

PONENCIAS MESA 4

Iván Castro Rivadeneyra

BIMSA PARA CIRT

Sería redundante describirles, en esta oportunidad, los resultados que obtuvo Bimsa el 2 de julio, ya que fueron públicos en su momento y conocemos de su precisión. Además, resultaría impropio enfatizar su exactitud, cuando Roy Campos ya nos explicará cuáles son los objetivos reales de un ejercicio de conteo rápido. Entonces, aunque Bimsa fue la empresa más precisa, no será este el tema que deseamos tratar.



Este es un mensaje para aquellas personas que en la mesa No.3 expresaron temores cuando se hablaba de las encuestas preelectorales y se consideraban filtros, ponderadores, el cuestionario, la selección de individuos, etc. Afortunadamente y desde el punto de vista estadístico, un conteo rápido es mucho más sencillo. Entonces, gracias a Dios estamos discutiendo solamente un conteo rápido.

¿EN QUÉ RADICA EL ÉXITO?

- Diseño Muestral
- Definición de estimadores
- Transmisión rápida y adecuada de información

BIMSA

DISEÑO DE MUESTRA

- Conglomerados y Estratos.
 - Secciones
 - Nivel de urbanización
- Selección de secciones
 - (PPT) usando la lista nominal
 - muestreo sistemático
 - dispersión geográfica

BIMSA ■

¿En qué radica el éxito de un conteo rápido? Básicamente en tres partes: primero, en el diseño de la muestra; después, en la definición de los estimadores (asunto que está íntimamente ligado al diseño muestral); y finalmente, y no por último menos importante, la transmisión rápida y adecuada de la información, la cual es un problema fundamentalmente de logística.

De esta manera, bajo la hipótesis de que el tercer punto se cumple en forma satisfactoria, centraremos la atención exclusivamente en la parte de diseño muestral y en la definición de estimadores.

Con respecto al diseño de muestra, creemos que es muy claro que, dada la estructura de la información, el diseño muestral debe estar definido a partir de conglomerados y estratos. Conglomerados, porque la preferencia individual es accesible a través de la suma de los votos a partidos en cada una de las casillas correspondientes a las secciones electorales de la muestra. Estratos, debido a las diferentes velocidades de transmisión de los datos de acuerdo al nivel de urbanización.

Resulta inmediato pensar que éste no es un problema de muestreo de secciones electorales, sino de votantes, es decir, al final, es el comportamiento del voto de las personas sobre lo que queremos inferir. Por lo tanto, debido a que esta información se presenta agregada, el tipo de muestreo requerido hace referencia a grupos o conglomerados de votantes.

De esta forma, el conjunto de secciones electorales constituye el conjunto de conglomerados que, a su vez, se identifican como las unidades primarias de selección. A este respecto, el definir distritos electorales como unidades primarias de selección asocia un beneficio relativo al control de la dispersión final de la muestra, pero a costa de un incremento importante en el efecto del diseño, por lo que resulta arriesgado.

¿Por qué además estratificar? Específicamente, la estratificación explícita de la población, a través del nivel de urbanización, ayuda en términos del equilibrio asociado al estimador, el cual está definido a partir de ponderadores que alivian el problema de desfase

en la accesibilidad de la información por estrato, situación que por ejemplo tiene el PREP, que inicia con la información de secciones urbanas y termina con las rurales. Lo anterior, aunado al hecho de que la estratificación en grupos homogéneos de poblaciones heterogéneas, conlleva un beneficio en términos de error de estimación.

El problema siguiente es especificar cómo se lleva a cabo la selección de secciones electorales. En ese sentido y hasta este momento, con la información de la votación del 2 de julio, es posible analizar (en términos del error cuadrático medio) cuál habría sido el mejor método de selección de secciones. Por ejemplo, valdría la pena averiguar si una selección PPT (con probabilidad proporcional al tamaño de la sección) es mejor a una selección aleatoria simple, como la literatura al respecto generalmente sugiere. En particular, Bimsa usó un modelo PPT. Dado que las diferencias de tamaño entre secciones no son marcadas, se podría esperar en realidad una diferencia sutil entre los dos métodos.

Otro aspecto a considerar es la manera en cómo se obtiene la muestra PPT. Bimsa, generalmente, hace uso de un método sistemático que permite una ordenación anticipada de la base de secciones completa, la cual induce una estratificación implícita en la selección. Cabe enfatizar que dicha estratificación implícita, aplicada correctamente, puede pagar enormes dividendos en términos de precisión.

El 2 de julio, Bimsa trató de lograr para la muestra una máxima dispersión a través de la ordenación de la base por criterios geográficos. Ésta trabajó correctamente, sin embargo se podría pensar que maximizar la dispersión geográfica no puede resultar siempre en lo óptimo. Al respecto, si finalmente se desea inferir sobre información de votación, la mejor estratificación implícita debería estar en términos de una estratificación política, basada en procesos anteriores.

ESTIMADORES

Estimador H.T. de porcentajes a candidatos

$$p(x_j) = \sum_{k=1}^K S_k \left(\frac{\sum_{i=1}^{n_k} p(x_{ijk})}{n_k} \right)$$

- $p(x_j)$: es la proporción de votos obtenida por el candidato j .
- $p(x_{ijk})$: es la proporción de votos para el candidato j obtenidos en la i -ésima sección del estrato k -ésimo.
- S_k : es el ponderador del k -ésimo estrato.
- n_k : es el número de secciones muestreadas en el k -ésimo estrato.
- K : es el número de estratos.

BIMSA ■

Todo lo anterior define los estimadores que a continuación se revisan. Si se observa la expresión entre paréntesis, se puede verificar que los estimadores de porcentajes dentro de cada estrato son el promedio simple de los porcentajes asociados a cada candidato en cada una de las secciones en el estrato.

El resultado anterior refleja el método de selección de secciones basado en un muestreo PPT. De hecho, la definición de estimadores y varianzas para este caso, son expresiones muy sencillas, lo que representa una ventaja más del muestreo PPT.

ESTIMADORES

$$p(x_j) = \sum_{k=1}^K S_k \left(\frac{\sum_{i=1}^{n_k} LN_{ik} p(x_{ijk}) LN_k / LN_{ik}}{n_k LN_k} \right)$$

BIMSA ■

Ahora bien, el doctor Edmundo Berumen nos hacía la observación de que se tendría que revisar con cuidado cómo se definen e interpretan esos porcentajes en cuanto al hecho de que el tipo de selección obedece a un diseño PPT. En ese sentido, el considerar la abstención como candidato obliga a calcular porcentajes sobre una base referida al total en lista nominal, el cual es también utilizado en la selección, evitando así cualquier incongruencia entre el método de selección y los estimadores.

Haciendo referencia a los estimadores tipo Horvitz-Thompson (H.T.), esta lámina presenta las probabilidades de inclusión de manera explícita (la expresión en rectángulos). Los estimadores tipo H.T. están definidos sobre totales y es por eso que, dado que se está definiendo un estimador para el porcentaje asociado a cada candidato (sobre la base de lista nominal), aparecen otros términos que se anulan de acuerdo a los círculos marcados, regresando así a la expresión original.

ESTIMADOR DE PORCENTAJES A CANDIDATOS DENTRO DE SECCIONES

$$p(x_{ijk}) = \frac{x_{ijk}}{\sum_{j=1}^J x_{ijk}}$$

x_{ijk} : Es el total de votos obtenidos por el candidato j en la i -ésima sección del estrato k -ésimo.

J : Es el total de candidatos incluyendo a los no registrados como otro único, a los votos nulos como otro único y la abstención como otro.

BIMSA ■

A continuación se observa la definición del porcentaje a cada uno de los candidatos, incluyendo a la abstención como tal.

Pensando en la forma de presentación de información, es conveniente subrayar que los porcentajes referidos a una base de lista nominal no son comparables directamente con los porcentajes oficiales, a excepción de la abstención. Una alternativa es hacer trabajar con estimaciones de totales, a los que se les pueden asociar pruebas de hipótesis de manera directa (según el material aquí presentado), orientadas al total de la diferencia entre los candidatos de la primera y segunda fuerza. Una segunda alternativa es hacer uso de estimadores de razón para ajustar la base de porcentajes a la estimación del total de votos. Sin embargo, esta opción llevaría, para la realización de pruebas de hipótesis, a la definición de expresiones aproximadas del cálculo de errores relacionados a estimadores de razón que en esta plática no se presentan.

ESTIMADORES

$$S_k = \frac{LN_k}{\sum_{k=1}^K LN_k}, \quad R = \frac{\sum_{k=1}^K LN_k R_k}{\sum_{k=1}^K LN_k}, \quad R_k = \frac{\sum_{i=1}^{n_k} TV_{ik} / LN_{ik}}{n_k}$$

LN_k : Es la lista nominal del k -ésimo estrato

LN_{ik} : Es la lista nominal de la sección i del k -ésimo estrato

R_k : Estimación de la participación en el estrato k .

TV_{ik} : Total de votos obtenidos en la sección i del estrato k incluyendo a los votos nulos y a candidatos no registrados.

R : Estimación de participación

BIMSA ■

En la presente lámina se puede observar la definición explícita del estimador de participación y de los ponderadores fijos de estratos, basados en la lista nominal.

ESTIMADORES

Prueba de hipótesis de diferencias significativas

$$\sigma^2_{1-2} = \sum_{k=1}^K S_k^2 \left(\frac{\sum_{i=1}^{n_k} (C_{ik} [I-2])^2}{n_k^2 - n_k} \right)$$

donde

$$C_{ik} [I-2] = A_{ijk}(1) - A_{ijk}(2), \quad A_{ijk}(1) = p(x_{ijk}) - \frac{\sum_{i=1}^{n_k} p(x_{ijk})}{n_k}$$

BIMSA ■

Con relación a la prueba sobre ganador, esta lámina presenta el estimador de la varianza de la diferencia entre porcentajes de primera y segunda fuerza. Una expresión muy similar se obtendría si se decidiera realizar la prueba sobre totales.

ESTIMADORES

Prueba de hipótesis de diferencias significativas

Finalmente, el estadístico de prueba (para la primera diferencia) es:

$$T = \frac{p(x_1) - p(x_2)}{\sigma_{1-2}}$$

Para el cual se calcula el valor p asociado a una prueba de hipótesis de una sola cola utilizando una distribución *t de student*, de manera que si éste es menor a 0.005 (lo que equivale a un nivel de confianza del 99.5%) se concluye una diferencia significativa.

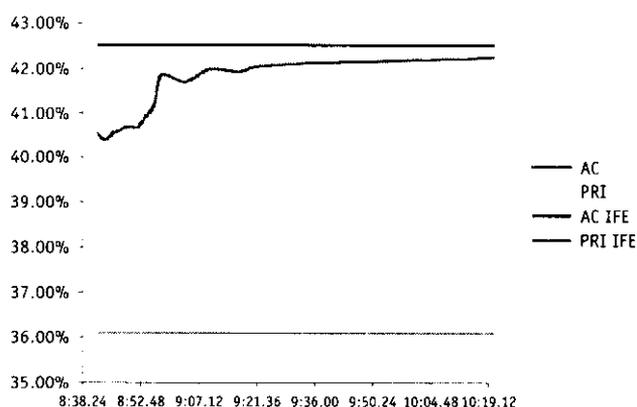
BIMSA ■

La varianza estimada se usa para llevar a cabo una prueba "t" orientada a verificar si la diferencia entre los porcentajes de primera y segunda fuerza puede considerarse estadísticamente significativa.

El nivel de confianza manejado por Bimsa fue del 99.5 por ciento. Dadas las consecuencias de un posible error en la interpretación de la prueba, y como la estadística clásica lo permite, se protegió la hipótesis nula de igualdad de porcentajes. Variar este nivel es una cuestión personal de evaluación de riesgo.

Este fue el comportamiento de los estimadores a lo largo del 2 de julio. Se puede observar que a partir de las 21:30 horas, aproximadamente, ya se tenía un comportamiento bastante estable.

EVOLUCIÓN DE ESTIMADORES CR



BIMSA ■

OBSERVACIONES PARA EXPERIENCIAS FUTURAS

- ¿Un esquema de selección irrestricta de secciones es mejor a una PPT? Cálculo de varianzas para datos reales.
- No cualquier submuestra es una muestra aleatoria de la población de interés, por lo que pruebas de dispersión son cruciales.
- Deberán manejarse esquemas de pronóstico cuando no se tenga la totalidad de la muestra por medio de técnicas que vayan más allá del muestreo.

BIMSA ■

¿Qué se tendría que revisar para ejercicios futuros? En primer lugar, ¿realmente un esquema de selección estricta de secciones es mejor a un PPT? Cuando los datos completos de la votación estén disponibles se podrán calcular las varianzas (que en su momento fueron teóricas) y medir los beneficios de cada diseño.

Además, y muy importante, se aprendió que no cualquier submuestra es una muestra aleatoria de la población de interés. Entonces, comprobar la aleatoriedad de las submuestras es crucial, lo cual se pueden llevar a cabo a través de una prueba de bondad de ajuste sobre la distribución de grupos de dispersión. Ésta verifica si realmente las submuestras con las que se cuenta en el momento de estimación son adecuadas para hacer inferencias.

En este mismo sentido se podrá pensar en métodos de pronóstico que salgan del esquema de muestreo y hagan uso de ajustes multivariados que anticipen el comportamiento de la votación en secciones para las cuales no se tenga información disponible.

Asimismo, se requiere probar estratificaciones más complejas, por lo que ya se mencionaba que no necesariamente las estratificaciones geográficas son las mejores.

OBSERVACIONES PARA EXPERIENCIAS FUTURAS

- ¿Un esquema de selección irrestricta de secciones es mejor a una PPT? Cálculo de varianzas para datos reales.
- No cualquier submuestra es una muestra aleatoria de la población de interés, por lo que pruebas de dispersión son cruciales.
- Deberán manejarse esquemas de pronóstico cuando no se tenga la totalidad de la muestra por medio de técnicas que vayan más allá del muestreo.

BIMSA ■

Con relación al cálculo de errores *a priori*, es muy arriesgado asegurar límites de error específicos (ya que estos dependerán de la variabilidad registrada en el proceso), por lo que lo más honesto es decir: dados los resultados de variabilidad en experiencias pasadas, las mejores estimaciones del error posible son tales.

Por otra parte, el único efecto negativo que podría tener el considerar la abstención como candidato en un proceso de estimación, está relacionado con la presencia de casillas especiales, cuando vota gente en la sección que no está en el listado nominal correspondiente, sin embargo, se podría considerar que dicho efecto es mínimo.

Finalmente, la incorporación de pruebas no paramétricas y bayesianas (como ya se vio) podrá mejorar estos sistemas.

A manera de conclusión se pueden subrayar los siguientes puntos:

Los diseños PPT de selección de secciones llevan a la definición de estimadores sencillos de porcentajes y estimadores de varianzas.

Para la prueba de diferencia significativa entre primera y segunda fuerza, se pueden usar expresiones exactas de los estimadores de varianzas para porcentajes a candidatos sobre la base de lista nominal o sobre totales de votos.

La incorporación de nuevas técnicas estadísticas encaminadas a esquemas de pronóstico, mejorarán las estimaciones a horas tempranas de la noche en el día del proceso electoral. ■