Introducción a la base de datos EQxIS, algunos conceptos básicos de estadística

Las principales características del sistema EQxIS son las siguientes:[1]

(i) EQxIS permite hacer desagregaciones por: quintiles de ingreso, género, raza/etnicidad, área urbana/rural y regiones;

(ii) Tiene varias opciones de presentación con: Mapas, gráficas, cuadros, brechas de acceso y sumarios de las estadísticas principales de un país;

(iii) Los cuadros, contienen varias medidas estadísticas: media, error estándar, coeficiente de variación, número de observaciones e intervalos de confianza;

(iv) Tiene una cobertura para varios países de América Latina y distintos años con los indicadores que forman parte de los Objetivos de Desarrollo del Milenio que se pueden calcular desde las encuestas de condiciones de vida.

Algunos conceptos básicos en estadística

Una revisión somera de algunos términos estadísticos puede ser útil para interpretar mejor los indicadores contenidos en EQxIS. Los indicadores en EQxIS se construyeron a partir de encuestas muestrales – la selección de una parte representativa de la población a ser encuestada. Sus estimaciones provienen de muestras de población y no de censos de toda la población.

Toda estimación a partir de muestras está sujeta a *errores muestrales*. Los censos no adolecen de esta falta, pero su tamaño los hace costosos y por tanto incluyen pocas preguntas sobre la población y las características de los hogares. El error muestral es el costo que hay pagar para tener información más oportuna y detallada. Cuando se

[1] La base de datos EQxIS fue creada por el programa MECOVI del Banco Interamericano de Desarrollo con apoyo inicial del Programa de Naciones Unidas para el Desarrollo. Todas las estimaciones están basadas en encuestas de hogares, y por tanto, algunos indicadores pueden carecer de suficientes observaciones.

selecciona una muestra se define su tamaño, número de hogares o personas a encuestar -- que identificamos como *n,* de tal forma que nos de una buena representación de la totalidad de la población, *N* número de hogares o personas. Ahora bien, puede haber una variedad de muestras de tamaño *n* diferentes una de la otra. La teoría estadística nos permite saber cuántas muestras diferentes puede haber, dado un tamaño de muestra *n* y un tamaño de población *N*.[2] Estimaciones de medias y promedios a partir de muestras pueden tomar valores relativamente distantes de las correspondientes media en la población y, por tanto, no ser representativas de ésta. Por supuesto que también se tienen estimaciones muestrales cuyos valores se encuentran bastante cercanos a aquellos de la población. Si bien se desconoce cuales son los valores medios en la población, la teoría estadística permite decir qué tan buena es una muestra para estimar los valores reales de la población que uno está intentando representar.

La teoría estadística nos enseña que conforme aumentamos el tamaño, *n*, de una muestra, su representación de la totalidad de la población aumenta, es decir, se incrementa su *representatividad*. En otras palabras, entre más grande es la muestra, sus estimaciones tenderán a ser mas precisas. La estimación de cualquier medida estadística con base en una muestra adolecerá de una desviación respecto del valor de dicha medida estadística calculada a partir de la población. La teoría estadística nos dice, entonces, que esta desviación se hará cada vez menor conforme el tamaño de la muestra se acerque al tamaño de la población (la diferencia entre *n* y *N* se haga cada vez más pequeña). Intuitivamente, el valor de una medida estadística, digamos la media, en una muestra puede diferir por que la muestra no haya incluido algunos casos existentes en la población. Si el tamaño de la muestra aumenta, la probabilidad de que dichos casos sean excluidos se reduce, y la precisión de la estimación mejora. Mientras más casos sean

---

[2] Del análisis de combinaciones, la fórmula de combinación está dada por:

$$C_{N,n} = \binom{N}{n} = \frac{N!}{n!(N-n)!},$$

En donde $C_{N,n}$ especifica la combinación de *N* elementos tomados *n* en *n*, i.e. que dice cuántas combinaciones diferentes de *n* elementos es posible obtener de una serie de *N* elementos.

incluidos más precisa será la estimación, acercándose cada vez más a su valor real y reduciendo la fluctuación de su valor en torno del valor real poblacional.

El número de observaciones contenidas en una encuesta indica ya el tamaño de la muestra. Comparando este número con el tamaño de la población, se puede tener una idea de qué tan buena es esa muestra. Por otra parte, la distancia media entre la estimación de la muestra y la estimación de la población, el *error estándar*, permite también darnos una idea de qué tan buena es dicha muestra.

A partir del concepto de error estándar se desarrolla el concepto de *coeficiente de variación*. Este se define como el resultado, coeficiente, de dividir error estándar entre su valor medio estimado.[3] Es fácil ver la utilidad de este concepto. Si se tiene un error estándar de más/menos 5 respecto de un valor medio estimado en 30, esta estimación será más robusta que otra que, teniendo el mismo error de más/menos, arroje un valor medio estimado de 10.

Tomando en cuenta que tanto el error estándar como el valor medio están calculados en las mismas unidades, el coeficiente de variación no depende de la unidad de medida y es por tanto, un indicador más confiable. El coeficiente de variación describe el tamaño relativo de una fluctuación y, por ende, proporciona una buena idea de que tan precisa es una estimación basada en una muestra.

Probablemente la herramienta más útil para saber que tan precisa es una estimación con respecto al valor real del indicador en la población, es el *intervalo de confianza*; especialmente cuando se trata de comparar entre estimaciones para diferentes sub-grupos. El intervalo de confianza se construye alrededor de una estimación muestral definiendo un rango, el intervalo, dentro del cual puede estar contenido el valor real estimado en la población. El rango especificado puede ser grande o pequeño. Mientras más grande sea el rango, mayor confianza podremos tener de que el valor real correspondiente a la población se encuentre dentro de ese rango. Pero, por el contrario, menor utilidad tendrá

---

[3] Formalmente, si σ es el error estándar de una variable, digamos, *X,* y μ es el valor medio, entonces el coeficiente de variación *CV* está dado por

$$VC = \frac{\sigma}{\mu}.$$

para la interpretación. Un intervalo de confianza del 95% construido en torno a una estimación muestral indica que la probabilidad de que dicho intervalo, que contiene la estimación muestral, contenga también el valor real correspondiente al total de la población es de 95%; es decir, que de cada 100 casos, en 95 de ellos la estimación real de la población estará dentro del intervalo definido.

Como señalamos anteriormente, las estimaciones de indicadores con base en muestras de población, pueden variar de una muestra a otra – aun cuando ambas seas muestras de la misma población. Algunas estimaciones muestrales estarán muy cercanas a aquellas de la población y otras no lo estarán tanto. Pero lo que es importante resaltar es que si bien las estimaciones puede diferir, el hecho de que provengan de la misma muestra hace que estén de alguna manera relacionadas.

Por ejemplo, cuando se calcula el valor de la media de una variable, nunca será más grande que el valor más alto en la población, o más pequeño que su valor mínimo. Además, también se puede decir que los valores extremos, aquellos que están distantes de la media de la población, tienen menos probabilidades de estar presentes en un intervalo de confianza dado. En el uso del intervalo de confianza se pueden cometer dos tipos de errores. Uno de ellos es rechazar una hipótesis que es verdadera. A este error se le llama error de rechazo, o error tipo I. El otro error es aceptar una hipótesis que es falsa. A este se le llama error de aceptación o error tipo II.

Supongamos que se tomó una muestra de una población y se quiere saber si la probabilidad de ser pobre de las familias que residen en dos poblados diferentes, ambos captados en la muestra, es la misma. Supongamos que en efecto la probabilidad de ser pobre es la misma en ambos poblados. Si la pobreza la medimos por el ingreso, estimaremos el ingreso medio de las familias en el poblado A y B a partir de la muestra tomada. Nuestras estimaciones estarán sujetas al error muestral, y nos darán una estimación de un ingreso medio X en el poblado A y un ingreso Y en el poblado B. Si nuestra estimación muestral de ingreso medio X e Y no son exactamente iguales, corremos el riesgo de cometer el error de concluir que la probabilidad de ser pobre es mayor en un poblado que en otro. Si definimos un intervalo de confianza, digamos del 99%, para cada una de las estimaciones y resulta que dichos intervalos no se traslapan, concluiremos erróneamente que dicha probabilidad es diferente. Para reducir este error,

sin embargo, podemos simplemente ampliar el intervalo de confianza en los dos poblados. Sin embargo, al aumentar el intervalo de confianza, aumentamos el riesgo de cometer el error de aceptación, el error Tipo II. En este contexto, y suponiendo ahora que la probabilidad de ser pobre realmente es diferente de un poblado a otro, al aumentar el intervalo de confianza estaríamos aumentando la probabilidad de que los intervalos de confianza se traslapasen y de concluir, erróneamente, que la probabilidad de ser pobre es la misma.

No es posible saber *a priori* cual es el intervalo de confianza que minimice el riesgo de cometer errores Tipo I y Tipo II. El punto es que la estimación del indicador para la población no se conoce. No obstante, lo que si se puede decir es que el riesgo de cometer cualquiera de estos dos errores se reduce cuando aumentamos el tamaño de la muestra. Intuitivamente, conforme aumentamos el tamaño de la muestra, es menos probable obtener muestras que contengan valores extremos, muy distantes de la media real, y por tanto nuestra estimación va a mejorar.

Una de las ventajas de EQxIS radica en que proporciona intervalos de confianza de las indicadores estimados para distintos subgrupos de la población. Supongamos que estamos estimando la tasa neta de asistencia a educación primaria para dos diferentes quintiles de ingreso, y que obtenemos unas tasas de 96% y 95%. ¿Significa esto que la tasa neta de asistencia es realmente más grande en un quintil que en otro? Atención. Esta diferencia puede ser consecuencia, solamente, de la forma en que la muestra captó estas poblaciones. Si el intervalo de confianza nos dice que con un 90% de confianza la tasa neta de asistencia de uno de los quintiles estará entre 95.2% y 96.8% y la del otro entre 94% y 96%, entonces, podremos rechazar, con un 90% de confianza estadística, la hipótesis de que las tasas de asistencia no son iguales.