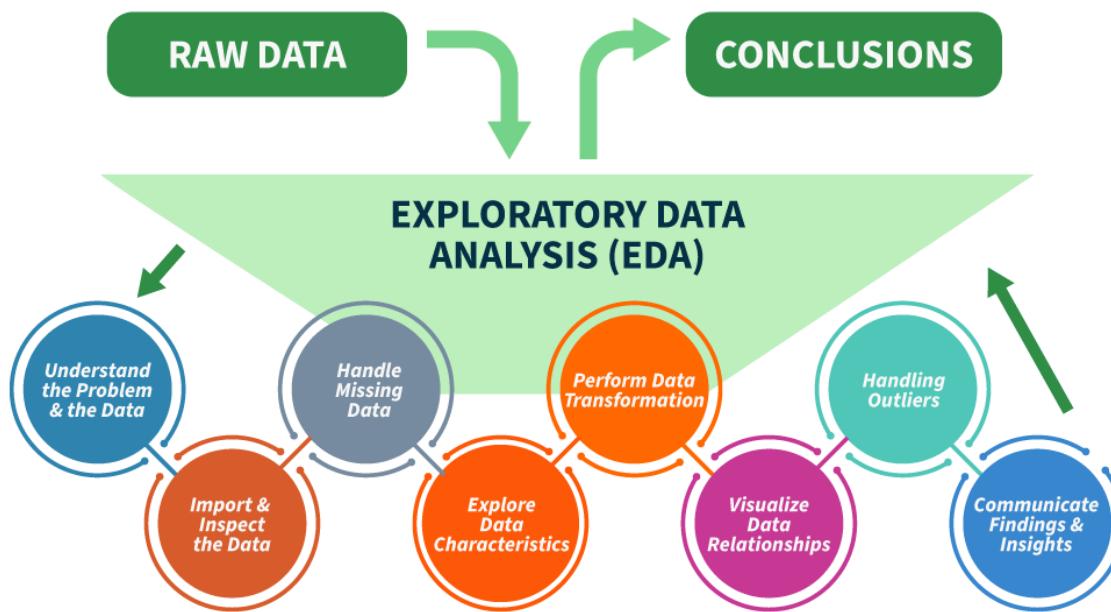


INTRODUCCIÓN A LOS DATOS



Profesor: Pedro Saa (pnsaa@uc.cl)

Año: 1-2025

OBJETIVOS DE APRENDIZAJE

- ▶ **O1:** Comprender los principios de tidy data
 - ▶ **O2.** Identificar y explicar los 4 principios del diseño experimental: controlar, replicar, bloquear y aleatoria
 - ▶ **O3:** Clasificar un estudio como observacional o experimental, y que tipo de conclusiones se pueden obtener (generalizables o causales)
-
- ▶ **O4:** Distinguir entre muestreo aleatorio, simple y estratificado
 - ▶ **O5.** Explicar la diferencia entre muestreo aleatorio y asignación aleatoria y cuál la ventaja que tiene cada una.

Tidy Data - Definiciones

Usualmente nos enfrentamos que los datos no se almacenan en formatos fácil de analizar, la personalización del formato dificulta la **reproducibilidad** del estudio

Cinética Pellet 5.2.25 Humedad=10% b.s																				
Pmuestra [s]	Nº	Tiempo [min]	Tiempo [hr]	PCO2 [kg]	Kg/g	V1	V2	g/g x 100	Corrección	Det. Astaxantina			mg/g	mg/g	vol. resuspensi	vol. aforado				
					CO2/Sub	Vial inicial [g]	Vial + Ex [g]			Ex [g]	Ex/SubM	Ex/SubM	Abs1	Abs2	Va	AxT [mg]	Ax/SubM	Ax/SubM		
P [bar]	550			0,00	95,9	0		0	0						0	0	5	5000		
Temp [°C]	40	1	5,0	0,08	95,85	0,050	18,6356	18,6433	0,0077	0,765	1,328	0,231	0,236	400	0,139	0,138	0,667	10	10000	
U1 [g/min]	10	2	10,0	0,17	95,8	0,100	19,1353	19,1401	0,0048	1,242	2,156	0,286	0,288	800	0,085	0,223	1,077	Materia prima		
U2 [g/min]	10	3	20,0	0,33	95,7	0,200	18,554	18,5609	0,0069	1,927	3,347	0,252	0,247	500	0,119	0,341	1,647	4	10000	
U3 [g/min]	10	4	42,0	0,70	95,5	0,400	18,6726	18,6795	0,0069	2,612	4,537	0,343	0,233	500	0,137	0,477	2,306		5000	
U4 [g/min]	9,09	5	60,0	1,00	95,3	0,600	19,0359	19,046	0,0101	3,616	6,279	0,274	0,295	300	0,226	0,701	3,390			
U5 [g/min]	11,1	6	120,0	2,00	94,7	1,200	18,7012	18,7095	0,0083	4,440	7,711	0,328	0,320	300	0,257	0,957	4,624			
U6 [g/min]	10	7	240,0	4,00	93,5	2,400	18,8974	18,9123	0,0149	5,920	10,282	0,626	0,622	300	0,495	1,449	7,001			
U7 [g/min]	10	8	362,0	6,00	92,3	3,600	18,7869	18,8016	0,015	7,380	12,818	0,511	0,495	200	0,599	2,044	9,875			
U8 [g/min]	9,84	9	480,0	8,00	91,1	4,800	18,6865	18,6964	0,0099	8,364	14,525	0,401	0,413	200	0,485	2,525	12,201			
U9 [g/min]	10,2	10	600,0	10,00	89,9	6,000	18,6959	18,7256	0,0297	11,314	19,649	0,681	0,674	150	2,151	4,661	22,525			
U10 [g/min]	10	Determinación de oleoresina y astaxantina en la muestra inicial								Humedad										
E	#####				g/g*100					mg/g	Ensayos humedad previa a la extracción peletización									
		Muestra [g]	Vial inicial [g]	Vial + Ex [g]	Ex/M	Abs1	Abs2	Va	AxT [mg]	Ax/M	Muestra	Mi [g]	Muestra [g]	Suma	Mf [g]	diff [g]	%H	Desvest		
Fresco	A1	0,0344	5,1479	5,1569	26,163	0,523	0,552	100,0	1,024	29,762	30,397	1	6,8612	0,2995	7,1607	7,1507	0,01	3,3388982	0,1693	
	A2	0,027	5,1283	5,1361	28,889	0,403	0,45	100,0	0,812	30,088			2	6,8153	0,313	7,1283	7,1171	0,0112	3,5782748	5%
	A3	0,0325	5,0626	5,0709	25,538	0,496	0,494	100,0	0,943	29,011			Ensayos humedad posterior a la extracción peletización							
	B1	0,0302	5,1407	5,1494	28,808	0,497	0,505	100,0	0,954	31,599			3	15,6972	0,3085	16,0057	15,9995	0,0062	2,0097245	0,309
	B2	0,0265	5,0969	5,1038	26,038	0,432	0,456	100,0	0,846	31,914			4	6,796	0,3052	7,1012	7,0964	0,0048	1,5727392	17%
	B3	0,0371	5,0851	5,0958	28,841	0,585	0,584	100,0	1,113	30,009								3,459		
Agotado	A1	0,0306	5,0628	5,0653	8,170	0,517	0,524	200,0	0,248	8,100	7,872									
	A2	0,0338	5,1483	5,1511	8,284	0,563	0,567	200,0	0,269	7,960										
	A3	0,0282	5,0968	5,0987	6,738	0,447	0,448	200,0	0,213	7,557										
Análisis Final											18									
Roleo.	RA Ax..	Pureza Ax.	Corrección.ax	corrección.ole	Re. Ax..	Pureza Ax. Cor														
46,8	15,3	4,1	0,21	0,58	32,5	7,7														

¿Qué problema presentan estos datos?

¿Qué tipo de problemas presentan estos datos?

D	F	F	G	H	I	J	K	M	N	O	P	Q	R	S	T
	Taller Técnicas Básicas			Taller Campamento		Taller Primeros Auxilios		Taller Orientacion		Taller Invernal					
email	C1	C2	Rec Marcha	Canoitas	C3	Roque	Aux 1	Mahuida	C4	C5	Maitenes	C6	C7	Union	Insasistencias
andresabudm@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
tjbosch@uc.cl			p	p	p	p	p		p	p	p	p	p	p	APROBADO
simon.contreras1@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
tafernandez@uc.cl			p	p	p	p	p		p	p	p	p	p	p	APROBADO
cgallegos@live.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
gleisner.pablo@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
jagrezv@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
lorebea7@gmail.com								p	p	p	p	p	p	p	APROBADO
gfhoch@uc.cl			p	p	p	p	p		p	p	p	p	p	p	APROBADO
msinfante@uc.cl			p	p	p	p	p		p	p	p	p	p	p	APROBADO
patricia.maublen@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
robclimber89@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
jdsalas@uc.cl			p	p	p	p	p		p	p	p	p	p	p	APROBADO
sanchezg.julian@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
cfsaravia@uc.cl			p	p	p	p	p		p	p	p	p	p	p	APROBADO
pausaravia.v@hotmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
ignacio@agroindustrias.cl			p	p	p	p	p		p	p	p	p	p	p	APROBADO
pablovidaless@yahoo.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
rodrigo.viverosa@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
dim.zogg@gmail.com			p	p	p	p	p		p	p	p	p	p	p	APROBADO
namocain@gmail.com			p	a	p	p	a		a	a	a	a	p	p	7
diegoasta10@gmail.com			a	p	p	p	a		p	p	p	p	p	p	3
tomasboricf@gmail.com			a	a	a	a	a		a	a	a	a	a	a	12
mjbravo4@uc.cl			p	p	p	a	a		a	a	a	a	a	a	9
jeferda@uc.cl			p	p	p	p	p		p	a	n	p	p	p	3
jdragovelasco@gmail.com			a	p	p	p	a		a	p	p	p	a	p	5
silvesru@gmail.com			a	p	p	p	p		a	a	n	p	p	p	5

Tenemos variables que tienen distinta codificación, y también tenemos datos vacíos

Tidy data: concepto de como trabajar adecuadamente los datos para que sean fáciles de analizar, este consiste en construir una **matriz de datos** en que cada fila sea una **observación** y cada columna una **variable**

C	E	F	G	H	I	J	K	L	M	N
Rut	Generacion	Semestre_cur	Tipo	Curso	Clase_teorica1	Clase_teorica2	Salida	Puntaje	Nota	Curso_completo
169382069	201802	201802	ALUMNO	CAMP	1	1	1	89%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	GDR	1	1	0	77%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	INV	1	1	1	100%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	OZN	1	1	1	77%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	PA	1	1	1	100%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	TM	1	1	2	81%	Aprueba Examen	1
189342799	201702	201801	RECUPERADOR	PA	1	1	0	100%	Aprueba Examen	0
186388135	201802	201802	ALUMNO	CAMP	0	1	1	89%	Aprueba Examen	1
186388135	201802	201802	ALUMNO	GDR	1	1	0	77%	Aprueba Examen	1
186388135	201802	201802	ALUMNO	INV	0	1	0		Sin Salida	0
186388135	201802	201802	ALUMNO	OZN	1	1	1	77%	Aprueba Examen	1
186388135	201802	201802	ALUMNO	PA	1	1	1	100%	Aprueba Examen	1
186388135	201802	201802	ALUMNO	TM	1	0	2	78%	Aprueba Examen	1
175157336	201801	201801	ALUMNO	CAMP	1	0	1		Sin Examen	0
175157336	201801	201801	ALUMNO	INV	0	0	0		Sin Salida	0
175157336	201801	201801	ALUMNO	OZN	0	0	0		Sin Salida	0
175157336	201801	201801	ALUMNO	PA	0	0	0		Sin Salida	0
175157336	201801	201801	ALUMNO	TM	1	1	1	47%	Reaprueba Examen	0
190810402	201602	201801	RECUPERADOR	OZN	0	1	1	77%	Aprueba Examen	1
190810402	201602	201801	RECUPERADOR	PA	1	1	0		Sin Salida	0
186365925	201702	201801	RECUPERADOR	CAMP	1	0	1	70%	Aprueba Examen	1
186365925	201702	201801	RECUPERADOR	INV	1	0	1	100%	Aprueba Examen	1
186365925	201702	201801	RECUPERADOR	OZN	0	1	1	84%	Aprueba Examen	1
186365925	201702	201801	RECUPERADOR	PA	1	1	1	53%	Reaprueba Examen	0

Es buena práctica al entregar los datos tener un diccionario de variables que expliquen cada variable y posiblemente como se obtuvieron estos datos

Existen dos grandes familias de variables están pueden ser **numéricas** o **categóricas** estas a su vez dependiendo de su naturaleza se pueden subdividir en variables **continuas, discretas, nominales o ordinales**

C	E	F	G	H	I	J	K	L	M	N
Rut	Generacion	Semestre_curso	Tipo	Curso	Clase_teorica1	Clase_teorica2	Salida	Puntaje	Nota	Curso_completo
169382069	201802	201802	ALUMNO	CAMP	1	1	1	89%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	GDR	1	1	0	77%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	INV	1	1	1	100%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	OZN	1	1	1	77%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	PA	1	1	1	100%	Aprueba Examen	1
169382069	201802	201802	ALUMNO	TM	1	1	2	81%	Aprueba Examen	1
189342799	201702	201801	RECUPERADOR	PA	1	1	0	100%	Aprueba Examen	0
186388135	201802	201802	ALUMNO	CAMP	0	1	1	89%	Aprueba Examen	1

Variable	Tipo	Descripción
RUT	Numérico Entero	ID alumno
Generacion	Numérico Entero	Año mes de ingreso
Semestre_Curso	Numérico Entero	Año mes del curso
Tipo	Categórico Binario	Describe si es alumno o recuperador
Curso	Categórico Bmultiple	ID Curso
Clase_Teorica1	Numérico Entero Binario	1 asiste 0 ausente
Clase_Teorica2	Numérico Entero Binario	2 asiste 0 ausente
Salida	Numérico Entero Binario	3 asiste 0 ausente
Puntaje	Numérico Entero	Porcentaje puntaje examen
Nota	Categorico	Descriptivo status
Curso_Completo	Numérico Entero Binario	1 completa curso 0 no

Es buena práctica al entregar los datos tener un diccionario de variables que expliquen cada variable y posiblemente como se obtuvieron estos datos

Principios del diseño experimental

En un **diseño experimental** el ingeniero cambia **deliberadamente** las variables del sistema o proceso y observa el resultado de la variable respuesta, luego hace una **inferencia** o **toma unas decisiones** sobre que variables son responsables de los cambios observados en la variable respuesta

El ingeniero entiende que su sistema presenta **variabilidad** natural y el **pensamiento estadístico** le provee **herramientas** que le permitan **factorizar** la variabilidad en la toma de decisiones o conclusiones de sus resultados

¿Cuáles son los principios del diseño experimental?

Aleatorizar

Reducir el sesgo de estudio o incorporar la variabilidad de la población en mi unidad experimental

Replicar

Existe una variación natural en distintas unidades experimentales, esta se reduce replicando el estudio o experimento

Fijar/Controlar

Fijar una variable de estudio que puede introducir variación a mi unidad experimental

Bloquear

Bloquear por variables que sospechas y **no tienes control** sobre ellas que pueden afectar tu estudio

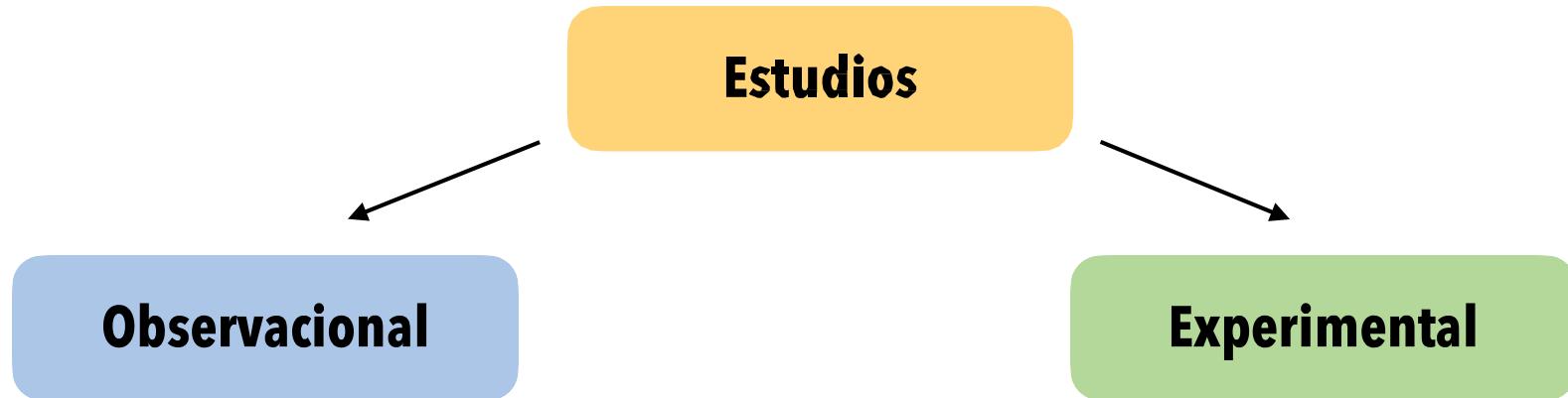
Se busca conocer el efecto de un nuevo sistema de aireación de CO₂ en la tasa de crecimiento de microalgas en piscinas abiertas. Tu sospechas que la fluctuación de temperatura durante el día puede impactar tus resultados

¿Cuál es la mejor estrategia?

1. Tomas 20 muestras para un mismo sector de una piscina aireada y otra no y luego comparas
2. Tomas una muestra para un mismo sector de una piscina aireada y otra no, por 20 días y luego comparas
3. Tomas 20 muestras al azar de una piscina aireada y otra no, luego comparas
4. Tomas 3 muestras al azar de una aireada y otra no, bloqueas por mañana y tarde, y lo haces al azar por 5 días

Tipos de estudio

Existen dos macro clasificaciones de tipos de estudios, estos son **observacionales** y **experimentales**, la diferencia entre ellos recae en como se reúnen los datos...



Reúnen datos de tal forma que no interviene en como el dato se origina

Se define un estudio **retrospectivo** cuando ocupa data reunida en el pasado

Se define un estudio **prospectivo** la data se reúne durante el estudio

Investigadores asignan de forma aleatorias a sujetos de estudio a distintos tipos de tratamiento

Existen dos macro clasificaciones de tipos de estudios, estos son **observacionales** y **experimentales**, la diferencia entre ellos recae en como se reúnen los datos...



En general solo se puede establecer **asociación** entre variables

Si el experimento está bien diseñado y controlado por las fuentes de sesgo, puede establecer **causalidad** entre las variables

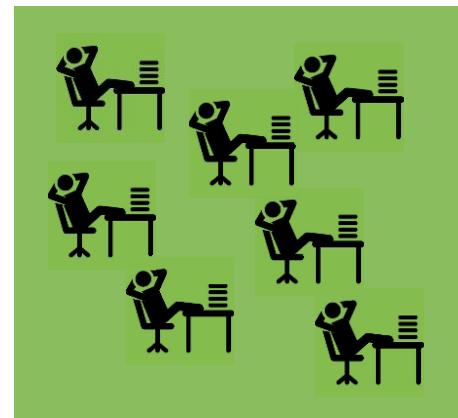
Caso de estudio

**Evaluar el efecto del ejercicio regular en
el peso de los sujetos de estudio**

¿Cómo diseñaría este estudio como observacional?

Primero tomamos 2 muestras al **azar** de la población:

- Personas que **declaran ejercitarse regularmente** según una frecuencia y entrenamiento específico
- Personas que no ejercitan y se **declaran sedentarias**.



Buscamos medir el **peso promedio** que tiene cada grupo y los **comparamos** entre sí mediante alguna herramienta estadística

¿Cómo diseñaría este estudio como experimental?

Primero tomamos una muestra al **azar** de la población, de esta muestra **asignamos de forma aleatoria** individuos que:

- Entrarán a un programa de **ejercitación específica**
- Aquellos que no ejercitarán y tendrán una actividad **sedentaria**

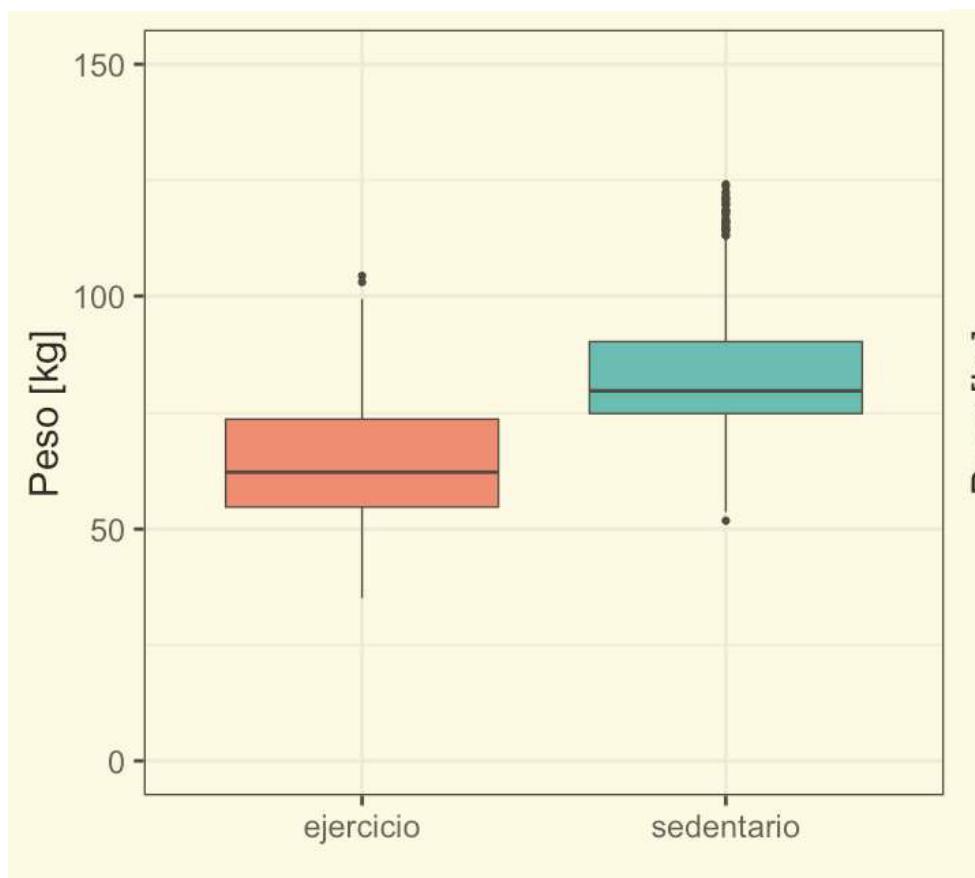


Buscamos medir el **peso promedio** que tiene cada grupo y los **comparamos** entre sí mediante alguna herramienta estadística

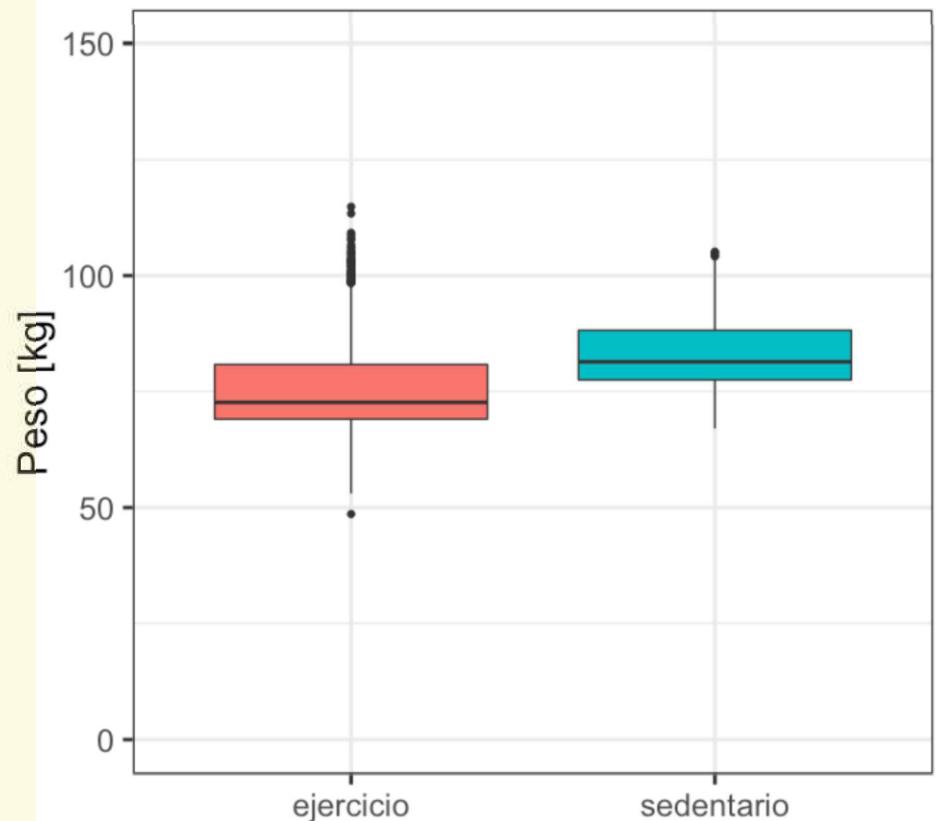
La **clave** que diferencia el estudio observacional con el experimental es que la decisión de pertenecer a un tratamiento no recae en el sujeto sino es **impuesto** por el investigador

¿Qué puedo concluir de los resultados?

Resultados Estudio Observacional



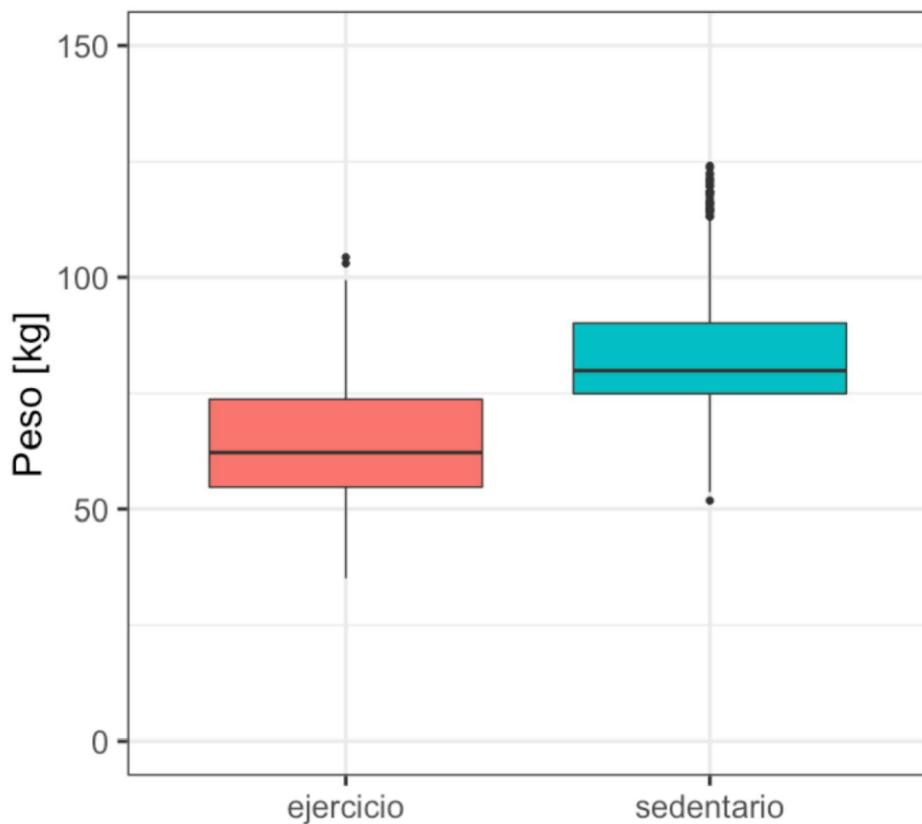
Resultados Estudio Experimental



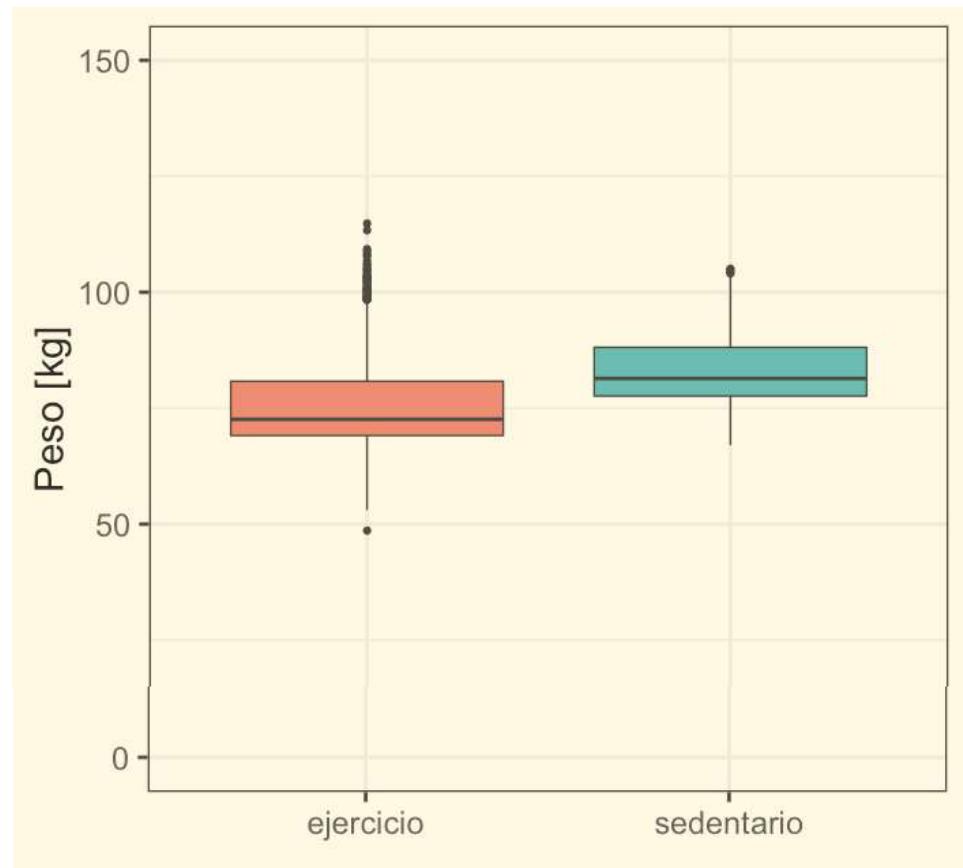
Existe una **asociación negativa** entre el peso de los sujetos y actividad física.
¿Por qué?

¿Qué puedo concluir de los resultados?

Resultados Estudio Observacional



Resultados Estudio Experimental



De existir una diferencia significativa entre grupos, podemos dar una explicación causal: **el ejercicio regular en promedio afecta negativamente el peso de los sujetos**

Caso de estudio

Análisis de un artículo de prensa sobre un estudio científico que relaciona el peso de mujeres con Hábitos alimenticios

Estudio: El cereal de desayuno mantiene las mujeres delgadas*

Mujeres que toman desayuno de cualquier tipo mostraron tener un IMC (Indice de Masa Corporal) inferior al promedio de mujeres que **no mencionan** tomar desayuno.

Este indice era aún mas bajo cuando mujeres **declaraban** tomar cereales en el desayuno

Estos resultados fueron recogidos y seguidos de un estudio mas grande hecho por NIH en que **encuestaron** a 2.379 mujeres en California, Ohio y Maryland entre las edades de 9 y 19 años

La encuesta se realizaba 1 vez al año y se preguntaba qué habían comido los últimos 3 días.

¿Qué tipo de estudio es este?

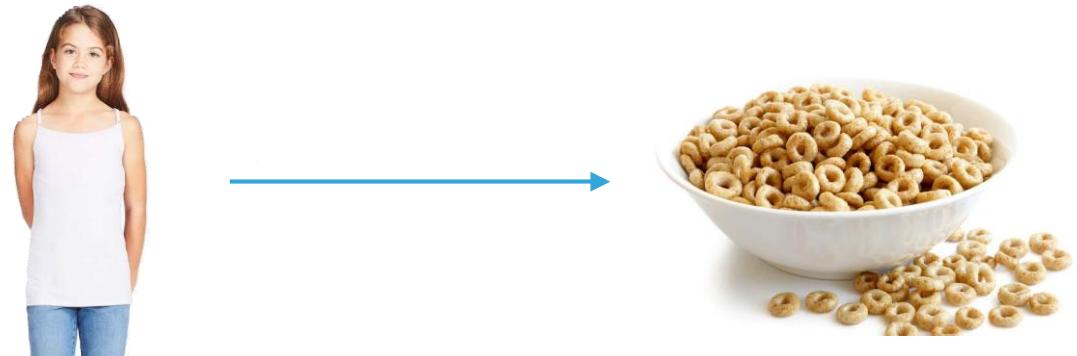
* Estudio liderado por Maryland Medical Research Institute con fondos de NIH (National Institute of Health) y fabricante cereales General-Mills (2008)

¿Qué puedo concluir de este estudio?

Tomar desayuno causa que las mujeres sean mas delgadas



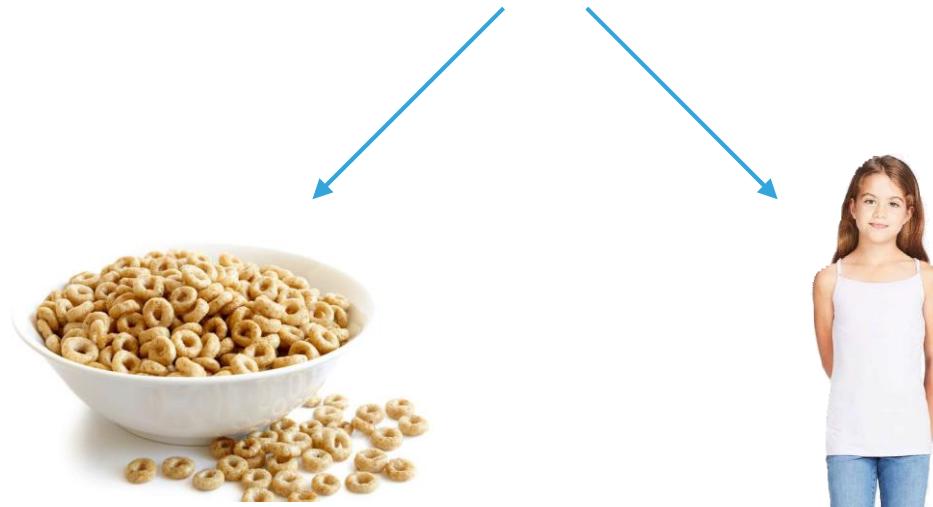
Ser delgada causa que las mujeres tomen desayuno



¿Qué puedo concluir de este estudio?

Puede existir una **tercera variable** que es responsable que las mujeres sean delgadas y a su vez que tomen desayuno

Factor o variable de confusión



Por ejemplo, tener una **consciencia de vida sana**, puede influenciar que las mujeres sean mas delgadas y a la vez alimentarse adecuadamente

En general un estudio define una **variable respuesta** y las **variables explicativas**, este simple etiquetado sirve para producir hipótesis de como una afecta a la otra pero no atribuye causalidad

Variables explicativas



Variable respuesta

" ... Se experimentó con 3 niveles de temperatura (30- 45 - 60°C) y 2 niveles de modificador (50 - 70%) a una misma presión (70 bar). Los resultados mostraron que al incrementar la temperatura en promedio empeoraba la recuperación de carotenoides y la actividad antioxidante de los extractos... Para el caso de 45°C y 50% de modificador la recuperación de carotenoides sobrepaso la recuperación con convencional con acetona...

¿Cuáles son explicativas?

¿Cuáles son respuesta?

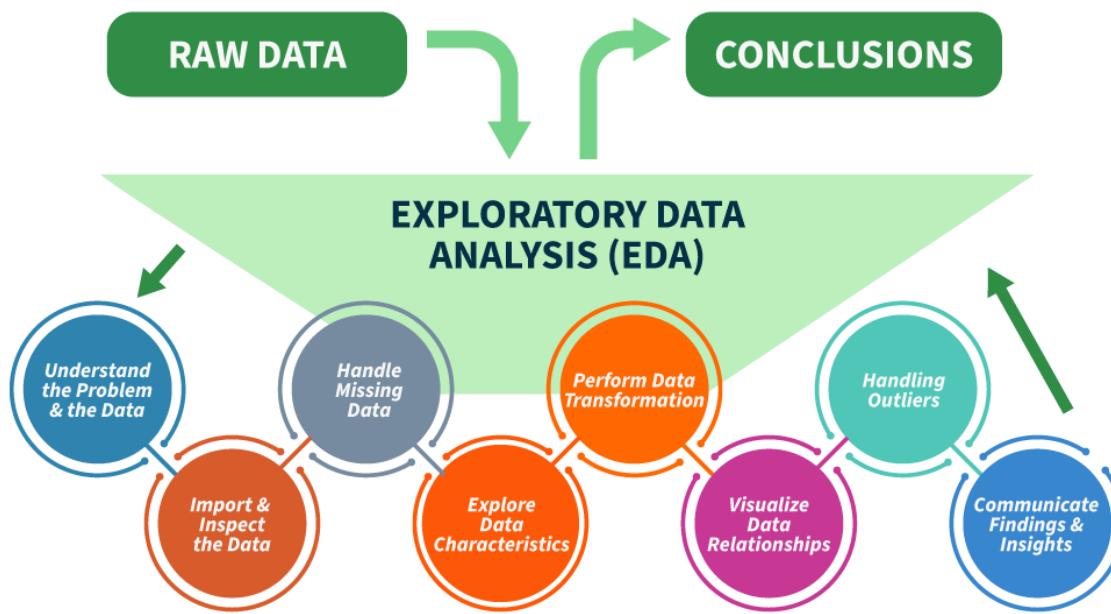
Nunca olvidar...

Correlación no implica causalidad

Resumen

- El diseño experimental tiene cuatro (4) principios fundamentales: **aleatorizar, replicar, fijar y bloquear.**
- Dependiendo de cómo se recolectan los datos y se llevan a cabo los **estudios** ellos pueden ser clasificados como **observacionales o experimentales.**
- Los **estudios experimentales** permiten explorar y entender las causas detrás de los objetos de estudio.
- En este curso nos ocuparemos primordialmente de aprender **metodologías efectivas para el diseño de estudios experimentales.**

INTRODUCCIÓN A LOS DATOS



Profesor: Pedro Saa (pnsaa@uc.cl)

Año: 1-2025