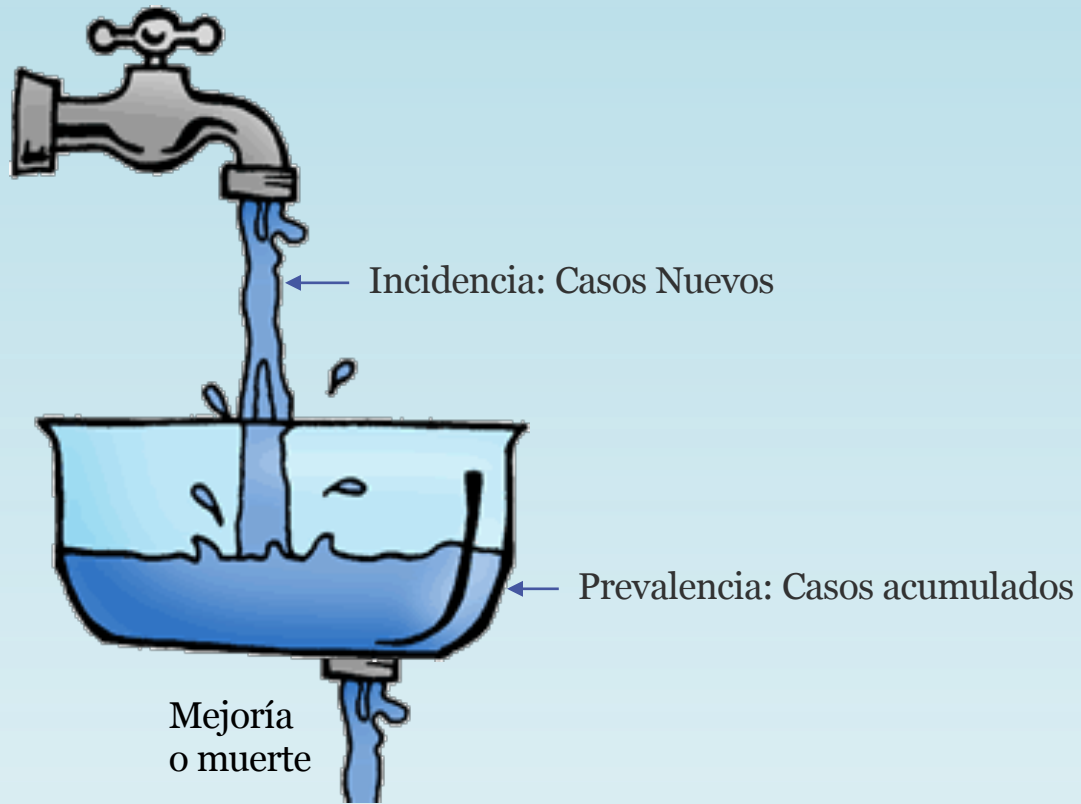


Diseños de Investigación (III) Estudios analíticos de cohorte

OSCAR URREJOLA ORTIZ

Temario

- Definición de incidencia
- Estudios de cohorte
- Medidas de asociación
- Medidas de correlación



Incidencia (Casos nuevos)



Definición

- Probabilidad o riesgo de enfermar
- Mide la ocurrencia de casos nuevos en una población en un período de tiempo

*Población
que está en
riesgo*

*Cantidad de tiempo que son
seguidos los miembros de la
población hasta que desarrollan el
evento*

INFECCIÓN POR VIH CHILE	CASOS NUEVOS 2015	POBLACIÓN 2015
HOMBRES	3.738	8.911.940
MUJERES	523	9.094.467
IGNORADO	30	-
TOTAL	4.291	18.006.407

Fuente: Laboratorio Referencia VIH, Instituto de Salud Pública. INE Proyección población

Tasa de incidencia (notificación) de
VIH
 $= (4.291 / 18.006.407) = 23,8$ por 100 mil habitantes

La probabilidad de contraer infección por VIH en Chile el año 2015 fue de 23,8 por 100.000 habitantes (aunque en rigor, es la probabilidad de ser diagnosticado).

El riesgo varía entre 41,94 en hombres a 5,75 en mujeres con una razón H:M = 7,3

Tabla 3: Casos confirmados de VIH en mayores a 13 años (inclusive) y tasas*, por región y año. Chile, 2010-2015.

Región	2010		2011		2012		2013		2014		2015	
	Casos	Tasa	Casos	Tasa	Casos	Tasa	Casos	Tasa	Casos	Tasa	Casos	Tasa
Arica y Parinacota	61	34,8	79	44,2	73	40,0	86	46,2	76	40,1	105	54,4
Tarapacá	81	34,8	86	36,0	79	32,2	96	38,1	104	40,2	128	48,3
Antofagasta	113	24,6	129	27,6	120	25,2	150	31,0	159	32,3	169	33,8
Atacama	28	12,1	55	23,4	36	15,1	45	18,5	57	23,1	57	22,8
Coquimbo	85	14,8	88	15,0	114	19,1	124	20,4	141	22,7	144	22,8
Valparaíso	251	17,5	275	19,0	291	19,8	348	23,4	374	24,9	380	25,0
Metropolitana	1.802	32,0	1.770	31,0	1.940	33,5	2.291	39,0	2.397	40,2	2.367	39,2
L. B. O'Higgins	71	10,0	87	12,1	104	14,2	96	13,0	96	12,8	100	13,2
Maule	62	7,6	78	9,4	91	10,9	118	13,9	92	10,7	118	13,7
Biobío	160	9,6	237	14,0	256	15,0	295	17,1	260	14,9	302	17,2
Araucanía	68	8,8	60	7,7	80	10,1	83	10,4	65	8,0	118	14,5
Los Lagos	113	17,5	131	20,0	121	18,2	176	26,2	142	20,9	185	27,0
Los Ríos	41	12,9	32	9,9	43	13,2	39	11,8	42	12,6	55	16,4
Aysén	12	14,7	9	10,9	6	7,2	14	16,5	13	15,1	16	18,3
Magallanes	20	15,3	23	17,4	35	26,3	43	32,0	47	34,7	47	34,5
Total	2.968	21,4	3.139	22,3	3.389	23,7	4.004	27,7	4.065	27,7	4.291	28,9

*Tasas por 100.000 habtes. en base a Proyección de Población para ≥ 13 años INE , 2016.

Fuente: Laboratorio Referencia VIH, Instituto de Salud Pública.

La incidencia se puede medir como:

Incidencia acumulada o
Proporción de
incidencia

Basada en las personas en riesgo

$$\frac{\text{Nº de casos nuevos de una enfermedad en un período de tiempo}}{\text{Población en riesgo de enfermar al inicio del período}}$$

Tasa de Incidencia o
densidad de incidencia

Basada en unidades de persona-tiempo en riesgo

$$\frac{\text{Nº de casos nuevos de enfermedad}}{\text{Persona – tiempo de observación en población en riesgo}}$$

Incidencia acumulada (o incidence proportion)

- La proporción de incidencia es una medida del riesgo de enfermedad o la probabilidad de desarrollar la enfermedad durante el período especificado.
- Como medida de incidencia, incluye solo nuevos casos de enfermedad en el numerador.
- El denominador es el número de personas en la población al inicio del período de observación.
- Debido a que todas las personas con nuevos casos de enfermedad (numerador) también están representadas en el denominador, el riesgo es también una proporción.

Incidencia acumulada

$$IA = \frac{N^{\circ} \text{ de casos nuevos de una enfermedad en un período de tiempo}}{\text{Población en riesgo de enfermar al inicio del período}}$$



Entre 1961 y 1973 se diagnosticaron 11 casos de cáncer cerebral en trabajadores del plástico en Suecia.



De acuerdo al censo, en 1960 habían 3076 trabajadores en la industria de plástico en Suecia.

$$IA = 11 / 3076 = 0,004$$

La probabilidad de desarrollar un cáncer cerebral en los trabajadores del plástico fue de 4 por cada 1000 trabajadores en 12 años de observación.

¿Que hubiera pasado si extendiendo el período de observación?

- Es la probabilidad de que una población sana desarrolle una enfermedad
- El numerador es parte del denominador
- El denominador es la población al inicio del período
- El resultado depende de la duración del período de observación
- Refleja el riesgo promedio de enfermar de un individuo en un período de tiempo
- Como toda probabilidad varía entre 0 y 1

Tasa de incidencia o densidad de incidencia

$$\text{Incidencia acumulada} = \frac{\text{Nº de casos nuevos de enfermedad}}{\text{Persona} - \text{tiempo de observación en población en riesgo}}$$

- Una tasa de incidencia describe qué tan rápido ocurre una enfermedad en una población.
- La fuerza que una enfermedad tiene para cambiar el estado de salud de una población al estado de enfermedad por unidad de tiempo en relación a la población susceptible en ese momento.
- Se basa en tiempo-persona.
 - Ventaja: permite registrar a las personas que entran y salen del estudio.
 - Inconveniente: supone que la probabilidad de enfermedad durante el período de estudio es constante, por lo que 10 personas seguidas durante un año equivalen a una persona seguida durante 10 años. Debido a que el riesgo de muchas enfermedades crónicas aumenta con la edad, este supuesto a menudo no es válido.

Tasa de incidencia o densidad de incidencia

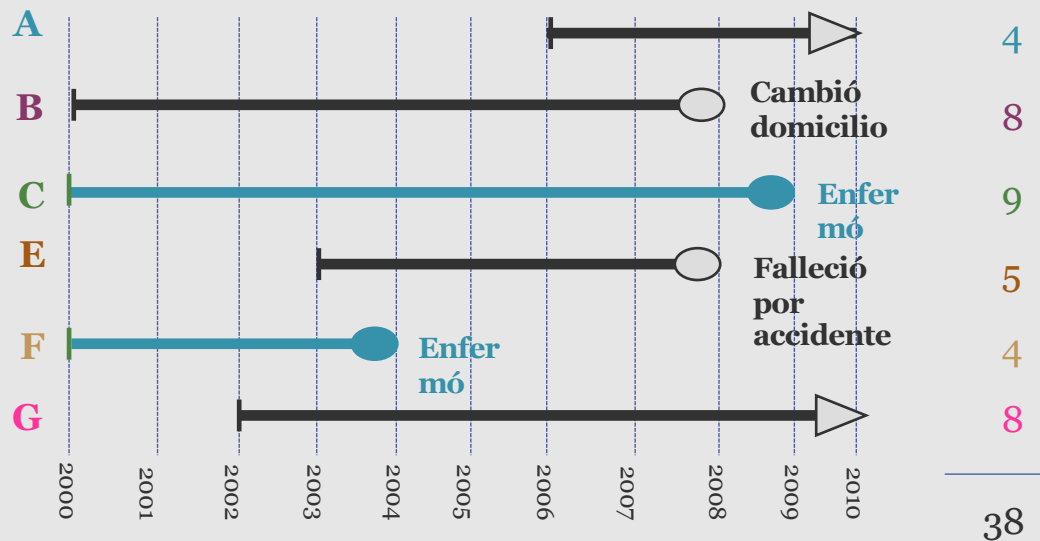
- El tiempo-persona es una jerga epidemiológica: reemplace "años-persona" por "personas por año".
 - Reportar los resultados como 2.5 nuevos casos de cáncer por cada 1.000 personas por año es más comprensible. También transmite el sentido de la tasa de incidencia como un proceso dinámico, la velocidad a la que ocurren nuevos casos de enfermedad en la población.
- Las tasas de incidencia mas usadas se calculan con un numerador de casos observados o informados, y un denominador basado en la población de mediados de año. Este tipo de tasa de incidencia es comparable a una tasa de tiempo por persona.
- Se expresa como "n" casos por 100, 1000 o 100.000 personas-año.
- Por ejemplo en el VIH, la enfermedad produce 23,8 casos por cada 100.000 personas-año, entonces en 2 años producirá ..?.. Casos

Tasa de incidencia: tiempo de observación de cada sujeto

- Seguimiento de 6 personas por diferentes períodos de tiempo
- Se produjeron 2 casos en 38 años-persona de seguimiento

$$TI = \frac{2}{38} = 0,053 \text{ casos por persona - año}$$

o 53 por 1000 persona - año

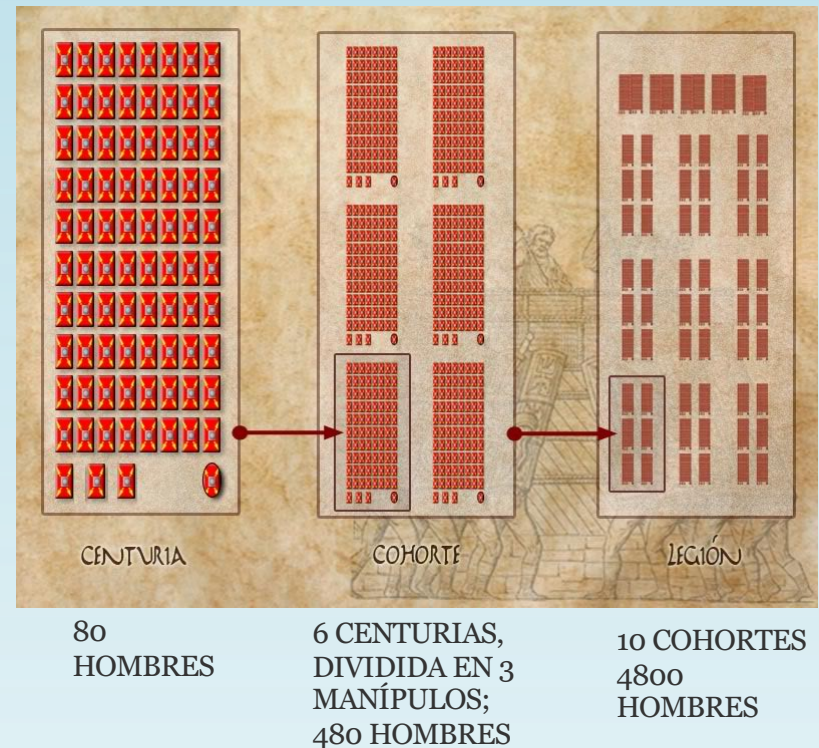


- El denominador es la suma de todos los periodos libres de enfermedad o evento que se está midiendo.
- Se usa en estudios de seguimiento
- Mide la velocidad de la enfermedad; cuantas personas enferman por unidad de tiempo (años, meses, días, horas)

La cohorte romana

Unidad táctica ejército romano:

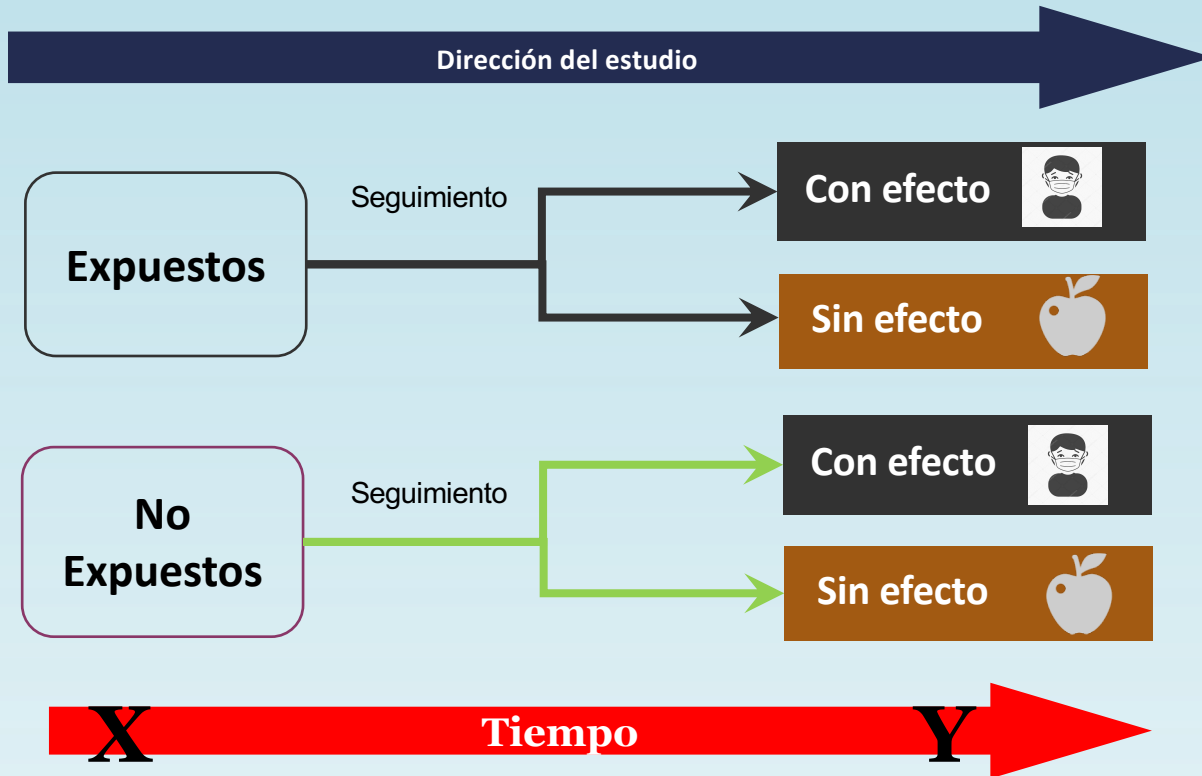
- La 1ª era siempre la mejor de una legión,
- la 6ª la componían los mejores hombres jóvenes,
- la 8ª eran tropas selectas
- la 10ª buenas tropas.



Aspectos Generales

- Se denomina así a un grupo de personas que son observadas en el tiempo.
- Para demostrar asociaciones (comprobar hipótesis):
 - Compara un grupo de personas expuestas a un factor en estudio, con un grupo de personas no expuestas, y los sigue en el tiempo para evaluar en cual grupo hay mayor frecuencia (incidencia) de enfermedad.
 - Si la diferencia en la frecuencia de aparición del efecto o enfermedad en ambos grupos es significativa, se comprueba la asociación.

Diseño básico en la lógica de Cohortes:



Clasificación de estudios

Criterios de ordenamiento



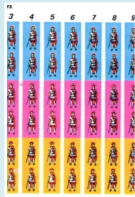
Clasificación de los estudios de cohorte



Según el tipo de población estudiada



Según el momento en que se realiza el estudio



Según el número de cohortes en seguimiento



Según el tipo de población estudiada



Según el momento en que se realiza el estudio



Según el número de cohortes en seguimiento

Tipo de población	Definida por:	Seguimiento	Medición de la frecuencia de la enfermedad	Ejemplo
Abierta/dinámica	Característica que se puede modificar	Integrantes entran y salen de la cohorte; ocurren pérdidas	Tasa de incidencia	• Residentes de una localidad • Personas que ingresan a una empresa u ocupación determinada
Fija	Un evento irremediable	No entran nuevos miembros; ocurren pérdidas	Tasa de incidencia	• Sometidos a un accidente • Intervenidos quirúrgicamente
Cerrada	Un evento irremediable	No entran, ni salen miembros; no ocurren pérdidas (de corta duración)	Incidencia acumulada	• Asistentes a un evento

Por tipo de población también se puede entender la cohortes basadas en un población que se sigue o aquellas basadas en una exposición que se estudia (a la que se agrega una cohorte de no expuestos)



Según el tipo de población estudiada

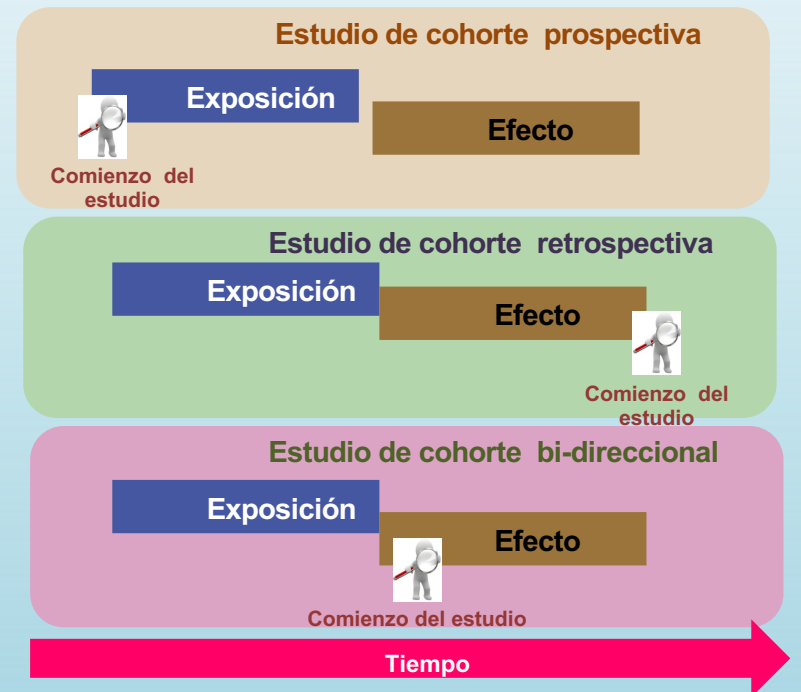


Según el momento en que se realiza el estudio



Según el número de cohortes en seguimiento

- Lo habitual es que el investigador sea contemporáneo al inicio del seguimiento: estudio concurrente (Cohorte prospectiva)
- Pero puede ocurrir que la cohorte se haya formado en el pasado, pero susceptible de ser reconstruida y seguida: estudio no concurrente o cohorte histórica (retrospectiva).
- También puede ser una cohorte bidireccional, es decir, parte con una Cohorte Histórica y sigue con una Cohorte concurrente.





Según el tipo de población estudiada



Según el momento en que se realiza el estudio



Según el número de cohortes en seguimiento

- **Cohorte única** (antes y después): grupo de individuos que comparten o compartieron una misma exposición; en ellos se mide el efecto.
- **Dos cohortes**: formato clásico. Dos grupos de individuos sanos que se diferencian en el factor de exposición.
- **Cohortes múltiples**: múltiples grupos con diferentes grados de exposición y se comparan con una cohorte control. Son útiles en la evaluación de relaciones de dosis-respuesta.

Estudio de Framingham

- 1948 : se enrolaron 5.209 hombres y mujeres, sanos, entre 30 y 62 años de la ciudad de Framingham, Massachusetts, con el objetivo de identificar factores de riesgo comunes para las enfermedades cardiovasculares.
- Al inicio se hace una evaluación física completa, exámenes de laboratorio y entrevistas sobre estilos de vida; cada participante se realiza un control cada 2 años, repitiendo estas mediciones.
- 1971 se enroló a una segunda generación: 5.124 hijos y cónyuges de los sujetos originales.
- Abril 2002 se enroló a una tercera generación: los nietos de la cohorte original

Qué tipo de cohorte es:

Tipo de población:

- Dinámica ✓✓
- Fija
- Cerrada

Según el momento del estudio:

- Prospectivo ✓✓
- Retrospectivo
- bidireccional

Según el número de cohortes:

- Única
- Dos
- Múltiples ✓✓

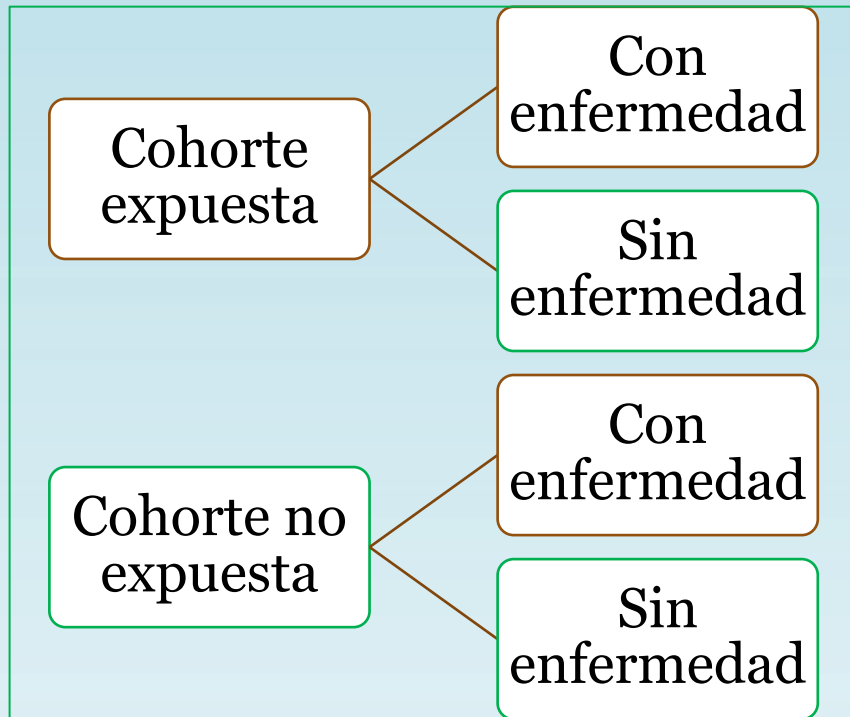
Resultados

- Permitió la identificación de los factores de riesgo más importantes para enfermedad cardiovascular: HTA, colesterol elevado, consumo de tabaco, obesidad, diabetes y sedentarismo.
- También proporcionó información sobre el rol de los triglicéridos y HDL, así como la edad, sexo y aspectos psicosociales.
- Ha permitido además desarrollar y evaluar tratamientos efectivos y estrategias de prevención.
- Se han publicado aproximadamente 1.200 artículos en las revistas científicas más importantes.

<http://www.framinghamheartstudy.org>

ANÁLISIS DE LAS COHORTES

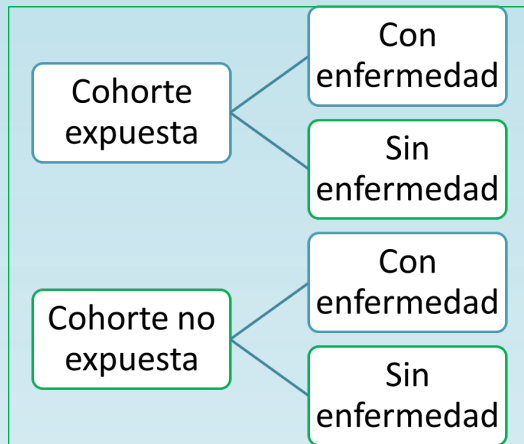
Análisis de los estudios de cohorte



- Cálculo del riesgo (tasas de incidencia) en expuestos y no expuestos:
 - Densidad de incidencia: suma todos los tiempos con los que contribuyeron los individuos estudiados (en cohortes abiertas y fijas)
 - Incidencia acumulada tanto en los expuestos y no expuestos (en cohortes cerradas)
- Cálculo de la medida de Asociación:
 - Riesgo Relativo (RR) o Rate Ratio (razón de las densidades de incidencia)
 - Cálculo medidas de Impacto
 - Riesgo Atribuible (RA)
 - Riesgo Atribuible porcentual (RA%)
 - Riesgo Atribuible poblacional (RAP)
 - Riesgo Atribuible poblacional porcentual (RAP%)

Análisis:

Cálculo de las medidas de riesgo y de la medida de asociación (RR)



	Enfermos	No enfermos	Total
Expuestos	a	b	a+b
No expuestos	c	d	c+d
Total	a + c	b+d	a+b+c+d

Medidas de riesgo:

Incidencia expuestos = enfermos sobre el total de expuestos $a/(a+b)$

Incidencia en no expuestos = enfermos sobre el total de no expuestos $c/(c+d)$

Medida de asociación: Riesgo Relativo o razón de riesgos

$$RR = \frac{\text{Incidencia expuestos}}{\text{Incidencia no expuestos}}$$

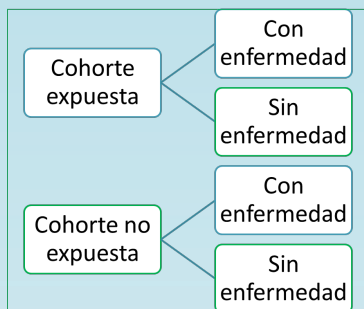
Interpretación:

Cuántas veces aumenta el riesgo en expuestos comparación con los no expuestos

Valor 1= no hay asociación; >1 factor de riesgo y < 1 protector

Para descartar error aleatorio se calcula intervalo de confianza, hab. al 95%

Cálculo de las medidas de impacto: Riesgos Atribuibles o diferencia de riesgo



	Enfermos	No enfermos	Total
Expuestos	a	b	a+b
No expuestos	c	d	c+d
Total	a + c	b+d	a+b+c+d

Medidas de riesgo:

Incidencia expuestos = enfermos sobre el total de expuestos $a/(a+b)$

Incidencia en no expuestos = enfermos sobre el total de no expuestos $c/(c+d)$

Medidas de Impacto: Riesgo Atribuible, Riesgo Atribuible Porcentual y Riesgo Atribuible Poblacional

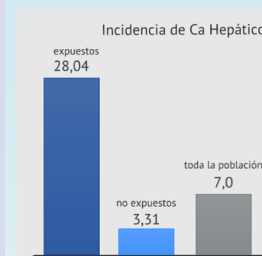
$$RA = Incidencia_{expuestos} - Incidencia_{no\ expuestos}$$

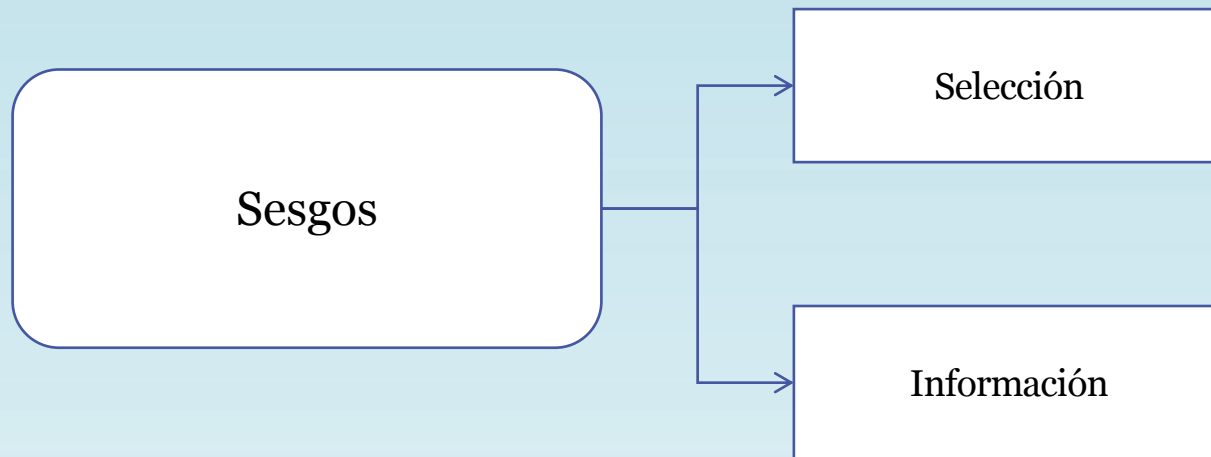
$$RA\% = \frac{I_{expuestos} - I_{no\ expuestos}}{I_{expuestos}} \times 100$$

$$RA\ Poblacional\% = \frac{I_{poblacional} - I_{no\ expuestos}}{I_{poblacional}} \times 100$$

Interpretación:

- RA: Cuál es el riesgo de los expuestos que se puede atribuir a la exposición
- RA% Qué proporción del riesgo de los expuestos se puede atribuir a la exposición
- RAP% Qué proporción del riesgo de la población se puede atribuir a la exposición





Sesgos en estudios de cohortes

- El sesgo más frecuente es la pérdida de Seguimiento
 - Aparece cuando los sujetos no pueden ser localizados o no quieren seguir participando en el estudio. Puede reducir el poder del estudio para detectar asociaciones que realmente están presentes.
 - Las pérdidas pueden ser dependiente o independientes de la condición de exposición, o del desenlace. Si las pérdidas están relacionadas con alguna característica de los participantes (misma exposición o el desenlace estudiado) determinarán de alguna manera los resultados obtenidos. (diferencial o no diferencial)
- Selección de población general como grupo de comparación para cohorte ocupacional, porque surge efecto de “trabajador sano”
- Clasificación errónea de la exposición (cohortes retrospectivas) o con definiciones amplias de exposición, por ejemplo usar un título de trabajo en vez de la exposición propiamente tal.
- Clasificación errónea de la enfermedad, por ejemplo cuando se agrupan diagnósticos y solamente algunos están asociados a la exposición del estudio.

Cuándo hacer un estudio de Cohorte

- Cuando investigadores quieren saber sobre múltiples efectos de una exposición o cuando la exposición es rara.
- Las cohortes retrospectivas se indican cuando el período de inducción y latencia son largos.
- Se ha usado mucho para examinar los efectos de la exposición a sustancias tóxicas ambientales y ocupacionales.

Ventajas

- Entre los diseños observacionales son lo que entregan la mejor información sobre temporalidad y riesgo.
- Mide la incidencia de una enfermedad (expuestos, no expuestos y total). Es decir mide directamente el riesgo.
- Es menos vulnerable a sesgos
- La exposición está bien definida
- Permite medir las medidas de asociación RR RA% y RAP%.
- Permite estudiar la historia natural de la enfermedad.
- Permite estudiar efectos de exposiciones de baja frecuencia.
- Permite estudiar más de un efecto o consecuencia para la salud.

Desventajas

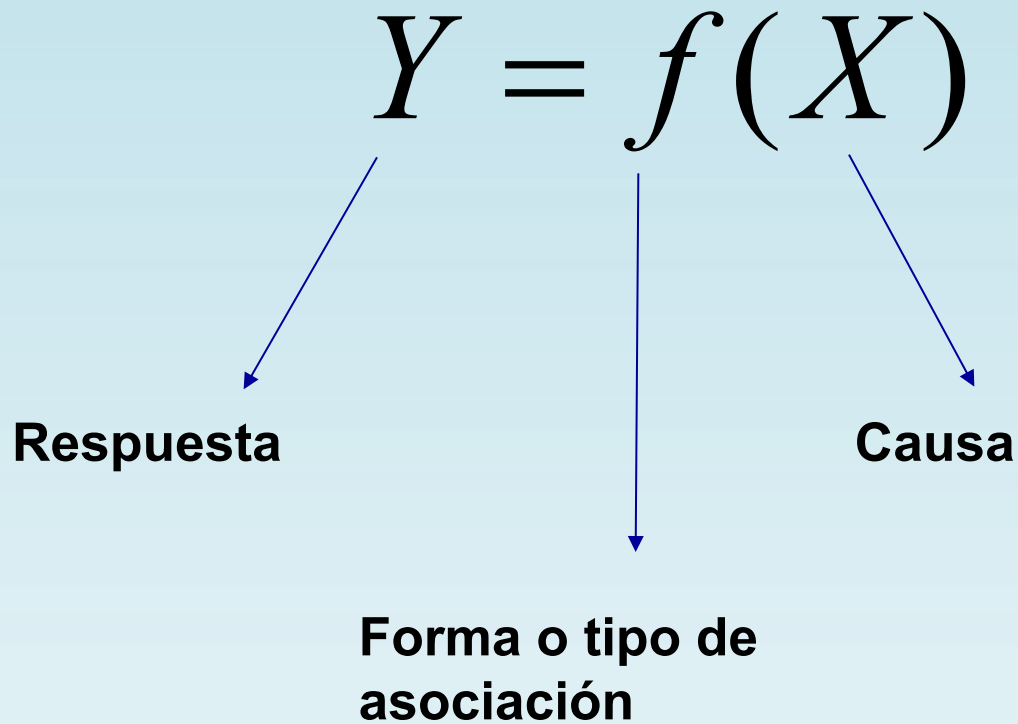
- Caros y consumidores de tiempo
- Complejos y difíciles de llevar a cabo
- Afectado por pérdida de sujetos.
- La exposición puede estar mal clasificada (en cohortes retrospectivas)
- Utilidad limitada para enfermedades de baja frecuencia.

Bioestadística asociada

Regresión lineal simple

Cuando se observa una causa, buscamos la o las causas que lo produjeron. Al simplificar esta estructura cognoscitiva podemos pensar que una respuesta es generada por una causa; lo que podemos representar, cuando causa y efecto son medibles numéricamente (principalmente continuas y SIEMPRE paramétricas), por una relación funcional:

Regresión lineal simple



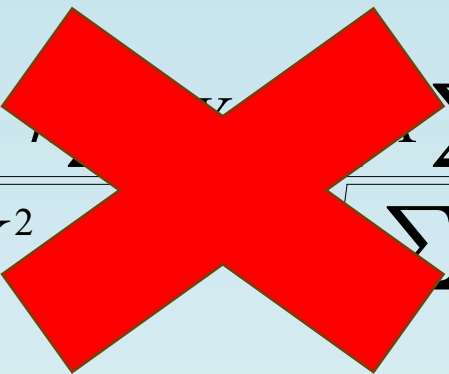
Regresión lineal simple

Particularmente, nos interesa modelar la respuesta cuando la relación funcional entre la respuesta y la causa es lineal, es decir, de la forma:

$$Y = \alpha + \beta \cdot X$$

Regresión lineal simple

Podemos definir el Coeficiente de Correlación entre X e Y como:

$$r_{xy} = \frac{\sum X \sum Y}{\sqrt{n \sum X^2} \sqrt{\sum Y^2 - (\sum Y)^2}}$$


r_{xy} mide el grado de asociación lineal entre X e Y, puede demostrarse que:

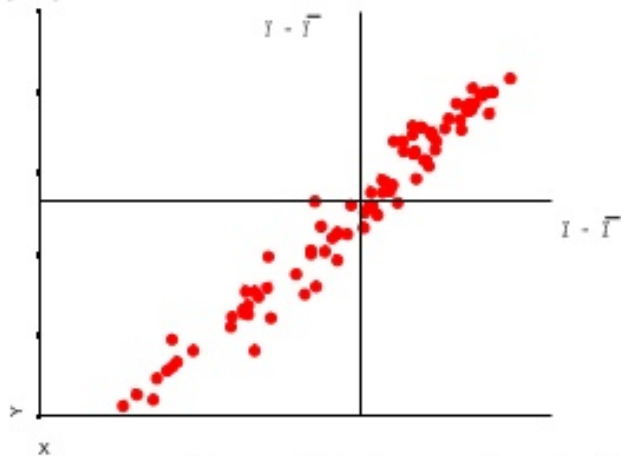
$$-1 \leq r_{xy} \leq 1$$

Regresión lineal simple

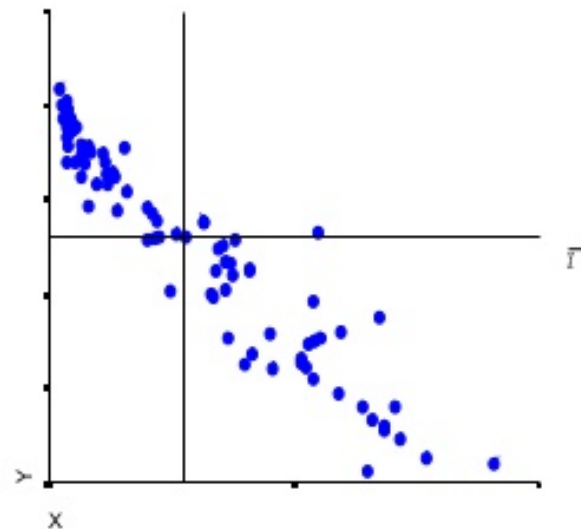
- r_{xy} tiende a 1 la asociación es directa
- r_{xy} tiende a -1 la asociación es inversa
- r_{xy} tiende a 0 no existe asociación lineal

Regresión lineal simple

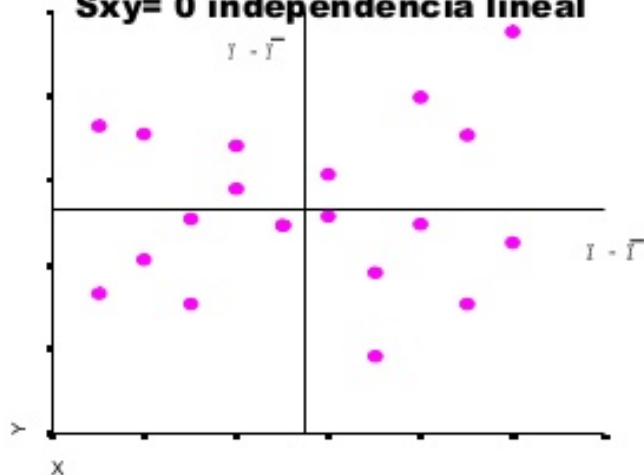
$S_{xy} > 0$ relación lineal directa o positiva



$S_{xy} < 0$ relación lineal inversa o negativa



$S_{xy} = 0$ independencia lineal



Regresión lineal simple

Cuando tengamos ajustado un modelo de regresión, es necesario conocer la calidad del mismo, para ello la variabilidad total de Y, que no depende del modelo ajustado, puede descomponerse del siguiente modo:

$$\sum (Y - \bar{Y})^2 = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \bar{Y})^2$$

$$\text{SCTotal} = \text{SCError} + \text{SCRegresión}$$

$$\text{Varianza Total} = \text{Varianza no explicada} + \text{Varianza explicada}$$

Regresión lineal simple

La descomposición de la variabilidad es posible resumirla en la conocida Tabla de Análisis de la Varianza (Tabla ANOVA)



Fuente de Variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
Regresión	1	$\sum (\hat{Y} - \bar{Y})^2$	$CM_{reg} = \frac{\sum (\hat{Y} - \bar{Y})^2}{1}$	$F = \frac{CM_{reg}}{CM_{res}}$
Residuo	n-2	$\sum (Y - \hat{Y})^2$	$CM_{res} = \frac{\sum (Y - \hat{Y})^2}{n-2}$	
Total	n-1	$\sum (Y - \bar{Y})^2$		

Regresión lineal simple

Asociada a la descomposición de la variabilidad y por ende a la calidad del modelo, se tiene la siguiente dócima:

$$H_0 : \mathfrak{R}^2 = 0$$

$$H_1 : \mathfrak{R}^2 > 0$$

Cuya estadística de prueba es:

$$F = \frac{CM_{reg}}{CM_{res}} \sim \text{F}(1, n-2)$$

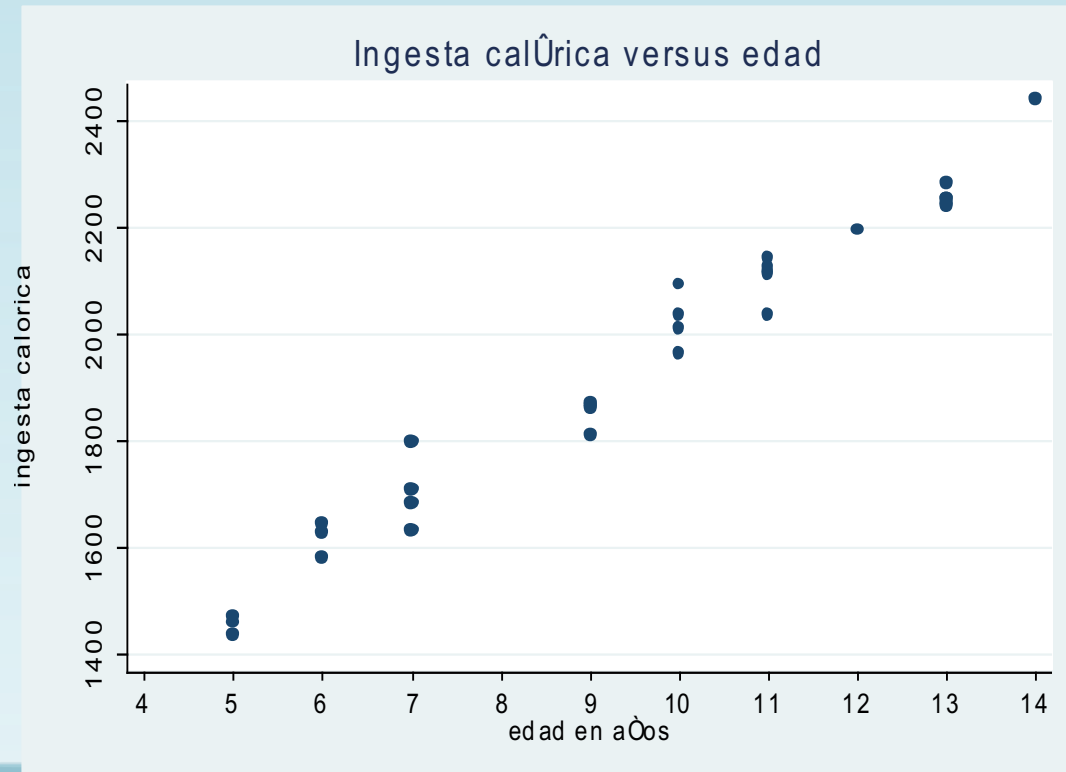
Regresión lineal simple

Ejemplo: Un nutriólogo, desea probar la hipótesis que afirma que la ingesta calórica diaria en niños varones no obesos entre los 5 y 15 años de edad aumenta con esta, para probar dicha hipótesis dispone de la siguiente información:

id	edad	cal	id	edad	cal
1	6	1628	16	13	2283
2	11	2126	17	7	1684
3	10	1963	18	10	2092
4	11	2035	19	7	1710
5	11	2112	20	14	2441
6	6	1581	21	7	1633
7	11	2143	22	6	1645
8	5	1436	23	9	1868
9	10	2009	24	12	2194
10	13	2238	25	13	2243
11	7	1797	26	9	1862
12	5	1460	27	9	1810
13	9	1867	28	13	2252
14	13	2251	29	5	1472
15	10	2035	30	11	2116

Regresión lineal simple

```
twoway (scatter calorías edad, sort)
```



```
. corr cal edad  
(obs=30)
```

	cal	edad
cal	1.0000	
edad	0.9858	1.0000

Regresión lineal simple

```
. reg cal edad
```

Source	SS	df	MS	Number of obs	=	30
Model	2208153.27	1	2208153.27	F(1, 28)	=	961.65
Residual	64294.1928	28	2296.22117	Prob > F	=	0.0000
Total	2272447.47	29	78360.2575	R-squared	=	0.9717
				Adj R-squared	=	0.9707
				Root MSE	=	47.919

cal	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
edad	99.42725	3.206252	31.01	0.000	92.85954 105.995
_cons	994.9363	31.48555	31.60	0.000	930.4411 1059.432

$$\text{Calorias} = 994.9363 + 99.42725 \cdot \text{Edad}$$

Regresión lineal múltiple

$$Y = f(X_1, X_2, \dots, X_p)$$


Respuesta


Causas


**Forma o tipo de
asociación**

Regresión lineal múltiple

- Cuando se decide que hay variables que no explican en un modelo, se puede optar por reducir el mismo. Sin embargo hay que docimar la capacidad de explicación del modelo reducido, esto se hace a través de la dócima de RAZÓN DE VEROSIMILITUDES (LIKELIHOOD RATIO TEST)

Ejemplo

- Un nutriólogo está interesado en explicar el IMC de un grupo de pacientes según: edad, ingreso, horas de ejercicio físico, horas de sueño, ingesta diaria de agua e ingesta diaria de pan.

Ejemplo

Estadística descriptivas

```
. tabstat edad ingreso hduerme hejer agua pan imc, stat(n min q max mean sd) col(stat)
```

variable	N	min	p25	p50	p75	max	mean	sd
edad	150	33	47	49	52	66	49.12	4.8076
ingreso	150	332.1	419	461.4	492.8	565	458.126	50.36931
hduerme	150	6	7	8	8	9	7.486667	.5877573
hejer	150	0	1	2	2	3	1.52	.5760313
agua	150	.5	1	1.2	1.4	1.9	1.212	.2985243
pan	150	1	2	3	3	5	2.586667	.8369541
imc	150	18.4	24.1	26.4	28.4	33.7	26.06667	3.024701

Ejemplo

MATRIZ DE CORRELACIONES

```
. pwcorr imc edad ingreso hduerme hejer agua pan, sig
```

	imc	edad	ingreso	hduerme	hejer	agua	pan
imc	1.0000						
edad	-0.0533 0.5174	1.0000					
ingreso	-0.1005 0.2208	0.2253 0.0056	1.0000				
hduerme	0.0658 0.4236	-0.0541 0.5112	-0.2152 0.0082	1.0000			
hejer	-0.7977 0.0000	0.0088 0.9147	0.1166 0.1553	-0.0785 0.3397	1.0000		
agua	-0.3412 0.0000	0.0364 0.6583	0.0459 0.5774	0.0047 0.9541	0.1196 0.1450	1.0000	
pan	0.6019 0.0000	-0.0827 0.3146	-0.0056 0.9461	0.0024 0.9771	-0.0245 0.7660	-0.1842 0.0241	1.0000

Ejemplo

EXPLORACION DE LA MULTICOLINEALIDAD

```
. pwcorr  edad ingreso hduerme hejer agua pan, sig
```

	edad	ingreso	hduerme	hejer	agua	pan
edad	1.0000					
ingreso	0.2253 0.0056	1.0000				
hduerme	-0.0541 0.5112	-0.2152 0.0082	1.0000			
hejer	0.0088 0.9147	0.1166 0.1553	-0.0785 0.3397	1.0000		
agua	0.0364 0.6583	0.0459 0.5774	0.0047 0.9541	0.1196 0.1450	1.0000	
pan	-0.0827 0.3146	-0.0056 0.9461	0.0024 0.9771	-0.0245 0.7660	-0.1842 0.0241	1.0000

**¡ PROBABLE
MULTICOLINEALIDAD !**

Ejemplo

CONFIRMACION DE LA MULTICOLINEALIDAD

```
. factor edad ingreso hduerme hejer agua pan,pc
(obs=150)
```

(principal components; 6 components retained)

Component	Eigenvalue	Difference	Proportion	Cumulative
1	1.42249	0.24599	0.2371	0.2371
2	1.17650	0.15329	0.1961	0.4332
3	1.02321	0.13323	0.1705	0.6037
4	0.88998	0.09924	0.1483	0.7520
5	0.79074	0.09367	0.1318	0.8838
6	0.69707	.	0.1162	1.0000

```
.display
1.42249/0.69707
```

2.0406702
(MULTICOLINEALIDAD BAJA)

Eigenvectors

Variable	1	2	3	4
edad	0.44889	-0.10041	-0.60833	0.39719
ingreso	0.58582	-0.31494	-0.06527	0.10332
hduerme	-0.41881	0.36780	-0.23519	0.69329
hejer	0.35019	0.12463	0.68468	0.42822
agua	0.30642	0.61708	0.14516	0.01489
pan	-0.25177	-0.59928	0.28375	0.40904

Criterio:

- <30 baja multicolinealidad
- 30 y 100 multicolinealidad moderada
- Sobre 100 alta multicolinealidad

Ejemplo

Estimación del modelo:

¡ Hay modelo !

```
. reg imc edad ingreso hduerme hejer agua pan
```

Source	SS	df	MS
Model	1358.30222	6	226.383704
Residual	4.87114551	143	.034063955
Total	1363.17337	149	9.14881456

Number of obs = 150
F(6, 143) = 6645.84
Prob > F = 0.0000
R-squared = 0.9964
Adj R-squared = 0.9963
Root MSE = .18456

imc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
edad	.0034429	.0032404	1.06	0.290	-.0029624	.0098481
ingreso	-.000092	.000317	-0.29	0.772	-.0007186	.0005345
hduerme	.0259068	.026389	0.98	0.328	-.0262562	.0780698
hejer	-4.022064	.0266561	-150.89	0.000	-4.074755	-3.969373
agua	-1.49267	.0519327	-28.74	0.000	-1.595325	-1.390015
pan	2.010894	.0184395	109.05	0.000	1.974445	2.047343
_cons	28.4669	.3089183	92.15	0.000	27.85626	29.07754

Ejemplo

Estimación del modelo:

```
. reg imc edad ingreso hduerme hejer agua pan
```

Source	SS	df	MS
Model	1358.30222	6	226.383704
Residual	4.87114551	143	.034063955
Total	1363.17337	149	9.14881456

¡ Estas variables no explican !

Number of obs = 150
F(6, 143) = 6645.84
Prob > F = 0.0000
R-squared = 0.9964
Adj R-squared = 0.9963
Root MSE = .18456

imc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
edad	.0034429	.0032404	1.06	0.290	-.0029624 .0098481
ingreso	-.000092	.000317	-0.29	0.772	-.0007186 .0005345
hduerme	.0259068	.026389	0.98	0.328	-.0262562 .0780698
hejer	-4.022064	.0266561	-150.89	0.000	-4.074755 -3.969373
agua	-1.49267	.0519327	-28.74	0.000	-1.595325 -1.390015
pan	2.010894	.0184395	109.05	0.000	1.974445 2.047343
_cons	28.4669	.3089183	92.15	0.000	27.85626 29.07754

Likelihood ratio test

```
. reg imc edad ingreso hduerme hejer agua pan
```

Source	SS	df	MS	Number of obs =	150
Model	1358.30222	6	226.383704	F(6, 143) =	6645.84
Residual	4.87114551	143	.034063955	Prob > F =	0.0000
Total	1363.17337	149	9.14881456	R-squared =	0.9964
				Adj R-squared =	0.9963
				Root MSE =	.18456

imc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
edad	.0034429	.0032404	1.06	0.290	-.0029624 .0098481
ingreso	-.000092	.000317	-0.29	0.772	-.0007186 .0005345
hduerme	.0259068	.026389	0.98	0.328	-.0262562 .0780698
hejer	-4.022064	.0266561	-150.89	0.000	-4.074755 -3.969373
agua	-1.49267	.0519327	-28.74	0.000	-1.595325 -1.390015
pan	2.010894	.0184395	109.05	0.000	1.974445 2.047343
_cons	28.4669	.3089183	92.15	0.000	27.85626 29.07754

```
. est store A
```

```
. reg imc hejer agua pan
```

Source	SS	df	MS	Number of obs =	150
Model	1358.22901	3	452.743004	F(3, 146) =	13368.87
Residual	4.94435819	146	.033865467	Prob > F =	0.0000
Total	1363.17337	149	9.14881456	R-squared =	0.9964
				Adj R-squared =	0.9963
				Root MSE =	.18403

imc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
hejer	-4.024967	.0263614	-152.68	0.000	-4.077066 -3.972868
agua	-1.491234	.0517365	-28.82	0.000	-1.593483 -1.388985
pan	2.009378	.0183264	109.64	0.000	1.973159 2.045598
_cons	28.7944	.0922879	312.01	0.000	28.61201 28.97679

```
. lrtest A
```

```
likelihood-ratio test  
(Assumption: . nested in A)
```

```
LR chi2(3) = 2.24  
Prob > chi2 = 0.5246
```

El modelo reducido
explica lo mismo que
el modelo completo

¿Y cuando la relación es no paramétrica?

- En este caso podemos acudir a una dócima denominada suma de rangos de Spearman
- Se utiliza tanto cuando las variables con cualitativas y también, siendo cuantitativas, se comportan de manera no paramétrica

```
. spearman mpg rep78
```

```
Number of obs = 69  
Spearman's rho = 0.3098
```

```
Test of Ho: mpg and rep78 are independent
```

```
Prob > |t| = 0.0096
```

```
. spearman mpg rep78 headroom trunk weight length  
(obs=69)
```

	mpg	rep78	headroom	trunk	weight	length
mpg	1.0000					
rep78	0.3098	1.0000				
headroom	-0.4683	-0.1583	1.0000			
trunk	-0.6456	-0.2235	0.6749	1.0000		
weight	-0.8567	-0.4138	0.5193	0.6621	1.0000	
length	-0.8463	-0.3854	0.5348	0.7342	0.9508	1.0000

GRACIAS

The image features the word "GRACIAS" in large, bold, sans-serif capital letters. Each letter is a different color: 'G' is orange, 'R' is maroon, 'A' is orange, 'C' is orange, 'I' is yellow, 'A' is maroon, and 'S' is yellow. Below each letter, a hand is visible, holding it up. There are seven hands in total, one for each letter. The hands are light-skinned and appear to be of different ages or sizes. The background is a light blue gradient, with a stylized ocean scene at the bottom. The ocean has horizontal wavy lines in shades of blue and white, suggesting water and sky. The overall composition is clean and modern.