

# Bioestadística

## Tema 6: Muestreo

- Parte de los conceptos de la teoría del muestreo han sido discutidos con anterioridad. Aquí los repasaremos y ampliaremos. Por ejemplo, hemos mencionado que las poblaciones están formadas por individuos, pero sería mejor denominarlas **unidades de muestreo** o **unidades de estudio**:
  - Personas, células, familias, hospitales, países...
- La población ideal que se **pretende estudiar** se denomina **población objetivo**.
  - No es fácil estudiarla por completo. Aproximamos mediante muestras que den idealmente la misma probabilidad a cada individuo de ser elegido.
  - Tampoco es fácil elegir muestras de la población objetivo:
    - Si llamamos por teléfono excluimos a los que no tienen.
    - Si elegimos indiv. en la calle, olvidamos los que están trabajando...
- El grupo que en realidad **podemos estudiar** (v.g. los que tienen teléfono) se denomina **población de estudio**.

## Fuentes de sesgo

- Las poblaciones objetivo y de estudio **pueden diferir** en cuanto a las variables que estudiamos.
  - El nivel económico en la población de estudio es mayor que en la objetivo,...
  - Los individuos que se eligen en la calle pueden ser de mayor edad (mayor frecuencia de jubilados p.ej.)...
  - En este caso, diremos que las muestras que se elijan estarán **sesgadas**. Al tipo de sesgo debido a diferencias sistemáticas entre población objetivo y población de estudio se denomina **sesgo de selección**.
- Hay otras fuentes de error/sesgo
  - **No respuesta** a encuestas embarazosas
    - Consumo de drogas, violencia doméstica, prácticas poco éticas,...
  - **Mentir** en las preguntas “delicadas”.
- Para evitar este tipo de sesgo se utilizan la técnica de **respuesta aleatorizada**.

## Técnicas de respuesta aleatorizada

- **Reducen la motivación para mentir** (o no responder) a las encuestas.
  - ¿Si digo la verdad, se me verá el plumero...?
- **¿Cómo se hace?**  
Pídele que lance una moneda antes de responder y...
  - Si sale **cara** que diga la “**opción comprometida**”
    - (no tiene por qué avergonzarse, la culpa es de la moneda)
  - Si sale **cruz** que diga la **verdad**
    - (no tiene por qué avergonzarse, el encuestador no sabe si ha salido cara o cruz)
- Aunque no podamos saber cuál es la verdad en cada individuo, podemos hacernos una idea porcentual sobre la población, viendo en cuánto se alejan las respuestas del 50%.

## Ejemplo: ¿Ha tomado drogas alguna vez?

Sin respuesta aleatorizada

100% No



Insineros!!



Con respuesta aleatorizada

40% No  
60% Sí

Diferencia entre los que han dicho sí y los que debían hacerlo por que así lo indicaba la moneda

¡No son mitad y mitad!

El porcentaje estimado de ind. que tomó drogas es:

$$p^* \approx \frac{0,6 - 0,5}{1 - 0,5} = 0,2 = 20\%$$

Los que deben decir la verdad

## Técnicas de muestreo

- Cuando elegimos individuo de una población de estudio para formar muestras podemos encontrarnos en las siguientes situaciones:
  - **Muestreos probabilistas**
    - Conocemos la probabilidad de que un individuo sea elegido para la muestra.
    - **Interesantes** para usar estadística matemática con ellos.
  - **Muestreos no probabilistas**
    - No se conoce la probabilidad.
    - Son muestreos que seguramente esconden sesgos.
    - En principio **no se pueden extrapolar** los resultados a la población.
      - A pesar de ello una buena parte de los estudios que se publican usan esta técnica. ¡Buff!
- En adelante vamos a tratar *exclusivamente* con muestreos con la menor posibilidad de sesgo (probabilistas): **aleatorio simple, sistemático, estratificado y por grupos.**

## Muestreo aleatorio simple (m.a.s.)

- Se eligen individuos de la **población de estudio**, de manera que todos tienen la misma probabilidad de aparecer, hasta alcanzar el **tamaño muestral** deseado.
- Se puede realizar partiendo de listas de individuos de la población, y eligiendo individuos aleatoriamente con un ordenador.
- Normalmente tiene un coste bastante alto su aplicación.
- En general, las técnicas de inferencia estadística suponen que la muestra ha sido elegida usando m.a.s., aunque en realidad se use alguna de las que veremos a continuación.

## Muestreo sistemático

- Se tiene una lista de los individuos de la **población de estudio**. Si queremos una muestra de un tamaño dado, elegimos individuos igualmente espaciados de la lista, donde el primero ha sido elegido al azar.
- CUIDADO: Si en la lista existen periodicidades, obtendremos una muestra sesgada.
  - Un caso real: Se eligió una de cada cinco casas para un estudio de salud pública en una ciudad donde las casas se distribuyen en manzanas de cinco casas. Salieron con mucha frecuencia las de las esquinas, que reciben más sol, están mejor ventiladas,...

## Muestreo estratificado

- Se aplica cuando sabemos que hay ciertos factores (variables, subpoblaciones o estratos) que pueden influir en el estudio y queremos asegurarnos de tener cierta cantidad mínima de individuos de cada tipo:
  - Hombres y mujeres,
  - Jovenes, adultos y ancianos...
- Se realiza entonces una m.a.s. de los individuos de cada uno de los estratos.
- Al extrapolar los resultados a la población hay que tener en cuenta el tamaño relativo del estrato con respecto al total de la población.

## Muestreo por grupos o conglomerados

- Se aplica cuando es difícil tener una lista de todos los individuos que forman parte de la población de estudio, pero sin embargo sabemos que se encuentran agrupados naturalmente en grupos.
- Se realiza eligiendo varios de esos grupos al azar, y ya elegidos algunos podemos estudiar a todos los individuos de los grupos elegidos o bien seguir aplicando dentro de ellos más muestreos por grupos, por estratos, aleatorios simples,...
- Para conocer la opinión de los médicos del sistema nacional de salud, podemos elegir a varias regiones de España, dentro de ellas varias comarcas, y dentro de ellas varios centros de salud, y...
- Al igual que en el muestreo estratificado, al extrapolar los resultados a la población hay que tener en cuenta el tamaño relativo de unos grupos con respecto a otros.
  - Regiones con diferente población pueden tener probabilidades diferentes de ser elegidas, comarcas, hospitales grandes frente a pequeños,...

## Estimación

- Un **estimador** es una cantidad numérica **calculada sobre una muestra** y que esperamos que sea una buena **aproximación** de cierta cantidad con el mismo significado en la población (**parámetro**).
- En realidad ya hemos trabajado con estimadores cada vez que hacíamos una práctica con muestras extraídas de una población y suponíamos que las medias, etc... eran próximas de las de la población.
  - Para la media de una población:
    - “El mejor” es la media de la muestra.
  - Para la frecuencia relativa de una modalidad de una variable:
    - “El mejor” es la frecuencia relativa en la muestra.
- Habría que precisar que se entiende por “el mejor estimador” pero eso nos haría extendernos demasiado. Ver libro.

## ¿Es útil conocer la distribución de un estimador?

- Es la **clave para hacer inferencia**. Ilustrémoslo con un ejemplo que ya tratamos en el tema anterior (**teorema del límite central**).
  - Si de una variable **conocemos  $\mu$  y  $\sigma$** , sabemos que para muestras “grandes”, la **media muestral** es:
    - aproximadamente normal,
    - con la misma media y,
    - desviación típica mucho menor (**error estándar**)
- $$EE = \frac{\sigma}{\sqrt{n}}$$
- Es decir si por ejemplo  **$\mu=60$  y  $\sigma=5$** , y obtenemos muestras de tamaño  **$n=100$** ,
    - La desv. típica de la media muestral (error estándar) es  **$EE=5/\text{raiz}(100)=0,5$**
    - como la media muestral es aproximadamente normal, el 95% de los estudios con muestras ofrecerían estimaciones entre  **$60\pm 1$**
    - Dicho de otra manera, **al hacer un estudio tenemos una confianza del 95%** de que la verdadera media esté a una distancia de  $\pm 1$ .

- En el ejemplo anterior la situación no era muy realista, pues como de todas maneras no conozco  $\sigma$  desconoceré el intervalo exacto para  $\mu$ .
- Sin embargo también hay estimadores para  $\sigma$  y puedo usarlo como aproximación.
- Para tener una idea intuitiva, analicemos el siguiente ejemplo. Nos servirá como introducción a la estimación puntual y por intervalos de confianza.

- Ejemplo: Una muestra de  $n=100$  individuos de una población tiene media de peso 60 kg y desviación 5kg.
  - Dichas cantidades pueden considerarse como aproximaciones (**estimaciones puntuales**)
    - 60 kg estima a  $\mu$
    - 5 kg estima a  $\sigma$
    - $5/\sqrt{n} = 0,5$  estima el error estándar (típico) **EE**
      - Estas son las llamadas estimaciones puntuales: un número concreto calculado sobre una muestra es aproximación de un parámetro.
  - Una estimación por **intervalo de confianza** es una que ofrece un intervalo como respuesta. Además podemos asignarle una probabilidad aproximada que mida nuestra confianza en la respuesta:
    - Hay una confianza del 68% de que  $\mu$  esté en  $60 \pm 0,5$
    - Hay una confianza del 95% de que  $\mu$  esté en  $60 \pm 1$ .
- Ojo: He hecho un poco de trampa. ¿La ves?



## Estimación puntual y por intervalos

- Se denomina **estimación puntual** de un parámetro al ofrecido por el estimador sobre una muestra.
- Se denomina **estimación confidencial** o **intervalo de confianza** para un **nivel de confianza  $1 - \alpha$**  dado, a un intervalo que ha sido construido de tal manera que con frecuencia  $1 - \alpha$  realmente contiene al parámetro.
  - Obsérvese que la probabilidad de error (no contener al parámetro) es  $\alpha$ .
    - En el siguiente tema se llamará prob. de error de tipo I o nivel de significación.
    - Valores típicos:  $\alpha=0,10$  ; **0,05** ; 0,01
  - En general el tamaño del intervalo disminuye con el tamaño muestral y aumenta con  $1-\alpha$ .
  - En todo intervalo de confianza hay una noticia buena y otra mala:
    - La buena: hemos usado una técnica que en % alto de casos acierta.
    - La mala: no sabemos si ha acertado en nuestro caso.

## Aplicación

### Descriptivos para Número de hijos

		Estadístico	Error típ.
Media		1,90	,045
Intervalo de confianza para la media al 95%	Límite inferior	1,81	
	Límite superior	1,99	
Media recortada al 5%		1,75	
Mediana		2,00	
Varianza		3,114	
Desv. típ.		1,765	
Mínimo		0	
Máximo		8	
Rango		8	
Amplitud intercuartil		3,00	
Asimetría		1,034	,063
Curtosis		1,060	,126

- Al final del tema 2 dejamos sin interpretar parte de los resultados que obteníamos con SPSS.
- ¿Sabrías interpretar lo que falta por sombrear?
- ¿Puedes dar un intervalo de confianza para la media al 68% de confianza?
- Observa la asimetría. ¿Crees probable que la asimetría en la población pueda ser cero ya que la obtenida en la muestra es aprox. 1?



## ¿Qué hemos visto?

- Sesgo de selección
  - Población objetivo
  - Población de estudio
- Otros sesgos
  - Técnica de respuesta aleatorizada
- Técnicas de muestreo
  - No probabilistas
  - Probabilistas
    - m.a.s.
    - Sistemático
    - Estratificado
    - Conglomerados
  - Estimación
    - Estimador
      - Estimación puntual
      - Error estándar
    - Estimación confidencial
    - Nivel de confianza  $1-\alpha$

