PONENCIAS MESA 3

Edmundo Berumen
BERUMEN Y ASOCIADOS PARA IFE

No voy a reseñar en este foro lo que constituye alguno de los distintos ejercicios de investigación electoral. Me voy a conformar, y será un éxito para mí, si simplemente logro dejar planteadas una o dos preguntas, que no espero que sean resueltas aquí, y con eso yo me doy por satisfecho. A lo mejor ustedes quedan frustrados, pero yo me daré por satisfecho.

Para mi beneficio más que para el de ustedes, escribiré algunas fórmulas sencillas, pero a mí siempre me ayuda a regresar a lo básico. Todos recuerdan el estimador de un total basado en una muestra aleatoria simple de tamaño **n** de una población de tamaño **N**, que me permito reescribir varias veces:

$$\hat{Y} = N * \bar{y}$$

$$= N \frac{\sum_{i=1}^{n} y_i}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\frac{1}{N}}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}$$

Pero cuando ya llegué a esta manera de reescribirla, al menos yo ya le entiendo un poquito más a la fórmula y digo: ¡Ah, qué fórmula tan inteligente! ¿Qué es lo que está haciendo? Bueno, pues lo que está haciendo es tomar una de las observaciones y la divide por **1** entre el tamaño **N** de la población, y ¿qué es eso? Ese término de la suma, por sí solo, está estimando lo que busco: el total de la característica de interés. La variable que observé, piensen en votos de la sección electoral **i** a favor de un partido, lo estoy dividiendo entre **1** sobre el total de los elementos en la población (entre el total de secciones, por ejemplo). En un muestreo aleatorio simple, cada extracción de una observación, de un elemento que voy a observar de la población de interés, es extraído del universo con esta probabilidad, **1** entre el total **N** de la población.

Entonces, este solo término estima lo que busco, con la simpleza de multiplicar lo observado en una sola unidad por el total de unidades en el universo. Pero como no tengo sólo una unidad medida, tengo **n** de ellas, estoy sumando cada una de estas estimaciones del mismo total y luego las promedio y, entonces, tengo un promedio de **n** estimaciones de lo mismo, que luego en modelos de muestreo más complicados generalizamos. Entonces, decimos: para estimar un total, lo único que tengo que hacer es dividir cada observación entre su probabilidad de selección, y luego promediar:

$$\hat{Y}_{ppt} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

$$p_i = \frac{c_i}{C},\ \textit{medida de tamaño}$$

Llegamos, entonces, a una clase de estimadores que toman las observaciones, dividen entre la probabilidad de selección, las suman y luego las promedian. Un subconjunto de esa clase de estimadores son los estimadores donde la selección de las unidades a medir es con probabilidad proporcional a una medida de tamaño, que llamamos selección con PPT, en el lenguaje del muestreo. Piensen que las medidas de tamaño son ciudadanos del padrón y que estamos seleccionando secciones electorales con probabilidad proporcional a los parámetros del padrón. La varianza muestral de esta clase de estimadores está dada por:

$$\hat{V}(\hat{Y}_{Appt}) = \frac{1}{\quad}\sum\ p_i\left(\frac{y_i}{\quad} - \hat{Y}_{Appt}\right)^2$$

$$\hat{Y}_{A\,ppt} = \frac{1}{n}\sum\frac{y_i}{p_i}$$

$$= \frac{C}{n}\sum_{i=1}^{n}\frac{y_j}{C_j}$$

y_i = *número de ciudadanos que votaron por el partido A*

Si se toma la esperanza de ese estimador, pues es simplemente la esperanza de cada uno de sus sumandos que es todos los valores posibles de la población, de la variable que estoy usando, por la probabilidad de haber seleccionado ese elemento, que en el muestreo con PPT es la medida de tamaño. Hace uno el álgebra y llegamos al total poblacional, entonces decimos que es un estimador insesgado. Todos contentos hasta ahí:

$$E(\hat{Y}_{A\,ppt}) = \frac{1}{n}\sum\left(\sum_{i=1}^{N}\frac{Y_i}{p_i}p_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{i=1}^{N}Y_i$$

$$= \frac{1}{n}\,n\,Y_A = Y_A$$

De la misma manera se estima el total de ciudadanos que votaron por cualquier partido, esto se convierte en la variable Xi digamos, y se toma el cociente de estas dos estimaciones para estimar el porcentaje de interés.

Sin embargo, algunos colegas utilizaron otro tipo de estimador para estimar el porcentaje de votos a favor de un partido:

$$p_{A\,ppt} = \frac{1}{n}\sum\frac{y_i}{v_i}$$

v_i = *número de ciudadanos que votaron*

Donde V_i difiere de C_i, de acuerdo a la tasa de participación en cada sección electoral:

Ahora viene la pregunta que quiero dejar como una de las cosas que tenemos pendientes. El estimador usado por algunos colegas es distinto del que detallamos al principio. Aquí el denominador es las personas que

votaron por el partido, que es la clase que decía Raúl Rueda. Cada término de la suma es el porcentaje a nivel de la sección electoral en muestra, es el porcentaje de votos para la Alianza por el Cambio, por el PRI, por el PRD, para cualquiera de estos, en una sección electoral en particular.

¿Qué es lo primero que notamos diferente de la clase que estábamos viendo antes? Que el denominador ya es una variable aleatoria, ya no es un parámetro como en el modelo más sencillo. Antes era una constante, que era la medida de tamaño con la cual seleccionaba la sección electoral.

Esta pequeña diferencia hace que esto ya no sea una constante, el número de ciudadanos en el padrón o lista nominal, sino una variable; es simplemente el número de ciudadanos que en la sección electoral i ejercieron su derecho a votar; fueron, hicieron su cola y votaron. No lo conocemos *a priori* como sí *a priori* conocemos la medida de tamaño, es una variable aleatoria.

¿Qué pasa si yo tomo la esperanza a esa clase de estimadores? Bueno, uno de los viejos trucos que nos enseñan cuando estudiamos estadística, ¿cómo tomo la esperanza de algo bivariado? primero tomo una esperanza condicional, condiciono sobre una de las variables para mantenerla fija, tomo la esperanza sobre la otra variable, y si las cosas me ayudan y se cancelan los términos problemáticos, al tomar la segunda esperanza todo mundo vive feliz.

Entonces, si yo tomo la esperanza de esta clase de estimadores, esto es la esperanza de cada uno de estos términos, la condicional, estoy condicionando en que es conocido el total de ciudadanos que ejercieron su derecho a voto en cualquiera de las secciones electorales, el resultado es el siguiente:

$$E(p_{A\ ppt}) = \frac{1}{n} \sum \left(\sum_{i=1}^{N} \frac{Y_i}{v_i} p_i \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{i=1}^{N} \frac{Y_i}{v_i} \frac{C_i}{C}$$

Donde V_i difiere de C_i de acuerdo a la tasa de participación en cada sección electoral.

Pero una de las inconveniencias del resultado es que los términos molestos no se cancelan.

Queda pendiente ahora tomar sobre estas sumas la esperanza que falta, que es sobre los posibles valores del número de ciudadanos que votan, es una esperanza que todavía está pendiente de tomar. Esto requiere algún tipo de modelo de distribución.

Aquí estamos corriendo riesgos fuertes, cuando estamos utilizando en nuestros sistemas de estimación de ejercicios de conteo rápido estimadores de esta clase, cuya esperanza es ésta: ¿Y luego cómo resolvemos la esperanza pendiente? La variable aquí es 1 sobre el número de ciudadanos que ejercieron su derecho a votar. ¿Cómo se distribuye eso? Podemos empezar a modelarlo, qué les parece que el número de ciudadanos que votaron en la sección electoral i, es una función muy sencilla que depende del tamaño del padrón de la sección electoral, el número de ciudadanos en esa sección por una tasa de participación.

$$v_i = \beta_i C_i$$

Le debemos sumar un término de error a ese modelo.

Si sustituimos este modelo en la expresión me permite cancelar las C_i pero deja el pendiente de: ¿Qué vale tu parámetro de tasa de participación a nivel de cada una de las secciones electorales? Ah, pues sigo simplificando, porque nos encanta simplificar. Bueno, pues vamos a usar otro modelo, vamos a suponer que la tasa de participación es pareja y que en cualquier sección electoral de este país la tasa de participación es la misma:

$$v_i = \beta C_i$$

Entonces, usemos este modelito y, entonces, ya las cosas son muy bonitas porque al sustituir se cancela lo problemático y ya puedo, entonces, decir que esta expresión se reduce, y regreso a un estimador que va a ser simplemente un promedio del total del porcentaje, del total del padrón, multiplicado por la tasa de participación.

Bueno, pero estas simplificaciones y este modelaje –insisto– han ignorado el término de error que estos modelos tienen. Yo anoche le preguntaba a algunos que estaban alrededor de la mesa mientras cenábamos, si una variable se distribuye como una normal, el término de error que me falta poner aquí, ¿cómo se distribuye a la inversa? Algún día lo sabía, ya se me olvidó, llamémosle una normal inversa.

Esto implicaría que agregáramos un modelo adicional sobre el término de error, que nos quedaría pendiente aquí de modelar y de estimar. Ese es uno de los pendientes que tenemos, desde mi punto de vista, para un ejercicio que consideramos que es sencillísimo de una estimación de conteo rápido usando esta clase de estimadores.

Los ejercicios que hicimos en el IFE no usaron esta clase de estimadores. Ya presentó Raúl Rueda la clase de estimadores que usamos en el ejercicio del conteo rápido para el IFE. A diferencia de estos, que son estimadores, que se llaman separados, estimadores de razón separados porque es el promedio de cocientes, donde numerador y denominador son variables, el que para el conteo del IFE, y que puso en una de sus láminas Raúl Rueda, es lo que se llama un estimador combinado, se estima todo, el numerador de manera independiente de la estimación de todo el denominador, por un lado estimo los votos a favor de la Alianza por el Cambio y después, de manera independiente, estimo el total de votos para cualquier partido, más no registrados, más nulos.

Hay varias ventajas que en la literatura son conocidas de los estimadores combinados sobre los estimadores separados cuando el tamaño de muestra a nivel de cada estrato es mayor a 20. No es necesario el recurrir a la demostración, eso está disponible en varios textos.

Sin embargo, a pesar de eso, varios colegas que usaron conteos rápidos usaron el estimador de razón separado, que es menos preciso.

Ahora, si yo logré transmitirles mi inquietud de trabajar sobre este tema, de usar esa categoría de estimadores y resolver lo que no está resuelto de un error adicional que ya proviene del modelo, ya no proviene del muestreo, yo me doy por satisfecho.

Pero quisiera retomar ahora un par de preguntas nada más de la relación de las empresas que hicimos el conteo para el IFE con el Comité Técnico con el que estuvimos trabajando.

Un resultado también harto conocido es que la varianza de un diseño que tiene fijación óptima es menor o igual que la varianza de un diseño que tiene fijación proporcional y es menor o igual que la varianza de un diseño que usa un muestreo aleatorio simple.

Para el tipo de problemas que estamos discutiendo, esta relación se sostiene, es más eficiente una fijación óptima que una proporcional, que un muestreo aleatorio simple.

En el conteo del IFE el diseño, que fue responsabilidad del Comité Técnico, y ya Roy Campos lanzó a la mesa una pregunta que ya no es tanto técnica, ya es de una naturaleza más importante, ¿qué tanta responsabilidad puede asumir la empresa sobre su trabajo si se le quita la responsabilidad de diseñar la muestra? Creo yo que es una pregunta muy interesante que vale la pena discutirse, pero en el terreno técnico la pregunta valedera es por qué voy a usar un sistema de selección de secciones electorales con muestreo aleatorio simple, si sé que son más eficientes los discutidos.

Además, si uso aleatorio simple con proporcional, también es sabido que si se estratifica y se usa sistemático, donde el sistemático en el orden de la lista tiene una estratificación implícita, eso es más eficiente que un muestreo aleatorio simple con afijación proporcional.

Yo creo que en ejercicios tan relevantes como estimar parámetros del resultado de una elección, deben tener especial cuidado de revisar y tomarnos libertades que en otras condiciones no nos las tomamos. Yo creo que el valorar la muestra seleccionada y tener la capacidad de decir: la tiro a la basura porque es mala, es una responsabilidad para este tipo de ejercicios; y selecciono otra y la valoro y digo: sigue siendo mala, la tiro, y así hasta que sale una que es "buena".

Hay varios criterios y varios parámetros para juzgar esto y no creo que es faltarle a la estadística el usar esos criterios para decir el ejercicio de estimación que voy a hacer va a ser con el uso de esta muestra. Eso yo creo que es una de las facetas que en ejercicios futuros valdría la pena discutir. Yo creo que los colegas que no trabajan para clientes como el IFE, en las condiciones de este ejercicio en particular, probablemente varios de ellos lo hicieron, probablemente se tomaron el tiempo de examinar las muestras seleccionadas y desechar varias para quedarse con algunas de las buenas.

En el IFE no hubo esa oportunidad. Se nos dio la muestra asignada. Ya se comentó que fue responsabilidad del Comité Técnico. Yo examiné la muestra particular que a nuestra empresa le tocó y a nivel global se veía muy bien, pero cuando la analizábamos por estado, por distrito electoral, y comparábamos las composiciones urbanas, mixta, rural, de las secciones electorales *versus* la clasificación del padrón, empezaban a surgir discrepancias serias y fuertes que pudieron haber impactado el proceso de estimación, que felizmente no fue el caso; pero fue un felizmente *a posteriori*, cuando se pudo haber prevenido y tener más certeza, un grado mayor de certeza, de que no podrían pasar cosas desagradables.

Creo, entonces, el segundo tema que quiero dejar asentado como tema a discusión es la pertinencia de que en ejercicios tan relevantes, por lo que dijo Raúl Rueda del costo político que tienen, se tome el tiempo para valorar muestras *a priori* y desechar las que nos pueden poner en riesgo el ejercicio de estimación. ■