LA MUESTRA COMO BOLA DE CRISTAL

Luis Zapata Escobar

1. Introducción

A raíz de la publicación de las preguntas del referéndum vinculante comprometido por el Presidente de la República y planteado desde la oficina de Coordinación Gubernamental, en este trabajo se discute el uso de información muestral de fines de junio de 2004 para hacer pronósticos de los resultados del referéndum, previos al 18 de julio de 2004 y comparar con los resultados oficiales, ¿Qué hacer con el gas? planteado especialmente con las preguntas 4 y 5 de la boleta del referéndum a la letra dicen:

a) "¿Esta Ud. de acuerdo con la política del Presidente Carlos Mesa de utilizar el gas como recurso estratégico para el logro de una salida útil y soberana al océano pacífico?"

b) "¿Esta usted de acuerdo con que Bolivia exporte gas en el marco de una política nacional que cubra el consumo interno, fomente la industrialización en el país, cobre impuestos, destine los recursos de la exportación del gas a educación, salud, caminos, empleo.?

Las primeras encuestas en el ámbito político y ciudadano realizadas por el Coordinador del Referéndum, registró entre el 90 al 95 % de apoyo al contenido de las preguntas. En mayo, dirigentes sindicales y líderes de los movimientos sociales desahuciaban los resultados del referéndum del 18 de julio, vaticinaron el rechazo por una abstención masiva.

En mayo, empresas especializadas en medir opinión, ajenas al Gobierno realizaron sondeos por encargo de canales de televisión. Los resultados de las encuestas mostraban un cambio paulatino de la población hacia el contenido de las preguntas. Programas especiales, mesas redondas, paneles de análisis, seminarios de evaluación, dirigieron en cierta medida la opinión pública y las encuestas registraron esos cambios. En algunos casos como en la red Unitel el sondeo de opinión abarco pregunta por pregunta.

Un ampliado de trabajadores realizado en la zona cocalera del Chapare, emitió un voto resolutivo aconsejando a la población, en especial campesina, apoyar con el Si a las preguntas 1, 2 y 3 y con el No a las preguntas 4 y 5. A este pronunciamiento siguieron otros en el mismo sentido, en especial en la zona minera y campesina de occidente.

A finales de junio, encuestas en muestras pequeñas de ciudades del eje central, trataron de medir los cambios que pudieron existir a raíz del pronunciamiento del Chapare enfocados en especial a las preguntas 4 y 5. Los registros comparados con resultados de 15 días antes, mostraban, en efecto, la fuerte influencia generada por aquella declaración, en occidente creció el apoyo al NO, en el oriente y el sur, creció el apoyo por el SI, hablando -claro esta- de la 4ª y 5ª preguntas.

Los financiadores de esas encuestas (canales de televisión, prensa y radio) comentan y difunden los resultados de los porcentajes a favor del SI y comienzan a especular sobre la negativa que podrá dar la población para las preguntas 4 y 5 y las consecuencias para el presidente y la política trazada por él para ¿qué hacer con el gas?

De los antecedentes, el problema planteado es saber antes del 18 de julio, ¿Qué estimador es más útil al momento de pronosticar la proporción de votantes por el SI? ¿El estimador máximo verosímil? (las empresas y los financiadores usan como recurso inmediato ya sea en forma consiente o no) o ¿el estimador de Bayes?

2. Objetivo

Comparar la construcción del estimador de Bayes versus el estimador máximo verosímil para la proporción del Si en cualquiera de las preguntas del referéndum, en especial para las preguntas 4 y 5 en base a la información muestral recogida entre el 27 de junio y 10 de julio por Marketing SRL y cedida gentilmente.

3. Marco conceptual.

3.1. Distribución inicial o a priori y distribución final o a posteriori.

Considerando que el problema central de la inferencia estadística consiste en seleccionar observaciones de una variable X con función de probabilidad f (x| θ) donde θ es un parámetro de valor desconocido dentro un espacio paramétrico Ω. La solución es intentar determinar dónde es probable que se encuentre el verdadero valor de θ partiendo de una muestra aleatoria de X. Esta solución consiste en construir una distribución de probabilidad para θ en el conjunto Ω antes de haber obtenido una muestra, esta función es la distribución inicial o a priori de θ, g (θ) que representa la verosimilitud relativa de que el verdadero valor de θ se encuentra en cada una de las regiones de Ω antes de muestrear f (x| θ)

Sin embargo, existe controversia en utilizar el método bayesiano, muchos estadísticos creen que en todo problema estadístico se puede elegir una distribución inicial para el parámetro θ, que es de naturaleza subjetiva basada en la experiencia y que esta distribución no es distinta de ninguna otra distribución de probabilidad utilizada. Por otro lado, hay estadísticos que afirman que en muchos problemas no es apropiado hablar de una distribución de probabilidad de θ, porque el verdadero valor de θ no es una variable aleatoria sino más bien un valor fijo desconocido y que se podría asignar una distribución inicial para θ únicamente cuando se tiene a mano información previa, mejor si es extensa con distribución experimental de frecuencias relativas con las que θ ha tomado posibles valores a lo largo del tiempo.

Lo interesante en el caso del referéndum, único para los estadísticos, no hay antecedentes para estudiar el comportamiento histórico de casos como el presente, no hay historia y menos distribuciones de frecuencia. El referéndum de 1938 tenia 10 preguntas y los temas tenían un fin diferente y una población votante de élite.

Para construir la distribución final o posteriori, asumimos que las n variables aleatorias X_1 X_2, X_3,--- X_n, de la muestra aleatoria f(x| Θ), tienen distribución de probabilidad conjunta:

$$f_n(x_1, x_2, ...x_n|\theta) = f(x_1|\theta)f(x_2|\theta)..... f(x_n|\theta)$$

Puesto que se asumió que la distribución inicial de θ es g(θ), la función de probabilidad conjunta f_n se debe considerar como la función conjunta condicional de X_1 X_2 ... X_n para un valor dado de θ. Al multiplicar esta función de probabilidad conjunta condicional por g(θ), se obtiene la función de probabilidad conjunta de (n+1) dimensiones, todas las X_i
más θ, de la forma

$$f_n(x_1, x_2, ...x_n|\theta)\, g(\theta) \qquad (1)$$

La función de probabilidad conjunta marginal de X_1 X_2 ...X_n se encuentra por integración

$$h_n(x_1, x_2, \dots\dots, x_n) = \int_{\Omega} f_n(x_1, x_2, \dots\dots, x_n|\theta)\, g(\theta) d(\theta)$$

Además, la función de probabilidad condicional de θ dado X_1=x_1 X_2 = x_2...X_n = x_n que denotaremos por g(θ | x) debe ser igual a la función de probabilidad conjunta marginal de X_1, X_2, ...X_n, θ dividido entre la función de probabilidad conjunta marginal de X_1 X_2 ...X_n, por tanto resulta que la distribución final o a posteriori es

$$g(\theta|x) = \frac{f_n(x_1, x_2, \dots\dots, x_n|\theta) g(\theta)}{h_n(x_1, x_2, \dots\dots, x_n)} \qquad para\ \ \theta \in \Omega$$

3.2 Estimador de Bayes

Usando la notación $\tilde{\theta}$ para el estimador de θ que es una función de los valores observados de la variable X, $\tilde{\theta}$ = t(x_1, x_2,...x_n). Se define como buen estimador aquel que se aproxima lo más cerca al valor verdadero de θ, es decir que la diferencia ($\tilde{\theta}$ - θ) ~ 0. Así para cada valor de θ ε Ω existe un valor L (θ, $\tilde{\theta}$) que mide la pérdida o el costo para el investigador cuando el verdadero valor del parámetro es θ y su estimador es $\tilde{\theta}$. En general a medida que aumenta la distancia ($\tilde{\theta}$ - θ) será mayor el valor de L ($\tilde{\theta}$, θ) o función de pérdida.

Si g(θ) es la función de probabilidad inicial de θ ε Ω el investigador elige un particular $\tilde{\theta}$ la pérdida esperada será

$$E\left[L\left(\theta, \tilde{\theta}\right)\right] = \int_{\Omega} L\left(\theta, \tilde{\theta}\right) g(\theta) d(\theta)$$

Supongamos se elige el estimador $\hat{\theta} = t^*(x1, x2, \cdots e_n)$, es un estimador de Bayes si la pérdida esperada es mínima.

$$E\left[L\left(\theta, \hat{\theta}\right)\right] = \min_{\theta \in \Omega} E\left[L\left(\theta, \tilde{\theta}\right)\right]$$

La función de pérdida más utilizada es la del error cuadrático medio $E\left[L\left(\theta, \hat{\theta}\right)\right] = (\theta - \hat{\theta})^2$ y el estimador de Bayes, en este caso hace mínimo la perdida esperada si $\hat{\theta} = t^*(x_1, x_2, \cdots x_n)$.es la esperanza de la distribución de θ

4. Distribución inicial y final de la proporción ττ

El valor ττ ,la proporción del SI en el referéndum, para las 5 preguntas, eran desconocidas y pueden tomar infinitos valores en el intervalo (0,1) es necesario asignar una función de probabilidad a priori.

4.1. Función inicial uniforme.

$$g(\pi) = \begin{cases} 1 & 0 \leq \pi \leq 1 \\ 0 & en\ otro\ caso \end{cases} \tag{2}$$

Las encuestas realizadas en las cuatro ciudades del eje cuyos resultados se informaron el 7 de julio, dan una pauta del comportamiento de los votantes. Con X_i = 1 para el Si, X_i = 0 para el NO entonces X_1, X_2, ...X_n constituyen n pruebas Bernoulli con parámetro ττ donde:

$$f(x|\pi) = \begin{cases} \pi^x(1-\pi)^{1-x} & x = 0{,}1 \\ 0 & en\ otro\ caso \end{cases} \tag{3}$$

Entonces la función de probabilidad conjunta de X_1, X_2, ...X_n puede escribirse

$$\begin{aligned} f(x_1, x_2, \ldots\ldots, x_n|\pi) &= f(x_1|\pi)\, f(x_2|\pi)\ \cdots\ f(x_n|\pi) \\ &= \pi^{\sum x_i}(1-\pi)^{n-\sum x_i} \end{aligned} \tag{4}$$

Haciendo y = Σx_i entonces $f(x_1, x_2, \ldots\ldots, x_n|\pi) = \pi^y(1-\pi)^{n-y}$ (5)

Aplicando la ecuación (1) se obtiene la función

$$f_n(x_1, x_2, \ldots\ldots, x_n|\pi)g(\pi) = \pi^y(1-\pi)^{n-y} * 1 \tag{6}$$

Al comparar este resultado con la densidad de la función de probabilidad Beta β(p,q)

$$f(u|p,q) = \frac{(p+q-1)!}{(p-1)!\,(q-1)!}u^{p-1}(1-u)^{q-1} \quad 0 \leq u \leq 1$$

$$con\ E(U) = \frac{p}{p+q} \quad y\ V(U) = \frac{pq}{(p+q)^2(p+q+1)}$$

Se observa que excepto por un factor constante la ecuación (6) tiene la misma forma que el modelo β(p, q) con p = y + 1, q = n-y+1 en consecuencia la función de probabilidad final de ττ es una distribución beta.

$$g(\pi|x) = \frac{\Gamma(\mathrm{n}+2)}{\Gamma(\sum x_i + 1)\Gamma(\mathrm{n}+1-\sum x_i)}\pi^{\sum x_i}(1-\pi)^{n-\sum x_i}$$

Considerando como función de perdida el error cuadrático medio, y la función inicial uniforme, el estimador de bayes en esta circunstancia es

$$\hat{\pi} = \frac{1+y}{n+2} = \frac{1}{n+2}\left(1 + \sum_{i=1}^{n} x_i\right)$$

4.1. Función inicial triangular.

$$g(\pi) = \begin{cases} 2(1-\pi) & 0 \le \pi \le 1 \\ 0 & en\ otro\ caso \end{cases} \quad (7)$$

De la ecuación 3 y 4 la función de distribución final de TT resulta ser

$$f_n(x_1, x_2, \ldots\ldots, x_n|\pi)g(\pi) = \pi^y(1-\pi)^{n-y} * 2(1-\pi)$$

$$= 2\pi^y(1-\pi)^{n-y+1}$$

Resultando también, excepto un factor constantes, la distribución semejante a un modelo beta de parámetros p = y + 1 y q = n - y + 2

$$g(\pi|x) = \frac{(\mathrm{n}+2)!}{(1+\sum x_i)!\,(\mathrm{n}+2-\sum x_i)!}\pi^{\sum x_i}(1-\pi)^{n+1-\sum x_i}$$

Y el estimador de Bayes en este caso tiene la forma

$$\hat{\pi} = \frac{1+\sum x_i}{n+3}$$

4.3 Función inicial: beta (p, q) y f(x| TT): binomial

$$g(\pi) = \frac{(p+q-1)!}{(p-1)!\,(q-1)!}\pi^{p-1}(1-\pi)^{q-1} \quad 0 \le \pi \le 1$$

En este caso se considera a X como resultado de n observaciones bernoulli, entonces

$$f(x|\pi) = \begin{cases} \binom{n}{x}\pi^x(1-\pi)^{n-x} & x = 0,1,2\cdots n \\ 0 & en\ otro\ caso \end{cases}$$

De la ecuación 3 y 4 la función de distribución final de **TT** resulta ser

$$f_n(x|\pi)g(\pi) = \frac{(p+q-1)!}{(p-1)!\,(q-1)!}\pi^{p-1}(1-\pi)^{q-1}\binom{n}{x}\pi^x(1-\pi)^{n-x}$$

$$= \binom{n}{x}\frac{(p+q-1)!}{(p-1)!\,(q-1)!}\pi^{x+p-1}(1-\pi)^{n-x+q-1}$$

Resultando también, excepto un factor constantes, la distribución semejante a un modelo beta de parametros p = y + 1 y q = n - y + 2

$$g(\pi|x) = \frac{(n+p+q-1)!}{(x+p-1)!\,(n-x+q-1)!}\pi^{x-p-1_i}(1-\pi)^{q-1}(1-\pi)^{n-x+q-1_i}$$

Y el estimador de Bayes en este caso tiene la forma $\hat{\pi} = \frac{x + p}{n + p + q}$

5. Resultados

Primera encuesta. La empresa que ha cedido los resultados para su análisis aplico la encuesta que figura en el anexo realizada en 120 hogares y selección aleatoria de un adulto por hogar.

Pregunta: "¿Usted esta de acuerdo con el referéndum?"
Cuadro 1

Personas en la muestra según aceptación del referéndum por ciudad del eje central				
Opinión	El Alto	La Paz	Cbba	Sta. Cruz
De acuerdo	64	78	76	80
Total desacuerdo	26	17	24	11
No responde	30	25	20	19

Como se observa, la proporción de personas que no responden están por encima del 10% de la muestra y los estimadores de bayes para los modelos uno y dos son ligeramente menores a los de máxima verosimilitud. Para la tercera posibilidad, asumiendo para p = número de no respuesta y q = 2 generan los siguientes estimadores.

Estimadores de la proporción de aceptación del referéndum según ciudad por método de cálculo				
	Máximo verosímil	Modelo uniforme	Modelo triangular	Modelo binomial
El Alto	53,3	53,2	52,8	61,8
La Paz	65,0	64,7	64,2	70,0
Cbba	63,0	63,1	62,6	67,6
Sta. Cruz	67,0	66,4	65,8	72,2

Notar que los estimadores máximo verosímiles son ligeramente superiores al 50% y menores que 70% contradiciendo el optimismo de personeros del Gobierno que afirmaron era del 90 al 95%.
Ahora considerando solamente las encuestas de los entrevistados inscritos en el padrón electoral la situación cambia

Muestra según ciudad por inscripción			Estimadores de proporción	
	Inscritos en el padrón	Personas que apoyan	Máximo verosímil	Modelo binomial
El Alto	109	60	55,0	63,8
La Paz	118	76	64,4	69,7
Cbba	114	78	65,8	69,8
Sta. Cruz	104	78	75,0	79,25

El grado de aceptación del referéndum es alto así como de las preguntas, una causa posible para este hecho, es la forma en que fueron planteadas. El cuadro siguiente muestra la equivalencia de las mismas

Preguntas del referéndum	1	2	3	4	5
Preguntas de la boleta Nº 1	6	8	7	9	11 a,b

Se ha cruzado la pregunta 3 con las preguntas 6,7,8,9,11, y 12 de la boleta. En algunos casos tambien no hay consistencia entre 6, 7, 8, 9, con la 12, “está de acuerdo con la política del presidente?" El NO se contradice con total acuerdo de las preguntas 6 y 7 y el NO de la 9.

Ciudadanos en la muestra, inscritos en el padrón que eligen Si en las preguntas equivalentes y pregunta 12 de la encuesta según ciudad							
Ciudad	Padrón	Preg1	Preg2	Preg3	Preg4	Preg5	Preg12
El Alto	11	95	98	102	74	83	78
La Paz	118	109	112	109	102	100	98
Cbba	114	104	101	105	83	89	82
Sta. Cruz	104	92	95	99	87	87	83

Estimadores de la proporción del SI según ciudad por número de pregunta.						
Ciudad	Preg1	Preg2	Preg3	Preg4	Preg5	Preg12
El Alto	87,4	89,5	92,3	72,1	79,0	75,5
La Paz	92,4	94,5	92,4	87,6	86,2	86,0
Cbba	91,17	88,9	91,9	75,7	80,1	75,0
Sta. Cruz	90,2	95,2	95,5	86,5	86,5	83,5

Segunda encuesta. 27 de junio. En la boleta se registra: 1. Zona de residencia. 2. Género. 3. Si esta inscrito en el padrón. Luego se pide llene fotocopia de la boleta a ser aplicada el 18-07. Los resultados son muy diferentes de la primera encuesta. n=120

Ciu	Pd	Preg 1			Preg 2			Preg 3			Preg 4			Preg 5		
		SI	NO	NR	**SI**	NO	NR	**SI**	NO	NR	**SI**	NO	NR	**SI**	NO	NR
E A	116	**67**	35	14	**72**	30	14	**68**	38	10	**30**	61	21	**29**	62	21
L P	118	**82**	24	12	**85**	20	13	**82**	19	17	**57**	46	15	**57**	44	17
Cbb	119	**84**	23	12	**83**	20	16	**80**	37	12	**59**	49	11	**62**	43	14
S C	115	**84**	32	13	**83**	16	16	**75**	24	16	**51**	46	18	**67**	33	15

En el siguiente cuadro se calcula y compara las predicciones de los resultados del referéndum considerando el giro que ha tomado, después de la declaración sindical de boicot al referéndum.

En el modelo binomial p=nº de respuesta de la encuesta anterior (ejl. El Alto, preg. 4: p=30). Entre los tres métodos, el de modelo binomial es el que utiliza información previa para la aproximación, teniendo en cuenta que en la muestra muchos de los encuestados no llenaron ninguna de las casillas (ni Si, ni NO)

Estimadores de la proporción del Si según ciudad por número de pregunta y método de cálculo										
Ciudad	Preg 1		Preg 2		Preg 3		Preg 4		Preg 5	
	Max. Ver	Binom.	Max. Ver	Binom.	Max. Ver	Binom.	Max. Ver	Binom.	Max. Ver	Binom.
E A	57,7	61,3	62,0	65,1	58,6	60,9	25,8	36,7	25,0	35,9
L P	69,5	71,2	72,3	73,7	69,5	72,2	48,3	53,3	48,3	54,0
Cbb	70,5	72,2	69,7	72,3	67,2	69,2	49,5	53,0	52,1	56,3
S C	60,8	63,8	72,2	74,4	65,2	68,4	44,3	51,1	58,2	62,12

Corte Nacional Electoral. Resultados oficiales de las proporciones del Si según Municipio por número de pregunta.					
Municipio	Pregunta 1	Pregunta 2	Pregunta 3	Pregunta 4	Pregunta 5
El Alto	62,16	66,72	61,38	31,53	33,13
La Paz	72,98	76,72	70,64	52,32	56,36
Cochabamba	69,41	74,82	70,94	51,12	56,06
Santa Cruz	65,19	73,26	65,03	47,16	59,73

Comparación con los resultados oficiales. De los resultados de la segunda encuesta, los estimadores máximo verosímiles están muy por debajo de los que proporciona la CNE, mientras que los estimadores de bayes son los más próximos, por defecto en las preguntas 1,2 y 3 y por exceso en la 4 y 5.

6. Conclusiones

1. El uso del método de Bayes para calcular estimadores de porporción, es una alternativa en situaciones de existir una proporción alta de No Respuesta, porque de estos grupos, a la hora de emitir su voto deben su voto deben tomar una decisión, que influye en los resultados finales.
2. Un estudio más detallado de las componentes para usar el modelo binomial como distribución inicial del parámetro π a ser estimado, implica decidir acerca de p y q.
3. Los estimadores de Bayes convergen a los de máxima verosimilitud según va creciendo el tamaño de la muestra pues estos últimos son muy sensibles al tamaño muestral.

Estudios de Marketing
CONSULTORES
¿Que hacer con el Gas?
Ciudadano, La información que usted proporcione a Marketing SRL es confidencial y solo para fines estadísticos. Agradecemos su colaboración por la sinceridad en las respuestas

1	Zona de residencia	
2	Género	1. Hombre
		2. Mujer
3	¿Esta usted inscrito en el padrón electoral?	1. Si
		2. NO
4	¿Esta usted informado sobre el referéndum?	1. Si continua con 5
		2. NO pasar a 6
5	¿Usted esta de acuerdo con el referéndum?	1. Total desacuerdo
		2. Total acuerdo
6	¿Esta de acuerdo en derogar la ley que ha permitido la llegada de empresas transnacionales para explotar el petróleo y gas en el país?	1. Total desacuerdo
		2. Mas o menos de acuerdo
		3. Total acuerdo
7	¿Esta de acuerdo que YPFB nuevamente sea la empresa que explote el petróleo y gas en el país?	1. Total desacuerdo
		2. Mas o menos de acuerdo
		3. Total acuerdo
8	¿Si Ud. fuese presidente ¿intentaría recuperar para el País la propiedad de los hidrocarburos?	1. Si
		2. NO
9	¿Esta Ud. de acuerdo con utilizar el gas para que Bolivia pueda negociar con Chile, una salida al mar?	1. Si Pasa a 11
		2. NO Pasa a 10
10	No esta de acuerdo por alguna de las sig. razones? a) Sadríamos perdiendo, el beneficiario es Chile b) Chile esta obligado a devolvernos territorio c) Chile exigirá que el gas salga por sus puertos	
11	Si Usted fuese Gobierno ¿qué acción tomaría respecto al gas? a. Abastecer las necesidades de los bolivianos y sí sobra vender al exterior b. Fomentar la industrialización en el país y exportar el resto c. Aumentar los impuestos para tener más recursos para el TGN y luego exportar	
12	¿Esta usted de acuerdo con la politica de Carlos Mesa con respecto de que hacer con el gas?	1. Si
		2. NO

Gracias por su colaboración
Fecha:_____

Índice