Facultad de Ciencias Exactas, Ingeniería y Agrimensura

FCEIA

UNIVERSIDAD NACIONAL DE ROSARIO

Probabilidad y Estadística

Unidad 2: Técnicas de recolección de datos

Lic. Maite San Martín

Marzo 2024

¿Qué es la estadística?

¿Qué es el método científico y cómo la estadística forma parte de ese círculo virtuoso de generación de conocimiento?

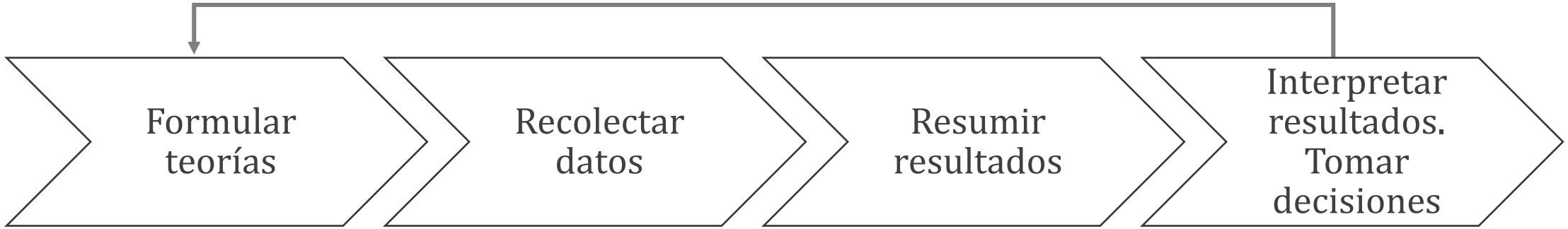
¿Qué es la estadística descriptiva?
¿Y la estadística inferencial?

Estadística y toma de decisiones… ¿test de hipótesis?

CONCEPTOS


parámetro
regla de decisión
p-value
probabilidad de error
variable
población
hipótesis
región de rechazo
muestra
estadística
resultados estadísticamente significativos
nivel de significación
valor crítico

Método científico



Estamos aquí

Selección de un diseño de investigación

- Experimentales
- No experimentales

Identificación de la población y diseño del plan de recolección de datos

- Población y Muestra
- Técnicas de muestreo
- Sesgos

Recolección de datos

No todos los datos se originan de la misma manera y es necesario, antes de consumir información, así como al momento de planificar un estudio, realizar(nos) algunas preguntas. La recolección de datos puede estar planificada de distintas formas, que dependen de muchos factores tales como el tiempo disponible, los recursos económicos, las posibilidades de acuerdo al objeto de estudio (por ejemplo, una enfermedad, un proceso productivo, un indicador económico, etc.). El tipo de estudio que se emplee va a condicionar el tipo de análisis posterior de los resultados.

Los estudios experimentales son aquellos en los que el investigador decide qué valores tomará alguna variable en particular (por ejemplo, si poner mucho o poco fertilizante en un estudio para evaluar el rendimiento de un cultivo), mientras que los estudios observacionales son aquellos en que el investigador no interviene de forma activa sino que simplemente observa las características (por ejemplo, si el cultivo rinde más en un clima seco o en un clima húmedo).

Recolección de datos

¿Cómo elegir un tipo de estudio?

¿Cuál es el tipo de estudio adecuado para responder a una pregunta en particular? Muchas veces, la respuesta en un ámbito estadístico es DEPENDE: depende de la pregunta que queramos responder. Por ejemplo, si interesa conocer la opinión de ciertas personas, describir sus estilos de vida, sus preferencias o describir variables demográficas como nacimientos, muertes o migraciones, es adecuado realizar estudios observacionales. En cambio, si interesa determinar la causa de un resultado o comportamiento, es decir, determinar si una variable puede provocar cambios en otra, será más conveniente un experimento.

Algunas definiciones

Un **experimento** es aquel estudio en el que el investigador impone algún tratamiento en forma aleatorizada sobre las unidades o sujetos a los fines de observar las respuestas.

Un **estudio observacional** es aquel en el que el investigador simplemente observa los sujetos o unidades y registra los datos para cada una de las variables de interés. El investigador no intenta manipular las unidades.

Cuando hablamos de **unidades** nos referimos a los objetos sobre los cuales se realizan las observaciones. Si las unidades son personas se las denomina **sujetos**.

Para resolver

Indicar y explicar cuál es el tipo de estudio más adecuado para responder a cada una de las siguientes preguntas:

1. ¿Están contentos los alumnos con el nuevo sistema de evaluación?
2. ¿El ausentismo de los alumnos es menor en verano que en invierno?
3. ¿El rendimiento de los alumnos en un examen es mejor si durante el mismo escuchan música de Vivaldi, en bajo volumen, en comparación con no escuchar nada?

Para resolver

Una educadora divide al azar un grupo de niñas y niños de preescolar en dos grupos con iguales capacidades iniciales (para ello les toma una prueba). En un grupo utiliza canciones para enseñarles a contar, y en el otro grupo utiliza el método tradicional.

a) ¿Es esto un estudio observacional o un experimento? Justificar.

b) ¿Por qué es importante aclarar que todos los alumnos tienen las mismas capacidades iniciales?

Estudios experimentales

En los estudios experimentales, tal como dijimos algunas páginas más atrás, el investigador participa de forma activa determinando los valores de algunas de las variables explicativas y observando las respuestas.

Los estudios experimentales pretenden estudiar el efecto del cambio de ciertas variables sobre las variables que expresan los resultados, llamadas **variables respuesta**. Las variables que se controlan en un experimento se llaman **variables explicativas** o factores. Un **tratamiento** es una combinación de niveles de varios factores.

En la mayoría de los ensayos clínicos, los pacientes son las unidades y las drogas son tratamientos. En la agricultura, las unidades experimentales son, frecuentemente, parcelas de terreno y los tratamientos pueden ser variedades de plantas, fertilizantes o pesticidas. En un experimento, los tratamientos no solo son observados, sino que son activamente impuestos a las unidades experimentales. Es la activa imposición de un tratamiento lo que hace que un estudio sea un experimento.

Estudios experimentales

Los experimentos bien planificados son diseñados siguiendo una serie de principios:

- **Control**: Se deben controlar los efectos de las variables de confusión. Es conveniente formar grupos de unidades homogéneas con respecto a los valores tomados por las variables de confusión, y luego adjudicar los tratamientos aleatoriamente entre los sujetos del mismo grupo. Dentro del grupo, la variable de confusión controlada afecta a ambos grupos de tratamientos en la misma medida, de modo que su efecto "se cancela" cuando se comparan los tratamientos.
- **Aleatorización**: La asignación aleatoria de las unidades a los grupos tiende a producir grupos de unidades experimentales que están balanceados con respecto a factores potenciales de confusión. La aleatorización asegura que el experimento no favorezca intencionalmente a un tratamiento u otro.
- **Repetición**: Es conveniente asignar al menos dos unidades a cada combinación de factores, para conocer la variabilidad natural de las respuestas de las unidades que reciben el mismo tratamiento.

Estudios experimentales

Para reducir los sesgos que pueden aparecer en la respuesta por distintos motivos, los sujetos en general no saben cuál es el tratamiento recibido (tratamiento activo o placebo). Un **experimento ciego** intenta eliminar cualquier sesgo de respuesta del sujeto debido a la información que recibe. El **sesgo de respuesta** es la distorsión en la respuesta que puede provenir del conocimiento que tenga el sujeto sobre el tratamiento bajo el cual está. Los sujetos deben ser "ciegos" con respecto de la medicina que toman, pero lo mismo debe ocurrir con quien las administra o quien registra los resultados. Si el que registra los resultados sabe si el paciente está tratado o recibió un placebo, puede, frente a un resultado dudoso dejarse llevar por sus prejuicios con respecto a la bondad del tratamiento. Sería un caso de sesgo del experimentador. El **efecto experimentador** es la distorsión que puede surgir por parte del experimentador, que generalmente tiende a defender su teoría o puede tener prejuicios sobre uno u otro tratamiento. Cuando ambos, experimentador y sujeto están ciegos con respecto a quien pertenece a cada grupo, el experimento se llama **doblemente ciego**.

Estudios observacionales

La observación es en general el primer paso: a partir de ella suelen abrirse diversos caminos para desarrollar nuevas teorías y modelos; la observación puede motivar la realización posterior de nuevos estudios. Desde un punto de vista estadístico, los estudios experimentales pueden resultar más valiosos que los estudios observacionales, pero algunas veces solo es posible realizar estos últimos. En muchas oportunidades incluso, no es ético realizar un estudio experimental. Por ejemplo, si se deseara estudiar la influencia del cigarrillo en las enfermedades pulmonares, no es posible forzar a 100 personas a fumar 3 paquetes de cigarrillos por día y a otras 100 personas a no fumar. Sí, en cambio, podrían estudiarse 200 personas considerando si cada una de ellas fuma, cuánto fuma y si tiene alguna afección pulmonar, o bien estudiarlas durante un período de tiempo y evaluar si aquellas que fuman o fumaron en algún momento de su vida desarrollaron problemas pulmonares a lo largo de los años.

Estudios observacionales

En los estudios observacionales se suelen realizar esfuerzos para identificar las **variables de confusión** y ajustarlas. Una variable de confusión es una variable cuyos efectos sobre la respuesta no pueden ser separados de los efectos que la variable explicativa tiene sobre la misma. Por ejemplo, supongamos el siguiente estudio: un investigador desea estudiar la posible influencia de la ingesta de fibra sobre las afecciones cardíacas. Si las personas que comen dietas ricas en fibra hacen más ejercicio que aquellos que no la comen, puede ocurrir que los efectos que se atribuyen a la dieta sean, en realidad la consecuencia del ejercicio. Aquí, las variables "dieta" y "cantidad de ejercicio" están confundidas. Sin embargo, los científicos podrían ajustar los resultados con respecto a las variables de confusión **siempre y cuando estas variables hayan sido medidas**. Es decir, es imposible incluir en el análisis variables que no se consideraron al momento de la recolección de datos.

Estudios observacionales:
estudios por muestreo

Supongamos que tenemos una olla de sopa: revolvemos con una cuchara, sacamos una porción -una muestra- y la saboreamos: si la cucharada ha sido tomada con la sopa bien mezclada, entonces podremos tener una buena idea del sabor de la totalidad de la sopa. Esta es la lógica de los estudios por muestreo: estudiar una parte de un todo, pero para poder sacar conclusiones sobre la totalidad.

Los estudios por muestreo son estudios observacionales que recurren a las técnicas de muestreo. Las técnicas de muestreo permiten obtener información sobre el total mediante el examen de sólo una parte de la población. De este modo, es posible describir las características de la muestra y, sobre todo, realizar inferencias: utilizar las estadísticas para estimar los parámetros de interés.

Estudios observacionales:
estudios por muestreo

¿Por qué trabajar con muestras?

Hay ocasiones en las que es imposible estudiar la población entera, por ser muy grande, por razones de tiempo y costo o porque el estudio lleva a la destrucción del material (por ejemplo, si quisiéramos analizar la vida útil de baterías: deberíamos usarlas hasta que se gasten, lo que hace inutilizable el producto). En estos casos se trabaja con una muestra.

El estudio de la población recibe el nombre de censo. El estudio de la muestra recibe el nombre de estudio por muestreo.

Conceptos básicos

La **población** es el conjunto de todas las unidades en las que estamos interesados, de las cuales queremos obtener conclusiones.

Una **muestra** es un subconjunto de la población para el cual tenemos (o planeamos tener) datos.

La **unidad de análisis** es el elemento mínimo de una población. Se refiere a qué o quién es objeto de interés en una investigación. Cuando las unidades de análisis son personas se las llama sujetos.

Conceptos básicos

Una **estadística** es una medida resumen que se calcula usando todas las unidades de la muestra extraída de cierta población.

Un **parámetro** es una medida resumen que se calcula a partir de todas las unidades de la población, y suele ser desconocido.

Conceptos básicos

Las **variables** son propiedades, atributos o características que forman parte del problema y a través de las cuales podremos explorarlo, describirlo o explicarlo. Las variables toman distintos valores según la unidad (por eso se las llama variables, porque varían de unidad a unidad).

Un **dato** es el valor que adopta una variable medida en una unidad de análisis.

Una **observación** refiere a todos los datos observados para unidad de análisis en particular.

Conceptos básicos

Supongamos que nos interesa estudiar cierta variable en una población en particular: por ejemplo, la situación laboral de lxs estudiantes de PyE de la FCEIA durante el 2024.

Supongamos que no podemos relevar ese dato en todxs lxs estudiantes, por lo que vamos a relevar información solo de algunos estudiantes de que cursan este año.

VARIABLE

POBLACIÓN

UNIDAD DE ANÁLISIS

MUESTRA

Un **parámetro** es una medida resumen calculada utilizando todas las unidades que componen la población.

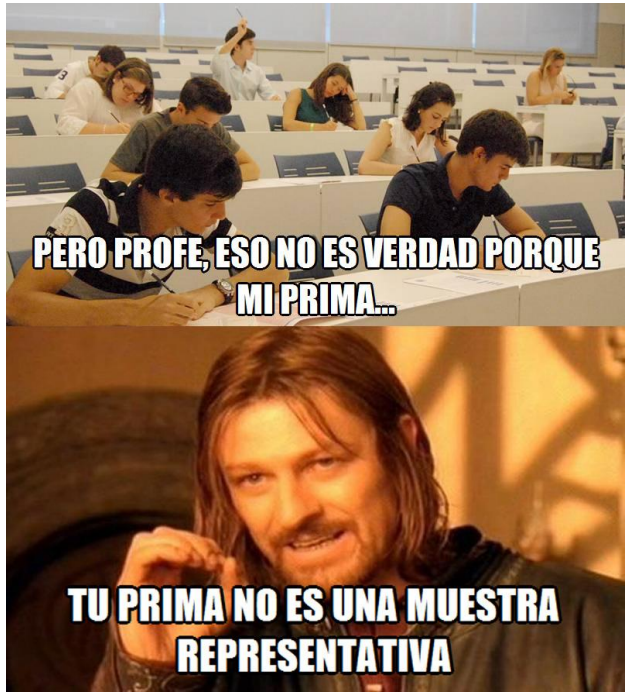
Una **estadística** es una medida resumen calculada utilizando todas las unidades que componen la muestra.

Estudios observacionales:
estudios por muestreo

¿Todas las muestras son buenas?

Una muestra "buena" debe ser **representativa** de la población. Esto significa, que todas las características importantes de la población tienen que estar en la muestra en la misma proporción que en la población. Asimismo, idealmente la muestra debe ser **aleatoria**, es decir, debe existir un mecanismo generador de aleatoriedad para seleccionar qué unidades de la población conforman la muestra.

Las **métodos probabilísticos de muestreo** son aquellos que permitirán obtener muestras representativas. Se trata de métodos aleatorios que asignan una probabilidad de selección conocida a cada unidad de análisis.



Estudios observacionales:
estudios por muestreo

Cuando hablamos de mecanismos generadores de aleatoriedad puede tratarse de un bolillero, de un sorteo, de tablas de números aleatorios o bien de mecanismos pseudoaleatorios (como las calculadoras o las computadora). En términos generales, cualquiera de ellos va a ser suficiente para que el azar garantice muestras de calidad.

En la técnica de muestreo más simple para tomar una muestra de alguna población, las N unidades que conforman la población se representan de alguna forma (con una bolilla por ejemplo), se mezclan y se seleccionan las n unidades que conformarán la muestra. Esta es llamada **muestra simple al azar**.

Llamamos con N al **tamaño poblacional**: cantidad total de unidades que componen a la población bajo análisis

Llamamos con n al **tamaño muestral**: cantidad total de unidades que componen a la muestra que se extrae

Estudios observacionales:
estudios por muestreo

Una muestra simple al azar de tamaño n es una muestra que se selecciona de una población de forma tal que asegura que cualquier muestra posible del mismo tamaño tiene la misma probabilidad de ser seleccionada.

La definición implica que cada individuo de la población tiene igual chance de ser seleccionado. Sin embargo, el hecho de que cada individuo tenga la misma chance de selección no es suficiente para garantizar una muestra simple al azar.

Un método comunmente empleado para seleccionar una muestra simple al azar es crear una lista llamada marco muestral, de objetos o individuos en la población. Se identifica a cada elemento de la lista con un número y se utiliza un generador de números al azar para seleccionar la muestra.

Estudios observacionales:
estudios por muestreo

Una muestra simple al azar de tamaño n es una muestra que se selecciona de una población de forma tal que asegura que cualquier muestra posible del mismo tamaño tiene la misma probabilidad de ser seleccionada.

La definición implica que cada individuo de la población tiene igual chance de ser seleccionado. Sin embargo, el hecho de que cada individuo tenga la misma chance de selección no es suficiente para garantizar una muestra simple al azar.

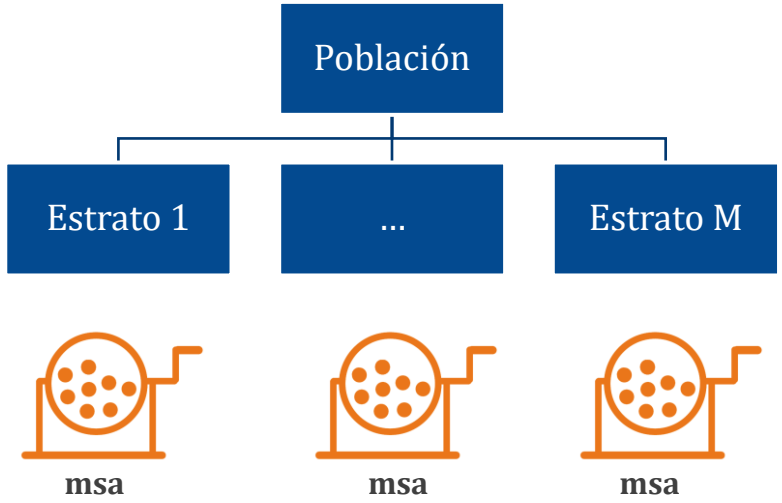
Un método comúnmente empleado para seleccionar una muestra simple al azar es crear una lista llamada marco muestral, de objetos o individuos en la población. Se identifica a cada elemento de la lista con un número y se utiliza un generador de números al azar para seleccionar la muestra.

Estudios observacionales:
estudios por muestreo

Las muestras simples al azar son muy útiles por su sencillez de selección y análisis. Sin embargo no siempre son las apropiadas, por lo que existen además otras técnicas de muestreo.

Muestreo estratificado

La variable de interés puede diferir sistemáticamente de acuerdo a si la unidad de análisis pertenece a algún estrato en particular. Por ejemplo, si quisiéramos medir los conocimientos en matemática de estudiantes de la LCC, cada año de estudio podría ser un estrato (los conocimientos de estudiantes de 1ro muy probablemente no sean los mismos que los de los estudiantes de 4to). En estos casos, se toma una muestra simple al azar en cada estrato.

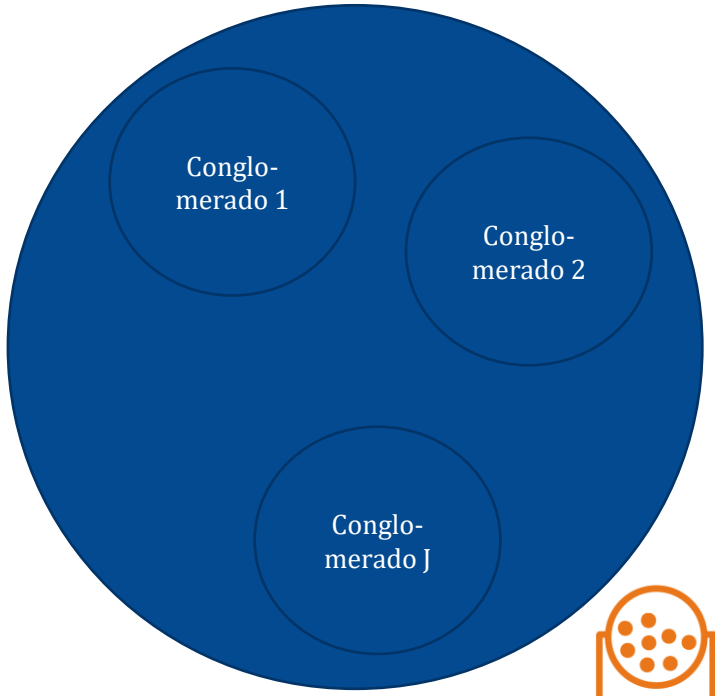


Estudios observacionales:
estudios por muestreo

Las muestras simples al azar son muy útiles por su sencillez de selección y análisis. Sin embargo no siempre son las apropiadas, por lo que existen además otras técnicas de muestreo.

Muestreo por conglomerados

Se reconocen como “mini poblaciones” en la población, grupos más o menos pequeños que conservan la variabilidad general presente en la población. Por ejemplo, sabiendo que las mujeres alcanzan su altura máxima a los 18 años (ya que después de la pubertad el crecimiento se detiene), si se quisiera analizar la altura de las estudiantes de la FCEIA, podrían seleccionarse solo algunas comisiones (conglomerados), dado que probablemente todas sean más o menos parecidas en cuanto a la altura de las estudiantes, reproduciendo la variabilidad poblacional.



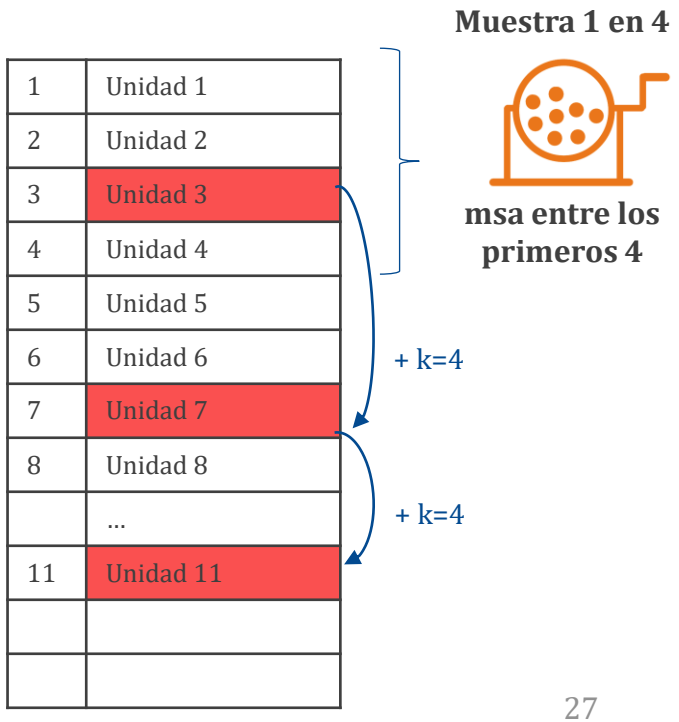
msa de conglomerados

Estudios observacionales:
estudios por muestreo

Las muestras simples al azar son muy útiles por su sencillez de selección y análisis. Sin embargo no siempre son las apropiadas, por lo que existen además otras técnicas de muestreo.

Muestreo sistemático

Las unidades de análisis se presentan en un orden determinado, que no tiene relación con la variable bajo estudio, y que puede facilitar la recolección de datos. En un muestreo sistemático 1 en k se toma una muestra simple aleatoria entre las primeras k unidades y luego se incluyen en la muestra una de cada k unidades.



Estudios observacionales:
estudios por muestreo

Al sacar una muestra podemos calcular distintas estadísticas para estimar el parámetro desconocido de interés. Sin embargo, si sacáramos otra muestra distinta es MUY probable que el valor de la estadística observado varíe también.

Estudios observacionales:
estudios por muestreo

La primera ventaja de las muestras aleatorias es que eliminan el **sesgo** del procedimiento de selección de una muestra. Aún así, suele no coincidir el resultado con el verdadero valor, debido a la variabilidad que resulta de la selección al azar. Este tipo de variabilidad es llamada variabilidad muestral.

Que la variabilidad muestral sea muy grande, significa que el valor del estadístico cambia mucho entre muestra y muestra. Por lo tanto no podemos creerle al resultado que obtenemos con una muestra en particular. Pero estamos salvados por una segunda ventaja que tienen las muestras aleatorias: la variabilidad de las estadísticas calculadas en distintas muestras de un mismo tamaño seguirá un patrón predecible.

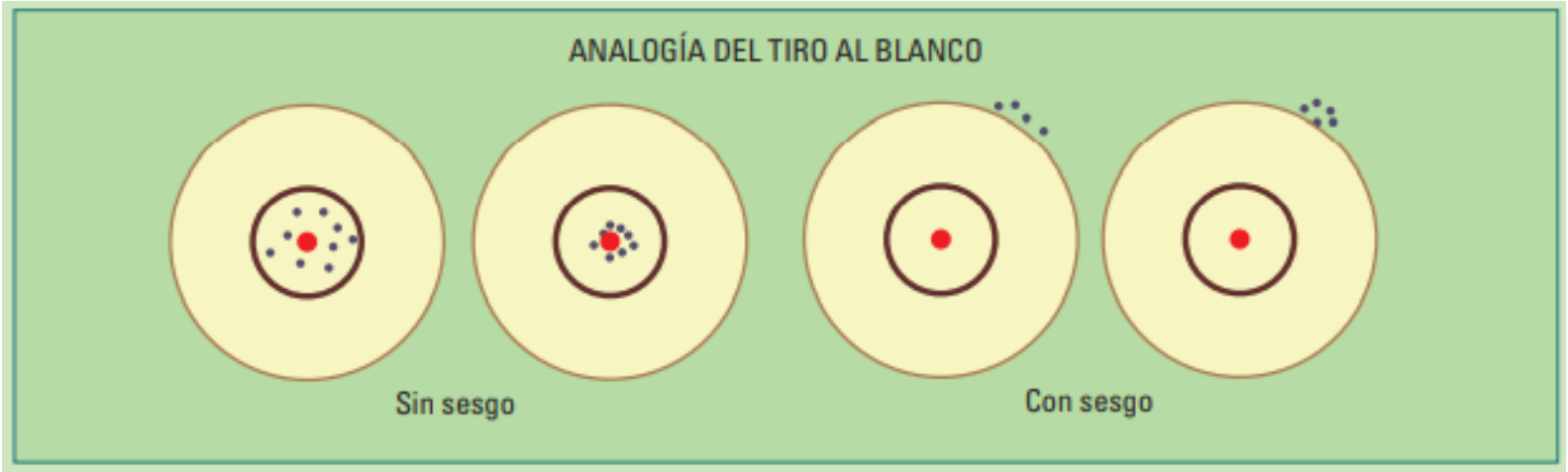
Estudios observacionales:
estudios por muestreo

Por más que una encuesta u otro tipo de estudio esté bien diseñada y bien conducida, dará el valor de un estadístico como estimación del parámetro poblacional. **Muestras diferentes darán valores diferentes y el error debido al muestreo estará siempre presente**. Pero podremos decir, con cierto grado de confianza, cuál va a ser la magnitud de ese error. Se trata de errores aleatorios, surgen de utilizar una muestra en vez de la población total.

Sin embargo, **podemos también encontrarnos con errores que no se deben al muestreo aleatorio**. Pueden ocurrir en cualquier encuesta e incluso en los censos. Un tipo particular de estos errores son los debido a la presencia de sesgos en el muestreo, en las respuestas y/o en su registro. Un respondente puede mentir respecto de su edad, de cuántas horas trabaja por día (puede pensar que trabaja poco y entonces las aumenta), de su salario (puede no querer que se sepa que gana mucho, o que gana poco) o puede haber olvidado cuantos paquetes de cigarrillos fumó la semana anterior.

Estudios observacionales:
estudios por muestreo

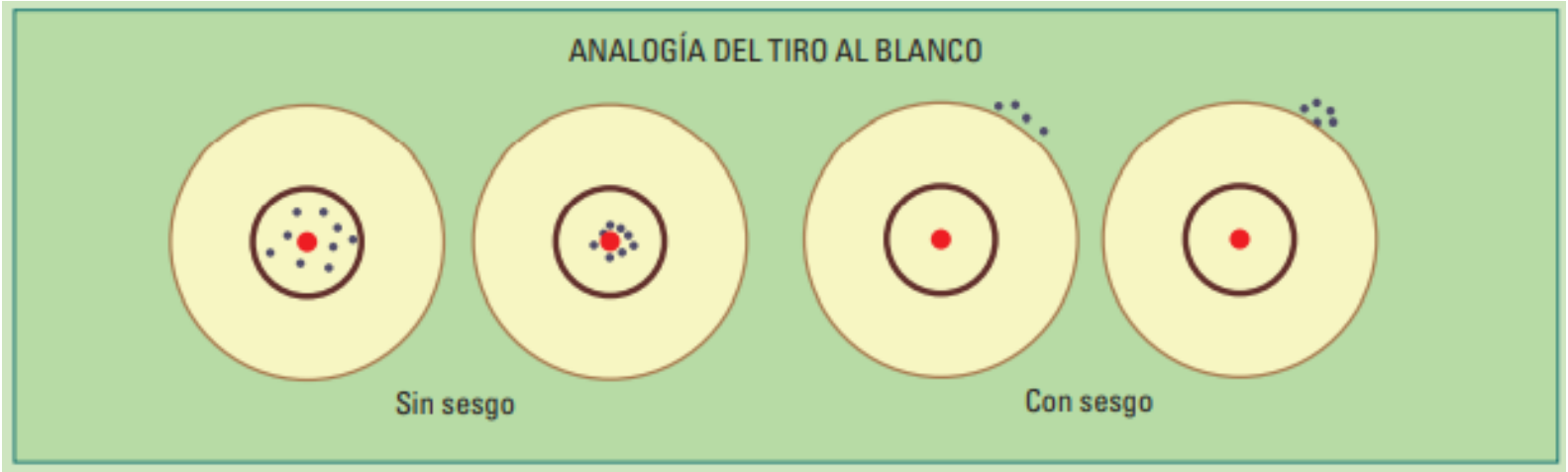
Podemos utilizar la analogía del juego del tiro al blanco para describir el efecto del sesgo y el tamaño de muestra en el error de muestreo. Supongamos que el centro del blanco(punto rojo de la figura) es el parámetro poblacional al que queremos acertar. Si estamos realizando un muestreo aleatorio, en cada muestra -es decir para cada tiro- obtendremos un punto cercano al centro. Algunas veces, el dardo caerá un poco arriba otras un poco abajo.



Estudios observacionales:
estudios por muestreo

El segundo esquema muestra puntos negros más concentrados que los del primer esquema, representando un aumento en el tamaño de las muestras y una reducción de la variabilidad de los resultados.

Si en cambio el procedimiento tiene sesgo, los valores estarán todos desviados en una misma dirección. Sin embargo, la distancia de los puntos negros al rojo no se reduce al reducirse la variabilidad cuando hay sesgo (parte derecha del esquema).



Estudios observacionales:
estudios por muestreo

En definitiva, ¿qué es el sesgo? Es un favoritismo de alguna etapa del proceso de recolección de datos beneficiando algunos resultados, perjudicando otros y **desviando las conclusiones en direcciones equivocadas**. Es habitual encontrarnos con muestras sesgadas:

- El **sesgo de selección** ocurre cuando la muestra refleja solo a una parte de la población que pretendía representar. Por ejemplo, una encuesta telefónica ignora a todos los sujetos que no tienen teléfono, o una encuesta que realiza entrevistas en hogares ignora a los que viven en la calle.

Casos particulares de muestras con sesgo de selección:

- **Muestra de conveniencia**. Exprimir las naranjas que se encuentran a la vista, en la parte de arriba del cajón, es un ejemplo de muestra de conveniencia. Obtener una muestra de esta manera es rápido y económico, pero la muestra no va a ser representativa de la población.

Estudios observacionales:
estudios por muestreo

En definitiva, ¿qué es el sesgo? Es un favoritismo de alguna etapa del proceso de recolección de datos beneficiando algunos resultados, perjudicando otros y **desviando las conclusiones en direcciones equivocadas**. Es habitual encontrarnos con muestras sesgadas:

- El **sesgo de selección** ocurre cuando la muestra refleja solo a una parte de la población que pretendía representar. Por ejemplo, una encuesta telefónica ignora a todos los sujetos que no tienen teléfono, o una encuesta que realiza entrevistas en hogares ignora a los que viven en la calle.

Casos particulares de muestras con sesgo de selección:

- **Muestras de respuesta voluntaria**. Surgen a partir de los individuos que se ofrecen voluntariamente a participar, por ejemplo cuando se pide a los oyentes de un programa de radio que voten por tal o cual cantante, llamando por teléfono o enviando un mensaje. Los participantes voluntarios, que por algún motivo decidieron participar, suelen tener opiniones más polarizadas.

Estudios observacionales:
estudios por muestreo

En definitiva, ¿qué es el sesgo? Es un favoritismo de alguna etapa del proceso de recolección de datos beneficiando algunos resultados, perjudicando otros y **desviando las conclusiones en direcciones equivocadas**. Es habitual encontrarnos con muestras sesgadas:

- El **sesgo de respuesta** o error de medición ocurre cuando el método de observación tiende a producir valores que difieren sistemáticamente del verdadero valor poblacional en alguna manera. Errores de medición ocurren cuando un instrumento está mal calibrado o cuando las preguntas en una encuestas están redactadas en una forma que influye la respuesta.

Estudios observacionales:
estudios por muestreo

Las diferentes palabras con las que se puede presentar una misma pregunta suele ser una fuente importante de sesgo en las respuestas. En un curso de manejo se proyectó una película sobre un accidente de tránsito a dos grupos de alumnos. Ambos grupos eran similares respecto de la edad y el género. Al finalizar la proyección se preguntó: al primer grupo ¿a qué velocidad piensa que los dos autos chocaron? El promedio de las respuestas fue de 50,9 km/h. Al segundo grupo: ¿a qué velocidad piensa que los dos autos se colisionaron? El promedio de las respuestas fue de 65,9 km/h.

El sesgo debido a la forma en que se presenta una pregunta puede ser intencional o no intencional. Las preguntas “¿No está usted harto de pagar impuestos para que todo siga igual de mal?” y “¿Le parece importante que se paguen impuestos para mejorar la educación, los servicios de salud y la seguridad?”, que apuntan al pago de impuestos, seguramente tendrán resultados muy diferentes. Ambas preguntas conllevan un sesgo intencional.

Estudios observacionales:
estudios por muestreo

En definitiva, ¿qué es el sesgo? Es un favoritismo de alguna etapa del proceso de recolección de datos beneficiando algunos resultados, perjudicando otros y **desviando las conclusiones en direcciones equivocadas**. Es habitual encontrarnos con muestras sesgadas:

- Algunas veces las personas que han sido seleccionadas para una encuesta son muy difíciles de localizar o simplemente se niegan a responder. Los individuos que no responden pueden ser muy diferentes de los que sí lo hacen. Este tipo de sesgo es llamado **sesgo de no respuesta**. La no respuesta puede distorsionar los resultados si los individuos que responden difieren de los individuos que no responden.

Es importante reconocer que los sesgos se generan por la forma en que la muestra es seleccionada o los datos son medidos. Un aumento del tamaño muestral no ayuda de ninguna manera a reducir el sesgo.