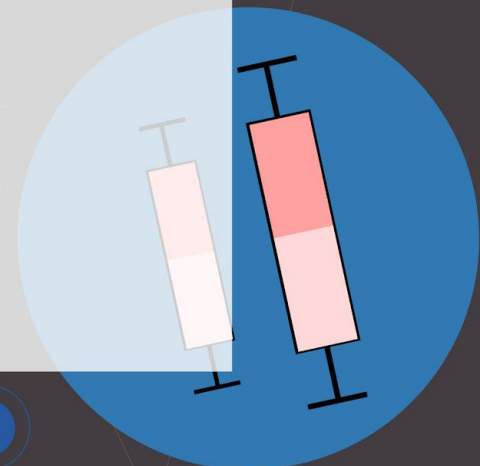


Herramientas Estadísticas y Forecast

Clase 2

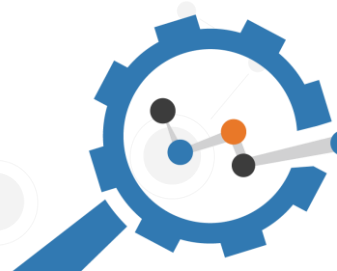


FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE



Contenido del curso

1. Introducción a la estadística y análisis descriptivo
2. Análisis descriptivo y gráfico
3. Probabilidad y Distribuciones
4. Muestreo
5. Inferencia estadística: Pruebas de hipótesis
6. Taller Práctico de inferencia
7. Introducción a los modelos estadísticos
8. Modelos predictivos I: Modelos de regresión lineal.
9. Modelos predictivos II: Regresión logística y otros modelos.
10. Modelos de Forecasting I.
11. Modelos de Forecasting II
12. Taller de Forecast

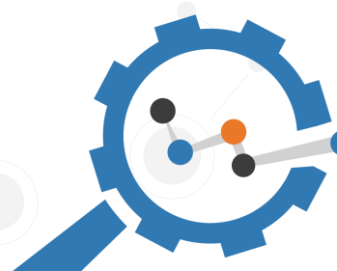


Introducción a la inferencia

Pregunta:

¿Cuántos de ustedes vive solo?

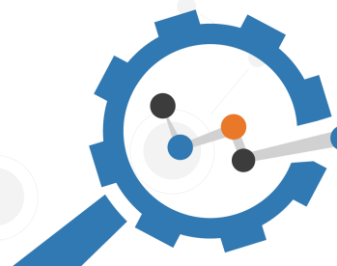
> Censo 2017 declara que en Santiago, el 17% vive en un hogar unipersonal.



Introducción a la inferencia

Denominaremos por **estimación** al proceso estadístico que busca valorizar un posible parámetro de la población a través de una muestra.

Las estimaciones no serán exactamente igual al parámetro poblacional, pero hay herramientas para lograr que sea lo más similar posible. A considerar el **Tipo de Muestra** y el **Nivel de Error**

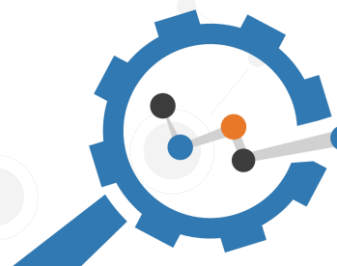


Muestreo

¿Qué es el muestreo?

El muestreo consiste en un procedimiento de selección de elementos de una población y en que se observa alguna característica de estos elementos, con el fin de estimar algo sobre la población entera

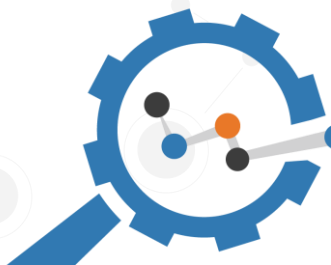
El objetivo del muestreo es hacer inferencia sobre una población con los datos observados en la muestra



Muestreo

Ejemplos:

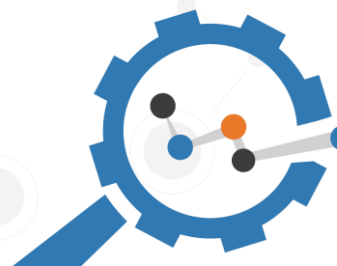
- Encuesta telefónica para estimar la opinión política antes de las elecciones.
- Muestra de clínicas para estimar la prevalencia de una enfermedad.
- Muestra de sitios para estimar la cantidad de petróleo en una región de un país.



Muestreo

Definiciones importantes:

- **Población objetivo:** Colección completa de todas las N unidades (personas, animales, terrenos, objetos, etc.) que se quieren estudiar para la generalización de las conclusiones.
- **Variables de interés:** Características propias de las unidades sobre la que se realiza la inferencia
- **Muestra:** es un subconjunto de la población objetivo.

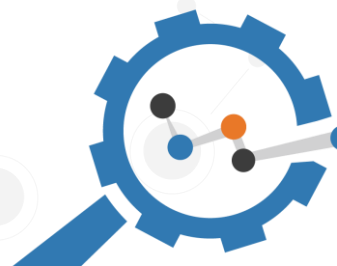


Muestreo

Ejemplo:

Se quiere estudiar el sueldo mensual promedio de los 658 empleados de la empresa ABC. Se seleccionó 150 empleados al azar sin remplazo y les pidieron el monto de su sueldo.

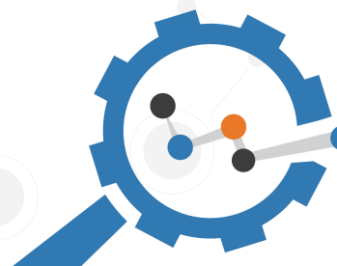
- Población objetivo: 658 empleados de la empresa ABC.
- Variable de interés: Sueldo mensual de los empleados
- Muestra: 150 empleados de la empresa ABC.



Diseño de muestreo

El diseño de muestreo es un procedimiento por el que se extrae la muestra, estos podemos clasificarlos en:

- Muestreo probabilístico
- Muestreo no probabilístico



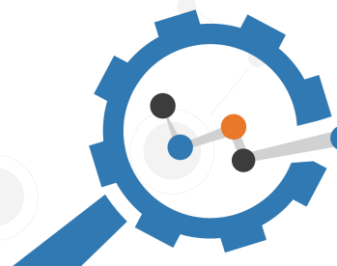
Muestreo probabilístico

Un muestreo se dice probabilístico si podemos calcular la probabilidad de extracción de una determinada muestra.

Es necesario que las unidades de la población objetivo sean identificables para que sea un muestreo probabilístico (para poder seleccionar las unidades). En particular, es necesario un **marco muestral**.

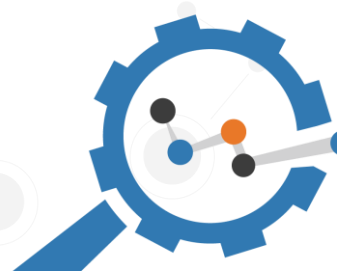
Los tipos de muestreo más usuales son:

- Muestreo aleatorio simple
- Muestreo aleatorio sistemático
- Muestreo estratificado
- Muestreo por conglomerados



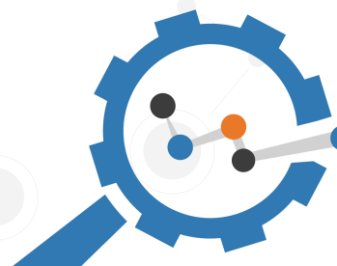
Marco muestral

El Marco muestral es una lista de elementos que componen la población objetivo que queremos estudiar y de la cuál se extra la muestra.



Muestreo probabilístico

Suponga que se desea estudiar la brecha salarial de los trabajadores hombres y mujeres dependientes en las empresas del país. ¿Cómo se pueden obtener 50 empresas para estudiar?

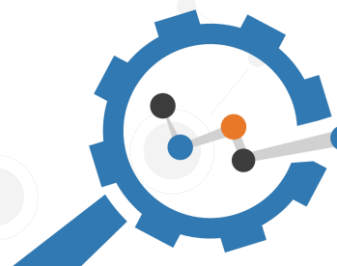


Muestreo aleatorio simple

Una muestra aleatorio simple (MAS) de tamaño n es un diseño de muestreo en el que se seleccionan n unidades de una población de tamaño N de modo que cada combinación de n unidades tiene la misma probabilidad de ser seleccionada.

Ejemplo:

Utilizando el directorio de todas las empresas del país, se seleccionan n al azar.

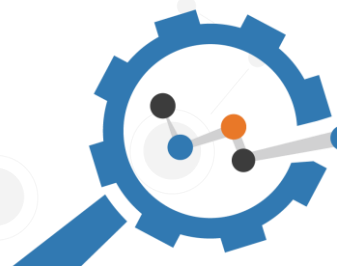


Muestreo aleatorio sistemático

Un muestreo aleatorio sistemático, consiste en enlistar a toda la población, elegir de forma aleatorio el primer individuo para la muestra y luego definir un intervalo de selección .

Ejemplo:

A través del directorio ordenado, se elige las empresas se seleccionan desde la 1 a la 101, de la 201 a la 301, ...



Muestreo estratificado

En un muestreo estratificado, la población se divide en estratos, (subgrupos de la población) y la muestra se seleccionan en cada estrato de manera independiente.

Este método es utilizado cuando podemos separar la población en grupos internamente homogéneos y heterogéneos entre ellos.

Ejemplo:

Se estratifica según el tamaño de las empresas, eligiendo de forma independiente algunas micro, pequeña, mediana y grandes empresas

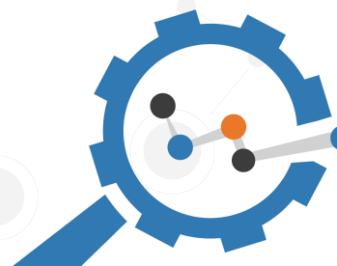
Muestreo por conglomerado

En un muestreo por conglomerados, la población se divide en grupos homogéneos, pero internamente heterogéneos.

Este muestreo es usualmente utilizado para reducir los costos del estudio a diferencia del muestreo estratificado, el cual tiene por objetivo incrementar la exactitud.

Ejemplo:

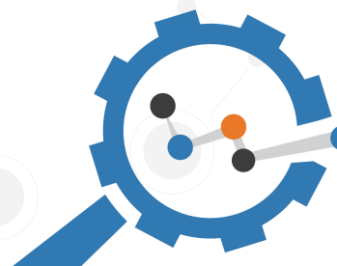
De los 135 sub rubros, se eligen 5 y se toman empresas de ellos.



Muestreo no probabilístico

Un muestreo se dice no probabilístico si no se puede calcular la probabilidad de extracción de una determinada muestra. Algunos tipos de muestreo no probabilístico son:

- Muestreo por conveniencia
- Muestreo por cuotas
- Muestreo bola de nieve



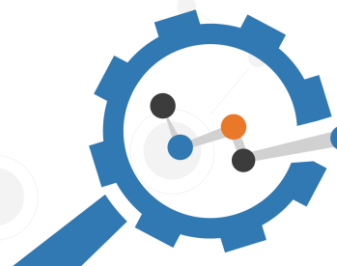
Muestreo no probabilístico

- **Muestreo por conveniencia:** Un médico elige a sus pacientes para hacer un estudio sobre una cierta enfermedad
- **Muestreo por cuotas:** La población está dividida en estratos. El investigador elige a los sujetos dentro de cada estrato
- **Muestreo bola de nieve:** Un pequeño grupo de personas reclutando a todos sus amigos en el estudio. El proceso sigue así hasta que el tamaño de muestreo se alcance.



Muestreo

La principal diferencia entre estos dos tipos de diseño es que el muestreo probabilístico no depende de la subjetividad del ser humano. La probabilidad de que una unidad de la población sea seleccionada no es sesgada por la opinión del investigador.



Tipos de errores

Error muestral: Error que se atribuye a la observación parcial de la población

Ejemplo:

En el estudio del sueldo promedio de la empresa ABC, con la muestra, se estima un sueldo promedio de 778.329 CLP al mes. Se tiene que el sueldo mensual promedio de la población es de 823.598 CLP al mes, la diferencia entre los \$778.329 y \$823.598 es el error muestral. Se atribuye a la selección de la muestra, es decir, solo se observaron 150 de los empleados.



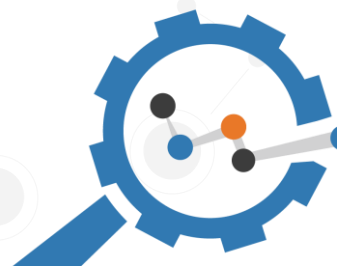
Tipos de errores

Error no muestral: error que no se atribuye a la selección de la muestra, es decir, al diseño muestral

Ejemplo:

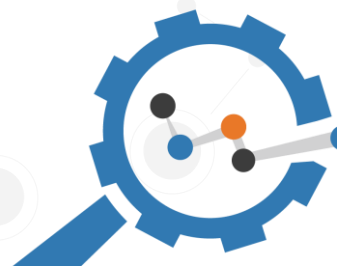
En el estudio de los sueldos promedio,

- Los ejecutivos decidieron no reportar su sueldo (error de no respuesta).
- Los encuestados olvidaron un componente de sueldo (ej. Bonificación, comisión, etc.).



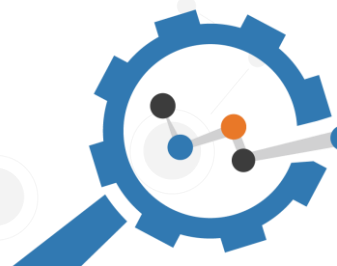
Factor de Expansión

El factor de expansión es la capacidad que tiene cada individuo seleccionado en una muestra probabilística para representar el universo en el cual esta contenido.



Muestreo

La compañía telefónica ABC lo contrata para un estudio que busca demostrar que poseen cobertura y una señal estable en todo Santiago, ¿cómo podría llevar acabo este estudio?



Diseño Muestral CASEN

Población objetivo

Esta constituida por los hogares que habitan las viviendas particulares ocupadas y las personas que residen en forma permanente.

Cobertura

La cobertura geográfica del estudio es Nacional. Sin embargo se excluyen algunas áreas alejadas y de difícil acceso.

Diseño de la muestra

Probabilístico, estratificado, por conglomerado y en múltiples etapas. La unidad última de selección es la vivienda, la unidad de interés es el hogar y la unidad de reporte son las personas miembros del hogar.

Diseño Muestral CASEN

Representatividad

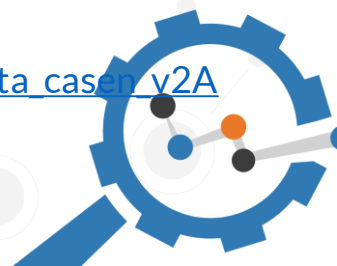
El diseño muestral es representativo para los siguientes dominios de estudio: nacional, regional, comunal y por zona urbano-rural.

Métodos de recolección de datos

Entrevista personal, aplicada por encuestador en cuestionario de papel.
Versión 2009 incluyó aprox. 330 preguntas .

Metodología encuesta CASEN:

http://observatorio.ministeriodesarrollosocial.gob.cl/layout/doc/casen/metodologia_encuesta_casen_v2A_O_504e0c95c24d3.pdf



Inmigrantes en Chile CADEM

Técnica

Encuesta web aplicada a través de nuestra plataforma Cadem Online, utilizando un cuestionario semi estructurado, con una duración aproximada de 10 minutos, a una base de datos propia.

Grupo Objetivo

Hombres y mujeres, inmigrantes de todas las nacionalidades, destacando especialmente Venezuela, Haití, Perú, Colombia, Argentina y Ecuador. Mayores de 18 años, pertenecientes a todos los NSE, residentes en las distintas zonas de Chile.

Muestra

793 Casos, sobre un Universo estimado (INE) de 1.251.225. Margen de error de +/- 3,5 puntos porcentuales al 95% de confianza.

Inmigrantes en Chile CADEM

Ponderación

Los datos fueron ponderados a nivel de sujeto por sexo y país, obteniendo una muestra representativa para el universo de estudio

Fecha de Terreno

Las encuestas fueron realizadas entre el Jueves 1 y el Viernes 9 de Agosto de 2019.

Fuente del estudio

<https://www.cadem.cl/wp-content/uploads/2019/10/Estudio-Inmigrantes-en-Chile-VClientes.pdf>



Estudio nacional de opinión pública N°84 - CEP

Universo que se incluye

Población de 18 años y más (urbana y rural) residente a lo largo de todo el país (se excluye Isla de Pascua).

Cobertura del Universo

Toda la población del país en base al Censo de 2017 (se excluye Isla de Pascua).

Muestra

1.496 personas fueron entrevistadas en sus hogares en 117 comunas del país.

El método de muestreo fue estratificado por región y zona urbana/rural, aleatorio y probabilístico en cada una de sus 3 etapas (manzana-hogar-entrevistado).

Estudio nacional de opinión pública N°84 - CEP

Nivel de precisión

El error muestral se estima en $\pm 3\%$ considerando una varianza máxima y un 95% de confianza.

Instrumento

Entrevistas individuales cara a cara, basadas en un cuestionario estructurado.

Fecha de terreno

La recolección de datos se efectuó entre el 28 de noviembre de 2019 y 6 de enero de 2020

https://www.cepchile.cl/cep/site/docs/20200116/20200116081636/encuestacep_diciembre2019.pdf

Actividad

Utilizando la Encuesta Nacional de Salud como marco muestral, realice un estudio sobre la proporción de Diabéticos utilizando una muestra de 500 personas, utilizando un muestreo aleatorio simple, y estratificado por fumador usando 250 personas por estrato, ¿Qué diferencias observa?, ¿Tiene sentido estratificar por fumador?

Realice un muestreo sistemático de tamaño 300 del peso, ¿qué método de muestreo utilizaría para mejorar esta estimación? Impleméntela.

Estudie la proporción de hipertensos utilizando una muestra de 500 personas utilizando un muestreo aleatorio simple, sistemático y estratificado por sedentarismo y estratificado por sexo. ¿Qué diferencias se observan?

Intervalos de confianza

Luego de tener información de una muestra, la estimación podrá fallar. La misión es utilizar herramientas estadísticas para asegurar su correcta similitud con la población. Para ello, el segundo punto a considerar es el **nivel de error** de una muestra.

El estudio de las brechas salariales comienza identificando el ingreso promedio de los varones, que en la última encuesta CASEN dicho ingreso promedio alcanza los \$436.688. ¿Será este el valor real del ingreso promedio en varones en Chile?

