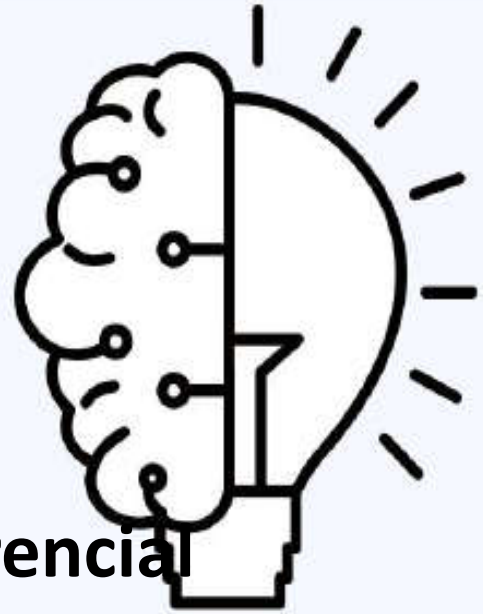
Ciencia de datos

Para el sector público de salud



Módulo 4 – **Modelos y Estadística inferencial**

Ricardo Aravena C.

En colaboración con DATA UC Estudios y Servicios Estadísticos

academia.opensaludlab.org

OPEN SALUD

Inferencia: proceso por el cual se derivan conclusiones a partir de premisas (ref. es.wikipedia.org)

Inferencia estadística: parte de la estadística que comprende los métodos y procedimientos que por medio de la inducción determina propiedades de una población estadística, a partir de una parte de ésta. Su objetivo es obtener conclusiones útiles para hacer deducciones sobre una totalidad, basándose en la información numérica de la **muestra** (ref. es.wikipedia.org).

Muestra: es un subconjunto de casos o individuos de una población. En diversas aplicaciones, interesa que una muestra sea representativa, y para ello debe escogerse una técnica de muestra adecuada que produzca una muestra aleatoria adecuada.

Si se obtiene una muestra sesgada, su interés y utilidad son más limitados, en función del grado de sesgos que presente.

Primeramente, es necesario diferenciar…

Big Data - Análisis de Datos

- Nos permite entender el comportamiento de la gente
- Nos muestra que es lo que la gente hace
- Potenciales sesgos – conectividad, uso de redes

Muestreo – Análisis de Datos

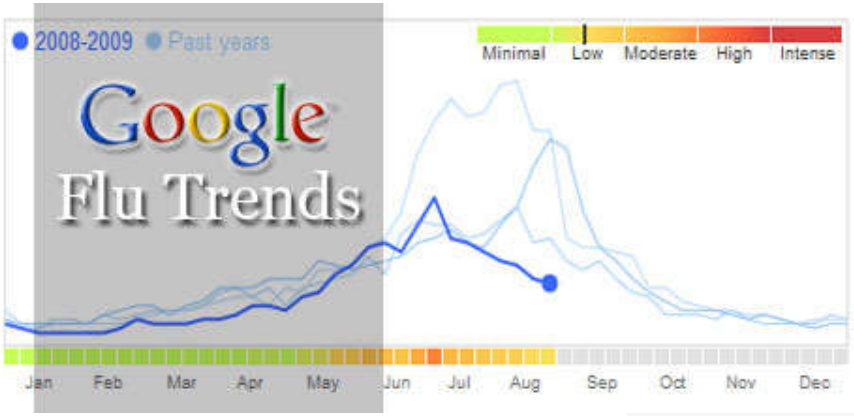
- Nos permite entender las actitudes de la gente
- Recoger opiniones de situaciones ficticias
- Comprender el por qué de ciertas cosas

Muestreo vs Big Data

No hay duda que, hoy en día, el disponer de "todos" los datos, es decir casi la población, hay pocos test estadísticos que son aplicables.

***Google Flu Trends*, aclamado algoritmo de Google, especializado en detectar los brotes de gripe por todo el mundo…**

Nature (2008), Detecting Influenza Epidemics using search engine query data.



Google Flu Trends - Summary

La detección temprana de la enfermedad, seguida de una respuesta rápida, puede reducir el impacto de la influenza estacional y pandémica...

Aquí presentamos un método para analizar las consultas de búsqueda de Google que están altamente correlacionada con el porcentaje de consultas médicas en las que un paciente presenta síntomas parecidos a la influenza…

Podemos estimar con precisión el nivel actual de actividad semanal de influenza en cada región de los Estados Unidos, con un retraso en la notificación de aproximadamente un día...

Cuando todos esperaban su confirmación **falló** – Gripe H1N1 del 2009.

Si entramos al sitio dice:

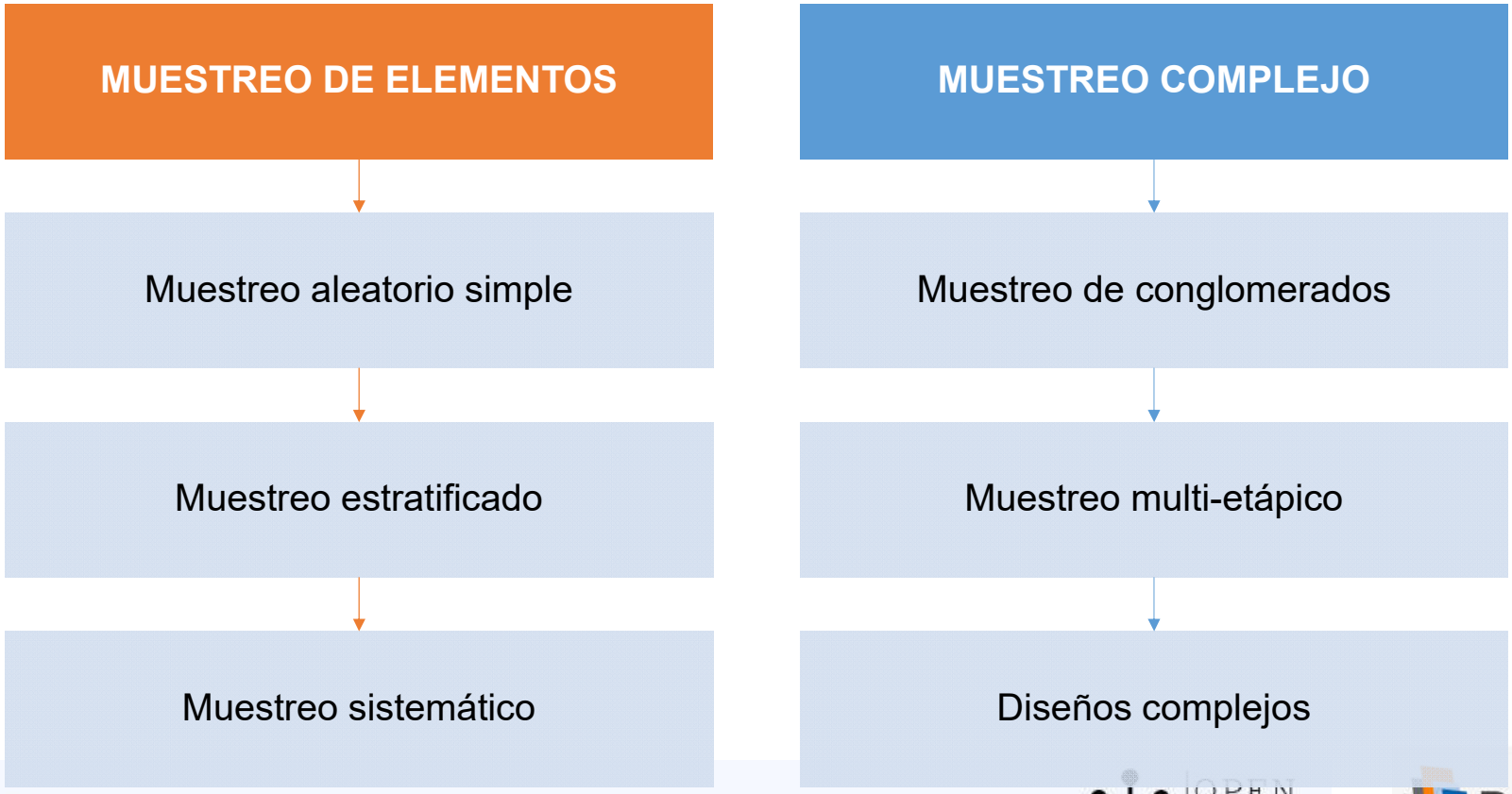
Google Flu Trends ya no publica estimaciones de la gripe en función de los patrones de búsqueda. Todavía es pronto para la predicción inmediata y herramientas similares para comprender la propagación de enfermedades como la gripe.

La recopilación de datos cesó el 2015 y el 2020 desapareció la página con las disculpas..

Pero….

Kogan, N. *et al* (2021) An early warning approach to monitor COVID-19 activity with multiple digital traces in near real time. Science Advances.

Técnicas de muestreo



Pero… de acuerdo al diseño muestral, la muestra puede resultar:

- **Auto ponderada.** Cada unidad muestral representa al mismo número de unidades en la población. En general, N/n (que corresponde al inverso de la probabilidad de selección). En este caso, no hay problema en realizar análisis.

- **No auto ponderada**. Significa que se aplicó un diseño complejo, en cuyo caso se debe obtener la probabilidad de selección de cada unidad y el inverso de éste se utiliza como factor de expansión para los cálculos.

Por ejemplo, en el informe: *Cálculo de factores de expansión* ***ENS2016-17***, pag.29 pueden encontrar la fórmula para la obtención de la probabilidad de selección:

$$P_{hijk}(l) = \mathrm{P}_{\mathrm{hij}}^{\mathrm{AJ}}(k) \cdot P_{hijk}(l|k)$$

$$= \left[\frac{1}{w''_{hijk}}\right] \cdot P_{hijk}(l|k)$$

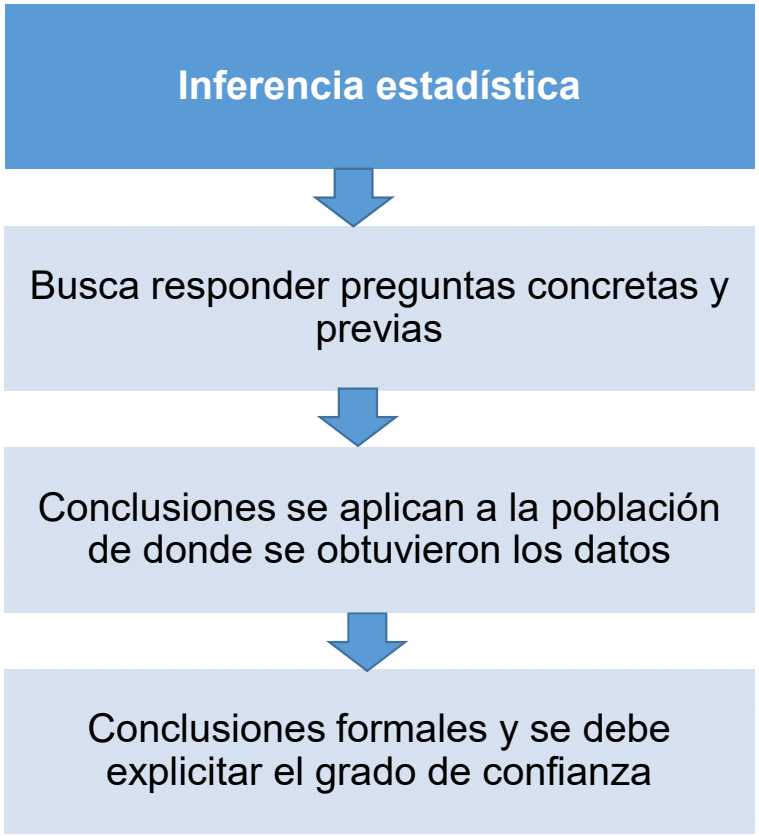
Y para realizar análisis estadístico (sea en R, SPSS, SAS, STATA, etc) existen módulos especiales para ello.

En R el paquete se llama "**survey**" el cual esta basado en el texto de T. Lumley (http://r-survey.r-forge.r-project.org/survey/)

Solo como ejemplo, la tabla adjunta muestra la variable: ENS2016 – **Depresión CIDI DSM IV 12m**, al obtener las tasas sin / con factor de expansión.

% Depresión	Sexo		
CIDI DSM IV	HOMBRE	MUJER	TOTAL
Sin FACTOR	2.20	7.49	5.58
Con FACTOR	2.15	10.07	6.19

Recuerden que el objetivo es definir políticas públicas, por tanto…


Inferencia estadística
Busca responder preguntas concretas y previas
Conclusiones se aplican a la población de donde se obtuvieron los datos
Conclusiones formales y se debe explicitar el grado de confianza

Problemas de salud en Chile –

Suponga que accedemos a una muestra aleatoria simple de tamaño 350, representativa de los habitantes ≥ 15 años de una comuna. En esta base se han registrado las siguientes características:

COLESTEROL - HDL - EDAD - FUMA - NEDU - SEXO - PAS - PAD

TALLA - CINTURA - CUELLO - PESO - DEPORTE - DIABETES

Para mayor detalle, descripción, códigos, etc. revisar base: ***Salud.xlsx***

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
num	COLESTEROL	HDL	EDAD	FUMA	NEDU	SEXO	PAS	PAD	TALLA	CINTURA	CUELLO	PESO	DEPORTE	DIABETES
1	162	43	21	No	> 12 años	Femenino	109	70	159	68	31	54	No	No
2	222	46	59	No	8 - 12 años	Masculino	149	90	171	88	41	89	No	No
3	255	87	51	No	8 - 12 años	Femenino	139	83	164	110	38	58	No	Si

Una condición básica de la estadística inferencial es …

¡no mirar los datos!

Por tanto, de acuerdo a los objetivos del estudio, es necesario tener a priori clara las hipótesis que se desean contrastar.

Hipótesis: es un enunciado no verificado, una vez refutado o confirmado dejará de ser hipótesis y sería un enunciado verificado.

La hipótesis es una conjetura científica que requiere una contrastación con la experiencia.

Hipótesis estadística: En un trabajo de investigación generalmente se plantean dos hipótesis mutuamente excluyentes: la hipótesis nula (**Ho**) y la hipótesis de investigación (**Ha** o **H1**). La hipótesis de investigación es una afirmación especial cuya validez se pretende demostrar, y si las pruebas empíricas no apoyan decididamente la hipótesis de investigación, ésta se abandona.

Algunas hipótesis involucran variables que pueden poseer una relación causal establecida. En ocasiones el investigador tendrá control o capacidad de observación sobre unas variables y sobre otras no

(var. independiente) X → Y (var. Dependiente)

En general, el planteamiento de hipótesis tiene relación con los objetivos planteados en el estudio. Además, y con certeza, para definir el estudio se realiza una revisión bibliográfica (estado del arte) y podría, por ejemplo, estar sustentado por otros estudios.

Encuesta Actividad Física y Deportes: sólo dos de cada 10 chilenos hace ejercicio

El estudio presentado en la Universidad San Sebastián por el Ministerio del Deporte, también mostró que hay una marcada brecha de género entre quienes realizan actividad física: mientras el 45,3% de los hombres es "activo", en las mujeres esa cifra se reduce a un 25,8%.

Lunes 7 de enero de 2019

¿Qué hipótesis se pueden plantear?

Hipótesis: prejuicio sobre un fenómeno.

1. Los hombres presentan un peso mayor que las mujeres.
2. Los diabéticos presentan un mayor nivel de colesterol que los no diabéticos.
3. Las mujeres tienen mejor nivel de HDL que los hombres.
4. Hay diferencias en el IMC entre fumadores y no fumadores.
5. Hombres fuman más que las mujeres.
6. **Menos de un tercio de los hombres práctica deportes.**
7. Y un largo etcétera.

Vamos a

Test de hipótesis

1. **Una muestra**
   a) Para una media
   b) Para una proporción
2. **Dos muestras Independientes**
   a) Comparación de medias
   b) Comparación de proporciones
3. **Asociación**
   a) Entre dos variables categóricas
   b) Entre dos variables continuas

Una muestra – Test para la media μ

1. Definir la hipótesis

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu \neq \mu_0$$
$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_a : \mu > \mu_0$$
$$H_0 : \mu \geq \mu_0 \quad \text{vs} \quad H_a : \mu < \mu_0$$

2. Escoger el estadístico (o prueba)
   a) Con σ conocido

$$Z_0 = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

   a) Con σ desconocido

$$T_n = \frac{\overline{X}_n - \mu}{S/\sqrt{n}} \sim \text{t-student}(n-1)$$

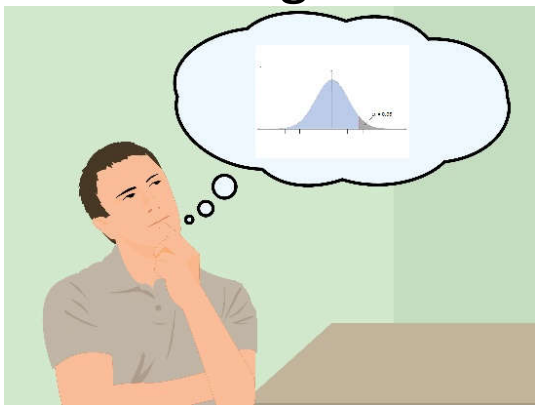
3. Evaluar y decidir
   a) Valor crítico
   b) Valor-p

¿Qué es el valor-p? Se define como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta. En términos simples, el valor *p* ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativo.

Si el valor *p* cumple con la condición de ser menor que un nivel de significancia impuesto arbitrariamente, este se considera como un resultado estadísticamente significativo y, por lo tanto, permite rechazar la hipótesis nula. *Ref: https://es.wikipedia.org/wiki/Valor_p*

Significación estadística versus Importancia

Si un investigador decide sobre la eficacia de un tratamiento, juzgando la importancia de los resultados en función de su significancia estadística, esta decisión puede ser incorrecta.



Es relevante, al iniciar una investigación, definir la magnitud del efecto para que el resultado sea relevante. Más aún, esta definición será crucial para la determinación del tamaño de muestra requerido.

Una muestra – Test para una proporción P

1. Definir la hipótesis

$$Ho: P = p_o \; vs \; Ha: P \neq p_o$$
$$Ho: P = p_o \; vs \; Ha: P > p_o$$
$$Ho: P = p_o \; vs \; Ha: P < p_o$$

2. Escoger el estadístico (o prueba)

$$Z_n = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

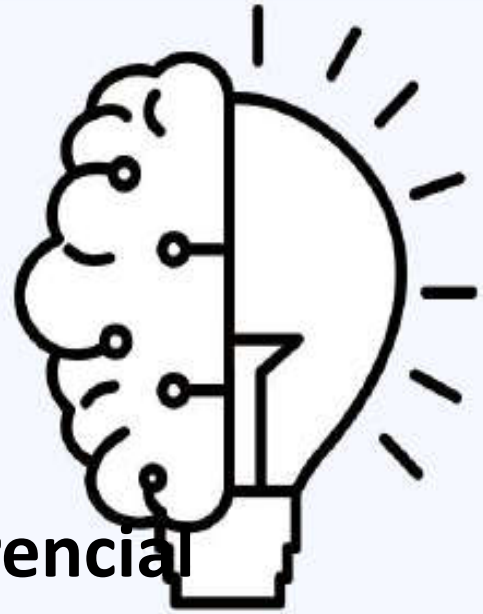
Con $\hat{p}$ proporción muestral.

3. Evaluar y decidir.

Vamos a

Ciencia de datos

Para el sector público de salud



Módulo 4 – **Modelos y Estadística inferencial**

Ricardo Aravena C.

En colaboración con DATA UC Estudios y Servicios Estadísticos

academia.opensaludlab.org

OPEN SALUD