Muestreo y Variación Aleatorios

Febrero 28, 2024

Prof. Sergio Béjar

Departamento de Estudios Políticos, CIDE

Objetivo para Hoy

Comenzar nuestra discusión sobre estadística inferencial con muestreo y variación aleatoria.

Llegamos a la Estadística Inferencial

Aquí comenzamos una de las partes más interesantes/importantes para hacer investigación aplicada.

- Ya no sólo estamos interesados en estadística descriptiva.

Queremos hablar de **estadística inferencial.**

¿Qué es la Población?

Términos importantes.

- **Población:** el universo de casos que queremos describir.
- **Parámetro poblacional:** el segmento desconocido de la población que queremos estimar.

Por ejemplo, las opinión de los mexicanos sobre el despliegue de tropas del ejército en las calles o el desempeño del Congreso.

Muestreo de la Población

Problema: no tenemos datos de los más de 120 millones de mexicanos.

- Por lo que tenemos que usar una **muestra**, que es un subconjunto de casos extraidos de la población.
- Si lo hacemos correctamente, la **muestra estadística** nos va a dar un estimado del parámetro poblacional.

Cuando tenemos el universo de todos los casos posibles lo llamamos **censo**.

- Ejemplos: Decisiones de la Suprema Corte, el número de crísis económicas en el mundo.

La Muestra Aleatoria

El *muestreo* es un concepto fundamental en la ciencia política y las relaciones internacionales aplicadas.

- Pero, siempre existe la posibilidad de hacerlo incorrectamente.
- Y un muestreo incorrecto tiene implicaciones potenciales muy importantes.

Un proceso de muestreo apropiado es llamado **muestreo aleatorio**.

- Una muestra aleatoria no tiene **sesgos de selección.**

Una muestra aleatoria sin sesgos de selección garantiza que no haya error sistemático en la muestra y por lo tanto minimiza la probabilidad de errores en nuestras inferencias.

La Encuesta de Literacy Digest (1936)

El *Literacy Digest* buscaba saber quién iba a ganar la elección presidencial en EEUU en 1936.

- Obutuvo nombres y direcciones de todos los usuarios telefónicos y dueños de coches.
- Combinó esos nombres con los de los suscriptores de la revista para crear una muestra de 10 millones de personas.
- Mandaron por correo los cuestioanrios y recibieron 2.4 millones respondidos.
- **Su conclusión:** Alf London iba a ganar de calle.

El resultado actual: Roosevelt ganó de calle y *El Literary Digest* quebró algunos años después.

La Encuesta de Literary Digest (1936)

El problema de la encuesta está en su **marco muestral**.

- Aunque el marco muestral que usaron no es controvertio en estos días, el problema radicó en que estaba compuesto por una mayoría de seguidores de Landon.

Enviar encuestas por correo tampoco ayuda porque genera potencialmente **sesgos de respuesta**.

- Los que *realmente* detestaban Roosevelt tenían más probabilidad de responder.

La Encuesta de Literacy Digest (1936)

En otras palabras, la muestra estadística era sistemáticamente diferente de los parámetros de población reales.

- Esto es un **sesgo** de libro de texto en la estadística inferencial.

¿Qué hacemos para no cometer ese error?

Obtención de Muestras Aleatorias

Las firma encuestadoras usan generadores de números aleatorios.

- Cada unidad elegible de la población es asignada con un número.
- El generador de números aleatoriamente selecciona n números entre la población.

Dado que el proceso es estrictamente aleatorio, cualquier diferencua entre la muestra y los que no son seleccionados también es aleatoria.

Error de Muestreo

La eliminación del sesgo *no* elimina completamente el error.

- El muestreo aleatorio introduce un **error de muestreo aleatorio** a propósito.
- No es perfecto, pero tener un error de muestreo sistemático es mucho peor.

El error de muestreo aleatorio lo podemos estimar.

Entendendo el Error de Muestreo Aleatorio (E.M.A.)

El parámetro poblacional que nos interesa se define así:

- Parámetro poblacional = Estadística de la muestra + E.M.A.
- "E.M.A." = error de muestreo aleatorio.

Hay dos factores que debemos considerar cuando medimos el E.M.A.

1. El tamaño de la muestra.
2. La variación en el parámetro poblacional.

$$\text{E.M.A.} = \frac{\text{Componente de variación}}{\text{Componente del tamaño de la muestra}}$$

Entendiendo el Error de Muestreo Aleatorio (E.M.A.)

Notemos que al incrementar el tamaño de la muestra **reducimos** el error de muestreo aleatorio.

- Sin embargo, el efecto no es lineal.
- El componente del tamaño de la muestra es la raíz cuadrada del número de observaciones en la muestra.

$$\text{Componente del tamaño de la muestra} = \sqrt{n}$$

Tamaño de la Muestra

Manteniendo todo lo demás igual, un incremento en el tamaño de la muestra de 100 a 400 reduce el error de muestreo solo el doble.

```
28/sqrt(100)
```

```
## [1] 2.8
```

```
28/sqrt(400)
```

```
## [1] 1.4
```

Implicación: hay que incrementar el tamaño de la muestra lo más que podamos.

Tamaño de la Muestra

Pero incrementar el tamaño de la muestra es muy costoso.

- Tan costoso que puede incentivar a usar un muestreo no aleatorio.

Por eso vemos que la mayoría de las encuestas tienen un tamaña de muestra entre 1,000 y 3,000.

Tamaño de la Muestra

```
(28/sqrt(100))/(28/sqrt(1000))
```

```
## [1] 3.162278
```

```
(28/sqrt(1000))/(28/sqrt(10000))
```

```
## [1] 3.162278
```

Variación de la Muestra

Si el componente de variación se incrementa, el error de muestreo aleatorio se incremeta.

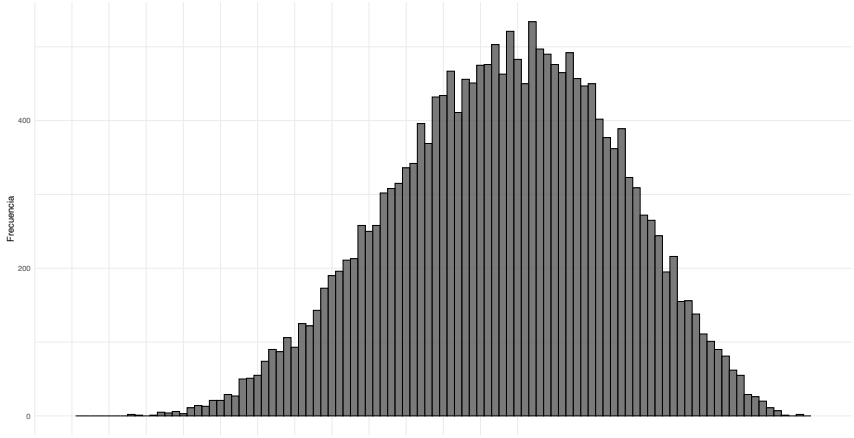
- Hay un estadístico aue mide esta variación: la **desviación estándar**.

Vamos a comparar los siguientes histogramas

- Histogram 1: alta desviación estándar
- Histogram 2: baja desviación estándar

Desviación Estándar Alta

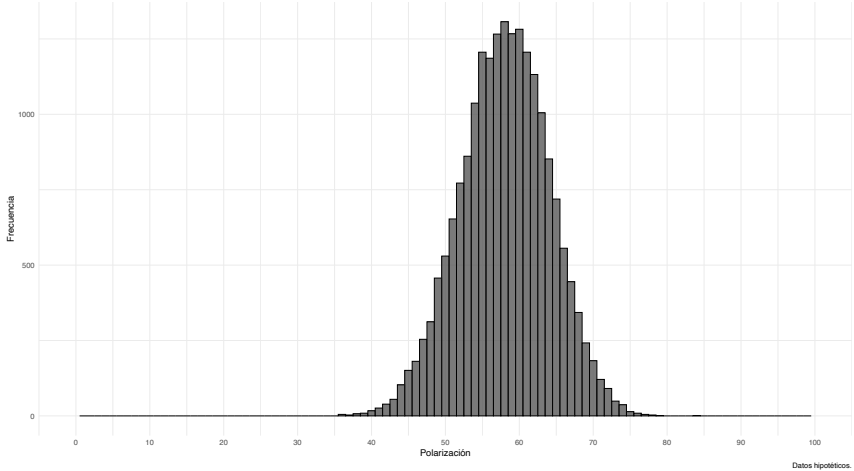
Datos simulados para tener una frontera de distribución normal (distribución beta) con media de 58 y desviación estándar de 15.



Datos hipotéticos.

Desviación Estándar Baja

Datos simulados para tener una frontera de distribución normal (distribución beta) con media de 34 y desviación estándar de 6.



Calculando la desviación Estándar

Vamos a calcular la desviación estándar

- Asumimos que conocemos los parámetros poblacionales, N y μ.

La notación anterior significa lo siguiente:

- N es el númro de casos en la población. n se refiere al tamaño de la muestra.
- μ es la medida de tendencia central en la población. $\bar{x}$ es la media muestral.

Las letras griegas se refieren a propiedades de la población no de la muestra.

Calculando la Desviación Estándar

1. Restamos μ de cada valor en la población.
2. Elevamos al cuadrado la diferencia para cada observación.
   - La suma de las desviaciones debe ser igual a 0.
3. Sumamos todas las desviaciones al cuadrado.
   - Esta es la suma de desviaciones al cuadrado.
4. Calculamos la media aritmética para la suma de desviaciones al cuadrado.
   - Esta es la **varianza**.
5. Sacamos la raíz cuadrada de la varianza.

Calculando la Desviación Estándar

Table 6-2 Central Tendency and Variation in Democratic Thermometer Ratings: Hypothetical Scenario B

Student	Democratic rating	Deviation from the mean	Squared deviation from the mean
1	25	–33	1,089
2	34	–24	576
3	50	–8	64
4	55	–3	9
5	56	–2	4
6	58	0	0
7	60	2	4
8	61	3	9
9	66	8	64
10	82	24	576
11	91	33	1,089
Summary information			
	Central tendency Summation of ratings = 638		*Dispersion* Summation of squared deviations = 3,484 Average of squared deviations (variance) = 316.7
	$N = 11$		
	$\mu = 58$		$\sigma = 17.8$

Error Estándar de la Media Muestral

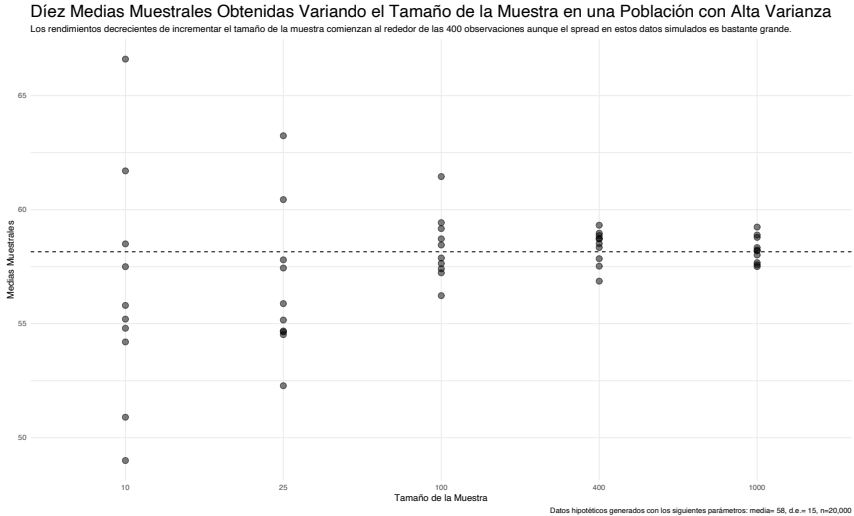
El muestreo aleatorio reduce el sesgo pero genera error aleatorio.

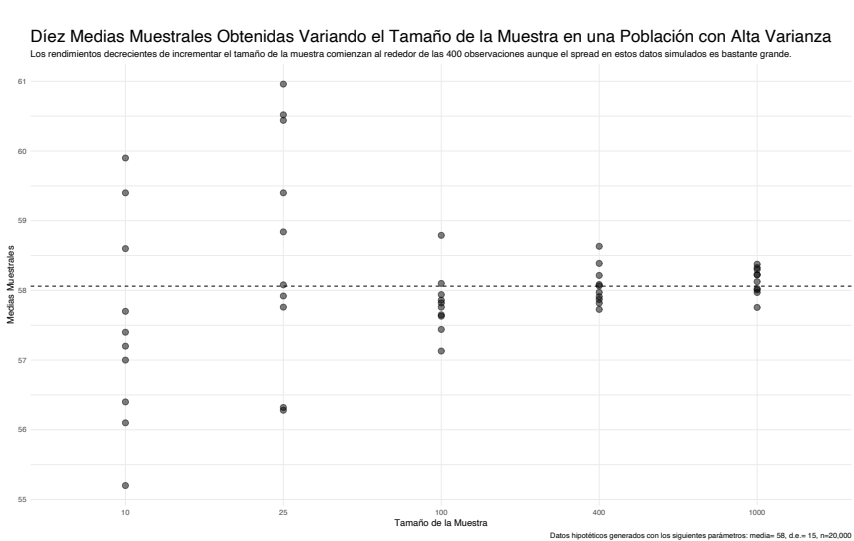
- Queremos eliminar el error aleatorio lo más posible.
- No es tan malo como el error sistemático pero sigue siendo ruido.

Y lo podemos hacer (de nuevo) incrementamdo el tamaño de la muestra.

- Consideremos los dos gráficos siguientes.
- μ es 58 in ambos paneles, pero la desviación estándar σ es mayor en el primero que en el segundo.

Es difícil eliminar la variación natural en la población, pero incrementando el tamaño de la muestra obtenemos estadísticas más creíbles.


Díez Medias Muestrales Obtenidas Variando el Tamaño de la Muestra en una Población con Alta Varianza
Los rendimientos decrecientes de incrementar el tamaño de la muestra comienzan al rededor de las 400 observaciones aunque el spread en estos datos simulados es bastante grande.
Medias Muestrales
65
60
55
50
10
25
100
400
1000
Tamaño de la Muestra
Datos hipotéticos generados con los siguientes parámetros: media= 58, d.e.= 15, n=20,000


Díez Medias Muestrales Obtenidas Variando el Tamaño de la Muestra en una Población con Alta Varianza
Los rendimientos decrecientes de incrementar el tamaño de la muestra comienzan al rededor de las 400 observaciones aunque el spread en estos datos simulados es bastante grande.
Medias Muestrales
61
60
59
58
57
56
55
10
25
100
400
1000
Tamaño de la Muestra
Datos hipotéticos generados con los siguientes parámetros: media= 58, d.e.= 15, n=20,000

Error Estándar de la Media Muestral

La fórmula para calcular el error de muestreo aleatorio es:

$$\text{Error Estándar de la Media Muestral} = \frac{\sigma}{\sqrt{n}}$$

Asumimos $\bar{x} = 59$, $\sigma = 24.8$, $n = 100$.

- Error Estándar $= 2.48$
- Podemos decir que el parámetro poblacional está entre 56.2 y 61.48
- Dado que $\mu = 58$, sabemos que es cierto.

Conclusión

Podemos obtener un estimado razonable de un parámetro poblacional si hacemos un muestreo aleatorio de la población.

- El estimador muestral es un buen pronóstico del parámetro poblacional cuando sabemos el valor de μ a priori.

Todavía no sabemos:

- ¿Qué tan probable es que $\bar{x}$ se encuentre a un error estándar de μ?
- ¿Quê pasa si no sabemos μ, pero tenemos una idea de lo que puede ser?

Table of Contents