

Analítica de los Negocios

Datos y Muestreo

Carlos Cardona Andrade

Plan para hoy

1. ¿Qué son los datos?
2. Muestreo




¿Qué son los datos?

¿Qué son los datos?

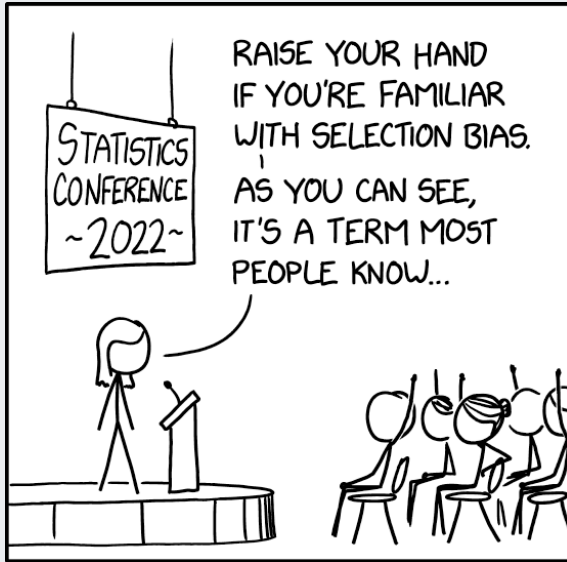
- Los datos representan **hechos sobre el mundo**, pero no son la **realidad misma**, sino una **forma de representación e interpretación**.
- Hoy llamamos “datos” a muchos elementos que antes no se consideraban como tal (nombres, fotos, videos, hábitos diarios).
- Los datos requieren **contexto** para ser significativos: vocabulario común, relaciones y convenciones interpretativas (es decir, metadatos).
- Existen múltiples formas de representar un mismo concepto, lo que hace necesaria una buena **arquitectura y gestión de datos**.

Datos vs Información

El modelo DIKW (Datos → Información → Conocimiento → Sabiduría) es común, pero presenta limitaciones:

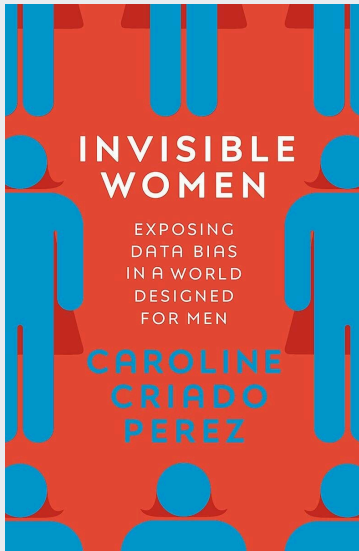
-  Asume que los datos existen por sí solos, cuando en realidad **los datos deben ser creados**.
 -  Plantea una secuencia lineal, sin considerar que se necesita conocimiento para generar datos.
 -  Separa datos e información, aunque están interrelacionados — los datos son una forma de información, y viceversa.
- ✓ Gestionar bien los datos requiere entender la relación compleja y cíclica entre datos, información y conocimiento.

Los límites de los datos



- ¡Los datos **nunca** son un reflejo perfecto del mundo!
- Es solo un subconjunto: no es el crimen, sino el **crimen reportado**
- La información es recopilada por humanos y procesada por máquinas: ¡las imprecisiones y errores son **inevitables**!
- Sean conscientes de los posibles sesgos (cognitivos y estadísticos)!

Mujeres invisibles



Caroline Criado Perez

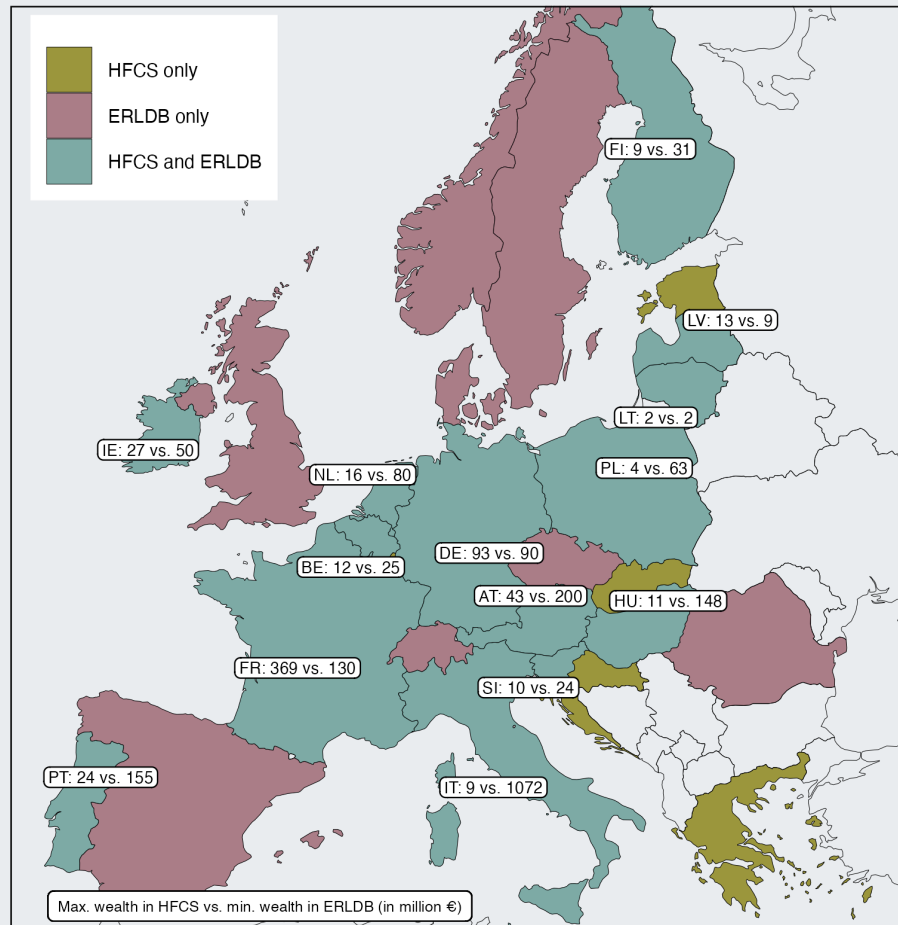
Exponiendo el sesgo de datos en un mundo diseñado para los hombres

Random House UK

ISBN: 978-1-78470-628-9

Diversas investigaciones, como las presentadas en el libro “Invisible Women” de Caroline Criado Perez, muestran cómo muchos de los datos utilizados en la toma de decisiones han sido históricamente recopilados sin considerar plenamente las diferencias de género. Esta **brecha de datos de género** puede tener implicaciones importantes en áreas como la medicina, el transporte o la planificación urbana. Para mejorar la equidad y la efectividad en las políticas públicas y el diseño de servicios, es fundamental incluir una mayor diversidad de perspectivas en los procesos de recolección y análisis de datos.

Ricos invisibles

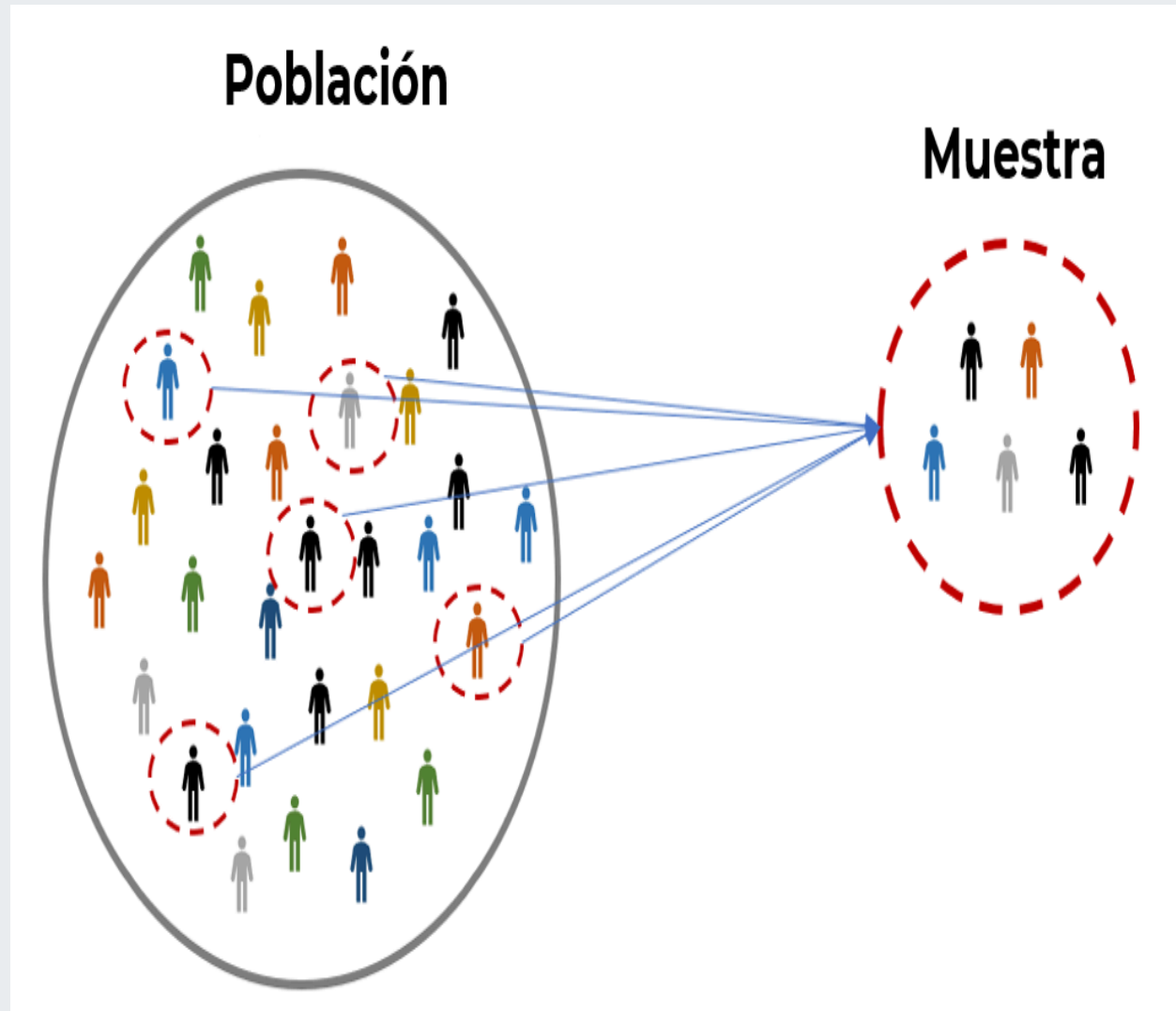


Los datos de encuestas se basan en muestras representativas extraídas de la población total. Sin embargo, la probabilidad de incluir a uno de los pocos hogares muy ricos en la muestra es infinitesimal. Además, la participación en encuestas es mayoritariamente voluntaria y la tasa de rechazo es más alta en la parte superior de la distribución. Esta **pobre cobertura de la cima** en las encuestas de riqueza e ingresos oculta la magnitud de la desigualdad.

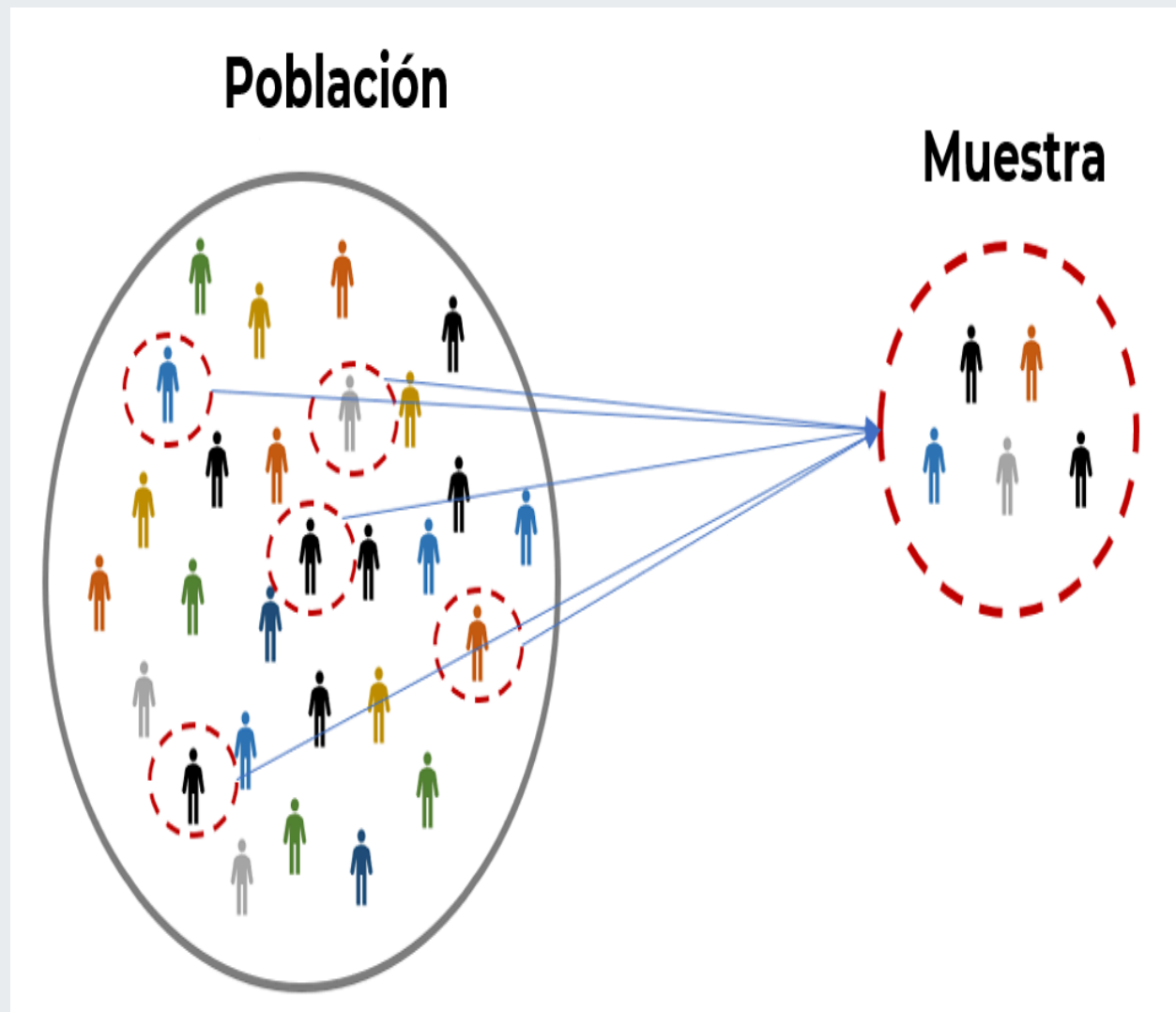
La figura muestra la brecha entre la observación más rica en los datos de encuestas de riqueza (HFCS) y la observación “más pobre” en listas nacionales de ricos elaboradas por revistas.

Muestreo

¿Por qué es importante el muestreo?



La incertidumbre importa!!



¿Por qué son útiles las muestras?

- Toma **menos tiempo** que seleccionar a cada ente de la población.
- Es **menos costoso** que seleccionar cada ente de la población.
- Es **imposible** recolectar datos de toda la población

Tipos de muestras

- No probabilística
 1. Muestreo por conveniencia
 2. Muestreo de respuesta voluntaria
- Probabilística
 1. Muestreo simple
 2. Muestre estratificado
 3. Muestreo por conglomerado
 4. Muestre en múltiples etapas

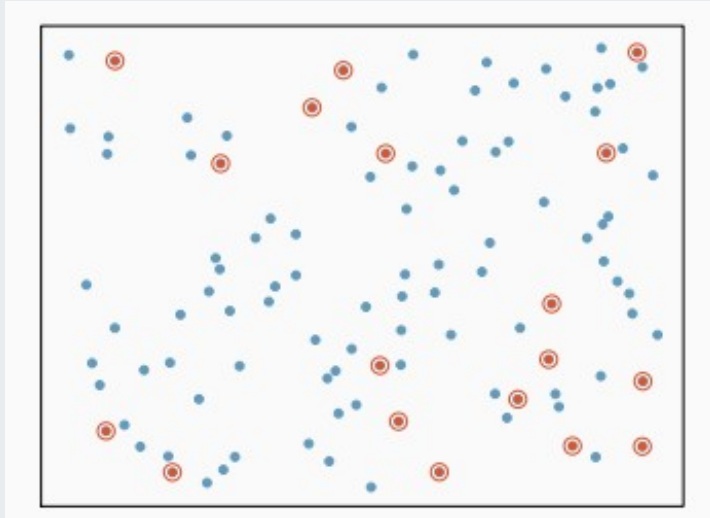
Malos ejemplos de muestreo

- Muestreo por conveniencia: consiste en tomar muestras de aquellos que son fácilmente accesibles.
 - Ejemplo: encuesta al “hombre en la calle” (barata, conveniente, popular en el “periodismo” televisivo).
 - Problema: los resultados pueden variar significativamente según el “cuándo y dónde” se realice la encuesta; falta de representatividad.

Malos ejemplos de muestreo

- Muestreo de respuesta voluntaria:
 - Ejemplo: encuestas en internet, encuestas telefónicas.
 - Solo las personas que visiten el sitio web o vean el programa serán muestreadas.
 - Las personas con opiniones fuertes son más propensas a participar.

Muestreo Simple



- Idea básica: colocar los nombres en una caja, agitar bien y sacar nombres al azar de la caja.
- Se necesita una lista de nombres de todos los sujetos de la población, llamada marco muestral.
- Todos los sujetos tienen la misma probabilidad de ser elegidos.

Muestreo Simple

- Pros: La composición de la muestra reflejará la composición de la población (edad/género/raza/ingresos...).
- Contras: La necesidad de un marco muestral lo hace poco práctico para poblaciones grandes

Muestreo Simple en R

```
1 library(tidyverse)
2 sample(x, size, replace = FALSE, prob = NULL)
```

- **x** - vector o conjunto de datos.
- **size** - tamaño de la muestra.
- **replace** - con o sin reemplazo de valores.
- **prob** - pesos probabilísticos

Muestreo Simple en R

```
1 # la muestra esta entre 1 y 5. El numero de muestras es 3
2 x <- sample(1:5, 3)
3 # veamos las 3 muestras
4 x
```

```
[1] 4 1 2
```

```
1 # la muestra esta entre 1 y 5. El numero de muestras es 6
2 x <- sample(1:5, 6)
```

Error in sample.int(length(x), size, replace, prob): cannot take a sample larger than the population when 'replace = FALSE'

```
1 x
```

```
[1] 4 1 2
```

```
1 #especificar replace=TRUE o T permitirá repeticiones de valores para que se
2 x <- sample(1:5, 6, replace=T)
3 x
```

```
[1] 1 1 3 1 3 4
```

Muestreo Simple en R - La Semilla

```
1 # establezcamos la semilla
2 set.seed(5)
3 #tomemos la muestra aleatoria con reemplazo
4 sample(1:5, 4, replace=T)
```

```
[1] 2 3 1 3
```

```
1 # cambiemos la semilla
2 set.seed(4)
3 sample(1:5, 4, replace=T)
```

```
[1] 3 3 3 4
```

```
1 set.seed(5)
2 sample(1:5, 4, replace=T)
```

```
[1] 2 3 1 3
```

Muestreo Simple en R

```
1 # creamos una lista de nombres y seleccionemos 1 al azar
2 sample(c('laura', 'diego', 'daniel', 'danna', 'manuela', 'ivan'), 1)
```

```
[1] "laura"
```

```
1 # Otro nombre
2 sample(c('laura', 'diego', 'daniel', 'danna', 'manuela', 'ivan'), 1)
```

```
[1] "laura"
```

```
1 # Otro nombre
2 sample(c('laura', 'diego', 'daniel', 'danna', 'manuela', 'ivan'), 1)
```

```
[1] "manuela"
```

```
1 # Otro nombre
2 sample(c('laura', 'diego', 'daniel', 'danna', 'manuela', 'ivan'), 1)
```

```
[1] "ivan"
```

Muestreo Simple en R - Los pesos

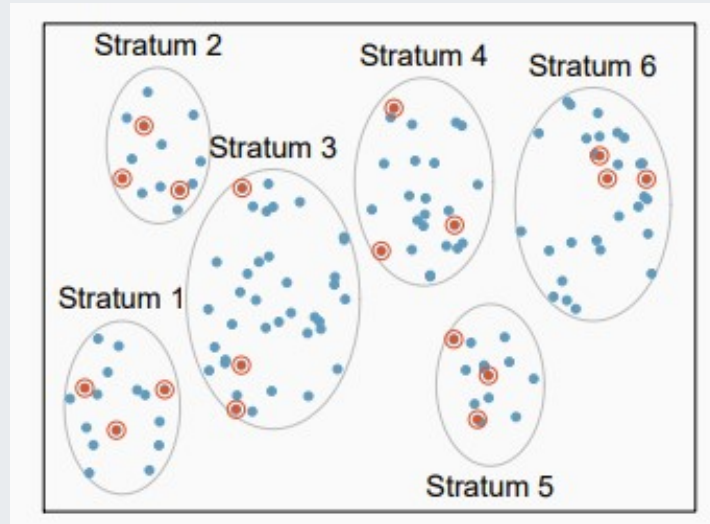
```
1 # creamos una probabilidad de 80% para Bueno y 20% para malo.  
2 sample (c('Bueno', 'Malo'), size=10, replace=T, prob=c(.80, .20))
```

```
[1] "Bueno" "Bueno" "Bueno" "Bueno" "Bueno" "Bueno" "Bueno" "Bueno" "Bueno" "Malo"  
[10] "Bueno"
```

```
1 # Establezcamos otras probabilidades  
2 sample (c('Bueno', 'Malo'), size=10, replace=T, prob=c(.60, .40))
```

```
[1] "Malo" "Malo" "Malo" "Bueno" "Bueno" "Bueno" "Bueno" "Bueno" "Bueno" "Malo"  
[10] "Bueno"
```

Muestreo Estratificado



La población se divide en grupos llamados **estratos**, y luego se elige una muestra aleatoria simple en cada estrato.

- Funciona mejor cuando los casos dentro de un estrato son similares, pero existen grandes diferencias entre los estratos.
- Inconvenientes: Se necesita un marco muestral para cada estrato, lo cual no es práctico para poblaciones grandes.

Muestreo Estratificado en R - Por tamaño

```
1 library(tidyverse)
2
3 # Creemos la población
4 data <- data.frame(grupo= rep(c('Profesores', 'Estudiantes', 'Admin', 'Invi
5 head(data)
```

```
      grupo felicidad
1 Profesores  95.08553
2 Profesores  85.31956
3 Profesores  84.40746
4 Profesores  89.90350
5 Profesores  90.74338
6 Profesores  89.21532
```

```
1 # Obtengamos la muestra estratificada
2 estrat_muestra <- data |>
3   group_by(grupo) |>
4   sample_n(size=15)
5
6 table(estrat_muestra$grupo)
```

Admin	Estudiantes	Invitados	Profesores
15	15	15	15

Muestreo Estratificado en R - Por proporción

```
1 # Obtengamos la muestra estratificada por proporción
2 estrat_muestra <- data |>
3   group_by(grupo) |>
4   sample_frac(size=.20)
5
6 # Veamos la frecuencia de personas para cada grupo
7 table(estrat_muestra$grupo)
```

Admin	Estudiantes	Invitados	Profesores
30	30	30	30

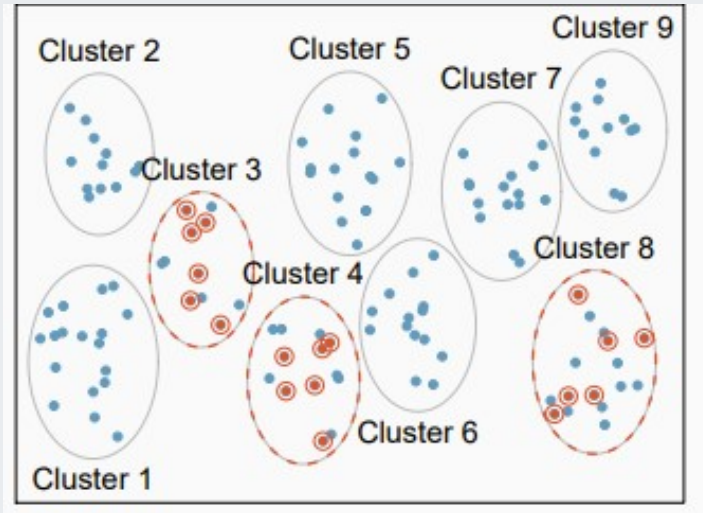
Muestreo por conglomerados

Muestreo por conglomerados en R

```
1 # Generemos la muestra
2 set.seed(123)
3 poblacion <- data.frame(
4   supermercado = paste("Supermercado", 1:1000, sep = "_"),
5   SatisfaccionCliente = rnorm(1000, mean = 75, sd = 10)
6 )
7
8 # Muestreo por conglomerado
9 supermercados_elegidos <- sample(poblacion$supermercado, size = 10, replace = FALSE)
10 muestra <- poblacion |>
11   filter(supermercado %in% supermercados_elegidos)
12
13 # Veamos la muestra
14 head(muestra)
```

	supermercado	SatisfaccionCliente
1	Supermercado_203	72.34855
2	Supermercado_225	71.36343
3	Supermercado_255	90.98509
4	Supermercado_354	76.16637
5	Supermercado_457	86.10277
6	Supermercado_554	77.49825

Muestreo en múltiples etapas



- Primera etapa: la población se divide en grupos, llamados **conglomerados**, y se elige una muestra de grupos.
- Segunda etapa: los grupos seleccionados se dividen a su vez en subgrupos, y se elige una muestra de subgrupos en cada grupo seleccionado.
- (Tercera etapa: ...)
- (Cuarta etapa: ...)

Muestreo en múltiples etapas

Muchas encuestas nacionales (como la Encuesta General Social) utilizan muestreo en cuatro etapas.

- departamentos → ciudades → barrios → hogares

Ventaja:

- Los sujetos seleccionados vivirán todos en las ciudades seleccionadas, y no estarán dispersos por todo el país, lo que puede reducir significativamente los costos de viaje de los entrevistadores.
- No es necesario elaborar un marco muestral para los subgrupos no seleccionados.

Muestreo en múltiples etapas en R

```
1 # Generación de los datos
2 set.seed(123)
3 barrios <- data.frame(
4   barrio = paste("Barrio", 1:500, sep = "_"),
5   IngresoMedio = rnorm(500, mean = 50000, sd = 10000)
6 )
7
8 hogares <- data.frame(
9   barrio = rep(sample(barrios$barrio, size = 500, replace = TRUE),
10                each = 20),
11   IDhogar = rep(1:20, times = 500),
12   TipoEmpleo = sample(c("Ocupado", "Desempleado"), size = 10000, replace =
13 )
14
15 head(hogares)
```

	barrio	IDhogar	TipoEmpleo
1	Barrio_13	1	Desempleado
2	Barrio_13	2	Ocupado
3	Barrio_13	3	Ocupado
4	Barrio_13	4	Desempleado
5	Barrio_13	5	Ocupado
6	Barrio_13	6	Ocupado

Muestreo en múltiples etapas en R

```
1 # Acá viene el muestreo por etapas
2 barrios_elegidos <- sample(barrios$barrio, size = 5, replace = FALSE)
3 hogares_muestra <- hogares |>
4   filter(barrio %in% barrios_elegidos) |>
5   group_by(barrio) |>
6   sample_n(10) |>
7   ungroup()
8
9 head(hogares_muestra)
```

```
# A tibble: 6 × 3
  barrio      IDhogar TipoEmpleo
  <chr>      <int>   <chr>
1 Barrio_196      11 Desempleado
2 Barrio_196      10 Desempleado
3 Barrio_196      12 Desempleado
4 Barrio_196      13 Ocupado
5 Barrio_196      17 Ocupado
6 Barrio_196      16 Desempleado
```

Problemas de muestreo - Sesgo de selección

Una tendencia sistemática por parte del procedimiento de muestreo a excluir un tipo de persona u otro de la muestra se llama sesgo de selección.

- Las personas sin dirección permanente son excluidas por las encuestas por correo.
- Aproximadamente 1/3 de los teléfonos residenciales no están listados. Los ricos y los pobres tienen más probabilidades de tener números no listados, por lo que el directorio telefónico tiende hacia la clase media.
- Se ha encontrado que las mujeres tienen más probabilidades de contestar el teléfono que los hombres. Las encuestas telefónicas a menudo incluyen más mujeres que hombres.
- Cuando un procedimiento de selección está sesgado, tomar una muestra grande no ayuda. Esto es solo repetir el mismo error a una mayor escala.

Problemas de muestreo - Sesgo de no respuesta

El sesgo de no respuesta causa problemas porque los que no responden pueden ser muy diferentes de los que sí responden.

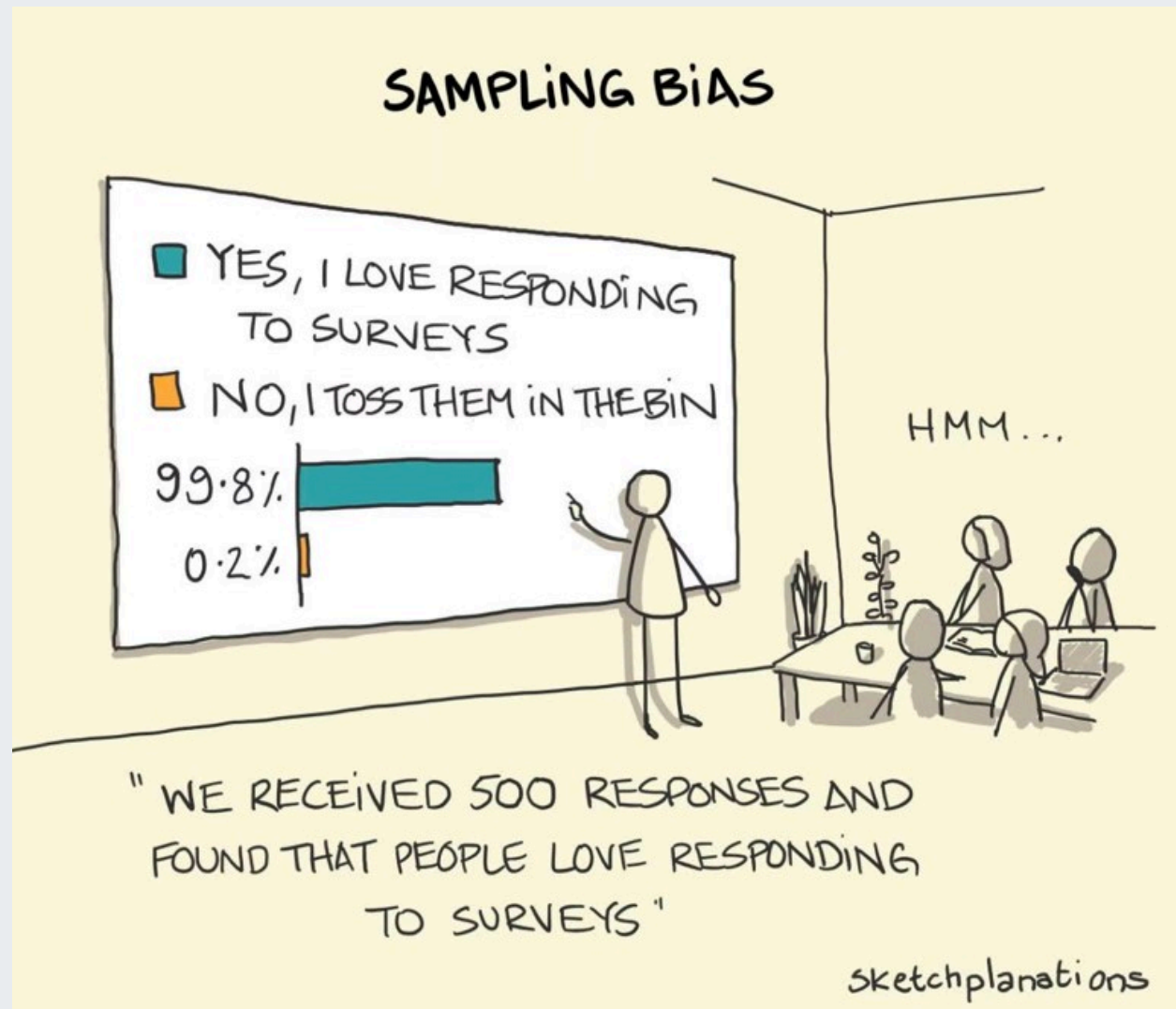
- Los que no responden pueden tener largas horas de trabajo, vivir solos con más probabilidad, o no molestarse en responder, etc.
- Cuando la tasa de respuesta es baja, no se puede tomar una nueva muestra para reemplazar a quienes no responden.
- Se debe intentar contactar a los que no responden, haciendo más llamadas/visitas, ofreciendo recompensas, etc.
- Siempre verifique la tasa de respuesta. Si es baja, el resultado de la encuesta podría no ser confiable.

Problemas de muestreo - Sesgo de respuesta

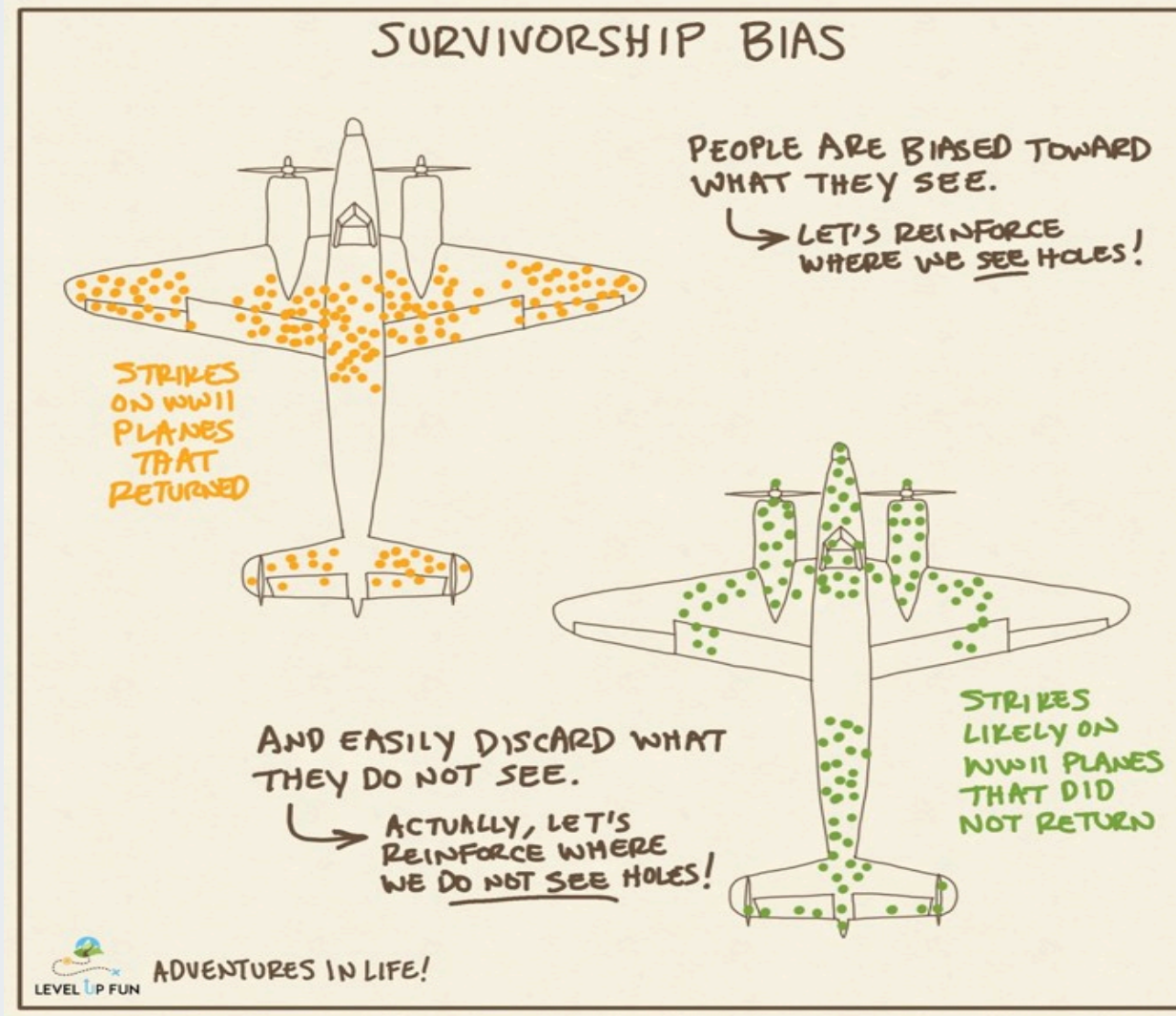
El sesgo de respuesta significa que las respuestas de los encuestados están influenciadas, en cierta medida, por la formulación de las preguntas e incluso por el tono o la actitud del entrevistador.

Solución: control del entrevistador y diseño adecuado de los cuestionarios.

Sesgo de muestreo

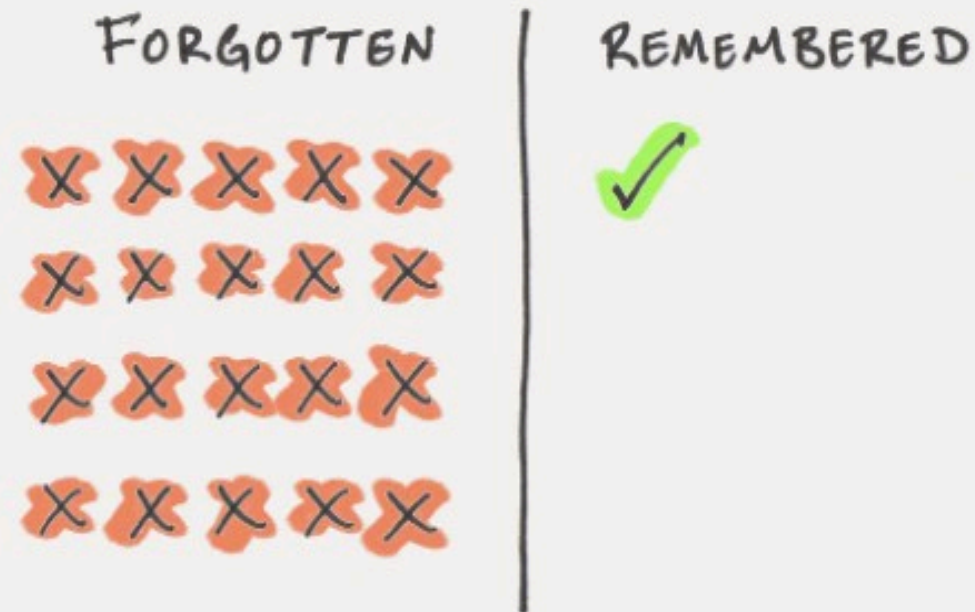


Sesgo de supervivencia



Sesgo de supervivencia

THE SURVIVORSHIP BIAS



JamesClear.com

