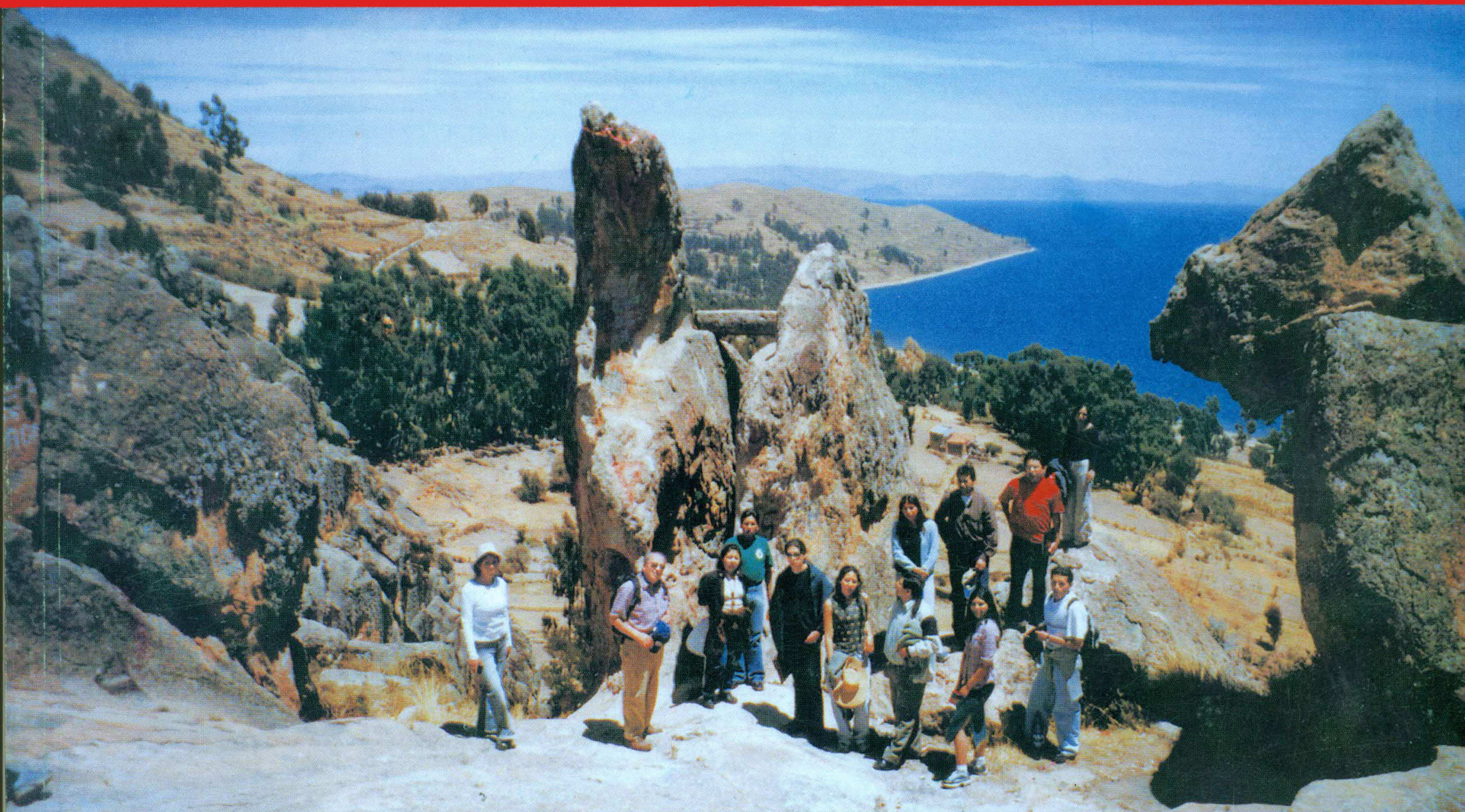
LA REVISTA DE LA CARRERA DE ESTADISTICA
VARIANZA

Universidad Mayor de San Andrés
Facultad de Ciencias Puras y Naturales



AÑO 4

Nº 4

Octubre 2005

REVISTA VARIANZA

Número 4, Año 04
Octubre d e 2005

DIRECCION:

José Aníbal Angulo A.
Director IETA

Colaboradores

Rubén Belmonte
David Barrera
Dindo Valdez
Aníbal Angulo
Nicolás Chávez
Jaime Pinto
Luis Zapata

IMPRESION

Casa Blanca
Calle Colón # 888
esq. Sucre
Telf.: 2282I67

PRESENTACIÓN

La presente revista, la cuarta de su serie se constituye en un referente de la actividad estadística de la carrera, la misma que desde la publicación de la primera revista ha concitado el interés tanto de docentes y estudiantes de la carrera así como de personas no vinculadas a la Universidad, de esta manera la presente revista constituye un aporte tanto a la comunidad estadística como a la sociedad en su conjunto.

A la espera que este trabajo colme las expectativas de los lectores de las anteriores tres publicaciones, mostrando temas de interés y principalmente actuales, también se espera observaciones o críticas de los lectores ávidos por mayores ampliaciones a los artículos aquí tratados.

A tiempo de felicitar al cuerpo de profesionales que hacen posible esta entrega, congratulo también al lector paciente y crítico

Lic. Raúl Delgado Alvarez
JEFE CARRERA DE ESTADÍSTICA

CARRERA ESTADISTICA
INSTITUTO DE ESTADISTICA TEÓRICA Y APLICADA
FACULTAD DE CIENCIAS PURAS Y NATURALES
UNIVERSIDAD MAYOR DE SAN ANDRES

La Paz - Bolivia
Edificio Antiguo - Planta Baja pb05
Telefax: 2442100

INDICE

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 muestrearf (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

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

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

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

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 π

El valor π ,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 = 1 para el Si, X, = 0 para el NO entonces X_1, X_2, —X_n constituyen n pruebas Bernoulli con parámetro π 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 0(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 P(p, q) con p = y + 1, q = n-y+1 en consecuencia la función de probabilidad final de π 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.

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

De la ecuación 3 y 4 la función de distribución final de π 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| π): 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 **π** 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 $\hat{\pi} = \dfrac{x+p}{n+p+q}$

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	70Preg 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 ora 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 negoc iar 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 su 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

METODOS DE SUPERVIVENCIA PARA ESTUDIAR LA DESERCIÓN

Rubén Belmonte Coloma

Introducción

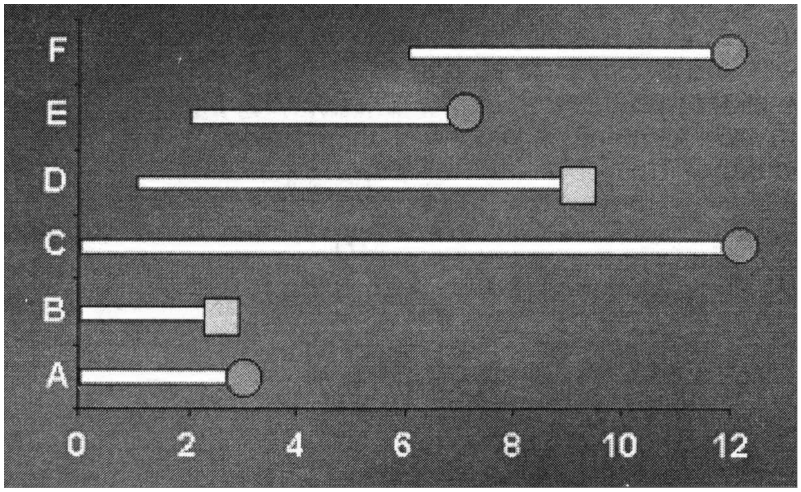
El análisis de supervivencia es una técnica estadística que tiene innumerables aplicaciones en el área de la salud, en el área actuarial, en la planificación industrial, para citar algunas. En este artículo se pretende establecer algunas bases para invitar a estudiar un fenómeno académico cuyo origen puede ser de preocupación y este es el de la deserción académica en la universidad.

Para nadie es desconocido el hecho de que la deserción de universitarios en el sistema público es un fenómeno de múltiples causas, económicas, sociales, principalmente, sin embargo este problema no es objeto de este artículo, se dan las bases del tratamiento técnico del tiempo de permanencia.

Se denomina análisis de supervivencia al conjunto de técnicas que permiten estudiar la variable "tiempo hasta que ocurre un evento" y su dependencia de otras posibles variables explicatorias. En este ejemplo, en el estudio de la deserción, el tiempo hasta que ocurre la deserción (tiempo de supervivencia) sin estudiar, por ahora, su dependencia a los factores extemos. Debido a que la variable tiempo es una variable continua podría ser, en principio, estudiada mediante las técnicas de análisis de la varianza o los modelos de regresión. Hay, sin embargo, dos dificultades importantes para este planteamiento. En primer lugar, en la mayor parte de los estudios la variable tiempo no tiene una distribución normal, más bien suele tener una distribución asimétrica y aunque podrían intentarse transformaciones que la normalizaran, existe una segunda dificultad que justifica un planteamiento específico para estas variables, y es que para observarlas se tiene que prolongar el estudio durante un período de tiempo suficientemente largo, en el cual suelen ocurrir pérdidas, que imposibilitan la observación del evento.

Existen tres motivos por los que pueden aparecer estas pérdidas, en primer lugar por fin del estudio. Supóngase, por ejemplo, que para evaluar se sigue en el tiempo, durante un año, a dos grupos de alumnos. A los de un grupo se les aplica algún método que incentive la permanencia en la universidad y a los de otro no, y se registró la duración del intervalo de tiempo entre la intervención pedagógica (o la entrada en el estudio, para el grupo no intervenido) y la deserción. Al final del estudio puede haber individuos que no hayan desertaron. Algunos de los individuos, y puede ser un número importante, cambian de carrera o simplemente están matriculados pero no asisten desaparecerán del estudio en algún momento del mismo por diversos motivos. En estudios de supervivencia en el área de la salud Aunque los ejemplos anteriores son del ámbito académico los más significativos están en el área de salud, estos mismos problemas aparecen en cualquier estudio que necesite un largo tiempo de observación.

Hay que tener en cuenta también que la variable es el tiempo hasta que ocurre un evento, y está definida por la duración del intervalo temporal entre los instantes en que empieza la observación y ocurre el evento. En los ejemplos citados, la observación no comienza en el mismo instante para todos los individuos. En algunos textos se denomina pérdida por la izquierda a esta no coincidencia de los tiempos en que comienza la observación, ya que, si el estudio está diseñado para acabar en un tiempo determinado, el efecto de esta no coincidencia es reducir, para los que empiezan más tarde, el tiempo de observación. En el esquema de la figura se detallan todas las posibles pérdidas. Evidentemente, se pueden evitar las pérdidas por la izquierda diseñando el estudio para que acabe, no en un tiempo establecido con carácter general, sino, para cada individuo, en un tiempo determinado después del inicio de la observación.



Esquema temporal de un estudio para observar tiempos de espera para un evento, por ejemplo supervivencia en una intervención quirúrgica. Con el círculo se representan las pérdidas y con el cuadrado las deserciones (ocurrencia del evento). El individuo A desaparece del estudio 3 años después del ingreso (sería una pérdida en sentido estricto). El B deserta a los 2,5 años (cinco semestres. El C sigue en la universidad hasta el final del estudio (sería una pérdida a los 12 semestres por fin del estudio). El D, al que ingresa después de un semestre, deserta en el 9, el tiempo de supervivencia sería 8 semestres (hay 1 semestre de pérdida por la izquierda). El E, al que se le ingresa en el segundo semestre, se pierde en el 7 (sería una pérdida a los 5 meses, ya que hay pérdida en sentido estricto y pérdida por la izquierda). El F, al que ingresa en el sexto semestre, sigue en la universidad al acabar el estudio, sería una pérdida a los 6 meses (existe pérdida por fin del estudio y pérdida por la izquierda).
Si se quisiera aplicar un modelo de regresión lineal a un estudio de este tipo, habría que eliminar del mismo las observaciones perdidas, ya que para ellas no se conoce el valor de la variable; sin embargo sí se tiene alguna información útil sobre la misma: se sabe que es mayor que el tiempo en el que se produjo la pérdida.

Distribución de la variable tiempo de espera

La variable tiempo de espera es una variable aleatoria continua y no negativa, cuya función de probabilidad puede especificarse de varias maneras. La primera es la habitual función densidad de probabilidad f(t), y relacionadas con ella, la función de supervivencia S(t) y la función de riesgo h(t).

La función densidad de probabilidad f(t) para una variable continua se define como una función que permite calcular la probabilidad de que la variable tome valores en un intervalo a través de la fórmula:

$$P(a < T < b) = \int_a^b f(t)dt \qquad 0 < t < \infty$$

La función de supervivencia S(t) se define como:

$$S(t) = P(T \geq t) = \int_t^{h(t)} f(U)dU$$

Por lo tanto, la función de supervivencia da la probabilidad complementaria de la habitual función de distribución acumulativa F(t) = P(T £ t), es decir S(t) = 1 - F(t).

Otro modo de expresar la probabilidad para la variable tiempo de espera es por medio de la función de riesgo h(t) que es la función de densidad de probabilidad de T, condicionada a que T ^{3} t. Por ejemplo, para la supervivencia al ingreso, la función de riesgo a los 2 años es la de densidad de probabilidad de desertar a los 2 años del ingreso. Esta probabilidad sería, realmente, la que en cada momento le importa al universitario.
Se puede demostrar que

$$h(t) = \frac{f(t)}{S(t)}$$

A veces se usa también la función de riesgo acumulada H(t), más difícil de interpretar, que se define como

$$H(t) = \int_{\square}^{t} h(x)dx$$

y que verifica

$$H(t) = -\ln\big(S(t)\big)$$

Es decir, las cuatro funciones están relacionadas; si se conoce una cualquiera de ellas, se pueden obtener las demás.
A pesar de que el tiempo es una variable continua, un observador sólo tiene acceso a valores discretos de la misma. Los datos observados para cualquiera de las experiencias descritas en la introducción son una serie de valores discretos. Conviene, por lo tanto, definir las funciones anteriores en el caso (práctico) de considerar a la variable tiempo como discreta, es decir, como un conjunto discreto de valores $t_1 < t_2 < \ldots$ El suponerlos ordenados de menor a mayor no representa ninguna pérdida de generalidad, de hecho es así como se observa el tiempo.

Para una variable discreta, la función densidad de probabilidad f (t) se define como:

$$f(t_i) = P(T = t_i), i = 1, 2, \ldots$$

y la función de supervivencia:

$$S(t_l) = \sum_{l/t_l \geq t_i} f(t_i)$$

La función de supervivencia da, por lo tanto, para cada valor t_i de T, la probabilidad de que la variable T sea mayor o igual que t_i (en este caso no es la complementaria de la función de distribución puesto que la probabilidad de que T sea igual a t_i, que en las variables discretas en general no es cero, está incluida en ambas funciones), aunque otros textos, justamente para que siga siendo la complementaria de la función de distribución la definen sin incluir el igual.

Las funciones de riesgo y riesgo acumulado para una variable discreta también son:

$$h(t_l) = \frac{f(t_l)}{S(t_l)} \qquad H(t_i) = -\operatorname{Ln} S(t_l)$$

El método Kaplan Meier

El método de Kaplan Meier es un método no paramétrico (no asume ninguna función de probabilidad) y por máxima verosimilitud, es decir se basa en maximizar la función de verosimilitud de la muestra. Una muestra aleatoria de tamaño n, extraída de una población, estará formada por k (k £ n) tiempos $t_1 < t_2 < \ldots < t_k$ en los que se observan eventos. En cada tiempo t_i existen n_i “individuos en riesgo” (elementos de la muestra para los que el evento puede ocurrir, o que T ³ t_i) y se observan d_i eventos. Además en el intervalo $[t_i,t_{i+1})$ se producen m_i pérdidas.

Se puede demostrar que la función de verosimilitud para toda la muestra es:

$$L = \prod_{i=1}^{k} {h_i}^{d_i} (1 - h_i)^{n_i - d_i}$$

Para construir esta función se ha asumido que la información contenida en las pérdidas es que, para cada una de ellas, el evento ocurre en un tiempo mayor que el tiempo en que se observa la pérdida. Maximizando esta función se encuentra que el estimador de la función de riesgo es

$$\hat{h}_i = \frac{d_i}{n_i} \; i = 1, 2, \ldots, k$$

y para la función de supervivencia, el denominado estimador producto límite o de Kaplan-Meier:

$$\hat{S}(t_i) = \prod_{i|t_i<t_i} \left(1 - \frac{d_i}{n_i}\right)$$

Se sigue en el tiempo a 12 individuos que ingresaron a la universidad y se encuentran los siguientes tiempos de supervivencia en años: 6*, 6, 6, 6, 10, 12*, 12, 15, 15*, 17, 22, 22, donde el asterisco indica deserción; es decir se perdieron 3 individuos en los tiempos 6, 12 y 15. La manera más cómoda de calcular los estimadores anteriores es disponer los datos en una tabla como la que sigue:

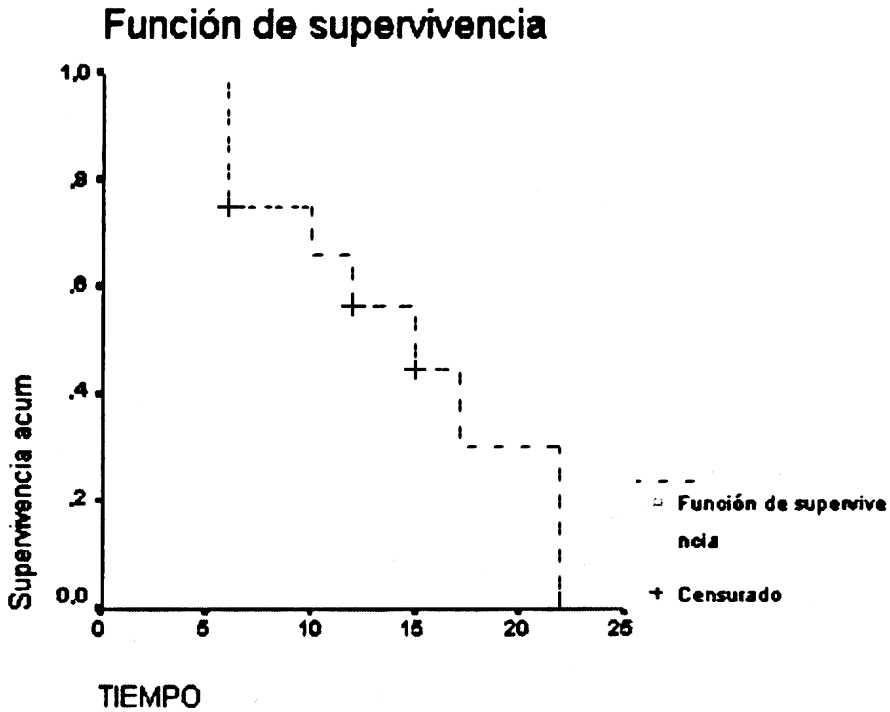
tiempo	ind. en riesgo	eventos	F. riesgo	F. supervivencia
6	12	3	3/12=0,25	1
10	8	1	1/8=0,125	0,750
12	7	1	1/7=0,143	0,656
15	5	1	1/5=0,2	0,562
17	3	1	1/3=0,333	0,450
22	2	2	2/2=1	0,300

Para analizar estos datos con un paquete estadístico, por ejemplo el SPSS, hay que introducir dos variables: el tiempo y el "status" con un código que indique si en ese tiempo se ha producido el evento o es una pérdida. La "salida" es:

Análisis de supervivencia por Tiempo

Tiempo	Estado	Supervivencia acumulada	Error Estándar	Sucesos acumulados	Renumeración
6	1			1	11
6	1			2	10
6	1	,7500	,1250	3	9
6	0			3	8
10	1	,6563	,1402	4	7
12	1	,5625	,1482	5	6
12	0			5	5
15	1	,4500	,1555	6	4
15	0			6	3
17	1	,3000	,1605	7	2
22	1			8	1
22	1	,0000	,0000	9	0

Numero de casos 12 Censurados: 3 (25,00%) Sucesos : 9
En la tercera columna (“Supervivencia acumulada”) aparece la función de supervivencia (S(t) en todos los tiempos en los que ocurren eventos. Esta función se suele representar en un¡ gráfica como



El SPSS también calcula y representa la gráfica de la función de riesgo acumulada (que en su versión en español denomina “Impacto”).
E.T.Lee, Statistical Methods for Survival Data Analysis Lifetime, Learning Publications. 1980. D.R. Cox D Oakes, Analysis of Survival Data, Chapman and Hall. 1984

Ida Tarbell, The Ways of Woman

Regocíjate, pues debajo de nubes y estrellas
nuestro planeta es más que maine o Texas.
Bendice los grandiosos hechos de tener
doce meses, nueve musas y dos sexos,
y en los señoríos de la tierra, infinidad de artes, climas, maravillas de ideas.

Phyllis McGinley, “In Praise of Diversity”

No obstante, la estadística no debe hacerse para demostrar una idea preconcebida
Florence Nihtingale, nota en Physique Sociales, de A. Quetelet

Ahora se había dado cuenta que no sabía nada fundamental y, como un monje agobiado con la conciencia del pecado, se lamentó: “¡Si, ¡y si tan sólo pudiera recordar la estadística!”

Casi todos los estados han devuelto sus censos. Le envío los resultados, con tinta negra, si están basados (en la medida de lo posible) en los datos reales, y con tinta roja si no son los datos regresados, aunque bastante conocidos. Con un pequeño margen para omisiones, son más de cuatro millones, aunque de hecho sabemos que las omisiones han sido muy grandes.
Thomas Jefferson, carta de David Humpreys

Personalmente, nunca me han interesado la ficción ni los cuentos. Lo que me gusta leer son hechos y estadísticas de cualquier tipo; aunque sean hechos acerca de cultivo de rábanos, me interesan. Precisamente ahora, por ejemplo, antes de que usted entrara señala una enciclopedia que se encuentra en un librero estaba leyendo un artículo sobre matemáticas perfectamente puras.

“Mi conocimiento matemático termina en “12 por 12”, pero disfruté inmensamente con ese artículo. No entendí una palabra, pero los hechos, o lo que el hombre cree que son los hechos, parecen siempre encantadores. Este matemático creía en sus hechos, también yo. Primero obtenga sus propios hechos y aquí la voz disminuye hasta casi imperceptible luego puede distorsionarlos tanto como desee”.

Mark Twain, citado por Rudyard Kipling, en from Sea to Sea
No hay remedio más efectivo para aplacar la rabiosa susceptibilidad de la gente que las cifras. Para estar segura, debe prepararse cuidadosamente, abarcar el caso y no sólo su cuarta parte, y debe reunirse por sí misma, no para comprobar alguna teoría. Este tipo de preparación puede encontrarla en el caso nacional.

Índice

PRUEBAS DE SIMULACION DE MONTECARLO DE LAS PRUEBAS DE RAIZ UNITARIA PARA UN PROCESO AUTOREGRESIVO AR(1)

David Barrera Ojeda

Proceso autoregresivo de orden “1”

Forma de dependencia más simple, consiste en relacionar X $_t$ con $X_{(t-i)}$ linealmente es decir:

$$X_t = \alpha + \beta * X_{t-1} + \varepsilon_t$$

donde Ω_t es ruido blanco con varianza Ω^2.

Ejemplo: X_t es el consumo del mes en alimentos. Cada mes se gasta una proporción fija de las existencias iniciales $(1\text{-}\Omega)*X_{t-1}$, y hay un consumo estable por termino medio igual a Ω pero que varía aleatoriamente de unos meses a otros según Ω_t .

algunas propiedades:

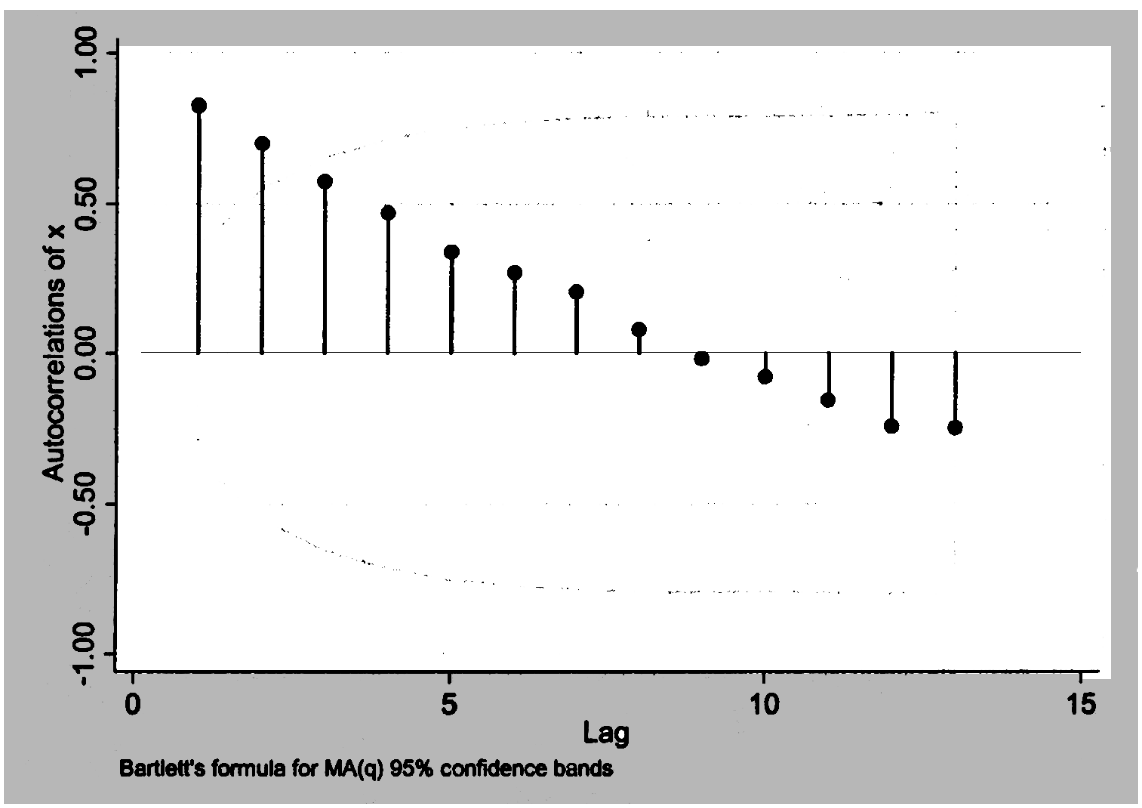
$$E(X_t) = \mu = \frac{\alpha}{1-\beta}$$

$$Var(X_t) = \sigma^2 = \frac{\sigma_\varepsilon^2}{1-\beta}, \text{ siempre que } |\beta| < 1$$

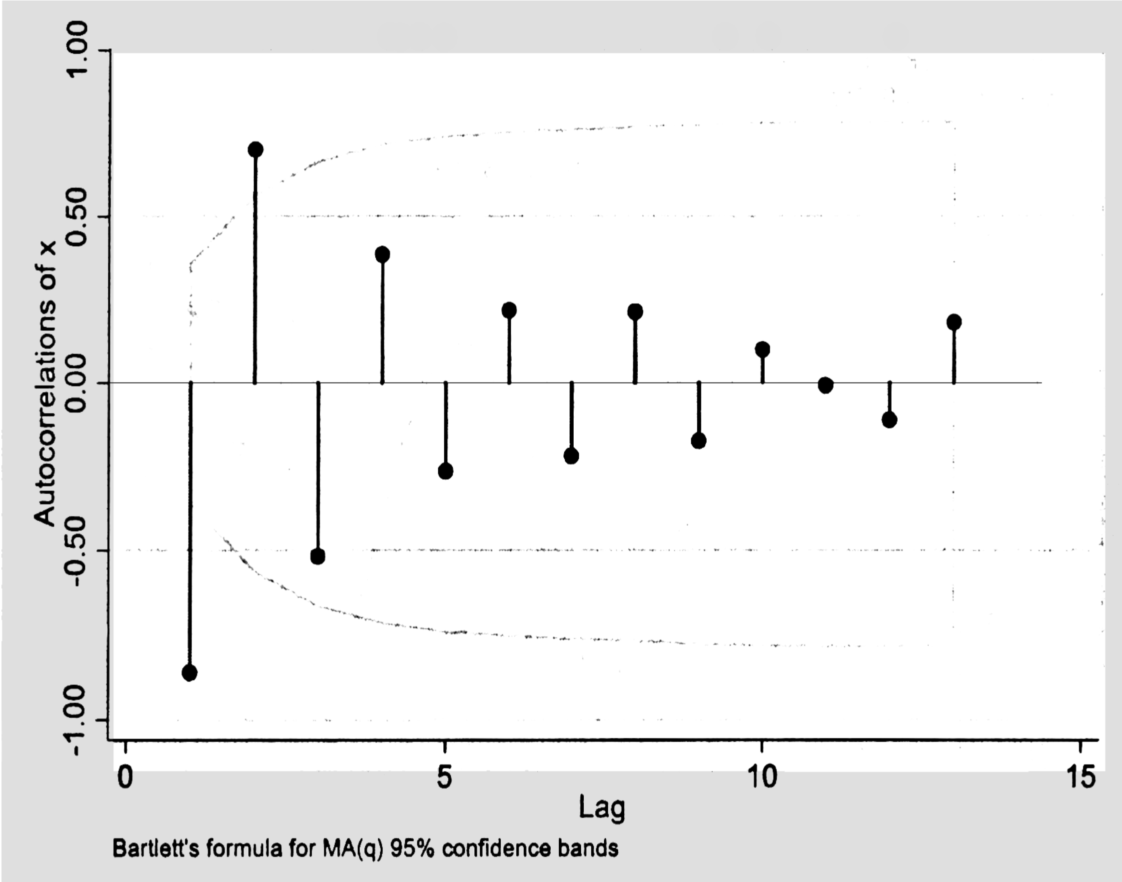
$$Cov(X_t, X_{t-k}) = \frac{\beta^k}{1-\beta} * \sigma_\varepsilon^2$$

Ejemplos: Para algunos valores Ω .

Cuando Ω=0.9



Cuando Ω=-0.9



En econometría es probable encontrarse con series no estacionarias, y es lo más común asumir que el proceso es AR(1).
En la práctica es clave primero identificar el proceso, es decir si se trata de efectivamente de un proceso AR(1), o AR(2), ARMA(P,Q), ARIMA(P,Q), para esto lo primero que se suele realizar es graficar el correlograma, luego hallar las funciones de autocorrelación, luego la autocorrelación parcial. Identificar el valor de P y el valor Q constituye todo un arte, identificar el proceso generador(Xt) de la mejor forma posible, si cuenta con o sin constante, o cuenta con tendencia y constante, luego estimar los parámetros por máxima verosimilitud o mínimos cuadrados ordinarios.

Una vez identificado el proceso se debe realizar las pruebas de hipótesis del modelo: Normalidad, media de los residuos igual a cero, varianza de los residuos igual a una constante, incorrelados los residuos.

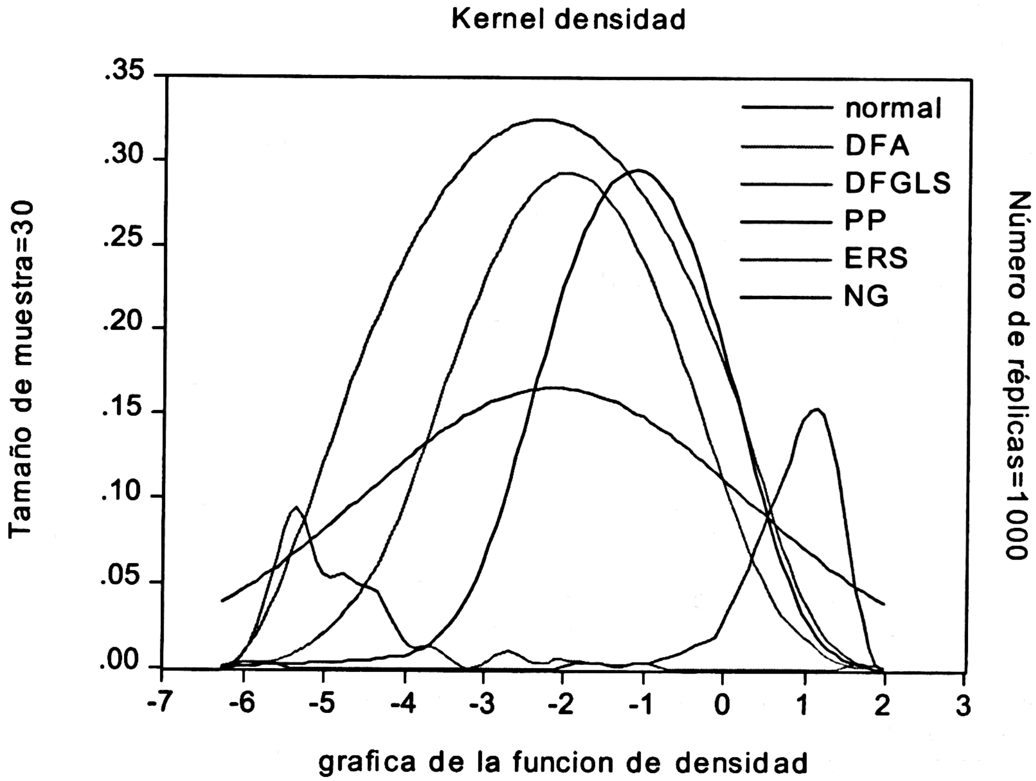
Una vez de estimar el proceso generador, digamos AR(1), se debe probar las pruebas de raíz unitaria, esto clave para realizar estimar los coeficientes de estructurales de un modelo de regresión o para estimar el modelo para fines predictivos.

La pruebas de simulación de Monte Cario, se usa el método de Matsumoto y Nishimura's para generar los números aleatorios. Con un número de réplicas de 1000, para dos tamaños de muestra y asumiendo que el modelo tiene el término constante.

Los test de raíz unitaria más conocidos son:
Augmented Dickey-Fuller, GLS Dickey-Fuller (Elliot, Rothenberg, y Stock), Phillips-Perron, Elliot-Rothenberg-Stock, Ng y Perron .

A.- El Modelo es $\mathbf{X_t = 3+0.9*X_{(t+1)}+\Omega_t}$

Para n=30



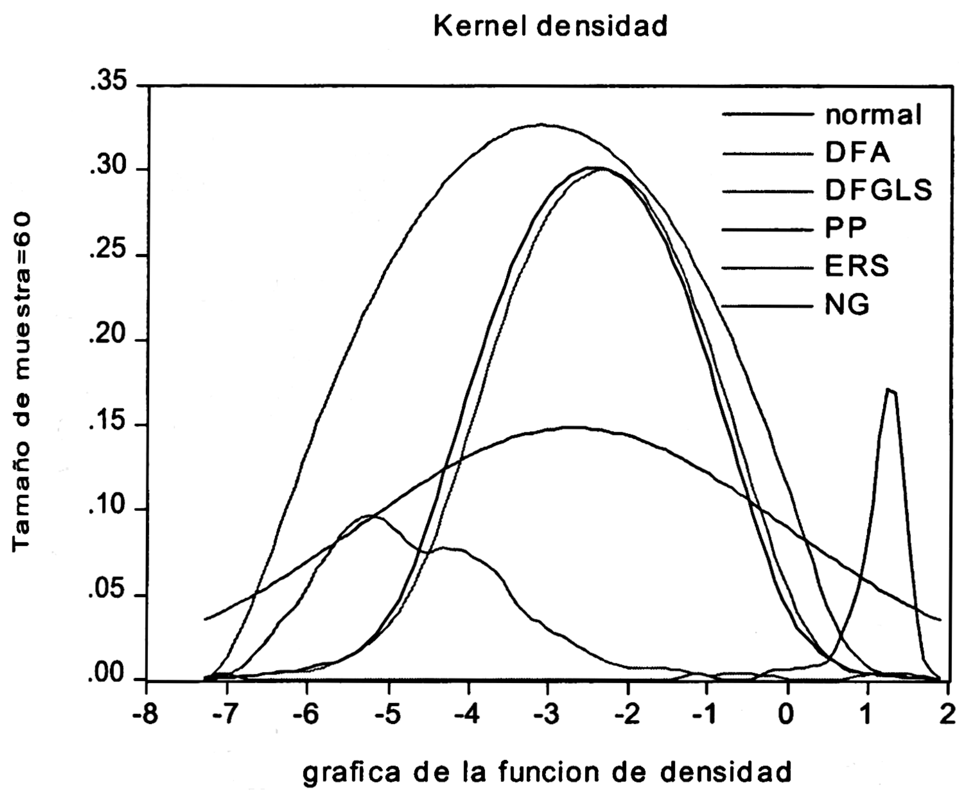
	ADF	DFGLS	PP	ERS	NG
Mean	-1.990	0.348	0.352	10.918	-4.861
Median	-1.937	0.311	0.299	8.802	-3.881
Maximum	-0.334	0.005	0.000	0.211	-41.032
Std.Dev.	0.759	0.250	0.241	9.463	4.750
Skewness	-0.254	0.512	0.468	2.634	-4.816
Kurtosis	2.757	2.208	2.225	12.542	35.359
CV	-0.381	0.719	0.684	0.867	-0.977
RI	0.951	0.376	0.369	8.012	3.770
Rango	3.631	0.903	0.899	63.406	40.428
RI. Rel	-0.845	0.989	1	0.993	-0.971
	DF	DGLS	PP	ERS	NG
Prob. Rech Ho.	0.110	0.000	0.090	0.000	0.010

Normalidad	DFA	DGLS	PP	ERS	NG
Estadísticos	Prob.	Prob.	Prob.	Prob.	Prob.
Geary	0.466	0.035	0.050	0	0
Shapiro F	0.727	0.045	0.013	0	0
Shapiro W	0.568	0.001	0.003	0	0

Conclusión:

Cuando el coeficiente de la pendiente(n) es aproximadamente 0.9, los test de raíz unitaria: DGLS, PP, ERS, NG, no son los más indicados, en tanto el test DFA es el más potente, puesto queasí lo demuestra el Kernel, y la estabilidad del estadístico está garantizado. Nótese que las pruebas de DGLS, ERS y NG indican que la probabilidad de rechazar la Ho(el proceso tiene raíz unitaria) es casi imposible, esto es inaceptable.

Para n=60



	ADF	DFGLS	PP	ERS	NG
Mean	-2.371	0.218	0.222	8.263	-6.099
Median	-2.280	0.182	0.189	7.545	-4.231
Maximum	-0.419	0.899	0.911	30.191	-0.318
Minimum	-4.977	0.000	0.000	0.152	-83.488
Std.Dev.	0.699	0.180	0.186	4.871	8.671
Skewness	-0.750	1.153	1.376	1.360	-7.394
Kurtosis	4.978	4.469	4.933	6.426	65.174
CV	-0.295	0.828	0.835	0.590	-1.422
RI	0.821	0.246	0.204	5.898	3.209
Rango	4.558	0.899	0.911	30.039	83.171
RI. Rel	-0.845	1.000	1.000	0.990	-0.992
	DF	DGLS	PP	ERS	NG
Prob. Rech Ho.	0.180	0.000	0.150	0.000	0.010

Normalidad	DFA	DGLS	PP	ERS	NG
Estadísticos	Prob.	Prob.	Prob.	Prob.	Prob.
Geary	0.000	0.000	0.000	0.000	0.000
Shapiro F	0.025	0.805	0.037	0.072	0.000
Shapiro W	0.001	0.000	0.000	0.000	0.000

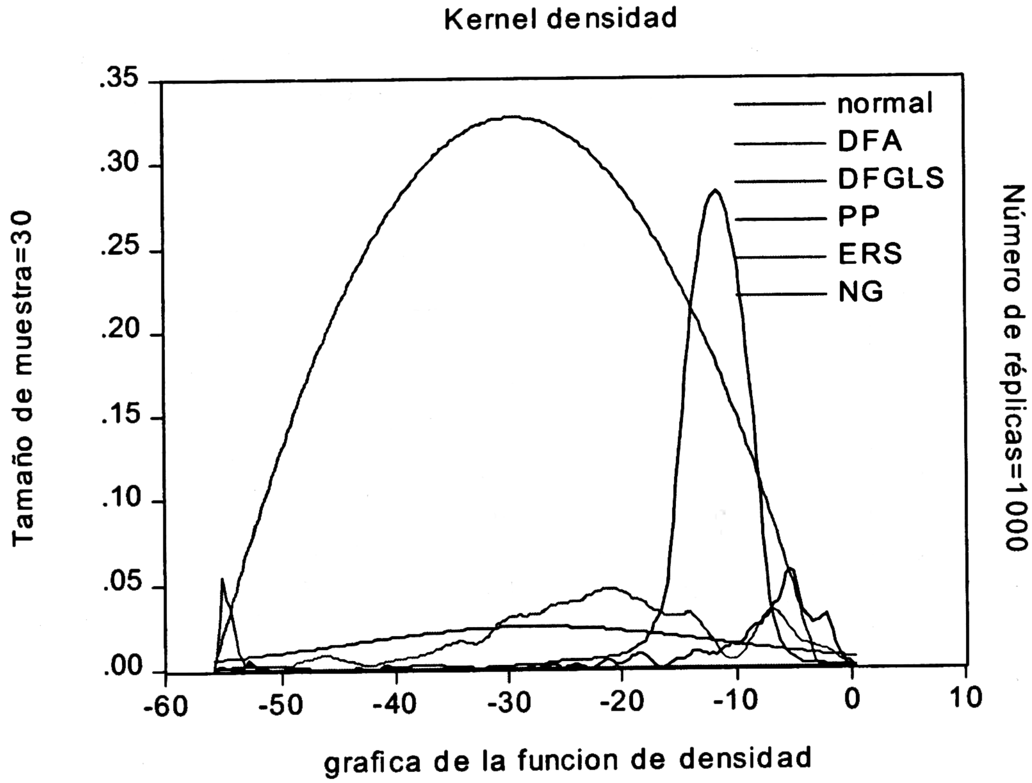
Conclusión:

Cuando aumenta el número de observaciones, el test de DFA sigue siendo el mejor, pese a la inestabilidad del estadístico, lo que había que rescatar es el test de Phillips Perron, está mejor que cuando n=30.

B.- El Modelo es $\mathbf{X_t = 3 - 0.9 * X_{(t-1)} + \Omega_t}$

Para n=30

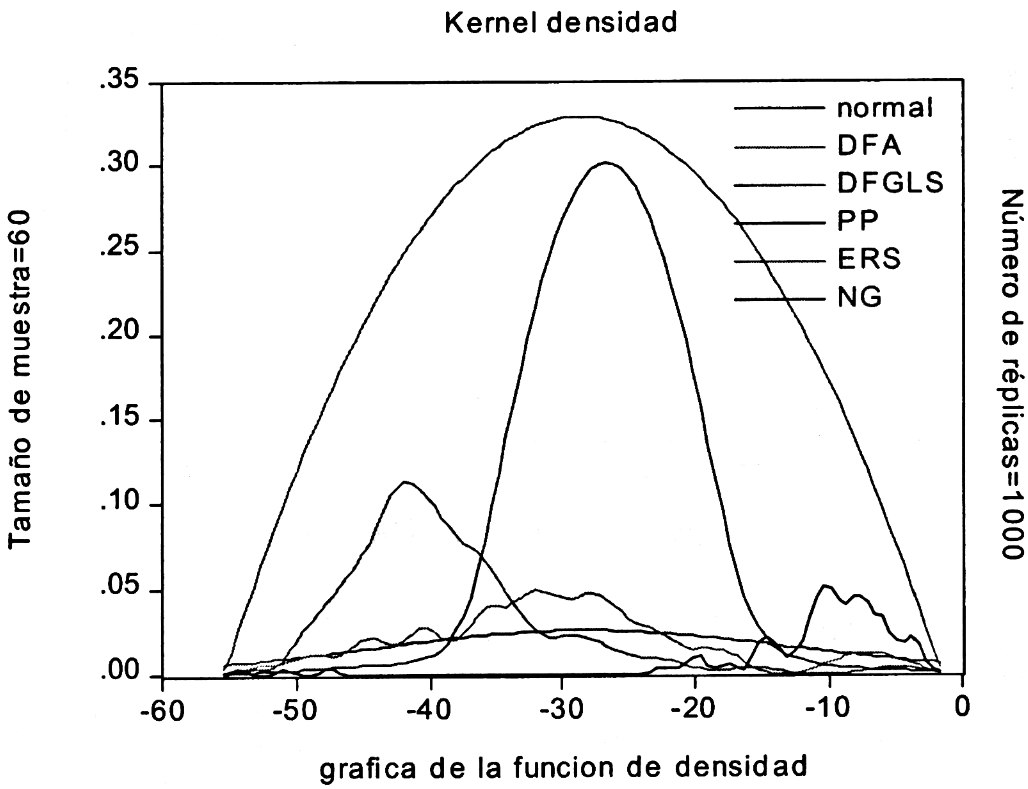
Nota.- En economía es raro que $\Omega<0$



	ADF	DFGLS	PP	ERS	NG
Mean	-21.829	0.004	0.000	26.939	-0.423
Median	-21.472	0.000	0.000	16.088	-0.047
Maximum	-1.974	0.296	0.000	781.126	3.894
Minimum	-53.567	0.000	0.000	1.219	-20.994
Std.Dev.	10.577	0.026	0.000	71.403	2.142
Skewness	-0.368	8.790	-2.339	8.344	-6.124
Kurtosis	3.114	87.231	6.470	77.717	51.918
CV	-0.485	6.981	0.370	2.651	-5.069
RI	13.230	0.000	0.000	12.203	0.573
Rango	51.593	0.296	0.000	779.906	24.888
RI. Rel	-0.929	1.000	1.000	0.997	-1.455
	DF	DGLS	PP	ERS	NG
Prob. Rech Ho.	0.980	0.000	1.000	0.000	0.005

Normalidad	DFA	DGLS	PP	ERS	NG
Estadísticos	Prob.	Prob.	Prob.	Prob.	Prob.
Geary	0.087	0.000	0.000	0.000	0.000
Shapiro F	0.216	0.000	0.000	0.000	0.000
Shapiro W	0.009	0.000	0.000	0.000	0.000

Conclusión:
La inestabilidad es marcada de todos los estadísticos, a pesar de este defecto en esta ocasión las pruebas de DFA y PP no son los más indicados, puesto que es casi seguro que vamos a rechazar Ho, y esto no es correcto.
Para n=60



	ADF	DFGLS	PP	ERS	NG
Mean	-30.906	0.000	0.000	9.377	-0.101
Median	-31.072	0.000	0.000	8.282	-0.073
Maximum	-3.956	0.003	0.000	32.488	3.292
Minimum	-53.129	0.000	0.000	0.318	-4.313
Std.Dev.	10.418	0.000	0.000	5.015	0.806
Skewness	0.396	9.698	-4.129	1.565	-0.840
Kurtosis	3.277	96.077	18.053	7.102	12.833
CV	-0.337	2.525	0.231	0.535	-7.959
RI	11.225	0.000	0.000	5.133	0.545
Rango	49.173	0.003	0.000	32.170	7.604
RI. Rel	-0.861	1.000	1.000	0.981	-7.448
	DF	DGLS	PP	ERS	NG
Prob. Rech Ho.	1	0	1	0	0.010

Normalidad	DFA	DGLS	PP	ERS	NG
Estadísticos	Prob.	Prob.	Prob.	Prob.	Prob.
Geary	0.169	0.000	0.000	0.000	0.000
Shapiro F	0.099	0.000	0.000	0.001	0.000
Shapiro W	0.072	0.000	0.000	0.000	0.000

Conclusión:

Definitivamente cuando $\mathbf{\Omega}$<0, la mayoría de las pruebas no son buenos, por las inestabilidad que muestra la mayoría, creo que el número de observaciones no es un factor que puede cambiar el curso del estadístico.

En definitiva, en este trabajo sólo se ha probado para un proceso AR(1), lo que tendríamos que probar que pasa para un proceso AR(p) para p>1, o un proceso ARMA, ARIMA, en definitiva la clave para aplicar correctamente es determinar el proceso, luego aplicar las pruebas de raíz unitaria. Si el proceso es estacionario sería el test de KPSS.

En el modelo con término constante y sin tendencia, el mejor es DFA, luego PP, y los otros, había que analizarlos en los modelos sin el término constante, y otro con constante y tendencia.

Bibliografía:
Enders, W. (1995). Applied Econometric Times Series.
John Wiley & Sons, Inc. United States.
Phillips, P.C.B. (1987), “Time Series Regression With a Unit Root”, Econometrica 55, 277-301.
White, H, 2001, Asymptotic Theory for Econometricians, Academic Press.
Software:
Macroeviews
Matlab

Índice

Los registros de nacimiento, mantenidos en orden para garantizar la condición de los | ciudadanos, pueden servir para determinar la población de un gran imperio sin recurrir a un censo de sus habitantes, operación laboriosa y difícil de realizar con exactitud. Pero para esto es necesario conocer la razón entre la población y los nacimientos anuales. El medio más preciso para esto consiste, primero, en elegir subdivisiones del imperio distribuidas de manera casi igual en toda la superficie, para obtener el resultado general independiente de las circunstancias locales; segundo, enumerar con cuidado a los habitantes de varias comunas en cada una de las subdivisiones, durante un tiempo determinado; tercero, determinar el número promedio de nacimientos anuales correspondiente al utilizar la cuantía de nacimientos durante varios años antes y después de este tiempo. Este número, dividido entre el número de habitantes, dará razón entre los nacimientos anuales y la población, de manera cada vez más confiable conforme aumenta la enumeración.

Pierre -Simón Laplace, Essai Philosophique sur les Probabilités

LA NAVAJA

Aníbal Angulo A.

El método de la **navaja** (Jackknife), usa el método de los grupos aleatorios permitiendo que las réplicas de los grupos se traslapen. La navaja fue introducido por Quenuilli (1949; 1956) como un método para reducir el sesgo; Tukey (1958) lo uso para estimar varianzas y calcular intervalos de confianza. El método navaja con una eliminación; Shao y Tu (1995) analizan otras formas de la navaja y dan los resultados teóricos.

Para una muestra aleatoria simple, sea $\hat{\theta}_{(j)}$ el estimador de la misma forma que $\hat{\theta}$, pero sin utilizar la observación j. Así, si $\hat{\theta} = \overline{y}$, entonces $\hat{\theta}_{(j)} = \overline{y}_{(j)} = \frac{\sum_{i \neq j} y_i}{n-1}$. Para una muestra aleatoria simple, definimos el estimador de navaja con una eliminación (llamado de esta forma pues eliminamos una observación en cada réplica) como

$$\hat{V}_{JK}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} \left(\hat{\theta}_{(j)} - \hat{\theta}\right)^2 \qquad (1)$$

¿Por qué el factor (n-1)/n, la estimación con reemplazo de la varianza de y .

Sea $\overline{y}_{(j)} = \dfrac{\sum_{i \neq j} y_i}{n-1} = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} y_i - y_j\right) = \dfrac{1}{n-1}\left(n\overline{y} - \overline{y} + \overline{y} - y_j\right) = \overline{y} - \dfrac{1}{n-1}\left(y_j - \overline{y}\right).$

Entonces $\displaystyle\sum_{j=1}^{n} \left(\overline{y}_{(j)} - \overline{y}\right)^2 = \frac{1}{(n-1)^2} \sum_{j=1}^{n} \left(y_j - \overline{y}\right)^2 = \frac{s_y^2}{n-1}.$

Veamos como se puede hacer uso de este método de la navaja para estimar la razón entre alumnos inscritos que no tienen comedor y aquellos que tienen en 10 facultades elegidas al azar, en este caso.

$$\hat{\theta} = \frac{\overline{y}}{\overline{x}}, \qquad \hat{\theta}_{(j)} = \hat{R}_{(j)} = \frac{\overline{y}_{(j)}}{\overline{x}_{(j)}},$$

y la varianza $\hat{V}_{JK}(\hat{R}) = \dfrac{n-1}{n} \sum \left(\hat{R}_{(j)} - \hat{R}\right)^2$

Para cada grupo de navaja, omitimos na observación. Así, $\overline{x}_{(1)}$ es el promedio de todas

las x, excepto x_1: $\overline{x}_{(1)} = \frac{\sum_{i=2}^{9} x_i}{9}$ (ver tabla).

En este caso $\hat{R} = 2{,}33012203,\quad \sum(\hat{R}_{(j)} - \hat{R})^2 = 0{,}09421389,\quad y\ \hat{V}_{JK}(\hat{R}) = 0{,}08479$

Tabla de replicaciones, que son las combinaciones tomadas con $m = 1$ y $g = 10$, de modo que se verifique $n = mg$.

Facultades	X	Y	$\overline{X}_{(j)}$	$\overline{Y}_{(j)}$	$\hat{R}_{(j)}$
1	136	374	157,888889	361,555556	2,28993666
2	167	498	154,444444	347,777778	2,25179856
3	150	150	156,333333	386,444444	2,47192608
4	108	216	161	379,111111	2,3547274
5	187	247	152,222222	375,666667	2,46788321
6	307	513	138,888889	346,111111	2,492
7	154	395	155,222222	359,222222	2,30434783
8	93	405	162,666667	358,111111	2,20150273
9	134	414	158,111111	357,111111	2,25860857
10	121	416	159,555556	356,888889	2,2367688

Extensión del método para el caso de Conglomerados.

Podría suponerse que bastará eliminar una unidad de observación a la vez, pero eso no servirá denada; pues se destruiría la estructura de conglomerado y daría una estimación de la varianza quesólo es correcta si la correlación entre las clases es igual a cero. En cualquier método de remuestreo y el él método de grupos aleatorios, conserve juntas las unidades de observación dentro de una unidad de observación dentro de la misma unidad primaria. Así, para una muestra por conglomerados, aplicaríamos el estimador de la varianza de navaja según (1) de modo que **n** sea la cantidad de unidades primarias y $\hat{\theta}_{(j)}$ la estimación de θ que se obtendría al eliminar todas las observaciones de la unidad primaria j.

En una muestra por conglomerados, estratificada y con varias etapas, la navaja se aplica por separado en cada estrato en la primera etapa de muestreo, eliminando una unidad primaria a lavez. Suponga que existen h estratos y que se eligen n_h unidades primarias para la muestra del estrato h, suponga que estas unidades primarias se eligen con reemplazo.

Para aplicar la navaja, eliminamos una unidad primaria a la vez. Sea $\hat{\theta}_{(hj)}$ el estimador de la mismaforma que $\hat{\theta}$ al omitir la unidad primaria j del estrato h.

Para calcular $\hat{\theta}_{(hj)}$, definimos una nueva variable de ponderación; sea

$$w_{i(hj)} = \begin{cases} w_i \\ 0 \\ \dfrac{n_h}{n_h - 1} w_i \end{cases}$$

toma el valor w_i, si la unidad de observación i no esta en el estrato h , el valor o , si la unidad de observación i está en la unidad primaria j del estrato h, y finalmente $\dfrac{n_h}{n_h - 1} w_i$ si la unidad de observación i está en el estrato h, pero no en la unidad primaria j.
Entonces usamos los pesos $w_{i(hj)}$ para calcular $\hat{\theta}_{(hj)}$:

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\hat{\theta}_{(hj)} - \hat{\theta})^2 \qquad (2)$$

Cuadro de trabajo para calcular la varianza

Unidad	M_i	y_1	y_2	$\overline{y}_i$	S_i^2	$M_i\overline{y}_i$	$M_i(M_i - m_i)\frac{S_i^2}{m_i}$	$M_i\hat{\overline{y}}_r$	$\left(M_i\overline{y}_i - M_i\hat{\overline{y}}_r\right)^2$
1	25	4	3	3,5	0,5	87,5	143,75	86,29032258	1,463319459
2	26	5	4	4,5	0,5	117	156	89,74193548	743,0020812
3	30	2	1	1,5	0,5	45	210	103,5483871	3427,913632
4	15	1	4	2,5	4,5	37,5	438,75	51,77419355	203,7526015
5	17	2	3	2,5	0,5	42,5	63,75	58,67741935	261,708897
6	20	3	5	4	2	80	360	69,03225806	120,2913632
7	27	4	2	3	2	81	675	93,19354839	148,6826223
8	32	5	6	5,5	0,5	176	240	110,4516129	4296,591051
9	10	2	4	3	2	30	80	34,51612903	20,39542144
10	15	1	6	3,5	12,5	52,5	1218,75	51,77419355	0,526795005
Sumas	217					749	3586		9224,327784
$\hat{\overline{y}}_r$	3.4516129					Varianza	35.86		

La varianza $s_r^2 = \dfrac{\sum\left(M_i\overline{y}_i - M_i\hat{\overline{y}}_r\right)^2}{n-1} = \dfrac{9224{,}327784}{9} = 1024.9$

Con esto la varianza $\hat{V}(\hat{\overline{y}}_r) = -\dfrac{1}{\overline{M}^2}\left[\left(1 - \dfrac{n}{N}\right)\dfrac{S_r^2}{n} + \dfrac{1}{nN}\sum_{i\epsilon S} M_i^2\left(1 - \dfrac{m_i}{M_i}\right)\dfrac{S_i^2}{m_i}\right]$

En el caso de la media $\overline{M} = 21.7$

Con estos valores la varianza $\hat{V}(\hat{\overline{y}}_r) = -\dfrac{1}{21.7^2}\left[\left(1 - \dfrac{10}{N}\right)\dfrac{1024.9253}{10} + \dfrac{35.86}{10N}\right]$

No conocemos N, el número de unidades en la población, aunque suponemos que es grande. Así, consideramos que la correción para las poblaciónes finitas al nivel primario es 1 y observamos que el segundo término de la varianza estimada será muy pequeño con respecto al primer término, en este sentido la varianza se reduce.

$\hat{V}(\hat{\bar{y}}_r) = -\frac{1}{21.7^2}\left[\frac{1024.9253}{10}\right]$, cuya raíz cuadrada es: $\sqrt{\hat{V}(\hat{\bar{y}}_r)} = 0.4665375$

La aplicación de la navaja para calcular la varianza de la media del numeró de hijos por hogar. Obtenemos $\hat{\theta} = \hat{y}_r = 749/217 = 3.4516$

En este ejemplo como no se conoce el tamaño de N en la población, calculamos la varianza con reemplazo.

En primer lugar, determinamos el vector de ponderaciones para cada una de las 10 iteraciones de la navaja. Sólo tenemos un estrato, de modo que $h = 1$ para todas las observaciones.

Para $\hat{\theta}_{(11)}$, eliminamos la primera unidad primaria. Así, los nuevos pesos para las observaciones en la primera unidad primaria son 0; los pesos en las restantes unidades primarias son los pesos anteriores multiplicados por $n_h/(n_h - 1) = 10/9$. Al usar los pesos de los datos anteriores construimos otra tabla que es la siguiente.

Tabla de trabajo para la construcción de los ponderadores

Unid	M_i	y_I	$M_i/2$	$\frac{M_i}{2}y_i$	$W(1,1)$	$W(1,2)$	...	....	$W(1,8)$	$W(1,9)$	$W(1,10)$
1	25	4	12,5	50	0	13,889	...	....	13,889	13,889	13,889
1	25	3	12,5	37,5	0	13,889			13,889	13,889	13,889
2	26	5	13	65	14,444	0			14,444	14,444	14,444
2	26	4	13	52	14,444	0			14,444	14,444	14,444
3	30	2	15	30	16,667	16,667			16,667	16,667	16,667
3	30	1	15	15	16,667	16,667			16,667	16,667	16,667
4	15	1	7,5	7,5	8,3333	8,333			8,333	8,333	8,333
4	15	4	7,5	30	8,3333	8,333			8,333	8,333	8,333
5	17	2	8,5	17	9,4444	9,444			9,444	9,444	9,444
5	17	3	8,5	25,5	9,4444	9,444			9,444	9,444	9,444
6	20	3	10	30	11,111	11,111			11,111	11,111	11,111
6	20	5	10	50	11,111	11,111			11,111	11,111	11,111
7	27	4	13,5	54	15	15			15	15	15
7	27	2	13,5	27	15	15			15	15	15
8	32	5	16	80	17,778	17,778			0	17,778	17,778
8	32	6	16	96	17,778	17,778			0	17,778	17,778
9	10	2	5	10	5,5556	5,556			5,556	0	5,556
9	10	4	5	20	5,5556	5,556			5,556	0	5,556
10	15	1	7,5	7,5	8,3333	8,333			8,333	8,333	0
10	15	6	7,5	45	8,3333	8,333			8,333	8,333	0
sumas	434			749	213,33	212,222			205,556	230	224,444

Observe que las sumas de pesos de navaja varían de una columna a otra, pues la muestra original no era autoponderada. Calculamos $\hat{\theta}$ como $\left(\sum w_i y_i\right)/\sum w_i$; para determinar $\hat{\theta}_{(11)}$, seguimos el mismo procedimiento pero utilizamos $w_{i(hj)}$ en vez de w_i.

Asi, se tiene las replicaciones finales, observare la última tabla para los resultados de los $\hat{\theta}_{(i,j)}$
Y los valores de $\hat{\theta}_{(1,1)} = 3.44528;\quad \hat{\theta}_{(1,2)} = 3.30886\,;\quad \hat{\theta}_{(1,10)} = 3.44799$

Resumen de los resultados de los estimadores $\hat{\theta}_{(i,j)}$

$\hat{\theta}_{(1,1)}$	$\hat{\theta}_{(1,2)}$	$\hat{\theta}_{(1,3)}$	$\hat{\theta}_{(1,4)}$	$\hat{\theta}_{(1,5)}$	$\hat{\theta}_{(1,6)}$	$\hat{\theta}_{(1,7)}$	$\hat{\theta}_{(1,8)}$	$\hat{\theta}_{(1,9)}$	$\hat{\theta}_{(1,10)}$	Var
3.44528	3.30886	3.76467	3.52223	3.53246	3.39890	3.515748	3.09743	3.47339	3.44799	
0.00004	.20378	0.09800	0.00498	0.006536	0.003104	0.004113	0.12544	0.00047	0.000013	0.263096

Cuya desviación estándar es la raíz cuadrada de $\sqrt{\hat{V}(\hat{\theta})} = 0.4866$

$$\hat{V}_{NAVA} = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} \left(\hat{\theta}_{(hj)} - \hat{\theta}\right)^2$$

Tabla de replicaciones eliminando un conglomerado cada vez

0	55,556	55,556	55,556	55,556	55,556	55,556	55,556	55,556	55,556
0	41,667	41,667	41,667	41,667	41,667	41,667	41,667	41,667	41,667
72,222	0	72,222	72,222	72,222	72,222	72,222	72,222	72,222	72,222
57,778	0	57,778	57,778	57,778	57,778	57,778	57,778	57,778	57,778
33,333	33,333	0	33,333	33,333	33,333	33,333	33,333	33,333	33,333
16,667	16,667	0	16,667	16,667	16,667	16,667	16,667	16,667	16,667
8,333	8,333	8,333	0	8,333	8,333	8,333	8,333	8,333	8,333
33,333	33,333	33,333	0	33,333	33,333	33,333	33,333	33,333	33,333
18,889	18,889	18,889	18,889	0	18,889	18,889	18,889	18,889	18,889
28,333	28,333	28,333	28,333	0	28,333	28,333	28,333	28,333	28,333
33,333	33,333	33,333	33,333	33,333	0	33,333	33,333	33,333	33,333
55,556	55,556	55,556	55,556	55,556	0	55,556	55,556	55,556	55,556
60	60	60	60	60	60	0	60	60	60
30	30	30	30	30	30	0	30	30	30
88,889	88,889	88,889	88,889	88,889	88,889	88,889	0	88,889	88,889
106,667	106,667	106,667	106,667	106,667	106,667	106,667	0	106,667	106,667
11,111	11,111	11,111	11,111	11,111	11,111	11,111	11,111	0	11,111
22,222	22,222	22,222	22,222	22,222	22,222	22,222	22,222	0	22,222
8,333	8,333	8,333	8,333	8,333	8,333	8,333	8,333	8,333	0
50	50	50	50	50	50	50	50	50	0
735	**702,222**	**782,222**	**790,556**	**785**	**743,333**	**742,222**	**636,667**	**798,889**	**773,889**

Bibliografía

Shao, J y d Tv 1995
The Jackkife and Boostrap
Macarthy, P,J 1969 Pseudo replicación Half Simples
Cochran W. G. Técnicas de Muestreo

Índice

VALIDACIÓN DE DATOS FALTANTES EN ENCUESTAS AGROPECUARIAS

Jaime Pinto A.

La generación de Bases de datos, para los análisis a realizarse, requiere considerar los muchos criterios que se plantean para depurar la base, que en algunos casos debido a diversos factores como una no adecuada supervisión o control de relevamiento de datos, critica y trascripción, hacen que haya datos faltantes en la base de datos.
Esta omisión se debe subsanar, mediante procedimiento de imputación que sugieren o indican cual pudiese ser el dato faltante.

La imputación generalmente se utiliza para asignar valores a los elementos faltantes. Frecuentemente se asigna un valor de reemplazo al valor faltante mediante un valor de otra persona en la encuesta, similar a la que no responde al elemento con respecto a otras variables. Al usar la imputación, hay que crear una variable adicional en el conjunto de datos que indique si la respuesta fue medida o imputada.

Algunos métodos de imputación son los siguientes:

IMPUTACIÓN DE LA MEDIA POR CELDA

Los productores que responden se dividen en clases (celdas) con base en variables conocidas como en los ajustes de clases de ponderación. Entonces, sustituimos el promedio de los valores de las unidades que responden y que están en la celda c, en cada valor faltante.
La imputación de la media por celda, supone que los elementos faltantes son faltantes completamente al azar dentro de las celdas.

Para este extractamos parte de la boleta y los datos obtenidos en una encuesta agropecuaria, realizada en una zona de producción.

Parte de la Boleta: ***ENCUESTA AGROPECUARIA***

IDENTIFICACIÓN.-

P1.- Uso: Agrícola (1) Pecuaria (2)

P2.- Unidad de Observación:Grande (1) No grande (2)

UNIDAD DE PRODUCCIÓN.-

P3.- Cual es la Superficie total de su Unidad de Producción Agropecuaria (UPA)?

SUPERFICIE		
Cantidad	Unidad	Uso de oficina

P4.- Cual es el Numero de Parcelas de su Unidad de Producción Agropecuaria?

P5.- Numero de personas que trabajaron en la Unidad de Producción Agropecuaria.?

Conjunto de datos para explicar el método:

Productor	P2	P3	P4	P5
	Unidad de Observación	**Sup. Total de su UPA (Has.)**	**Numero de Parcelas de la UPA**	**Nro. de personas que trabajaron en la UPA**
1	2	0.25	1	2
2	2	0.90	4	
3	2	0.25	3	
4	2	0.25	3	2
5	2	0.8	3	1
6	1	4.00	9	10
7	1	6.00	7	14
8	2	0.30	6	1
9	2	0.25	4	2
10	2	0.10	5	
11	2	0.85	7	1
12		0.25	2	3
13	2	0.90	2	1
14	1	5.00	12	10
15	1	4.00	14	16

Para esto construimos cuatro celdas usando las variables Número de Parcelas y Unidad de Observación (Grande y No Grande)

NUMERO DE PARCELAS

UNIDAD DE OBSERVACIÓN	Menor e igual a 5	Mayor a 5
Grande	Productores	Productores 7, 6, 14, 15
No Grande	Productores 1, 12, 13, **3**, 4, 5, **2**, 9,**10**,	Productores 8,11

Revisando la Base de Datos los Productores 2, 3 y 10 no tienen el valor para los "Numero de personas que trabajaron en la Unidad de producción Agropecuaria", recibirán el promedio de los restantes productores donde se encuentran, es decir de la columna "Menor e igual a 5 Parcelas" La media para cada celda después de la imputación es igual a la media de quienes respondieron.

Realizando operaciones:

Promedio = (Suma de datos de la variable, de los que respondieron) / Total de personas que respondieron.

Promedio = (2+3+1+2+1+2) 1 6 = 11/ 6 = **1,83**

Los valores a colocar en la variable "Numero de personas que trabajan en la Unidad de Producción Agropecuaria" para los Productores 2, 3 y 10 son en cada caso " **2 Personas**".

Imputación Deductiva.-

Algunos valores se pueden asignar en la edición de datos, mediante las relaciones lógicas entre las variables.

En la base o conjunto de datos la persona 12 no respondió si es "Productor Grande", pero como respondió que su Superficie Total de Unidad de Producción Agropecuaria es 0,25, la respuesta de Unidad de Observación debe colocarse 2 .(por corresponder a Productor No Grande).

Hago mención a otros métodos de Imputación, que también se utilizan como:
Imputación Hot-deck

En esta Imputación, al igual que en la imputación de la media por celda y los métodos de ajuste con ponderación, las unidades de la muestra se dividen en clases. El valor de una de las unidades de la clase y que responde se sustituye en cada respuesta faltante. Con frecuencia, los valores de un conjunto de elementos faltantes relacionados entre si se toman del mismo donante para preservar algunas de las relaciones multivariadas.
El nombre hot deck proviene de los días en que los programas de computadora y los conjuntos de datos se perforaban en tarjetas.

Imputación cold-deck.

En la imputación cold-deck, los valores se asignan a partir de una encuesta anterior o de otras informaciones, como datos históricos. Hay poca teoría para este método
Como en el caso de la imputación de hot-deck, la imputación de cold-deck no garantiza la eliminación del sesgo de selección.

Imputación por regresión.

La imputación por regresión, predice el valor faltante usando una regresión del elemento de interés sobre las variables observadas para todos los casos. Una variante es la imputación por regresión estocástica, donde el valor faltante se reemplaza mediante el valor predicho a partir del modelo de regresión, más un término de error generado aleatoriamente.

[Kennedy] leía una de cada 50 cartas de las treinta mil que llegaban semanalmente a la Casa Blanca, al igual que un resumen estadístico de todo el lote, aunque él sabía que con frecuencia era tan organizado y no representativo como las estacas de Pensylvania Avenue.

Theodore Sorensen , Kennedy

Índice

CONCEPTOS ACERCA DE LOS INDICADORES ESTADISTICOS

Dindo Valdez Blanco

Los indicadores estadísticos

Los indicadores son expresiones que se utilizan campo estadístico para referirse a una relación, esta puede ser:

- De una variable respecto a sí misma, en el tiempo
  Ejemplo: El número anual de egresados en la facultad
- Entre una constante y una variable
  Ejemplo: El número de facultades y la cantidad anual de egresados en cada facultad
- Entre dos variables
  Ejemplo: El número anual de egresados y de nuevos estudiantes en la facultad

En el primer caso se denominan indicadores simples y en el segundo y tercer caso, indicadores compuestos.
Un indicador debe ser: pertinente, relevante y factible de calcular con la información disponible. Los indicadores, pueden estar referidos a un instante en el tiempo, asemejándose a una fotografía; a un lapso de tiempo siendo semejantes entonces a una película (conjunto de fotografías); o bien pueden referirse a dos o más subconjuntos de un sistema completamente predefinido (Ej. Dos o más universidades).

Los indicadores en la estadística estatal

En el ámbito estatal, los indicadores oficiales deberían ser comparables con indicadores de otros países y entre las regiones del mismo país, para que nos permita conocer de manera conjunta y comparativa la situación de cada estado o región. Esto con el único fin de mejorar la toma de decisiones y la elaboración de políticas regionales y mundiales.
A fin de que el conjunto de indicadores sea comparable, la información necesaria para su cálculo no sólo debe estar disponible, sino que también, debe ser confiable. Para que un indicador sea comparable (estándar) los países deben utilizar una metodología de cálculo común.
Una de las aplicaciones de los indicadores, que resultan de gran utilidad, es el poder medir el comportamiento de una o más acciones producto de la ejecución de un plan, programa o procedimiento, a través de un número relativo o absoluto.

Ejemplos:

Muchos estudios han determinado la correlación inversa existente entre niveles de pobreza y educación.

En la salud, se observa que las personas con mayores niveles de educación previenen mejor las enfermedades, planifican mejor su fertilidad y hacen un mayor uso de atenciones médicas preventivas.

Tipos de indicadores

En este sentido, pueden construirse tres tipos de indicadores:

- **Estratégicos:**
  Para evaluar y controlar los programas estratégicos establecidos por la alta dirección que están relacionados con la ejecución de los objetivos de la empresa o institución.
- **Programáticos:**
  Para evaluar los resultados obtenidos en la aplicación e implementación de los programas establecidos por la empresa o institución.
- **Operativos:**
  Para medir los resultados, producto de la ejecución de los programas implementados.

Aspectos para la construcción de los indicadores

Para la construcción de indicadores, en esta clase de aplicación, es conveniente tomar en consideración los siguientes aspectos:

- Tener siempre como marco de referencia los objetivos correspondientes a cada nivel: Estratégico, Programático y Operativo.
- A partir de los objetivos, identificar las variables, que se consideran en principio como las-que explican en mayor proporción el comportamiento del sistema bajo análisis. Por variable-se debe de entender a un hecho, suceso o resultado susceptible de cuantificarse, producto de una acción.
- Establecer el tipo de unidad de medida a utilizar por cada variable, así como el o los instrumentos para la recolección, tabulación y procesamiento de la información que se genere al respecto.
- Determinar la periodicidad con que se requerirá la información.
- Identificar las fuentes de la información o determinar la necesidad de llevar a cabo un muestreo, para obtener los datos que alimentarán a los indicadores.
- Establecer la forma de desagregación de los datos para que éstos puedan ser utilizados hacia arriba y hacia abajo de la estructura jerárquica de la empresa.

La construcción de estos indicadores comienza con la búsqueda de relaciones que expliquen en forma simple o compuesta el comportamiento del sistema bajo análisis, de tal manera que a partir del conjunto de relaciones se puedan apreciar las tendencias a utilizar para detectar desviaciones que impliquen la toma de decisiones para la consolidación, corrección y adecuación de los planes y programas orientados al cumplimiento de los objetivos de la empresa o institución de que se trate.

Luego de su construcción, es necesario poner a prueba los indicadores propuestos, para determinar la bondad de estos en la explicación del comportamiento del sistema bajo análisis.

Sistema de Información Estadística

De esta manera los indicadores identificados y probados, pasarán a constituir la base de un Sistema de Información Estadístico (llamado también Cuadro de Mando), que es parte fundamental del control para la gestión del gobierno y las instituciones o empresas públicas y privadas

Por tal razón la construcción y operacionalización de un Sistema de Información Estadístico apoyado por medios electrónicos resulta hoy en día sumamente necesario para que en todos y cada uno de los niveles de trabajo se tenga esta herramienta de apoyo para la toma de decisiones óptimas con un mayor grado de certidumbre.

En la actualidad, la construcción de un Sistema de Indicadores Estadísticos se constituye en un instrumento necesario para llevar a cabo de la mejor manera posible los objetivos de las empresas e instituciones.

Bibliografía

Cálculo de indicadores sociales, Ing. Luis Alcántara, (página web del ine España).

Naturaleza de los Indicadores de Salud, OPS/OMS, Bolivia - SNIS Ministerio de salud y Deportes, Bolivia.

Cálculo de los Indicadores Socio - Demográficos, INE, Bolivia

Índice

EL PROCESO GENERAL DE NACIMIENTO Y MUERTE

NICOLAS CHAVEZ QUISBERT

Definición

Sea $\{ N(t) \quad , \quad t \geq 0 \}$ un proceso puntual discreto en el espacio y continuo en el tiempo, tal que cumple los siguientes axiomas:

1.- $P[N(t+\Delta t) - N(t) = 1 / N(t) = n] = \lambda_n \Delta t + 0(\Delta t)$
2.- $P[N(t+\Delta t) - N(t) = -1 / N(t) = n] = \mu_n \Delta t + 0(\Delta t)$
3.- $P[N(t+\Delta t) - N(t) = 0 / N(t) = n] = 1-(\lambda_n + \mu_n) \Delta t + 0(\Delta t)$
4.- $P[N(t+\Delta t) - N(t) = m / N(t) = n] = 0(\Delta t)$ para $m \neq 0,1,-1$
5. - $P[N(0) = n_0] = \delta\, n, n_0$

donde:

$$\delta\, n, n_0 \begin{cases} 1 & \text{sí} \quad n = 0 \\ 0 & \text{sí} \quad n \neq n_0 \end{cases}$$

Donde:
λ_n : Tasa de nacimiento o de llegadas
μ_n: Tasa de muerte o de servicios
Δt : Incremento en el tiempo, infinitésimal.
$0(\Delta t)$: Función del tiempo.
$\delta_{n.\, n0}$: Función Delta de Kroneker, donde "n_0" representa la población en un instante 0.

ECUACIONES DE ESTADO DE CHAPMAN-KOLMOGOROV

$$P_n(t+\Delta t) = P_n(t)\, P[N(t+\Delta t) - N(t) = 0 / N(t) = n] + P_{n-1}(t)\, P[N(t+\Delta t) - N(t) = 1 / N(t) = n-1] + P_{n+1}(t)\, P[N(t+\Delta t) - N(t) = -1 / N(t) = n+1] + 0(\Delta t)$$

$$= P_n(t)\, [1-(\lambda_n + \mu_n) \Delta t + 0(\Delta t)] + P_{n-1}[\lambda_{n-1}\Delta t + 0(\Delta t)] + P_{n+1}(t)\, [\mu_{n+1} \Delta t + 0(\Delta t)] + 0(\Delta t)$$

$$= P_n(t) - P_n(t)\, (\lambda_n + \mu_n) \Delta t + P_{n-1}(t)\, \lambda_{n-1} \Delta t + P_{n+1}(t)\, \mu_{n+1} \Delta t + 0(\Delta t)$$

Aplicando ecuaciones diferenciales a ambos termino

$$\lim_{\Delta t \to 0} \frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = -P_n(t)(\lambda_n + \mu_n) + P_{n-1}(t)\lambda_{n-1} + P_{n+1}(t)\mu_{n+1} + 0\Delta t$$

$$P_n`(t) = -P_n(t)(\lambda_n + \mu_n) + P_{n-1}(t)\lambda_{n-1} + P_{n+1}(t)\mu_{n+1}$$

Para n =1,2,3.......donde $P_n`(t)$ es una ecuación diferencial

$P_0(t+\Delta t) = P_0(t)\, P[\, N(t+\Delta t) - N(t) = 0 \,/\, N(t) = 0] + P_1(t)\, P[\, N(t+\Delta t) - N(t) = -1 \,/\, N(t) = 1] + 0(\Delta t)$

$= P_0(t)\, [1 - (\lambda_0 + \mu_0)\, \Delta t + 0(\Delta t)\,] + P_1(t)[\mu_1\, \Delta t + 0(\Delta t)\,] + 0(\Delta t)$

$= P_0(t) - P_0\, \lambda_0\, \Delta t + P_1(t)\, \mu_1\, (\Delta t) + 0(\Delta t)$

μ_0 es cero ya que no hay población y por lo tanto no hay quien muera.

Aplicando ecuaciones diferenciales a ambos miembros se tiene:

$$P_0`(t) = -P_0(t)\lambda_0 + P_1(t)\mu_1 \qquad \text{para} \quad n = 0$$

$$\lim_{\Delta t \to 0} \frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -P_0(t)\lambda_0 + P_1(t)\mu_1 + 0$$

a) Proceso de nacimiento y muerte lineal

Se dan con las siguientes tasas:

$\lambda_n = n\lambda$

con λ y μ constantes

$\mu_n = n\mu$

b) Proceso de inmigración y muerte lineal

Ocurre con las siguientes tasas:

$\lambda_n = \lambda$

con λ constante

$\mu_n = n\mu$

c) Proceso de nacimiento lineal y de emigración

Ocurre con la siguientes tasas:

$\lambda_n = n\lambda$

con μ constante

$\mu_n = \mu$

d) Proceso de inmigración y emigración

Ocurre con la siguiente tasa:

$\lambda_n = \lambda$

$\mu_n = \mu$

DISTRIBUCIÓN LÍMITE DE PROBABILIDADES

Si se supone la existencia de distribución límite, tenemos:

$$\lim_{t \to \infty} P_n(t) = P_n$$

Por lo que Pn para n = 0,1,2....., formara una distribución probabilística discreta tal que cumple la siguiente propiedades:

1) $P_n \geq 0$

2) $\sum_{n=0}^{\infty} P_n = 1$

De cumplir lo anterior de que

$$\lim_{t \to \infty} p_n^{'}(t) = 0$$

Por que el límite de un escalar es cero

ECUACIONES DE EQUILIBRIO

Aplicando límites al infinito en forma asintótica a las ecuaciones de estado del proceso general de nacimiento y muerte se obtienen las ecuaciones de equilibrio.

Las ecuaciones de equilibrio son:

$P'_n(t) = -P_n(t)(\mu_n + \lambda_n) + P_{n-1}(t)\lambda_{n-1} + P_{n+1}(t)\mu_{n+1}$

$P'_0(t) = -P_0(t)\lambda_0 + P_1(t)\mu_1$

$0 = -P_n(\mu_n + \lambda_n) + P_{n-1}(t)\lambda_{n-1} + P_{n+1}(t)\mu_{n+1}$

$0 = -P_0\lambda_0 + P_1\mu_1$

Como:

$$\sum_{n=o}^{\infty} P_n = 1$$

Por ser una función probabilística, la solución del sistema es el siguiente:

Para n=0 $\qquad 0 = -P_0\lambda_0 + P_1\mu_1$

$\mu_1 P_1 = P_0\lambda_0 \qquad P_1 = \dfrac{\lambda_0}{\mu_1} P_0$

Para n=1 $\qquad 0 = -(\lambda_1 + \mu_1) P_1 + \lambda_0 P_0 + \mu_2 P_2$

$$0 = -\lambda_1 P_1 - \mu_1 P_1 + \lambda_0 P_0 + \mu_2 P_2$$
$$= -\lambda_1 P_1 - \lambda_0 P_0 + \lambda_0 P_0 + \mu_2 P_2$$
$$= \mu_2 P_2 = \lambda_1 P_1$$

$$P_n = \left(\frac{\lambda_0 \lambda_1 \ldots\ldots\ldots\ldots\ldots\ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots\ldots\ldots\ldots\ldots\ldots \mu_n} \right) P_0$$

$$P_2 = \frac{\lambda_1}{\mu_2} P_1 \qquad P_2 = \frac{\lambda_1}{\mu_2} \left(\frac{\lambda_0}{\mu_1} P_0 \right) = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} P_0$$

Por ser p_n una función probabilística se tiene que :

$\sum_{n=o}^{\infty} p_n = 1$ aplicando sumatoria al infinito.

$$P_0 = \frac{1}{\sum_{n=0}^{\infty} \left(\frac{\lambda_0 \lambda_1 \ldots\ldots\ldots\ldots\ldots\ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots\ldots\ldots\ldots\ldots\ldots \mu_n} \right)}$$

Se obtiene la probabilidad de que el sistema este vacío, para que exista dicha probabilidad, la sumatoria del denominador debe converger.

Para que esto ocurra S debe ser finito.

$$S = \sum_{n=0}^{\infty} \left(\frac{\lambda_0 \lambda_1 \ldots\ldots\ldots\ldots\ldots\ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots\ldots\ldots\ldots\ldots\ldots \mu_n} \right) \langle \infty$$

Para el proceso de inmigración y emigración con tasa $\lambda_n = \lambda$ y $\mu_n = \mu$ dicha sumatoria se reduce a:

$$S = \sum_{n=0}^{\infty} \left(\frac{\lambda \; \lambda \; \ldots\ldots\ldots\ldots\ldots\ldots \lambda}{\mu \; \mu \; \ldots\ldots\ldots\ldots\ldots\ldots \mu} \right) = \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu} \right)^n$$

$$\text{que} \quad \text{converja} \quad a = \begin{cases} \dfrac{1}{1 - \lambda/\mu} & \left|\dfrac{\lambda}{\mu}\right| \langle 1 \\ \infty & \left|\dfrac{\lambda}{\mu}\right| \geq 1 \end{cases}$$

Por lo tanto se garantiza la existencia de distribución límite de probabilidades, siempre que la tasa de nacimiento o llegada "λ" sea menor a la tasa de muerte o servicio "μ" es decir $\lambda<\mu$

BIBLIOGRAFIA

KARLIN SAMUEL, TAYLOR HOWARD M. (1988) , A FIRST COURSE IN STOCHASTIC PROCESSES, ACADEMIC PRESS, LONDON.

NAYARAN BHAT U. (1984) ELEMENTS OF APPLIED STOCHASTICN PROCESSES, JOHN WILEY & SONS CANADA.

PARZEN EMANUEL (1972) , PROCESOS ESTOCASTICOS, PARANINFO, MADRID.

PRIESTLEY M. B. (1981), SPECTRAL ANALISIS AND TIMES SERIES, ACADEMIC PRESS, LONDON.

Índice

No tengo muchos recuerdos de ese día, excepto la rapidez de Jonson: al destacar el doctor Beattie que había visto a los carruajes 1 y 1000 (el primero y el último), como si fuera algo notable, Jonson replicó: “¿y eso que?.

Hay la misma posibilidad de ver esos que cualesquier otros dos”. Evidentemente tenía razón, aunque ver los dos extremos (cada uno de los cuales es, en cierto grado, más llamativo que los demás) pueden llamar más la atención que ver cualquier otro par.

James Boswell, The Life of Samuel Johonson

Una de las cosas que ella (mamá) me enseño debería ser evidente para todos, aunque muchos cocineros aún no se dan cuenta de ello. Prepare primero la comida que tarde más en cocinarse.

Pearl Bailey, La cocina de Pearl

MODELOS ASIMETRICOS EN SERIES DE TIEMPO

Juan Carlos Flores

Los modelos ARCH, GARCH, EGARCH están diseñados para modelar y predecir la varianza de una variable dependiente. En cada caso la varianza de la variable dependiente esta especificada para depender de valores pasados de la variable dependiente usando formulas de variables exógenas o independientes.

1. Modelo ARCH

Consideremos el modelo AR un proceso auto regresivo de orden p de una variable Y_t esta dado por

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \cdots \dots \dots \dots \dots \dots \dots \phi_p Y_{t-p} + \mu_t \qquad (1.1)$$

μ_t, es un ruido blanco, es decir $E(\mu_t) = 0$ y

$$E(\mu_t, \mu_T) = \begin{cases} \sigma^2;\ \forall t = T \\ 0\ en\ otro\ caso \end{cases} \qquad (1.2)$$

este proceso tiene covarianza estacionaria, tal que las raíces de

$$1 - \phi_1 z + \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

están fuera del circulo unitario.

Aunque (1.2) la varianza no condicional de μ_t es la constante σ^2, la varianza condicional de μ_t puede cambiar con el tiempo. Un acercamiento es describir el cuadrado de μ_t como un proceso AR(m):

$$u_t^2 = \zeta + \alpha_1 \mu_{t-1}^2 + \alpha_2 \mu_{t-2}^2 + \cdots + \alpha_m \mu_{t-m}^2 + \omega_t \qquad (1.3)$$

donde w_t es un nuevo ruido blanco $E(w_t) = 0$

$$E(\omega_t, \omega_T) = \begin{cases} \lambda^2;\ \forall t = T \\ 0\ en\ otro\ caso \end{cases}$$

donde u_t es el error de la predicción Y_t la expresión (1.3) implica que las proyecciones lineales de los errores al cuadrado de la estimación de Y_t en el anterior p-esimo error cuadrado estimado, esta dado por

$$E\left(\mu_t^2 \ / \ \mu_{t-1,}^2 \mu_{t-2}^2, \dots\right) = \zeta + \alpha_1 \mu_{t-1}^2 + \alpha_2 \mu_{t-2}^2 + \cdots + \alpha_m \mu_{t-m}^2 \qquad (1.4)$$

y un proceso ruido blanco μ_t que satisface (1.3) es descrito como un proceso autoregresivo condicional heterocedastico de orden m, denotado por $\mu_t \approx ARCH(\mathrm{m})$ estudiado y premio Nobel por Engle.

Es a menudo conveniente usar una representación alternativa para un proceso ARCH(m) proceso que impone las suposiciones ligeramente más fuertes sobre la dependencia de serie μ_t. Puesto que

$$\mu_t = \sqrt{h_t}.\nu_t \tag{1.5}$$

donde $\{\nu_t\}$ son independientes e idénticamente distribuidas, la sucesión con media cero y varianza unitaria. $E(\nu_t) = 0 \quad E(\nu_t^2) = 1$
si h_t evoluciona según

$$h_t = \zeta + \alpha_1 \mu_{t-1}^2 + \alpha_2 \mu_{t-2}^2 + \cdots + \alpha_m \mu_{t-m}^2 \tag{1.6}$$

entonces, implica que

$$E(\mu_t^2 \ / \ \mu_{t-1}, \mu_{t-2}, \dots) = \zeta + \alpha_1 \mu_{t-1}^2 + \alpha_2 \mu_{t-2}^2 + \cdots + \alpha_m \mu_{t-m}^2 \tag{1.7}$$

Como μ_t esta generado por (1.5) y (1.6), entonces μ_t sigue un modelo ARCH(m) en la proyección lineal (1.4) que es también una esperanza condicional.

2. Modelo GARCH

Consideremos $u_t = \sqrt{h_t}.\nu_t$ donde v_t es i.i.d. con media cero y varianza unitaria donde h_t se desarrolla como

$$u_t = \zeta + \alpha_1 \mu_{t-1}^2 + \alpha_2 \mu_{t-2}^2 + \cdots + \alpha_m \mu_{t-m}^2$$

En forma general, podemos imaginar un proceso para el cual la varianza condicional dependa de un número infinito de rezagos μ_{t-j}^2

$$h_t = \zeta + \pi(L)\mu_t^2 \tag{2.1}$$

donde $$\pi(L) = \sum_{j=1}^{x} \pi_j L^j$$

Una idea natural es dar parámetros $\pi(L)$ como el cociente de dos polinomios de orden finito

$$\pi(L) = \frac{\alpha(L)}{1 - \delta(L)} = \frac{\alpha_1 L^1 + \alpha_2 L^2 + \cdots + \alpha_m L^m}{1 - \delta_1 L^1 - \delta_2 L^2 - \cdots - \delta_r L^r} \tag{2.1}$$

donde asumimos que las raíces de $1 - \delta(z) = 0$ están fuera del círculo de la unitario.
Si (2.2) es multiplicado por $1 - \delta(L)$, el resultado es

$$[1 - \delta(L)]h_t = [1 - \delta(1)]\zeta + \alpha(L)u_t^2$$

o $$h_t = k + \delta_1 h_{t-1} + \delta_2 h_{t-2} + \cdots + \delta_r h_{t-r} + \alpha_1 \mu_{t-1}^2 + \alpha_2 \mu_{t-2}^2 + \cdots + \alpha_m \mu_{t-m}^2 \tag{2.3}$$

para $k \equiv [1 - \delta_1 - \delta_2 - \cdots - \delta_r]\zeta$, la Expresión (2.3) es el modelo GARCH(*r,m*) de u_t, propuesto por Bollerslev (1986).

3. Modelo EGARCH

Considerando $u_t = \sqrt{h_t}.v_t$, donde es i.i.d y v_t con media cero y varianza unidad, Nelson (1991) propuso siguiente modelo de la evolución de la varianza condicional de u_t

$$\log h_t = \zeta + \sum_{j=1}^{x} \pi_j\{|v_{t-1}| + E|v_{t-1}| + Xv_{t-1}\}$$

El modelo de Nelson se refiere al modelo GARCH exponencial, o EGARCH. SÍ $\pi_t > 0$, el modelo de Nelson implica que una desviación $|v_{t-j}|$ de su valor esperado hace que la varianza de ut, para ser más grande, un efecto similar a la idea de la especificación de *GARCH*.

Una parametrización natural es modelar $\pi(L)$ como el cociente de dos polinomios de orden finito como en la especificación de *GARCH (r,m)*:

$$\begin{aligned}\log h_t &= k + \delta_1 \log h_{t-1} + \delta_2 \log h_{t-2} + \cdots \\ &+\delta_r \log h_{t-r} + \alpha_1\{|v_{t-1}| - E|v_{t-1}| + Xv_{t-1}\} \\ &+\alpha_2\{|v_{t-2}| - E|v_{t-2}| + Xv_{t-2}\} + \cdots \\ &+\alpha_m\{|v_{t-m}| - E|v_{t-m}| + Xv_{t-m}\}\end{aligned} \tag{3.1}$$

El modelo de EGARCH se puede estimar por la máxima verosimilitud, especificando una densidad para v_t. Nelson propuso el usar de la distribución generalizada del error, normalizada para tener variación cero del medio y varianza unidad:

$$f(v_t) = \frac{v\, exp[-(1/2)|v_t/\lambda|^v]}{\lambda \cdot 2^{|(v+1)v|}\Gamma(1/v)} \tag{3.2}$$

Aquí $\Gamma(\bullet)$ está la función gamma, A es una constante dada como

$$\lambda = \left\{\frac{2^{(.v)}\Gamma(1/v)}{\Gamma(3/v)}\right\}^{1/2}$$

y v es un parámetro positivo. Para v = 2, la constante λ = 1 y la expresión (3.2) resulta ser la densidad normal estándar. La esperanza del valor absoluto de v_i, esta dado por

$$E|v_t| = \frac{\lambda 2^{1/v}\Gamma(2/v)}{\Gamma(1/v)}$$

Para el caso de v = 2, éste se convierte en

$$E|v_t| = \sqrt{2/\pi}$$

La señorita Pérez comentó que imaginaba que los maridos del la Ciudad de la Paz eran cariñosos y que estaba preparando un cuestionario para circularlo entre los hombres jóvenes de La Paz, para intentarlo averiguar sus preferencias en cuanto al matrimonio.
“pero los paceños no responderán los cuestionarios”, aseguró Luis.
“¿No lo harán?”, preguntó la señorita peréz decepcionada.
“No”, repuso Luis, “no lo harán. Como país , no importan las encuestas”
N.N.

Cuando la estadística no se basa en cáculos precisos, confunde en vez de guiarnos. La mente se deja llevar fácilmente por la falsa apariencia de xactitud que la estadística mantiene en sus errores, y confiadamente adopta errores escondidos bajo la forma de una verdad matemática.

Alexis de Tocqueville, Democracy in America

“Pero los promedios no son reales”, objeto Milo, “sólo son imaginarios”.
“Es posible”, coincidió el niño; sin embargo, el pequeño: “pero a veces son muy útiles. Por ejemplo si no tuvieras dinero, pero estuvieras con otras cuatro personas que contasen con diez dólares cada una, entonces, cada uno de ustedes tendría en promedio ocho dólares, ¿no es así?.
“Creo que si”, dijo Milo en voz baja.
“Bueno, piensa en lo bien que te va gracias a los promedios”, dijo convencido el niño y éste continuó con la siguiente explicación: “y piensa en el pobre grangero cuando no ha llovido en todo el año. De no ser por una precipitación anual promedio de 37 pulgadas en esa parte del país, toda su cosecha se perdería”.
Todo esto era muy confuso para Milo, ya que siempre había tenido problemas en la escuela con este tema.

“Hay más ventajas”, continuó el niño, “por ejemplo, si una rata fuese atrapada por nueve gatos, en promedio, cada gato obtendría el 10 por ciento de rata, mientras que ésta obtendría el 90 por ciento de cada gato. Si fueras una rata, verías cómo todo es más agradable”, le dijo el niño a Milo de forma irónica.
Norton Jester, The Phantom Tollbooth.

Índice

LA URBANIZACION EN BOLIVIA

Augusto S. Soliz Sánchez

Uno de los fenómenos sociales más notables del siglo XX es el acelerado proceso de urbanización que se ha venido experimentando en los países menos desarrollados, en general, y en los países de América Latina, en particular, debido al rápido crecimiento de la población urbana. El rápido crecimiento de la población urbana en las últimas cinco décadas ha concentrado el interés de las disciplinas sociales por las innegables implicaciones que el fenómeno tiene en los cambios que se dan en las estructuras económicas, sociales y demográficas de los mencionados países. La urbanización cambia la estructura económica dando preponderancia a las actividades no-agrícolas, concentra geográficamente la actividad económica y modifica la distribución por sectores de la población ocupada. La proporción de mano de obra dedicada a la agricultura se reduce y aumenta la población ocupada en la industria manufacturera y en el sector de los servicios. Entre los cambios que la urbanización introduce en la estructura social se puede mencionar el proceso de transición de una sociedad tradicional, típica de las zonas rurales, a una sociedad moderna, que caracteriza a las zonas urbanas. Asimismo, la composición de los hogares tiende a cambiar con el crecimiento de la proporción de hogares nucleares y la reducción de la proporción de hogares extendidos.
Entre los cambios demográficos asociados a la urbanización se puede mencionar el mayor crecimiento de la población urbana en comparación con el crecimiento de la población rural. Entre 1970 y 1990 la tasa media de crecimiento anual de la población urbana en el mundo alcanzó al 3.8 por ciento mientras que la tasa de crecimiento de la población rural llegó solamente al 1.3 por ciento[1]. Por otra parte, la urbanización se relaciona con el comportamiento diferencial de la fecundidad y la mortalidad. En numerosos estudios demográficos, particularmente en los países menos desarrollados, se ha observado que los niveles más bajos de fecundidad y mortalidad están relacionados con altos niveles de urbanización.

MEDICION DE LA URBANIZACION

Definición de población urbana

En la medición de la urbanización es fundamental la definición de población urbana. La definición de población urbana no es uniforme en todos los países y muchas veces cambia en un mismo país a través del tiempo.

En el caso de Bolivia, la definición de población urbana ha experimentado diversos cambios en el tiempo. En el Primer Censo Decenal del año 1900 la población urbana fue definida como la que vivía en centros poblados de 200 o más habitantes; en el

Censo Demográfico de 1950 se consideró como población urbana a la que habitaba en cabeceras de cantón, es decir, las capitales de las unidades político-administrativas menores. Finalmente, en los censos nacionales de población y vivienda de 1976, 1992 y 2001 se define como población urbana a la que vive en localidades de 2,000 o más habitantes.

[1] United Nations, Department of Economic and Social Information and Policy Análisis, World Urbanization Prospects, The 1992 Revisión, New York, 1993, p. 3

En el presente estudio se entiende como población urbana a la que vive en localidades de 2,000 o más habitantes. En adelante, estas localidades se denominan también ciudades, centros urbanos o simplemente localidades urbanas. Además, para mejorar la comparabilidad histórica de la información, se han reclasificado los datos de los censos de 1900 y 1950 de acuerdo a la definición de población urbana de los censos de 1976, 1992 y 2001. Esto significa que los porcentajes de población urbana en 1900, 1950, 1976, 1992 y 2001 son comparables para el estudio de la urbanización en Bolivia porque corresponden a la proporción de la población total que vive en localidades de 2,000 o más habitantes.

DEFINICIÓN DE URBANIZACIÓN

Las significativas diferencias que existen entre la población que vive en el campo y la que habita en las ciudades han sido estudiadas desde tiempos muy antiguos pero la clasificación urbano-rural, que constituye la base para el estudio de la urbanización, fue introducida en el campo de las estadísticas de población recién en el siglo XIX[2]. En la actualidad, pese a su complejidad, existe consenso en definir el proceso de urbanización como el crecimiento del porcentaje de población urbana como proporción de la población total de una determinada unidad territorial. La unidad territorial de análisis puede ser el mundo, un continente, una región, un país o una unidad político-administrativa de éste. En el presente caso, la unidad de estudio es el territorio y la población de Bolivia.

INDICADORES DEL NIVEL DE URBANIZACIÓN

NIVEL DE URBANIZACIÓN

El nivel de urbanización se refiere al número absoluto o relativo de población que se define como urbana. En ese sentido, los indicadores que se utilizan para medir el nivel de urbanización son el porcentaje de población urbana, la relación población urbana - población rural y el tamaño medio de la población de las ciudades o centros urbanos. En el presente estudio de la urbanización en Bolivia se utiliza solamente el porcentaje de población urbana.

PORCENTAJE DE POBLACIÓN URBANA

El porcentaje de población urbana es la relación por cociente entre la población urbana y la población total. El porcentaje de población urbana es el indicador más utilizado para medir el nivel de urbanización porque es fácil de calcular e interpretar y los datos
básicos para su cálculo casi siempre están disponibles. No obstante su sencillez,
cuando el porcentaje de población urbana se utiliza con fines comparativos, histórica o internacionalmente, se deben tener en cuenta la definición de población urbana y los límites de los centros urbanos.

INDICADORES DEL CRECIMIENTO DE LA URBANIZACIÓN

El crecimiento de la urbanización se refiere al cambio del nivel de urbanización durante un determinado periodo de tiempo. En el presente caso, el crecimiento de la urbanización en Bolivia se mide mediante la estimación de los puntos de variación anual y la tasa media de crecimiento anual del porcentaje de población urbana.

---

[2] U.S. Bureau of the Census, The Methods and Materials of Demography, by Henry S. Shryock, Jacob S. Siegel, and Associates, U.S. Government Printing Office, Washington, D.C., 1971, Volume 1, p. 151

INDICADORES DE LA CONCENTRACION DE LA POBLACION URBANA

ÍNDICE DE CONCENTRACIÓN DE GINI

El índice de concentración de Gini se utiliza como un indicador de la concentración de la población urbana. El valor del índice puede variar entre cero y uno. Si el valor del
índice de Gini fuera cero, la distribución de la población urbana sería uniforme entre todas las localidades urbanas, esto es, no existiría concentración de la población urbana; por el contrario, si el índice fuera uno, la población urbana estaría concentrada en una sola localidad urbana.

LOS DATOS BÁSICOS

Los datos básicos para examinar el proceso de urbanización en Bolivia provienen de los cinco censos nacionales levantados en el país entre 1900 y 2001. También se utilizan los resultados de la Encuesta Nacional de Población y Vivienda de 1988 por su enumeración completa de los centros urbanos y porque fue la primera en mostrar que el porcentaje de población urbana había superado al porcentaje de población rural en ese año. En consecuencia, las fuentes de datos que se utilizan en este capítulo son las siguientes:

- Primer Censo Decenal de 1900
- Censo Demográfico de 1950
- Censo Nacional de Población y Vivienda de 1976
- Encuesta Nacional de Población y Vivienda de 1988 (ENPV 88)
- Censo Nacional de Población y Vivienda de 1992
- Censo Nacional de Población y Vivienda de 2001

Por otra parte, conviene señalar que los datos que se utilizan corresponden a la población nominalmente censada, sin correcciones del error de omisión censal. Esto podría tener como posible consecuencia una sobreestimación del porcentaje de población urbana. En efecto, la experiencia en casi todos los países ha mostrado que la subenumeración censal puede afectar en mayor grado a la población rural. Sin embargo, en el caso de Bolivia, es muy difícil cuantificar la omisión censal a nivel de áreas urbano rural por la falta de información adecuada.

LA URBANIZACIÓN EN BOLIVIA EN EL PERIODO 1990-2001

a) El crecimiento de la población de Bolivia por áreas urbano-rural

Cuadro 2

CRECIMIENTO DE LA POBLACIÓN* DE BOLIVIA POR AREAS URBANO-RURAL, 1900-2001

PERIODO	CRECIMIENTO ABSOLUTO (miles)			CRECIMIENTO RELATIVO (%)		
	Población total	Población urbana	Población rural	Población total	Población urbana	Población rural
1900-1950	1148.4	482.9	665.4	73.8	214.0	50.0
1950-1976	1909.2	1217.7	692.0	70.6	171.8	34.7
1976-1992	1807.3	1769.0	38.3	39.2	91.9	1.4
1992-2001	183.6	1470.4	383.2	28.9	39.8	14.1

El estudio del crecimiento de la población de Bolivia por áreas urbano-rural se basa en fas definiciones censales de población urbana y población rural que se utilizan desde 1976. En este sentido, se considera población urbana a la que vive en localidades de 2,000 o más habitantes y población rural a la que habita en localidades de menos de 2,000 habitantes o vive en forma dispersa. Como se señala en una sección anterior del presente estudio, a fin de mejorar la comparabilidad histórica de los datos censales, los datos de la población urbana del Censo Decenal de 1900 y del Censo Demográfico de 1950 se han reclasificado de acuerdo a la definición de 1976. El cuadro 1 presenta la población de Bolivia por áreas urbano-rural para el periodo 1900-2001, según las definiciones señaladas.

En el citado cuadro se observa que la población boliviana ha sido predominantemente rural en la mayor parte del periodo que cubren los censos de población del siglo XX y que, a partir de 1988, tiende a ser predominantemente urbana. En el inicio del periodo, esto es, en el año 1900, la población rural en Bolivia constituía aproximadamente el 85 por ciento de la población total del país. En 1950, el mencionado porcentaje disminuyó a cerca del 75 por ciento y en 1976 a 58 por ciento. Hacia 1988, por primera vez, se observa que el porcentaje de la población rural es inferior al 50 por ciento de la población nacional, lo que

Cuadro 1

POBLACIÓN* DE BOLIVIA POR AREAS URBANO-RURAL, 1900-2001

(Miles de habitantes)

Año	Población total	Población urbana	Población rural	(%)		
				Población total	Población urbana	Población rural
1900	1,555.8	227.4	1,328.4	100.0	14.6	85.4
1950	2,704.2	708.6	1,995.6	100.0	26.2	73.8
1976	4,613.4	1,925.8	2,687.6	100.0	41.7	58.3
1988	6,405.1	3,286.3	3,118.8	100.0	51.3	48.7
1992	6,420.7	3,694.8	2,725.9	100.0	57.5	42.5
2001	8,274.3	5,165.2	3,109.1	100.0	62.4	37.6

* Población nominalmente censada, sin corrección por omisión

- Población urbana: localidades de 2,000 o más habitantes

- Población rural: localidades de menos de 2,000 habitantes y población dispersa

significa que la población urbana pasa a ser mayor que la población rural del país. La tendencia creciente del porcentaje de población urbana continúa en 1992 y 2001. Según el Censo de 1992, la población urbana constituye el 58 por ciento de la población total y según el Censo de 2001 el 62 por ciento. Por consiguiente, durante el periodo 1900-2001 la población de Bolivia presenta, por una parte, una tendencia clara de disminución de la población rural y, por otra, en el sentido contrario, un mayor crecimiento de la población urbana.

b) Crecimiento absoluto y crecimiento relativo de la población urbana y de la población rural

En el cuadro 2 se presentan las cifras del crecimiento absoluto y relativo de la población urbana y de la población rural de Bolivia durante los periodos intercensales 1900-1950, 1950-1976, 1976-1992 y 1992-2001. En este cuadro se observa la notable diferencia que existe entre el crecimiento de la población urbana y el crecimiento de la población rural durante todo el periodo 1900-2001. En el periodo intercensal 1900-1950 el crecimiento absoluto de la población rural es mayor que el crecimiento absoluto de la población urbana pero en cifras relativas el crecimiento de ésta es mayor que el de la primera. En términos absolutos, el crecimiento de la población urbana alcanzó aproximadamente a 480,00 personas y el crecimiento de la población rural a cerca de 670,000 personas. En cifras relativas, con respecto al año 1900, el aumento de la población urbana en el periodo 1900-1950 llega a algo más del 200 por ciento y el aumento de la población rural alcanza aproximadamente al 50 por ciento en el mismo periodo.

En el periodo intercensal 1950-1976 la población urbana creció más que la población rural, tanto en términos absolutos como en términos relativos. En términos absolutos, la población que vive en localidades de 2,000 o más habitantes aumentó en poco más de 1.2 millones de personas, cifra que, en términos relativos, representa aproximadamente el 170 por ciento de su tamaño en 1950. La población que vive en localidades de menos de 2,000 habitantes o en forma dispersa creció en este periodo en cerca de 700,000 personas en cifras absolutas, es decir, menos que la población urbana y este crecimiento de la población rural equivale, en términos relativos, a un aumento de cerca del 35 por ciento con respecto a 1950.

Las mayores diferencias en el crecimiento de la población boliviana por áreas urbano-rural se presentan en el periodo intercensal 1976-1992. El crecimiento absoluto de la población urbana llegó a cerca de 1.8 millones de personas en tanto que el crecimiento absoluto de la población rural alcanzó a 38,000

habitantes. En cifras relativas, la población urbana creció en algo más del 90 por ciento con relación a su tamaño de 1976 y la población rural aumentó en poco más del uno por ciento. Esta notable diferencia se

debe, aparentemente, a un significativo error de omisión del censo de 1992 en el área rural[3]. De hecho, la estimación de la población rural de la Encuesta Nacional de Población y Vivienda de 1988 proporciona una cifra mayor que la población rural nominalmente censada en 1992.

Por último en el periodo intecensal 1992-2001, tanto el crecimiento absoluto como el crecimiento de la población urbana es notablemente mayor que el crecimiento de la población rural.

c) Crecimiento anual de la población urbana y de la población rural

El cuadro 3 presenta las tasas medias de crecimiento anual de la población urbana y de la población rural y los periodos de duplicación respectivos que corresponden a los periodos intercensales 1900-1950,1950-1976, 1976-1992 y 1992-200

[3] En la preparación del Censo de 1992 era de conocimiento público la resistencia al levantamiento censal de parte de la Confederación Sindical Unica de Trabajadores Campesinos de Bolivia (CSUTCB).

Cuadro 3
TASA MEDIA DE CRECIMIENTO ANUAL Y PERIODO DE DUPLICACIÓN DE LA POBLACIÓN DE BOLIVIA POR AREAS URBANO-RURAL, 1900-2001

PERIODO	TASA MEDIA DE CRECIMIENTO ANUAL (%)			PERIODO DE DUPLICACIÓN (Años)		
	Población total	Población urbana	Población rural	Población total	Población urbana	Población rural
1900-1950	1.11	2.29	0.81	62	31	86
1950-1976	2.05	3.84	1.14	34	18	71
1976-1992	2.11	4.16	0.09	33	17	770
1992-2001	2.74	3.62	1.42	29	19	49

En el mencionado cuadro 3 se observa que el ritmo de crecimiento de la población urbana es mayor que el ritmo de crecimiento de la población rural y mayor que el crecimiento de la población del país en los cuatro periodos intercensales considerados. En el periodo 1900-1950 la tasa media de crecimiento anual de la población urbana, estimada en 2.28 por ciento, es poco más del doble de la tasa de crecimiento de la población nacional (1.11 %) y cerca del triple de la tasa de crecimiento de la población rural (0.81 %). Entre 1950 y 1976, esto es, en el segundo periodo intercensal, por una parte, se mantiene la diferencia entre las tasas de crecimiento del país y de la población urbana y, por otra, aumenta la diferencia entre

las tasas de crecimiento de la población urbana y de la población rural. En efecto, la tasa media de crecimiento anual de la población de las localidades de 2,000 o más habitantes en el citado periodo alcanza a 3.84 por ciento, valor que casi duplica la tasa de crecimiento de la población del país

(2.05 %) y equivale a más del triple de la tasa de crecimiento de la población rural
(1.14 %) en el mismo periodo.

En el periodo intercensal 1976-1992, la diferencia entre el ritmo de crecimiento de la población nacional y el ritmo de crecimiento de la población urbana se mantiene sin gran variación, es decir, la tasa media de crecimiento de la población urbana en el periodo 1976-1992 (4.14 %) es prácticamente el doble de la tasa de crecimiento de la población del país; pero la diferencia entre el ritmo de crecimiento de la población urbana y el ritmo comparable de la población rural aumenta de manera muy significativa: la tasa media de crecimiento anual de la población de las localidades urbanas es equivalente a 46 veces la tasa media de crecimiento de la población rural del periodo 1976-1992.

Por último, en el periodo 1992-2001, las diferencias entre las tasas de crecimiento anual disminuyen pero la tasa de crecimiento de la población urbana sigue siendo más alta que la del país y más del doble de la tasa de crecimiento de la población rural

d) Periodo de duplicación de la población

Con las tendencias de las tasas medias de crecimiento anual de la población estimadas para los cuatro periodos intercensales, el tiempo de duplicación del tamaño de la población urbana es significativamente menor que el periodo de duplicación de la población rural. De acuerdo con su tasa media de crecimiento anual del periodo 1900-1950, la población urbana se hubiera duplicado aproximadamente cada 31 años; con su tasa de crecimiento del periodo 1950-1976 cada 18 años, con su tasa de crecimiento del periodo 1976-1992 cada 17 años y con su tasa de crecimiento del 1992-2001 cada 19 años. La población rural, con su tasa de crecimiento anual del periodo 1900-1950, se hubiera duplicado aproximadamente en 86 años; con su tasa de crecimiento del periodo 1950-1976 se duplicaría en 61 años; con su casi nula tasa de crecimiento del periodo 1976-1992, el tiempo de duplicación de la población rural aumenta a 770 años y, por último, con su tasa de crecimiento del periodo 1992-2001, la población rural de Bolivia se duplicaría cada 49 años.

TENDENCIA DE LA URBANIZACION EN BOLIVIA

a) Crecimiento del nivel de urbanización: el porcentaje de población urbana

El proceso de urbanización, como se ha señalado al tratar su medición, se refiere al crecimiento del porcentaje de población urbana como proporción de la población total de una determinada unidad territorial. En el cuadro 1 se presenta el comportamiento del porcentaje de población urbana como indicador del nivel de urbanización en Bolivia durante el periodo 1900-2001.

En el citado cuadro se observa la tendencia creciente del nivel de urbanización en Bolivia en términos del porcentaje de población urbana. En el año 1900, el porcentaje de población que vive en localidades de 2.000 y más habitantes alcanza solamente el 14 por ciento de la población total del país, después de 50 años o sea, en 1950, el porcentaje de población urbana casi se duplica y llega al 26 por ciento de la población nacional: en 1976 el porcentaje de población urbana vuelve aumentar alcanzando el 42 por ciento; en 1992, la población urbana constituye el 57 por ciento de la población total del país y, finalmente, en el año 2001, llega a ser el 62 por ciento.

b) Crecimiento anual de la urbanización

El crecimiento de la urbanización se refiere al cambio del nivel de urbanización durante un determinado periodo de tiempo. En el caso de Bolivia, el crecimiento de la urbanización se mide mediante los cambios en el porcentaje de población urbana. Los cambios en el porcentaje de población urbana en los periodos intercensales se miden con la tasa media de crecimiento anual bajo el supuesto de que el porcentaje sigue una tendencia exponencial. Además, el crecimiento de la urbanización en términos del porcentaje de población urbana se mide en puntos de variación anual como se presenta en el cuadro 4.

Cuadro 4
INDICADORES DEL RITMO DE LA URBANIZACIÓN EN BOLIVIA, 1900 - 2001

PERIODO	Puntos de variación anual del porcentaje de población urbana	Tasa media de crecimiento anual del porcentaje de población urbana
1900-1950	0.23	1.17
1950-1976	0.59	1.78
1976-1992	1.01	2.05
1992-2001	0.53	0.88

Del citado 4 se puede señalar que, en general, el crecimiento de la urbanización en Bolivia sigue una tendencia creciente en el periodo 1900-2001. En términos de la variación anual del porcentaje de población urbana se observa que el crecimiento de la urbanización entre 1900 y 1950 aumenta en cerca de un cuarto punto por año; luego llega a mas de medio punto en el periodo de 1950- 1976; alcanza a más de un punto por año entre 1976 y 1992 y, finalmente, en el último periodo intercensal alcanza a 0.53 punto por año.

De manera similar, la tendencia al aumento de la urbanización de la población boliviana se puede ver en los valores de la tasa media de crecimiento anual del porcentaje de población urbana. De hecho, la tasa de crecimiento anual del porcentaje de población urbana aumenta de 1.17 por ciento a 1.78 por ciento entre los periodos 1900-1950 y 1950-1976; luego, llega a 2.05 por ciento entre los periodos 1950-1976 y 1976-1992 y, por último, disminuye a 0.88 por ciento entre los periodos 1976-1992 y 1992-2001. Por consiguiente, se puede concluir que el crecimiento de la urbanización de la población de Bolivia es bastante bajo entre los años 1900 y 1950 del siglo XX; aumenta significativamente en el periodo intercensal 1950-1976; llega a su valor más alto entre los periodos 1950-1976 y 1976-1992 y luego disminuye entre los periodos 1976-1992 y 1992-2001.

LA URBANIZACIÓN DE LA POBLACIÓN DE BOLIVIA EN EL CONTEXTO REGIONAL Y MUNDIAL

Cuadro 5

POBLACION URBANA DEL MUNDO, DE PAISES DESARROLLADOS, DE PAISES MENOS DESARROLLADOS DE AMERICA LATINA Y EL CARIBE Y DE BOLIVIA, 1950 - 2000

(Porcentajes)

REGIONES, PAISES	1950	1975	1990	2000
Mundo	29.7	37.8	43.2	47.4
Países Desarrollados	54.9	69.9	73.7	76.1
Países Menos Desarrollados	17.8	26.7	34.7	40.5
América Latina y el Caribe	41.4	61.2	71.1	75.4
America del Sur	42.9	64.0	74.6	79.8
Bolivia*	26.2	41.7	57.5	62.4

* Porcentajes estimulados en base a los censos nacionales de 1950, 1976, 1992 y 2001

El nivel de urbanización de la población, por las innegables implicaciones que tiene en las estructuras económicas, sociales y demográficas, presenta notables diferencias entre las regiones y países del mundo. Por consiguiente, es relevante conocer la posición en que se encuentra el nivel y la tendencia de la urbanización en Bolivia. En tal sentido con fines comparativos, en el cuadro 5 se presentan las cifras de los niveles de urbanización de la población del mundo, de los Países Desarrollados, de los Países Menos Desarrollados, de América Latina y el Caribe, de América del Sur y de Bolivia, en el periodo 1950-2000.

Según el cuadro 5, en primer lugar, se puede señalar que el nivel de urbanización de la población boliviana en el periodo de 1950-2000 es más alto que el nivel de urbanización de la población mundial, excepto en el inicio del periodo, y de la población de los Países Menos Desarrollados. En 1950 el porcentaje de población urbana del mundo (30%) es ligeramente mayor que el porcentaje de población urbana de Bolivia (26%) pero entre 1975 y 2000 el porcentaje de la población urbana de Bolivia es cada vez más alto que el porcentaje de población urbana del mundo. De hecho, en 1975 el porcentaje de población urbana en Bolivia es cuatro puntos más alto que el porcentaje correspondiente a la población mundial y esta diferencia aumenta a poco más de 17 puntos en el año 2000. En forma similar, la urbanización de la población boliviana se encuentra en un nivel cada vez más alto con respecto al nivel de urbanización de los Países Menos Desarrollados. En 1950 el porcentaje de población urbana de Bolivia es aproximadamente ocho puntos más alto que el porcentaje de población urbana de los Países Menos Desarrollados y en el año 2000 la citada diferencia aumenta a 24 puntos.

En segundo lugar, en comparación con el nivel de urbanización de los Países Desarrollados, la urbanización de la población boliviana es significativamente más baja aunque la diferencia tiende a disminuir durante el periodo en estudio En 1950, el porcentaje de población urbana de los Países Desarrollados alcanza a 55 por ciento, cifra que equivale a poco más del doble del porcentaje de población urbana de Bolivia (26%). Esta diferencia de cerca de 30 puntos disminuye notablemente en el año 2000, año en el que el porcentaje de población urbana de los Países Desarrollados llega a 76 por ciento y el porcentaje comparable de Bolivia a 65 por ciento. Por consiguiente, la diferencia entre los niveles de urbanización de los Países Desarrollados y de Bolivia se reduce a 11 puntos.

En tercer lugar, en el contexto regional, el nivel de urbanización de la población de Bolivia se encuentra por debajo de los niveles de urbanización de América Latina y de América del Sur. En efecto, según el cuadro 5, el nivel de urbanización de la población boliviana, en términos del porcentaje de población urbana, es menor que los niveles comparables de América latina, de América del Sur. Las diferencias, en general, aumentan entre 1950 y 1975 y, luego, tienden a disminuir entre los años 1975 y 2000. En comparación con la región, en 1950, el porcentaje de población urbana de América Latina alcanza a 41 por ciento y el porcentaje de población urbana de Bolivia a 26 por ciento lo que significa que el nivel de urbanización de América Latina en este año se encuentra 15 puntos por encima del nivel de urbanización de Bolivia. En 1975, cuando los porcentajes de población urbana de América Latina y Bolivia alcanzan a 61 por ciento y 42 por ciento, respectivamente, la diferencia entre los niveles de urbanización de la región y del país aumenta a casi 20 puntos. Finalmente, en el año 2000, cuando el porcentaje de población urbana de la región latinoamericana se estima en poco más del 75 por ciento y el porcentaje correspondiente a la población boliviana se estima en casi 65 por ciento, la diferencia entre los niveles de urbanización de la región y del país disminuye a cerca de 10 puntos.

LA CONCENTRACION DE LA POBLACION URBANA EN LAS GRANDES CIUDADES

Una de las notables características que se ha observado en el proceso de urbanización, particularmente en varios países de América Latina, es la alta concentración de la población urbana en las grandes ciudades. En algunos países esta concentración se ha dado en las capitales nacionales como es el caso de México, Argentina, Chile y Perú. Por la relevancia que este fenómeno tiene, existe un marco interés por medir el nivel de concentración de la población urbana en estrecha relación con la distribución de las localidades urbanas según él numero de habitantes.

En el presente caso se utiliza el índice de concentración de Gini como un indicador de la concentración de la población urbana. Como se señala en otra sección del estudio, el valor del índice puede variar entre cero y uno. Si el valor del índice de Gini fuera cero, la distribución de la población urbana sería uniforme entre todas las localidades urbanas, lo que significaría que no existe concentración de la población urbana; si el valor del índice fuera uno, la población urbana estaría concentrado en una sola localidad urbana.

En el cuadro 6 se presenta los valores calculados del índice de concentración de Gini como indicador de la concentración de la población urbana de Bolivia en el periodo 1900-1992.

Cuadro 6
INDICE DE CONCENTRACION DE GINI, POBLACION URBANA DE BOLIVIA, 1900-2001

Año	Indice de Gini
1900	0.50
1950	0.75
1976	0.65
1992	0.79
2001	0.78

Del cuadro 6, se puede señalar que el índice de Gini muestra, en general, una tendencia al aumento de la concentración de la población urbana de Bolivia en todo el periodo 1900-2001. Entre 1900 y 1950 el índice de Gini aumenta de 0.50 a 0.75; entre 1950 y 1976 disminuye de 0.75 a 0.65 y entre 1976 y 1992 vuelve a aumentar de 0.65 a 0.79 y mantiene casi el mismo valor en el año 2001. Por lo tanto, se puede concluir señalando que la población urbana de Bolivia, según los valores del índice de Gini tiende a concentrarse en pocas ciudades, es decir, en las grandes ciudades.

BIBLIOGRAFIA:

- Arriaga, Eduardo E., "Selected Measures of Urbanizaron" en Goldstein, Sidney y Sly, David F. (eds.), The Measurement of Urbanization and Projection of Urban Population, International Union for the Scientific Study of Population Working Papers, No. 2, Dohain, Belgium: Ordina Editions, 1975
- Oficina Nacional de Inmigración, Estadística y Propaganda Geográfica, Censo General de la Población de Bolivia, Septiembre 1º de 1900, tomos I y II, Editorial Canelas S.A., Cochabamba, Bolivia, 1973
- Instituto Nacional de Estadística, Bolivia: Características de la Población, Censo 2001, Serie I, Resultados Nacionales, Volumen 4, Septiembre, 2002
- Instituto Nacional de Estadística, Censo Nacional de Población y Vivienda 1992, Bolivia, Resultados Finales, Mayo, 1993
- United Nations, Department of Economic and Social Affairs, Population División, World Urbanization Prospects, The 1996 Revisión, New York, 1998
- Soliz Sánchez, Augusto S., La Población de Bolivia, OIE-Bolivia, La Paz, Octubre, 2001

Índice