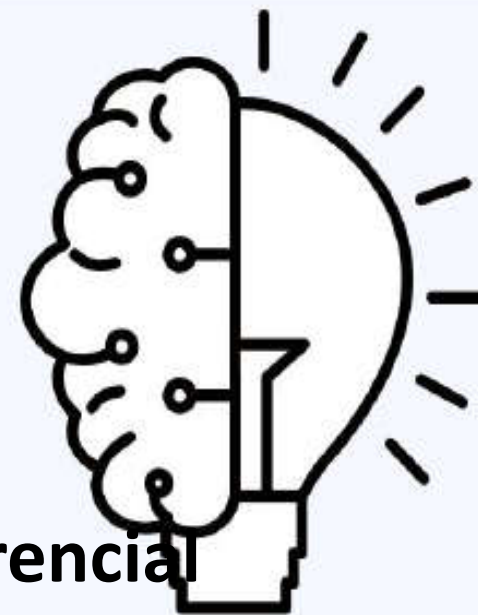


Ciencia de datos

Para el sector público de salud



Módulo 4 – Modelos y Estadística inferencial

Ricardo Aravena C.

2da Parte

En colaboración con



Primeramente, quién soy

Ricardo Aravena Cuevas (mail: ricardo.aravena@uc.cl)

Estadístico y Magister en Estadística UC.

Profesor de la Práctica Asociado – Facultad de Matemáticas, UC.

Director Académico **Diplomado en Estadística UC**

Asesor/Consultor

Time Ibope-Adimark-Cadem, CNED, CSE, Inacap, **SAS**, Dicom, CCS, AACH, BCI, BChile, Metro,..
lcmer – Fiscalización-Mintratel, DCV, C13, Megavisión, Chilquinta, Aguas Andinas, Essbío,..

Profesor Magister: UChile (MGPP, Geografía), UC (MGPI, C.Civil), UTalca (MGSS),...

Profesor Asociado UC – Cs. Políticas, Comunicaciones, Medicina, Ingeniería, Economía...

Y por que no, también soy Director de **ABStat Consultores Asociados SpA**



Volviendo a lo que habíamos quedado:

¿Qué hipótesis se pueden plantear?

Hipótesis: prejuicio sobre un fenómeno.

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	120	82	164	110	38	58	No	No

¿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.

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 - con μ_0 valor conocido.

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_a : \mu \neq \mu_0$$

$$H_0 : \mu \leq \mu_0 \quad \text{vs} \quad H_a : \mu > \mu_0$$

$$H_0 : \mu \geq \mu_0 \quad \text{vs} \quad H_a : \mu < \mu_0$$

2. Escoger el estadístico (o prueba)

a) Con σ conocido

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

a) Con σ desconocido

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

3. Evaluar y decidir

- a) Valor crítico
- b) Valor-p

3. Evaluar y decidir

- a) Valor crítico
- b) Valor-p

DOI: 10.1097/GME.0000000000001103
© 2018 by The North American Menopause Society

Older women do not have seasonal variations of vitamin D levels: a study from a southern country

Mariá S. Vallejo, MD,¹ Juan E. Blümel, MD,² Pablo Lavín, MD,³ Claudio Torres, MT,¹ Alejandro Araos, MD,¹ and Carlos Sciaraffia, MD¹

Received January 9, 2018; revised and accepted February 14, 2018.

From the ¹Clínica Quilín, Faculty of Medicine, University of Chile, Santiago de Chile, Chile; ²Department of Internal Medicine (South Campus), Faculty of Medicine, University of Chile, Santiago de Chile, Chile; and ³Department of Obstetrics and Gynecology (South Campus), Faculty of Medicine, University of Chile, Santiago de Chile, Chile.

Funding/support: None.

Financial disclosure/conflicts of interest: None reported.

Address correspondence to: Juan E. Blümel, MD, PhD, Department of Internal Medicine (South Campus), Faculty of Medicine, University of Chile, Orquídeas 1068, Department 302, PO BOX 7510258, Providencia, Santiago de Chile, Chile; E-mail: juan.blumel@redsalud.gov.cl

Examen	Unidad de Medida	Resultado	Intervalo de Referencia
25-OH-Vitamina D	ng/mL	16 ↓	[20 - 50]

Abstract

Objective: The aim was to study whether the seasonal variation of vitamin D [25(OH)-D or calcidiol] is similar or different in younger and older women living in a southern country.

Methods: Measurement of serum 25(OH)-D concentration in 739 Chilean women aged 20 to 87 years, residents of Santiago (latitude: 33.4° South) who, during a routine gynaecological checkup, agreed to be evaluated.

Results: The mean serum concentration of 25(OH)-D for the group was 24.1 ± 10.5 ng/mL. In women 20 to 39 years, the mean was significantly different from the mean of the ≥ 60 years old group (25.8 ± 10.6 ng/mL vs 23.9 ± 11.1 ng/mL; $P < 0.02$). Globally, 38.4% of participants had vitamin D deficiency and 36.1% insufficiency. A deficiency was present in 28.4% of the 20 to 39 years old, and in 43.9% in the ≥ 60 years old group ($P < 0.004$). In the whole group, a lower proportion ($P < 0.0001$) of vitamin D deficiency cases in the youngest women occurred during the summer (23.7%) in comparison to the winter (47.7%). It was observed that the proportion of participants in the 20 to 39 years old group with vitamin D deficiency fell from 48.9% in winter to 4.9% in summer ($P = 0.0001$). In the older groups, this change (less deficiency) is progressively smaller, 51.2% to 27.6% ($P = 0.0020$) in women 40 to 59 years old, and it does not happen in women ≥ 60 years (40% with vitamin D deficiency).

Conclusions: Serum vitamin D deficiency [25(OH)-D or calcidiol] is highly prevalent in Santiago, especially in older women (≥ 60 y) throughout the year. In contrast, in younger women (< 40 y), the vitamin D deficiency tends to disappear during summer. More epidemiological studies and targeted prevention actions on vitamin D deficiency are warranted.

Key Words: 25 (OH)-D – Aging – Calcidiol – Seasonal variation – Vitamin D – Women.

3. Evaluar y decidir

- a) Valor crítico
- b) Valor- p

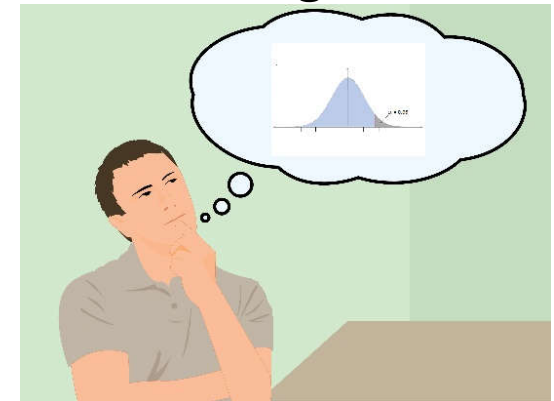
¿Qué es el valor- p ? Se define como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta. En términos simples, el valor p ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativo.

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

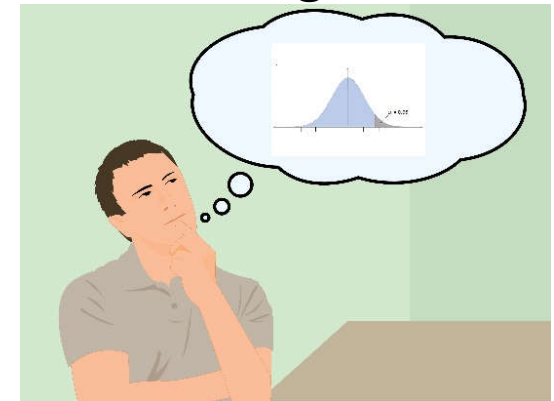
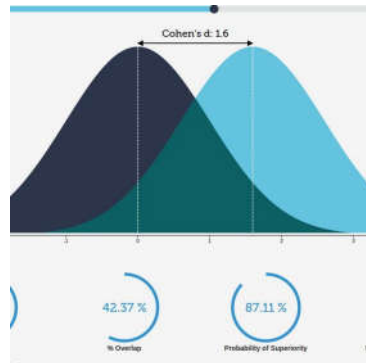
Significación estadística versus Importancia

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.



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 (D de Cohen). Esta definición será crucial para la determinación del tamaño de muestra requerido.

Realizaremos algunos test sobre una media, así que vamos a



Una muestra – Test para una proporción P

1. Definir la hipótesis – con p_o valor conocido.

$$H_o: P = p_o \quad vs \quad H_a: P \neq p_o$$

$$H_o: P = p_o \quad vs \quad H_a: P > p_o$$

$$H_o: P = p_o \quad vs \quad H_a: P < p_o$$

2. Escoger el estadístico (o prueba)

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

Con \hat{p} proporción muestral.

3. Evaluar y decidir.

Realizaremos algunos test sobre una proporción, así que vamos a



Dos muestras independientes – test de medias

1. Definir la hipótesis

$$H_0: \mu_1 = \mu_2 \quad vs \quad H_a: \mu_1 \neq \mu_2$$

$$H_0: \mu_1 = \mu_2 \quad vs \quad H_a: \mu_1 > \mu_2$$

$$H_0: \mu_1 = \mu_2 \quad vs \quad H_a: \mu_1 < \mu_2$$

2. Escoger el estadístico (o prueba)...

Si σ_X y σ_Y son desconocidos pero iguales:

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t - \text{Student}(n + m - 2)$$

$$\text{con } S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}$$

en R Test t

En R es el Test T de Welch

Si σ_X y σ_Y son desconocidos:

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \sim t - \text{Student}(\nu)$$

con

$$\nu = \left[\frac{(S_X^2/n + S_Y^2/m)^2}{\frac{(S_X^2/n)^2}{n-1} + \frac{(S_Y^2/m)^2}{m-1}} \right]$$

3. Y para decidir sobre las varianzas

$$\frac{[(n-1) S_X^2 / \sigma_X^2] / (n-1)}{[(m-1) S_Y^2 / \sigma_Y^2] / (m-1)} = \frac{S_X^2}{S_Y^2} \cdot \frac{\sigma_Y^2}{\sigma_X^2} \sim F(n-1, m-1)$$

4. Evaluar y decidir basado en el valor-p.

En R se usan

→ Test de igualdad de varianzas: ***var.test()***

→ Test de igualdad de medias: ***t.test()***, basta con indicar si las varianzas son iguales o desconocidas con la opción ***var.equal*** = T o F

Realizaremos algunos test sobre medias, así que vamos a



Dos muestras independientes – Test de comparación de proporciones

1. Definir la hipótesis

$$H_0: P_1 = P_2 \quad vs \quad H_a: P_1 \neq P_2$$

$$H_0: P_1 = P_2 \quad vs \quad H_a: P_1 > P_2$$

$$H_0: P_1 = P_2 \quad vs \quad H_a: P_1 < P_2$$

2. Escoger el estadístico (o prueba)

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

Con \hat{p}_1 y \hat{p}_2 proporciones muestrales.

3. Evaluar y decidir.

Realizaremos algunos test sobre proporciones, así que vamos a



Asociación entre dos variables categóricas

Por ejemplo, el fumar y el género ¿están asociados?

Para la muestra de $n=350$, se clasifican según género (masc/fem) y status de fumar (si/no).

La tabla resultante es

Género	NO fuma	SI fuma
Femenino	138	63
Masculino	92	57

Hipótesis: H_0 : Género y Status de fumar SON independientes

H_a : Género y Status de fumar NO SON independientes

Test Chi-cuadrado... vamos  El test puede extenderse a **tablas de $I \times J$**

Respondiendo a algunas consultas sobre el **Diplomado en Estadística UC - 2021**, mayor información sobre este (fechas, formato y otros)

Diplomado en Estadística Semipresencial (19 abril 2021)

<https://educacioncontinua.uc.cl/40673-ficha-diplomado-en-estadistica-mencion-metodos-estadisticos>

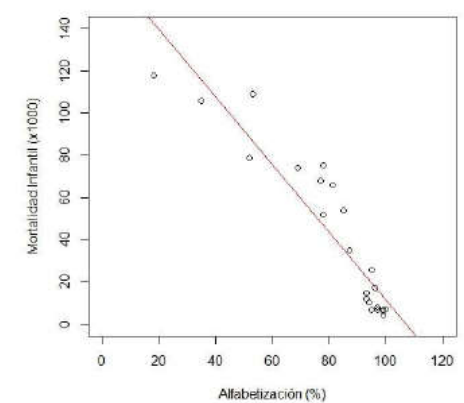
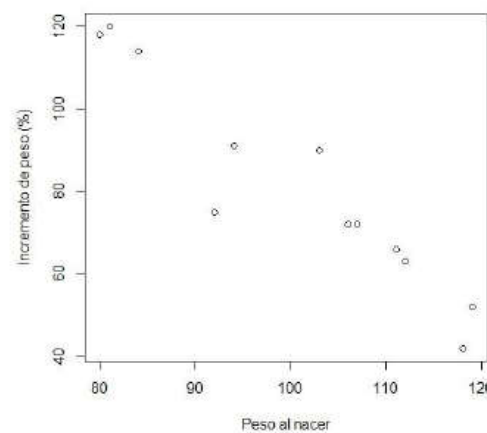
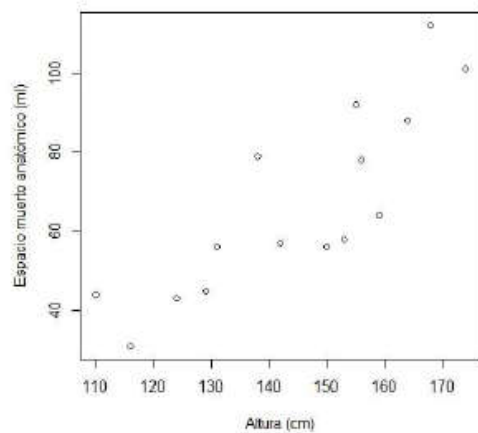
Diplomado en Estadística Online (20 abril 2021)

<https://educacioncontinua.uc.cl/41673-ficha-diplomado-en-estadistica-mencion-metodos-estadisticos>

Modelos lineales

$Y = \begin{cases} X_1 \\ X_2 \\ \dots X_k \end{cases}$, Donde Y = variable respuesta (continua) y las X 's son covariables (continuas o categóricas) que “explican” a Y .

Ejemplos – Modelos “simples”



Modelo de Regresión Lineal Simple

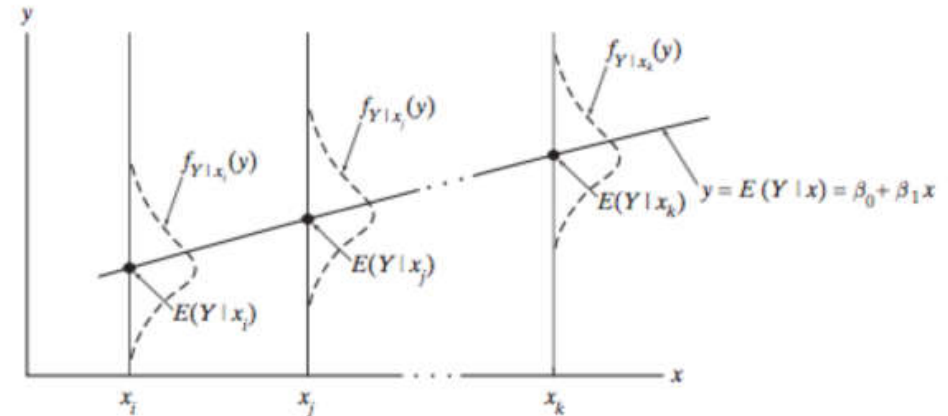
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots, n$$

Supuestos:

1. Linealidad: La media condicional de Y sobre x es lineal

$$y = E(Y | x) = \beta_0 + \beta_1 x$$

2. Homocedasticidad: La varianza asociada a $f_{Y|x}(y)$ es la para todo x e iguala σ^2 .
3. Independencia: Las distribuciones condicionales son variables aleatorias independientes para todo x .



Vamos a



Significancia de la(s) variable(s) explicativa(s)

Prueba T

La prueba de hipótesis para estudiar la inclusión de cada X_j en el modelo es:

$$H_0 : \beta_j = 0$$

para cada $j=1, \dots, k$, donde k es el número de variables explicativas y se asume bajo ciertos supuestos que:

$$T_{0j} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-(k+1)}$$

Se rechaza H_0 si valor-p asociado al test t es menor que α .

Por ejemplo, el modelo ajustado para EMA en función de la altura:

(EMA) y la altura:

$$\widehat{EMA} = -82,49 + 1,03 \cdot altura$$

Las tablas resumen que obtenemos son las siguientes:

	Estimación ($\hat{\beta}$)	Error estándar (se)	Estadístico t (t-value)	Valor-p (p-value)
Intercepto	-82,49	26,3	-3,14	0,0078
Altura	1,03	0,18	5,73	0,00006

	gl	SC	MC	F	valor-p
Altura	1	5607,4	5607,4	32,814	0,00006
Residuos	13	2221,5	170,9		



Modelo de Regresión Lineal Múltiple

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i \quad i = 1, \dots, n$$

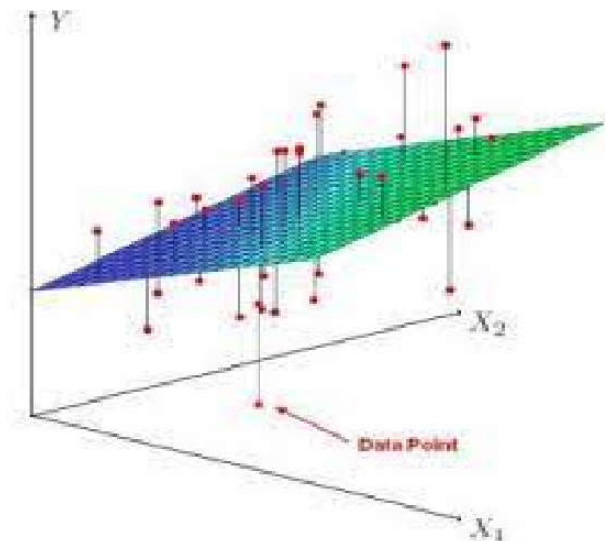
donde Y es la variable dependiente, X_j , $j = 1, \dots, k$ son las covariables del modelo, y los β_j son coeficientes constantes del modelo, y las ε_i son variables aleatorias tales que cumplen con:

$$E(\varepsilon_i) = 0$$

$$\text{Var}(\varepsilon_i) = \sigma^2$$

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$$

Vamos a



¿Cuán bueno es el modelo?

Coeficiente de determinación R^2 :

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^n (y'_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SCE}{SCT}$$

Coeficiente de determinación R^2 ajustado:

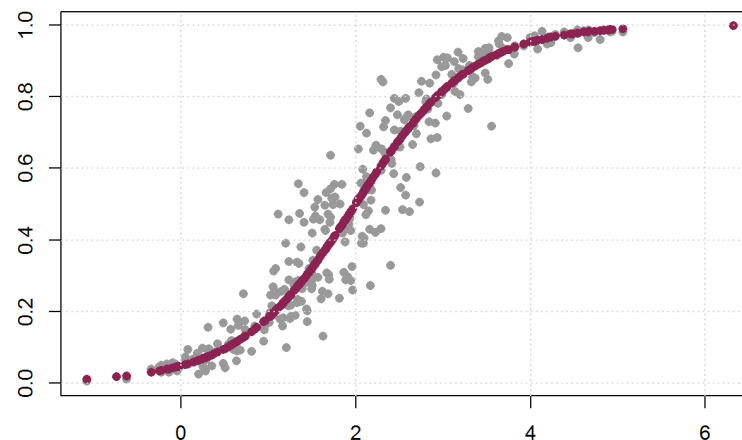
$$r^2 = 1 - \frac{s_{Y|x}^2}{s_Y^2} = 1 - \frac{(n-1) SCE}{(n-2) SCT} = \bar{R}^2$$

Ambos se interpretan como la proporción de variación total que es explicada por el modelo de regresión lineal.

Otros modelos

- Si la variable respuesta es dicotómica, por ej: $Y = 1$ si presenta la enfermedad o $Y=0$ si esta sano
- Si la variable respuesta es un “conteo”, por ej: $Y = \text{número de hijos}$

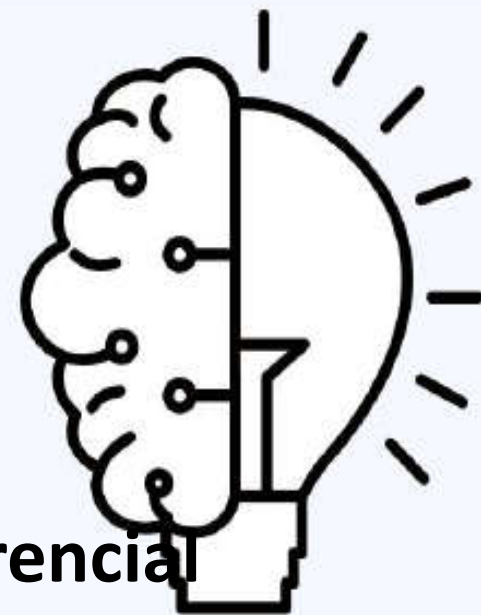
En el primer caso se utiliza un **Modelo de Regresión Logística**,



Mientras que el segundo caso es un **Modelo de Regresión Poisson**.

Ciencia de datos


Para el sector público de salud



Módulo 4 – Modelos y Estadística inferencial

Ricardo Aravena C.

2da Parte

En colaboración con  **DATA UC**
Estudios y Servicios Estadísticos