Analítica de datos

# Recolectando datos: muestreo, encuestas

Pontificia Universidad JAVERIANA Bogotá

Profesor: Nicolás Velásquez

# Los datos se recolectan de toda una población o una muestra

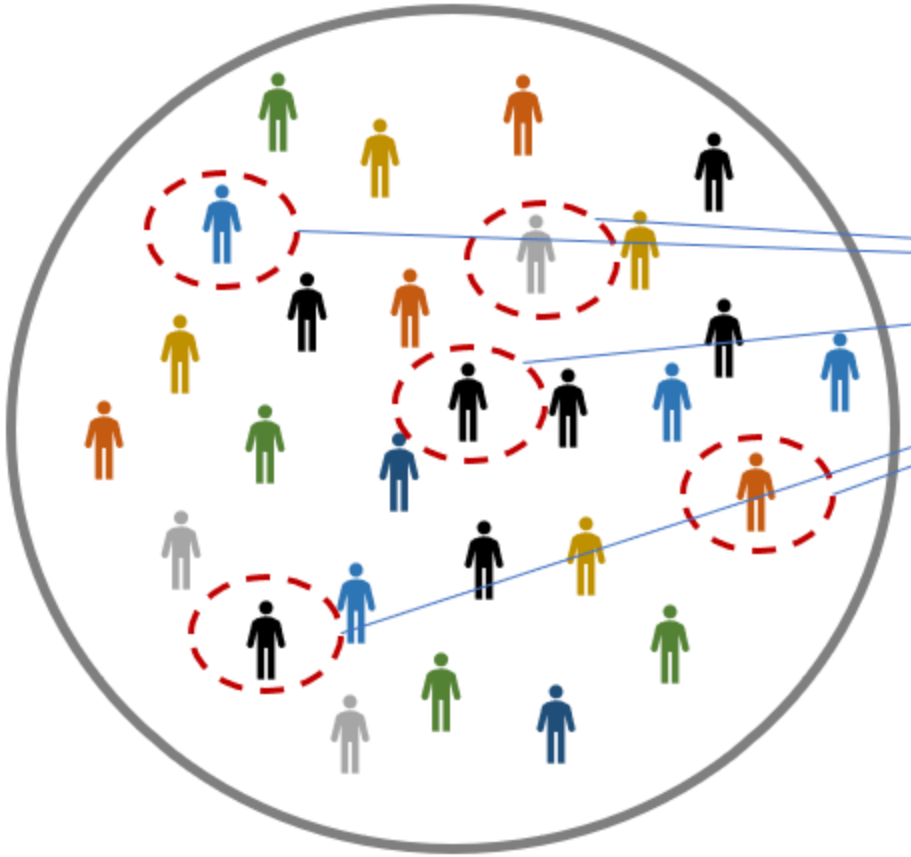
## POBLACIÓN

La **población** contiene todos los entes o individuos de interés que queremos estudiar.

## MUESTRA

Una **muestra** contiene solamente una porción de la población de interés.

## Población



Todos los entes o individuos

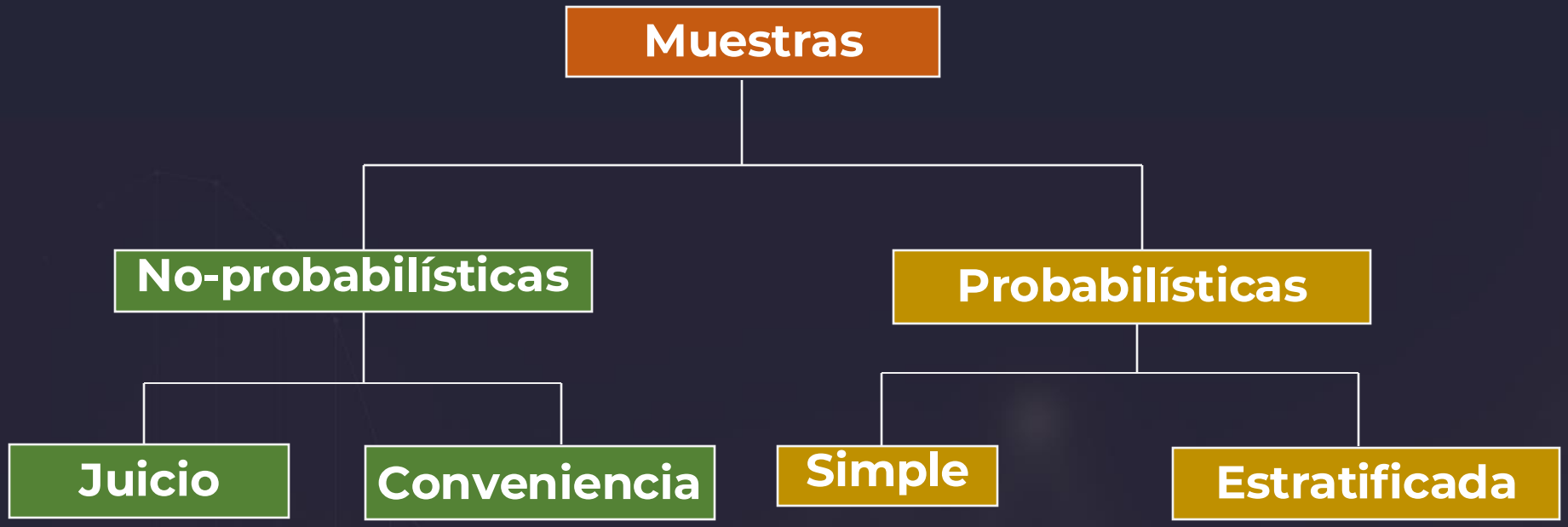
## Muestra

Una porción de la población

# Se recolectan datos mediante muestreo cuando....

- **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



# Tipos de muestras:
# Muestras no-probabilísticas

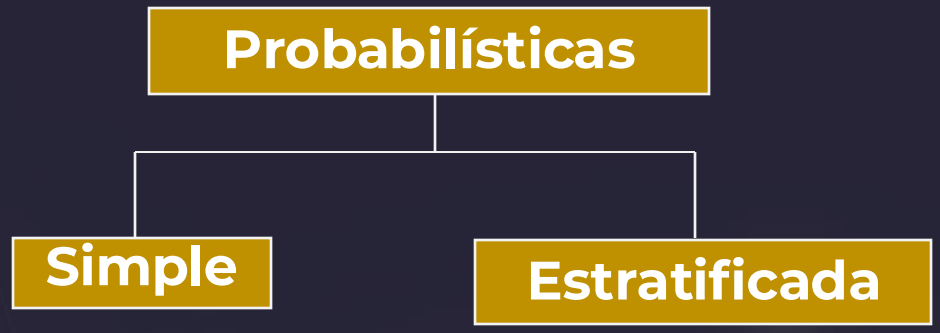
- En una muestra no-probabilística, los entes incluidos se escogen sin considerar su representatividad.

  - En **muestreo de conveniencia**, los entes son seleccionados porque son fáciles de conseguir, baratos.

  - En **muestreo de juicio,** se obtienen datos de expertos en el área.

# Tipos de muestras:
# Muestras probabilísticas

- En una **muestra probabilística**, los entes en la muestra son seleccionados en base a probabilidades



# Muestras probabilísticas:
# Muestreo Simple

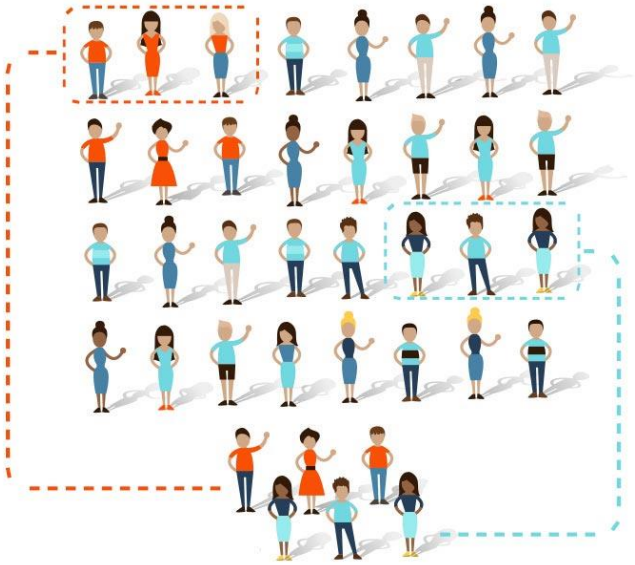
- Cada individuo o ente de una población tiene la **misma probabilidad** de ser seleccionado.

- La selección se puede hacer **con reposición** o **sin reposición.**

- Muestras se generan basándose en alguna tabla de números aleatorios o generadores de números aleatorios.

# Muestras probabilísticas: Muestreo estratificado

- Dividir población en dos o más subgrupos (llamados estratos) de acuerdo a alguna característica en común.
- Se toma una muestra aleatoria simple de cada sub-grupo, con el tamaño de la sub-muestra proporcional al tamaño del estrato en la población.
- Común cuando se muestrean poblaciones de votantes, haciendo la estratificación por grupo socioeconómico, departamento, etc..



| REGION DE MARKETING | POBLACIÓN | | MUESTRA ESTRATIFICADA PROPORCIONAL | MUESTRA ESTRATIFICADA PROPORCIONAL |
|---|---|---|---|---|
| | FRECUENCIA | PORCENTAJE | FRECUENCIA | PORCENTAJE |
| ZONA 1 | 18000 | 33% | 396 | 33% |
| ZONA 2 | 600 | 1% | 12 | 1% |
| ZONA 3 | 12000 | 22% | 264 | 22% |
| ZONA 4 | 24000 | 44% | 528 | 44% |
| **TOTAL** | **54600** | **100%** | **1200** | **100%** |

# Muestras probabilísticas: Comparando métodos

- Muestreo simple:
  - Fácil de usar.
  - Podría, al final, no representar correctamente a la población.
    -> Más probable si.....

- Muestreo estratificado:
  - Asegura que el muestreo sea representativo.

# Tamaño de la muestra

Si queremos una muestra representativa de Javeriana con aprox. 8500 estudiantes de pregrado, 4500 estudiantes de posgrado, 4200 profesores entre planta y cátedra y más de 1500 empleados administrativos como población.

- **Si la muestra es demasiado pequeña**, podría incluir una cantidad desproporcionada de encuestados atípicos (o “outliers”, en inglés), que generen valores anómalos. Esto puede producir una distorsión de los resultados, lo que no le permitirá obtener una visión realista de la población del estudio.
- **Si la muestra es demasiado grande**, la investigación se vuelve demasiado compleja, costosa y lenta. Si bien los resultados serán más precisos, sus beneficios no superarán los costos de realizar el estudio.

https://www.qualtrics.com/blog/calculating-sample-size/

https://es.surveymonkey.com/mp/sample-size-calculator/

# ENCUESTAS

# Evaluando una encuesta

- ¿Está basada en muestreo probabilístico?

- Hay que preocuparse por:

  - Error de cobertura
  - Error de no-respuesta
  - Error de medición
  - (Error de muestreo)

# Tipos de errores que llevan a “sesgos”

- **Error de cobertura:**

*Algunas personas son excluidas del muestreo, y como consecuencia el muestreo no es representativo de la población.*

- **Error de no-respuesta:**

*Algunas personas (seleccionadas aleatoriamente) no responden la encuesta. Esto es un problema porque…*

- **Error de muestreo:**

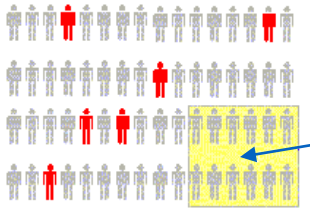
*Siempre presente, mientras tomemos una muestra de la población y no toda la población.*

- **Error de medición:**

*Debido a alguna debilidad en el diseño de alguna pregunta y/o errores del encuestado al momento de responder. Puede ser sistemático (problema grande) o aleatorio (problema).*

# Tipos de errores *(continuación)*

- Error de cobertura



**Excluidos del muestreo**

- Error de no-respuesta

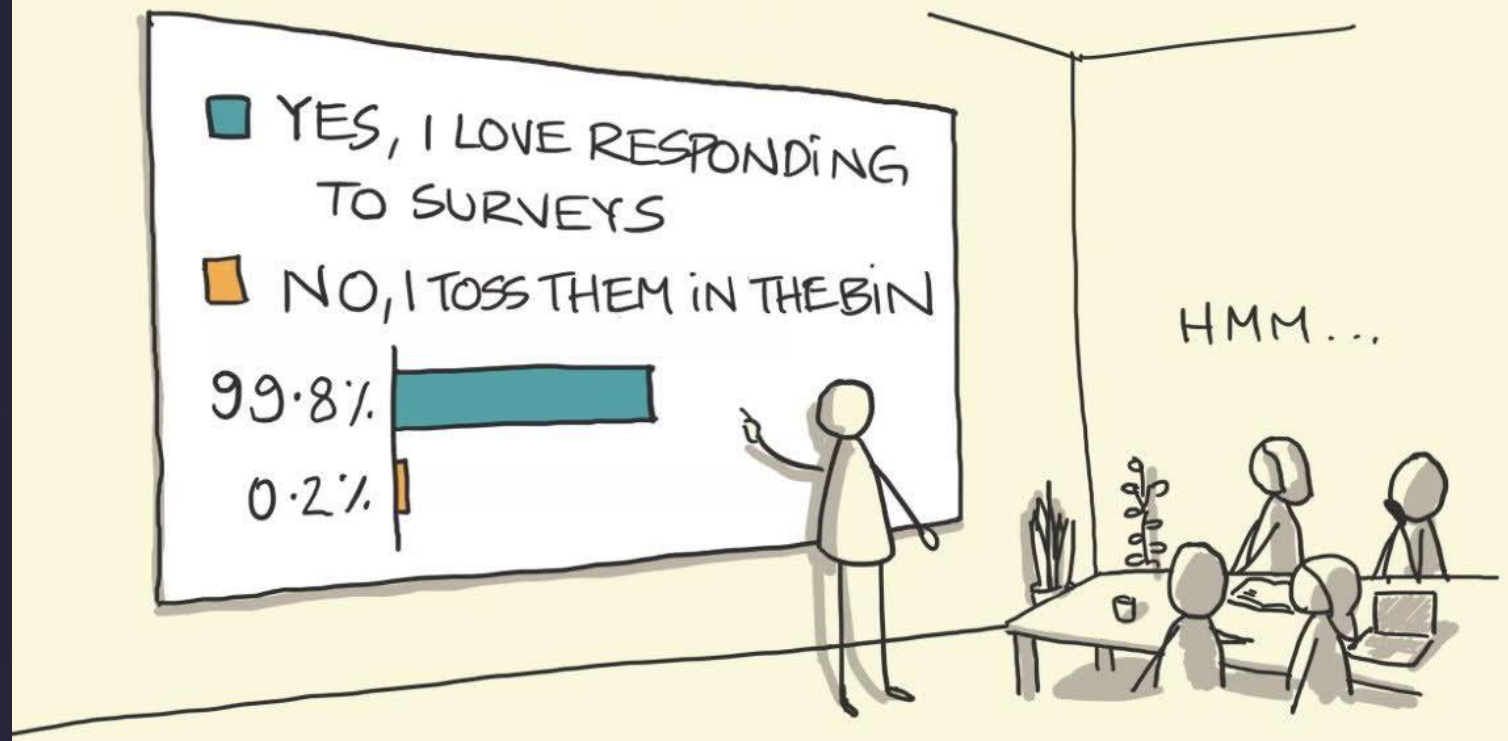
- Error de muestreo

**Siempre presente**

- Error de medición

**Pregunta "mal" diseñada o que sugiere alguna respuesta.**

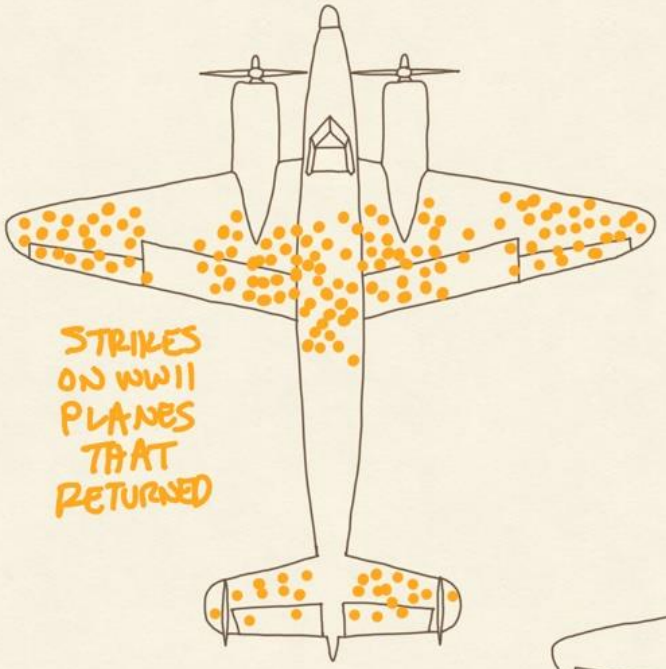
# SAMPLING BIAS




"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"

sketchplanations

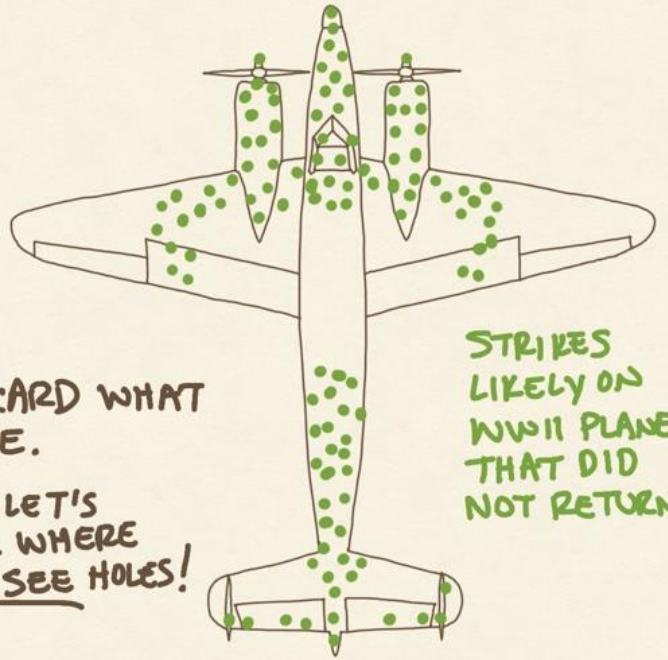
# SURVIVORSHIP BIAS



STRIKES ON WWII PLANES THAT RETURNED

PEOPLE ARE BIASED TOWARD WHAT THEY SEE.

↳ LET'S REINFORCE WHERE WE SEE HOLES!



STRIKES LIKELY ON WWII PLANES THAT DID NOT RETURN

AND EASILY DISCARD WHAT THEY DO NOT SEE.

↳ ACTUALLY, LET'S REINFORCE WHERE WE DO NOT SEE HOLES!

LEVEL UP FUN ADVENTURES IN LIFE!

# Cuestiones éticas en las encuestas

- El error de cobertura puede ser utilizado por el diseñador para sesgar los resultados de la encuesta.

- Siempre se debe mencionar el error de muestreo: incluir márgenes de error y no tratar la información como si reflejara precisamente a la población.

- Error de medición:
  - Selección/redacción de preguntas que sugieran cierta respuesta.

# Diseñando una buena encuesta

- **Cuide el orden de las preguntas**:
  - Preguntas sencillas al inicio, preguntas más difíciles al final.
  - Preguntas "delicadas" al final.

- **Verificar redacción de preguntas** (y posibles respuestas):
  - Evitar conceptos técnicos.
  - Definir conceptos si es necesario.
  - Preguntas lo más sencillas posibles.
  - Definir temporalidad.
  - Posibles respuestas deben ser exhaustivas.

# Diseñando una buena encuesta

- **Haga un pre-test/piloto:**
  - identificar preguntas que no son claras, percepción del tiempo que tomará responder la encuesta.
  - Pregunte a los encuestados que piensen en voz alta al momento de responder. ¿Qué errores podríamos evitar?

- Haga la encuesta lo más **corta** posible. ¿Qué errores podríamos evitar?
  - Solamente incluir preguntas que son relevantes para la pregunta que queremos responder con la encuesta.
  - Utilizar "branching" (ramificaciones) y filtros.

# Error de medición: sesgo de "deseabilidad social"

- Básicamente sucede cuando al encuestado le da vergüenza responder de cierta forma a cierta pregunta (presión social para responder de cierta forma):

  - ¿Sí o No a la paz?
  - ¿Es usted racista?
  - ¿Ha consumido drogas ilícitas?
  - ¿Ha sobornado a un funcionario público?
  - ¿Ha robado en su lugar de trabajo?

- Si el sesgo/presión para responder de cierta forma es significativo, esto resulta en error de medición (y en particular, en un error de medición sistemático).

# Error de medición: sesgo de "deseabilidad social"

- Hay varias alternativas que pueden disminuir el sesgo:
    - Encuestas telefónicas automatizadas.
    - Encuestas por internet.
    - Métodos para asegurar anonimidad: respuesta en privado y urnas.
- Método de respuesta aleatoria.
- Método de lista (Item Count Technique).

**Figura 2:** Modelo de preguntas con respuesta aleatorizada de Warner

A continuación, contesta la opción **A** si tu cédula de identidad termina en 1 o 2, en caso contrario contesta la opción **B.**

| opción **A** | opción **B** |
|---|---|
| Nunca hice trampa en un examen de la universidad | Alguna vez hice trampa en un examen de la universidad |
| SI ◯ | NO ◯ |

A continuación, contesta la opción **A** si tu cédula de identidad termina en 8 o 9, en caso contrario contesta la opción **B.**

| opción **A** | opción **B** |
|---|---|
| Nunca he consumido drogas | He consumido drogas en alguna ocasión |
| SI ◯ | NO ◯ |

# Método de lista (ICT)

**Table 1.** Experimental Design

Panel A: Comparison of Direct Report and Veiled Report treatments

| Direct Report | Veiled Report |
|---|---|
| • I remember where I was the day of the *Challenger* space shuttle disaster.<br>• I spent a lot of time playing video games as a kid.<br>• I would vote to legalize marijuana if there was a ballot question in my state.<br>• I have voted for a political candidate who is pro-life. | • I remember where I was the day of the *Challenger* space shuttle disaster.<br>• I spent a lot of time playing video games as a kid.<br>• I would vote to legalize marijuana if there was a ballot question in my state.<br>• I have voted for a political candidate who is pro-life.<br>• I consider myself to be heterosexual. |
| Please fill in the bubble that corresponds to the total number of statements above that apply to you.<br>0 1 2 3 4 | Please fill in the bubble that corresponds to the total number of statements above that apply to you.<br>0 1 2 3 4 5 |
| Do you consider yourself to be heterosexual?<br>Yes No | |

Coffman et. al. (2017), "*The Size of the LGBT population and the Magnitude of Antigay Sentiment are Substantially Underestimated*". Management Science.

# Método de lista (ICT)

- Definamos:

  - #D - promedio de ítems seleccionados como verdaderos en la encuesta con la pregunta directa (4 ítems en el ejemplo).

  - #V - promedio de ítems seleccionados como verdaderos en la encuesta con la pregunta indirecta (5 ítems en el (ejemplo).

  - %D - proporción/porcentaje de personas que respondieron no a la pregunta directa sobre orientación sexual.

# Método de lista (ICT)

- Entonces:

- La proporción estimada (%) de las personas cuya respuesta verdadera es: #V (5 ítems) - #D (4 ítems).

- El sesgo se puede medir comparando la proporción estimada (%) anteriormente y %D.


Recursos:

- https://www.questionpro.com/blog/es/preguntas-para-una-encuesta/
- https://www.ugr.es/~diploeio/documentos/tema2.pdf