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 $\hat{\theta} = \bar{y}$, entonces $\hat{\theta}_{(j)} = \bar{y}_{(j)} = \frac{\sum_{i \neq j} y_i}{n-1}$. 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 $\bar{y}_{(j)} = \frac{\sum_{i \neq j} y_i}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i - y_j\right) = \frac{1}{n-1}\left(n\bar{y} - \bar{y} + \bar{y} - y_j\right) = \bar{y} - \frac{1}{n-1}\left(y_j - \bar{y}\right).$

Entonces $\sum_{j=1}^{n}\left(\bar{y}_{(j)} - \bar{y}\right)^2 = \frac{1}{(n-1)^2}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{s_y^2}{n-1}.$

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.

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

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

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

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 de nada; pues se destruiría la estructura de conglomerado y daría una estimación de la varianza que sólo 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 misma forma 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 $$S_r^2 = \frac{\sum\left(M_i \overline{y}_i - M_i \hat{\overline{y}}_r\right)^2}{n-1} = \frac{9224{,}327784}{9} = 1024.9$$

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

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

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

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 numero 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_1	$M_i/2$	$\frac{M_i}{2} y_i$	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, observar 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