Siegel y Castellan (1995).

Schader y Schmid (1989), hacen notar que con las condiciones anteriores, la máxima diferencia en valor absoluto entre la función de distribución binomial exacta y la aproximación normal, presenta variaciones cercanamente lineales en función de p. De esto se deduce que la aplicación de cualquiera de ellas no permite controlar el error de aproximación de manera uniforme. En efecto, si $p = 0.5$, para cumplir la condición $np(1-p) \geq 9$, n debe ser mayor o igual a 36 . En este caso el valor absoluto del error es inferior a 0.00075384; mientras que si $p = 0.01$, según la misma condición $n \geq 909$, y el error llega a ser 0.02113783. Con errores un poco más grandes, se observa un comportamiento similar al utilizar la primera condición.

El objetivo de este artículo es presentar propuestas para los valores del parámetro n en función de p de tal manera que la magnitud del error de aproximación sea controlada cerca de valores δ prefijados. El trabajo fue realizado durante el desarrollo del curso de estadística no paramétrica en el segundo semestre de 1998, con el grupo de estudiantes cuya participación motivó un curso diferente y creativo.

2. Procedimiento

Para algunos valores prefijados ($\delta = 0.01, 0.005, 0.002$) y para p variando desde 0.01 hasta 0.5 con incrementos de 0.01, se calculó el mínimo valor del parámetro n de tal manera que la máxima diferencia en valor absoluto entre la función de distribución binomial exacta con parámetros n y p y la función de distribución normal aproximada, fuera menor que δ.

Para cada uno de los márgenes de error trabajados se buscó también identificar un modelo que permitiera estimar para cada p, el valor de n necesario para mantener el máximo error de aproximación dentro del margen establecido.

$$\min\Big\{n : \max_{k=0,1,\ldots,n} \Big| F_{n,p}(k) - \Phi\Big(\frac{k+0.5-np}{\sqrt{np(1-p)}}\Big) \Big| < \delta\Big\}$$

donde $F_{n,p}(k)$ representa la función de distribución binomial con parámetros n y p, calculada en el punto k y $\Phi(x)$ representa la función de distribución normal estándar calculada en el punto x.

La tabla 1 muestra los principales resultados obtenidos. En ella aparecen los tamaños de muestra para una distribución binomial que se requieren si se necesita que la aproximación normal (con corrección de Yates) arroje máximo un error inferior al indicado. También se presentan los valores de $np(1-p)$. Ellos permiten examinar los valores de p que satisfacen la segunda condición mencionada antes.

Tabla 1. Valores mínimos de n para obtener un error máximo en la aproximación normal de la distribución binomial con parámetros n, p

	Margen de error					
	0.010		0.005		0.002	
p	n	np(1-p)	n	np(1-p)	n	np(1-p)
0.50	4	1.0000	6	1.500	14	3.50
0.49	4	0.9996	7	1.749	16	4.00
0.48	4	0.9984	7	1.747	19	4.74
0.47	4	0.9964	8	1.993	22	5.48
0.46	4	0.9936	8	1.987	32	7.95
0.45	4	0.9900	11	2.723	47	11.63
0.44	5	1.2320	12	2.957	67	16.51
0.43	5	1.2255	15	3.677	89	21.81
0.42	5	1.2180	20	4.872	117	28.50
0.41	8	1.9352	25	6.048	147	35.56
0.40	8	1.9200	28	6.720	183	43.92
0.39	10	2.3790	36	8.564	226	53.77
0.38	11	2.5916	45	10.602	269	63.38
0.37	14	3.2634	52	12.121	322	75.06
0.36	17	3.9168	61	14.054	378	87.09
0.35	18	4.0950	69	15.698	438	99.65
0.34	21	4.7124	80	17.952	506	113.55
0.33	24	5.3064	94	20.783	579	128.02
0.32	28	6.0928	106	23.066	659	143.40
0.31	30	6.4170	120	25.668	746	159.57
0.30	34	7.1400	134	28.140	843	177.03
0.29	38	7.8242	152	31.297	948	195.19
0.28	43	8.6688	171	34.474	1061	213.90
0.27	48	9.4608	189	37.252	1186	233.76
0.26	54	10.3896	212	40.789	1323	254.55
0.25	60	11.2500	236	44.250	1475	276.56
0.24	67	12.2208	263	47.971	1638	298.77
0.23	70	12.3970	291	51.536	1818	321.97
0.22	82	14.0712	323	55.427	2019	346.46
0.21	90	14.9310	358	59.392	2242	371.95
0.20	100	16.0000	399	63.840	2486	397.76
0.19	111	17.0829	442	68.024	2762	425.07
0.18	122	18.0072	489	72.176	3067	452.69

Tabla 1. Valores mínimos de n para obtener un error máximo en la aproximación normal de la distribución binomial con parámetros n, p (Continuación).

	Margen de error					
	0.010		0.005		0.002	
p	n	np(1-p)	n	np(1-p)	n	np(1-p)
0.17	136	19.1896	547	77.182	3412	481.43
0.16	151	20.2944	607	81.581	3801	510.85
0.15	167	21.2925	680	86.700	4247	541.49
0.14	192	23.1168	763	91.865	4758	572.86
0.13	215	24.3165	854	96.587	5353	605.42
0.12	242	25.5552	967	102.115	6042	638.04
0.11	273	26.7267	1099	107.592	6871	672.67
0.10	311	27.9900	1259	113.310	7859	707.31
0.09	365	29.8935	1454	119.083	9076	743.32
0.08	424	31.2064	1698	124.973	10598	780.01
0.07	500	32.5500	2012	130.981	12556	817.40
0.06	601	33.8964	2431	137.108	15168	855.48
0.05	757	35.9575	3017	143.307	18841	894.95
0.04	973	37.3632	3898	149.683	24352	935.12
0.03	1334	38.8194	5365	156.121	33562	976.65
0.02	2054	40.2584	8301	162.700	51952	1018.26
0.01	4288	42.4512	17108	169.369	.	

Se puede observar que si el margen de error de aproximación debe ser 0.01, el criterio $np(1-p) > 9$ exige un valor de n demasiado grande si $p \geq .27$ y uno insuficiente si $p < 0.27$ y esta insuficiencia crece cuando p se acerca a 0.01. Si $p > 0.5$ la situación es simétrica y por lo tanto la regla $np(1-p) > 9$ mantiene el error por debajo de 0.01 solo si $0.27 \leq p \leq 0.73$. Cuando el margen de error es 0.005 puede verse que la misma regla solo funciona si $0.38 \leq p \leq 0.62$. Y para un error igual a 0.002 el intervalo de utilidad de la regla en cuestión es solo para $0.45 < p < 0.55$.

Si no se aplica la corrección de Yates, los tamaños requeridos para controlar el error máximo son todavía mayores.

La solución al problema de controlar el margen de error que aparece en la Tabla 1, se facilita en la práctica con los modelos descritos a continuación.

En la tabla 2 se muestran los coeficientes de modelos de regresión cuadrática en donde se tomó $np(1-p)$ como variable dependiente y p como variable independiente. Las ecuaciones obtenidas permiten estimar el valor de n mínimo para controlar el error de aproximación con precisión suficiente. El coeficiente de determinación encontrado fue superior a .999 en los tres casos.

$$n = \frac{b_0 + b_1 p + b_2 p^2}{p(1-p)}$$

Tabla 2. Coeficientes de modelos de regresión cuadrática para $np(1-p)$ en función de p para cada margen de error.

	Margen de error		
	0.010	0.005	0.002
b_0	44.1454644	177.0499937	1105.687260
b_1	−177.8572335	−711.4000926	−4429.251942
b_2	182.4136993	719.1980885	4443.666363

A partir de los modelos encontrados, se puede calcular n con las siguientes ecuaciones:

$$\text{error} = 0.01, \qquad n = \frac{44.1454644 - 177.8572335p + 182.4136993p^2}{p(1-p)}$$

$$\text{error} = 0.005, \qquad n = \frac{177.0499937 - 711.4000926p + 719.1980885p^2}{p(1-p)}$$

$$\text{error} = 0.002, \qquad n = \frac{1105.687260 - 4429.251942p + 4443.666363p^2}{p(1-p)}$$

Es posible encontrar algunas simplificaciones a los modelos encontrados sin sacrificar grandemente las estimaciones de n. Por ejemplo, para la primera ecuación (error $= 0.01$), se encontró que

$$n \sim 44\frac{1 - 4p(1-p)}{p(1-p)}$$

aproxima bastante bien el valor de n, siempre que los resultados no den menores que 4.

3. Conclusiones

Si el propósito es controlar la máxima diferencia entre la función de distribución binomial exacta y la aproximación normal con corrección de Yates, las reglas recomendadas tradicionalmente para el valor mínimo del parámetro n presentan inexactidutes importantes, generadas por la falta de regularidad en el control del error máximo. Para los valores de p cercanos a 0.5, las reglas, particularmente $np(1-p) > 9$, sobreestiman el valor de n necesario. Pero si p toma valores extremos (cerca de la unidad o de cero), el valor de n calculado a partir de ellas es insuficiente. El caso es más crítico con la primera regla enunciada al comienzo, siendo precisamente la más conocida y aplicada. Las propuestas presentadas ofrecen un control uniforme del margen de error de aproximación en el sentido de poder afirmar que si se calcula n con las fórmulas encontradas, el error será cercano al indicado para cada una.

Referencias

[1] Schader, Martin, Schmid, Friedrich, *Two rules of thumb for the approximation of the binomial distribution by the normal distribution.* The American Statistician **43** (1989), (23–24), no. 1.

[2] Siegel, Sidney, Castellan, J., *Estadística no paramétrica aplicada a las ciencias de la conducta.* Trillas, México (1995).