Principios de Muestreo

Olivier Nuñez

8 de junio de 2018

- **Reducción de costes/tiempo**: en muchos casos, es demasiado caro y poco factible obtener datos de cada unidad de la población (un censo).
- **Mayor fiabilidad**: mejor control de la calidad del proceso de recogida.

¿Qué es el muestreo?

Algunas definiciones

- *"Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population."* (**S. Thompson, 1992**).

- *"We are all accustomed to the idea of sampling in everyday life. The housewife visually samples the quality of the fruit she intends to buy. (...) If the greengrocer puts the best on display and sells us inferior qualities, we protest at the biased sample or change our supplier. (...) The notion of bias is not long the notion of sampling itself."* (**A. Stuart, 1983**).

Objetivo del muestreo

Representatividad «al mejor precio»

Las técnicas de muestreo permiten:

- Evitar que se excluya sistemáticamente parte de la población.
- Maximizar la representatividad (muestra = población a pequeña escala).
- Reducir los costes de recogida de datos.
- Controlar el **error muestral**.

Error muestral

Variación entre muestras

- Es consecuencia de nuestra observación parcial de la población.
- Corresponde a la variación de las conclusiones entre muestras.
- Es un error de naturaleza aleatoria que decrece a medida que aumenta el tamaño muestral.
- No debe ser confundido con el **sesgo** (o **error sistemático**)

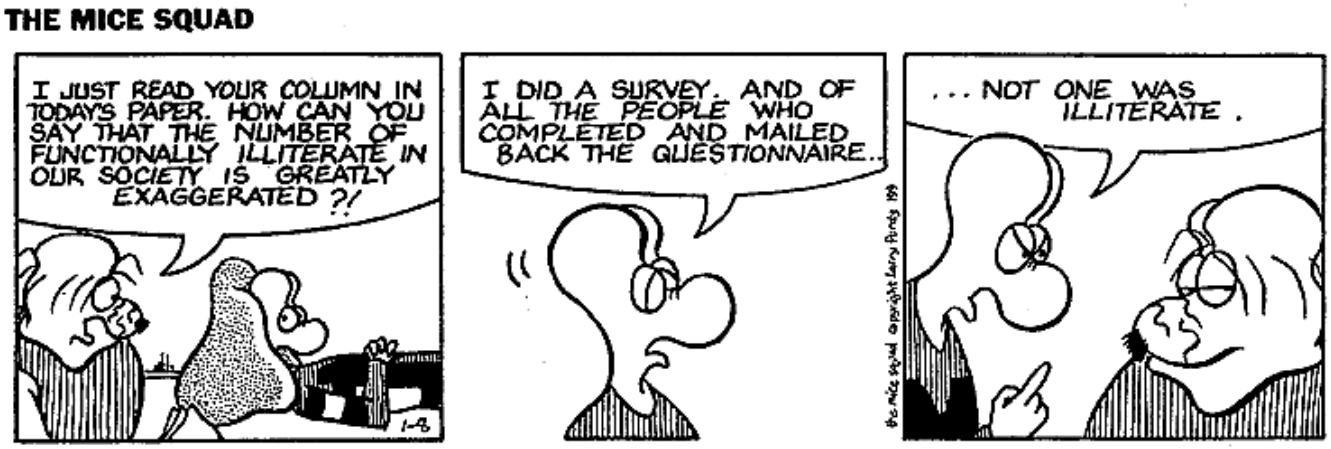
Error sistemático

Sesgo de selección y de información

- Distorsión en las conclusiones que ocurre de manera sistemática (afecta a todas la muestras).
- Se puede clasificar según el sesgo ocurre a nivel del muestreo o en la medición:
  - **Sesgo de selección**: ocurre cuando la selección de los individuos está condicionada por la característica que queremos medir.
  - **Sesgo de información**: sesgo que ocurre en la medición de la característica de interés. Sesgo de memoria, efecto del entrevistador o del cuestionario, ...
- Una selección **aleatoria** de las unidades permite evitar el sesgo de selección.

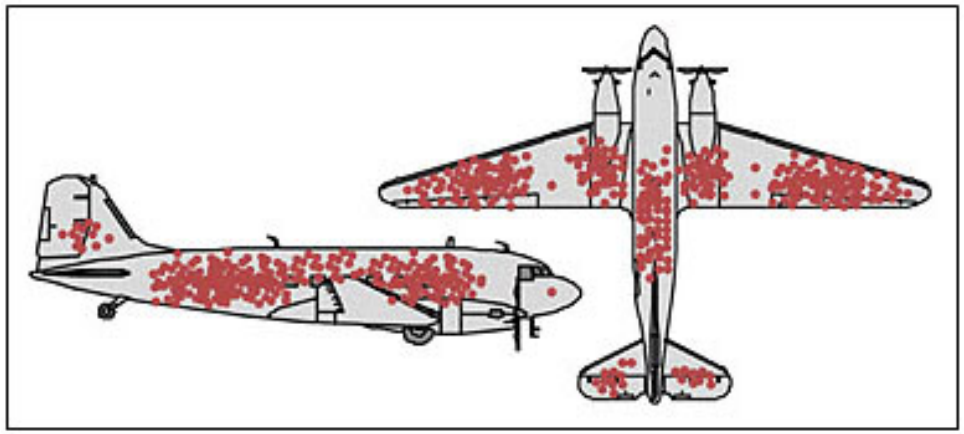
Sesgo de selección

No respuesta



Sesgo de selección

Sesgo del superviviente



Credit: Cameron Moll

Durante la Segunda Guerra Mundial, el estadístico Abraham Wald fue invitado a ayudar a los británicos a decidir dónde añadir armadura a sus bombarderos. Después de analizar los aviones que volvieron, recomendó reforzar los sitios donde no hubo impacto!

- **Población diana o general:** totalidad de unidades (individuos, empresas, productos, ...) sobre la cual nos interesa sacar conclusiones.
- **Población base o elegible:** conjunto de unidades sobre las cuales es posible obtener información (criterios de inclusión/exclusión). Si N es su tamaño, se puede identificar (al menos teóricamente) cada unidad por un número de 1 a N.
- **Muestra:** subconjunto observado de n unidades ($n \ll N$) de la población elegible.

Muestreo probabilístico

Imprescindible para evaluar el error muestral

- Con un muestreo **probabilístico**, cada unidad de la población tiene una probabilidad conocida (y no nula) de estar incluida en la muestra.
- Los muestreos **no-probabilístico** (de cuota, de conveniencia, ...)
  - Más fáciles de implementar y evitan la no respuesta.
  - Pero esta conveniencia es en detrimento del control del error muestral.
- A continuación, sólo se considerarán muestreos probabilísticos.

Definición: Es el diseño de muestreo más sencillo y se caracteriza por el hecho de que cada posible muestra de n unidades tiene la misma probabilidad de ser seleccionada.

Ventajas y limitaciones:

- Es un procedimiento **equiprobabilístico**: todas las unidades de la población tienen la misma probabilidad n/N de ser elegidas. Con lo cual, la media y la proporción poblacional se estiman directamente mediante la media y la proporción muestral.

- Limitaciones:
  - Requiere de una lista previa de las unidades de la población.
  - Poco eficiente (algunas muestras pueden ser poco representativas).

Muestreo aleatorio simple

Implementación

Se obtiene una muestra aleatoria simple de la siguiente manera:

1. Asignar un número de 1 a N a cada unidad de la población elegible.
2. Elegir n de estos números mediante el uso de algún proceso aleatorio (tablas o generador de números aleatorios)
3. Las unidades correspondientes a los números elegidos se toman como muestra.

Muestreo aleatorio con reposición

Variante del muestreo aleatorio simple

- Las unidades son seleccionadas una tras otra, devolviendo cada unidad antes de que se seleccione la siguiente.
- Con este diseño, una unidad puede aparecer más de una vez en la muestra!
- **Propiedades**:
  - Muestreo equiprobabilístico.
  - Pero, menos eficiente que el muestreo simple (muestras extremas).
  - Equivalente a un muestreo simple en una población de tamaño infinito.

«If we know nothing of the structure of the population apart from its size, we cannot do better than take a simple random sample»

Sin embargo, en general sabemos algo más sobre la población:

- La población se puede dividir en sub-grupos homogéneos (cada uno llamado estrato) y se espera que la variable de interés (ejemplo: sueldo) varíe entre los estratos (ejemplo: sexo).

- La población puede ser dividida en unidades primarias (ejemplo: hospitales) que muestran características similares y que son más fáciles/económicos de muestrear que las unidades elementales (ejemplo: pacientes).

Muestreo estratificado

Aumentando la representatividad de las muestras

- **Definición:** Para asegurar la representatividad de determinados sub-grupos o estratos de la población, se seleccionan por separado una sub-muestra dentro de cada estrato.

- **Principios de estratificación**: los estratos han de ser
  - Mutuamente excluyentes y exhaustivos.
  - Internamente homogéneos con respecto a la variable de interés y, por tanto, heterogéneos entre sí.
  - Definidos en función de variables fáciles de medir y relevantes para el objeto de estudio (tipo de municipios, área geográfica, ...).
  - En número reducido (rara vez resulta eficiente utilizar más de 5 estratos) y tamaño no muy pequeño.

Muestreo estratificado

Implementación

- La población de N unidades se divide en K estratos de tamaños $N_1, N_2, \ldots, N_K$.
- El tamaño muestral n se distribuye entre los estratos siguiendo un procedimiento de asignación:
  - **Asignación uniforme**: se selecciona el mismo número de unidades en cada estrato: $n_k = n/K$.
  - **Asignación proporcional**: se distribuye el de acuerdo al peso (tamaño) del estrato en la población: $n_k = n\,(N_k/N)$.
  - **Asignación óptima**: tiene también en cuenta la variabilidad en el seno de cada estrato.
- La selección dentro de cada estrato suele realizarse por muestreo aleatorio simple o sistemático

Muestreo estratificado

Propiedades

- La asignación proporcional produce un muestreo equiprobabilístico y por la tanto la media y proporción poblacional se estiman mediante la media y la proporción muestral.
- Para cualquier otra asignación, la estimación de parámetros poblacionales requiere de la inclusión de pesos (inverso de la probabilidad de selección).
- Para un mismo tamaño muestral, el muestreo estratificado facilita estimaciones más precisas (con menor error muestral) que el muestreo aleatorio simple.

Muestreo por conglomerados

Reduciendo los costes del muestreo

- **Principios:**
  - Suponemos que las unidades de la población están agrupadas por conglomerados (barrio, empresa, ...).
  - Idealmente cada conglomerado es una imagen a pequeña escala de la población.
- **Ventajas:**
  - No requiere de una lista con todas las unidades de la población.
  - Mejora la operatividad ya que las unidades muestrales están concentradas en los conglomerados.

Muestreo por conglomerados

Implementación

Para obtener un muestreo equiprobabilístico:

- Supongamos una población de N unidades agrupadas en M conglomerados de tamaños $N_1, N_2, \ldots, N_M$.
- Muestrear m conglomerados con probabilidad proporcional a su tamaño.
- Obtener una muestra aleatoria de tamaño n/m dentro de cada conglomerado seleccionado.