

Las 7 etapas de un proyecto con datos (y un ejemplo)

El recorrido de un proyecto de análisis de datos y un pequeño ejemplo

La estadística y las 6+1 etapas

El papel de la estadística y las 7 etapas

El papel de la Estadística

Observa la Realidad



Gráficos y Resultados



Experto
Tiene una **pregunta**

NUMBER	DATE	DISCOUNT	AMOUNT	MEMBERS	DATE	DISCOUNT	AMOUNT	MEMBERS	DATE	DISCOUNT	AMOUNT	MEMBERS
1	20	20	100	14	10	10	100	5	12	10	100	14
2	10	20	100	14	10	10	100	5	12	10	100	14
3	10	20	100	14	10	10	100	5	12	10	100	14
4	20	20	100	14	20	20	100	5	12	10	100	14
5	20	20	100	14	20	20	100	5	12	10	100	14
6	20	20	100	14	20	20	100	5	12	10	100	14
7	20	20	100	14	20	20	100	5	12	10	100	14
8	20	20	100	14	20	20	100	5	12	10	100	14
9	20	20	100	14	20	20	100	5	12	10	100	14
10	20	20	100	14	20	20	100	5	12	10	100	14
11	20	20	100	14	20	20	100	5	12	10	100	14
12	20	20	100	14	20	20	100	5	12	10	100	14
13	20	20	100	14	20	20	100	5	12	10	100	14
14	20	20	100	14	20	20	100	5	12	10	100	14
15	20	20	100	14	20	20	100	5	12	10	100	14
16	20	20	100	14	20	20	100	5	12	10	100	14
17	20	20	100	14	20	20	100	5	12	10	100	14
18	20	20	100	14	20	20	100	5	12	10	100	14
19	20	20	100	14	20	20	100	5	12	10	100	14
20	20	20	100	14	20	20	100	5	12	10	100	14
21	20	20	100	14	20	20	100	5	12	10	100	14
22	20	20	100	14	20	20	100	5	12	10	100	14
23	20	20	100	14	20	20	100	5	12	10	100	14
24	20	20	100	14	20	20	100	5	12	10	100	14
25	20	20	100	14	20	20	100	5	12	10	100	14
26	20	20	100	14	20	20	100	5	12	10	100	14
27	20	20	100	14	20	20	100	5	12	10	100	14
28	20	20	100	14	20	20	100	5	12	10	100	14
29	20	20	100	14	20	20	100	5	12	10	100	14
30	20	20	100	14	20	20	100	5	12	10	100	14
31	20	20	100	14	20	20	100	5	12	10	100	14
32	20	20	100	14	20	20	100	5	12	10	100	14
33	20	20	100	14	20	20	100	5	12	10	100	14
34	20	20	100	14	20	20	100	5	12	10	100	14
35	20	20	100	14	20	20	100	5	12	10	100	14
36	20	20	100	14	20	20	100	5	12	10	100	14
37	20	20	100	14	20	20	100	5	12	10	100	14
38	20	20	100	14	20	20	100	5	12	10	100	14
39	20	20	100	14	20	20	100	5	12	10	100	14
40	20	20	100	14	20	20	100	5	12	10	100	14
41	20	20	100	14	20	20	100	5	12	10	100	14

Tabla de Datos



Respuestas Conclusiones

Etapas de un proyecto con Datos

LA ETAPA DE ANÁLISIS DE DATOS (ETAPA4 + ETAPA5 + ETAPA 6) ESTÁ DIRECTAMENTE RELACIONADA CON UN SÓLO OBJETIVO ALINEADO CON EL PROBLEMA DEFINIDO EN LA ETAPA 1

ETAPA1: EL PROBLEMA

- Identificar el problema
- Definir un objetivo de estudio

ETAPA2: RECOLECCIÓN

- Estrategia de muestreo
- Diseño experimental

ETAPA3: LIMPIEZA

- Igualar formatos
- Eliminar información errónea

ETAPA4: EXPLORACIÓN

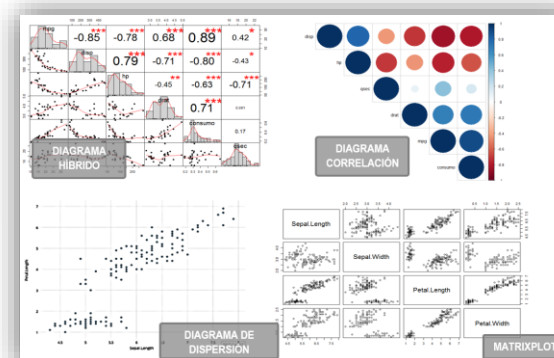
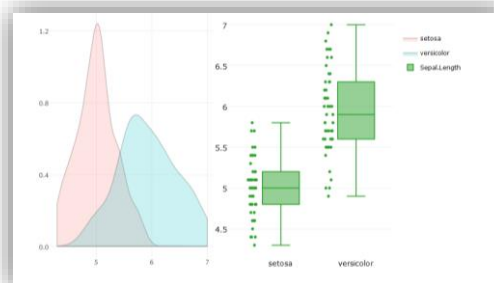
- Descripción de variables una a una
- Buscar diferencias entre grupos
- Forma de la distribución y proporciones

ETAPA5: ANÁLISIS

- Comparación de grupos
- Comparación de proporciones
- Modelos predictivos
- Relación entre variables

ETAPA6: CONCLUSIÓN

- Informe de resultados
- Interpretación y síntesis de resultados
- Objeciones y siguientes pasos



Contraste de Hipótesis
Correlación
Comparación de medias
Modelos predictivos
Clustering

Tabela con los valores de centralidad y dispersión (media [sepal.Length] y el p-valor de la prueba Wilcoxon para comparar grupos)

	Setosa	Versicolor	p-valor	Setosa	Virginica	p-valor	Versicolor	Virginica	p-valor
sepal.Length	5.00 [4.80 - 5.20]	5.50 [5.40 - 6.00]	0.000121	5.00 [4.80 - 5.20]	6.50 [6.20 - 6.90]	2.20E-08	5.90 [5.80 - 6.30]	6.50 [6.20 - 6.90]	5.87E-04
sepal.Width	3.40 [3.20 - 3.60]	3.80 [3.70 - 4.00]	2.14E-01	3.40 [3.20 - 3.60]	3.80 [3.60 - 4.00]	7.10E-04	3.80 [3.70 - 4.00]	3.80 [3.60 - 4.00]	0.000172
petal.Length	1.50 [1.40 - 1.60]	4.30 [4.20 - 4.40]	2.20E-08	1.50 [1.40 - 1.60]	5.00 [4.80 - 5.20]	2.20E-08	4.30 [4.20 - 4.40]	5.00 [4.80 - 5.20]	2.20E-08
petal.Width	0.20 [0.20 - 0.30]	1.30 [1.20 - 1.50]	2.20E-08	0.20 [0.20 - 0.30]	2.00 [1.80 - 2.30]	2.20E-08	1.30 [1.20 - 1.50]	2.00 [1.80 - 2.30]	2.20E-08

Diferencias significativas entre todos las especies y todas las variables.

Las características que mejor ayudan a diferenciar poblaciones de especies son las dimensiones de los pétalos para todas las especies.

Las peor característica es el ancho del sépalos.

Y la longitud del pétalo funciona muy bien para diferenciar setosa del resto. No tan bien para diferenciar versicolor de virginica.

Los conceptos base

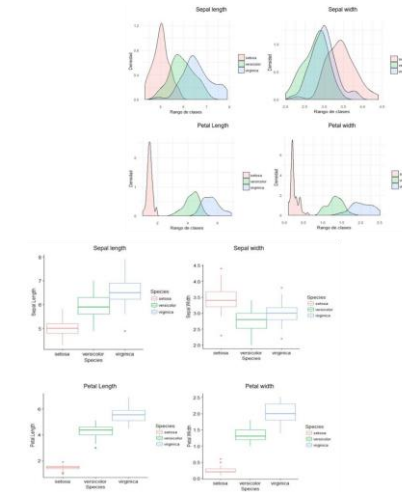
- **Tabla de Datos:**
variables, estructura de datos
- **Estadística Descriptiva:**
características, tablas y gráficos
- **Estadística Inferencial:**
contraste de hipótesis, probabilidad, p-valor, test estadístico, problemas tipo

Medimos y observamos datos mediante variables

Fisher's Iris Data

Largo de sépalo	Ancho de sépalo	Largo de pétalo	Ancho de pétalo	Especies
5.0	2.0	3.5	1.0	<i>I. versicolor</i>
6.0	2.2	4.0	1.0	<i>I. versicolor</i>
6.0	2.2	5.0	1.5	<i>I. virginica</i>
6.2	2.2	4.5	1.5	<i>I. versicolor</i>
4.5	2.3	1.3	0.3	<i>I. setosa</i>
5.0	2.3	3.3	1.0	<i>I. versicolor</i>
5.5	2.3	4.0	1.3	<i>I. versicolor</i>
6.3	2.3	4.4	1.3	<i>I. versicolor</i>
4.9	2.4	3.3	1.0	<i>I. versicolor</i>
5.5	2.4	3.7	1.0	<i>I. versicolor</i>
5.5	2.4	3.8	1.1	<i>I. versicolor</i>
4.9	2.5	4.5	1.7	<i>I. virginica</i>
5.1	2.5	3.0	1.1	<i>I. versicolor</i>
5.5	2.5	4.0	1.3	<i>I. versicolor</i>
5.6	2.5	3.9	1.1	<i>I. versicolor</i>
5.7	2.5	5.0	2.0	<i>I. virginica</i>
6.3	2.5	4.9	1.5	<i>I. versicolor</i>
6.3	2.5	5.0	1.9	<i>I. virginica</i>
6.7	2.5	5.8	1.8	<i>I. virginica</i>
5.5	2.6	4.4	1.2	<i>I. versicolor</i>
5.7	2.6	3.5	1.0	<i>I. versicolor</i>
5.8	2.6	4.0	1.2	<i>I. versicolor</i>
6.1	2.6	5.6	1.4	<i>I. virginica</i>

Estadística Descriptiva



Interpretar la tabla de datos con gráficos y Tablas

Estadística Inferencial

```
call:
glm(formula = lifeExp ~ log(gdppercap) + continent, data = data2007)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-19.4917   -2.3146   -0.0432    2.5498   14.8818

Coefficients:
(Intercept)      20.1376      4.0332      4.993 1.79e-06 ***
log(gdppercap)    4.6308    0.5274    8.780 6.14e-15 ***
continentAmericas 11.6942    1.6546    7.068 7.46e-11 ***
continentAsia     10.1144    1.4761    6.852 2.31e-10 ***
continentEurope   11.2682    1.8936    5.951 2.14e-08 ***
continentAfrica   12.9293    4.5211    2.860 0.00491 **
```

Muestra

Población

Interpretar resultados y extraer conclusiones

Un pequeño ejemplo

Un ejemplo para ilustrar lo que hemos visto

Problema, Recolección y limpieza

Problema

Mejorar el tratamiento del dolor lumbar.

El 85% de la población los sufre a lo largo de la vida



Recolección



Grupo Control

Tratamiento Convencional
VS
Tratamiento Estándar



Grupo Investigación

Mediciones
la mejora
antes y
después de
los
tratamientos

Problema, Recolección y limpieza

La tabla de datos

Nombre	Tipo	¿Qué característica observa?
<i>Id Paciente</i>	<i>Cuantitativa Discreta</i>	<i>Identificador del Paciente. Número identificado.</i>
<i>Sexo</i>	<i>Cualitativa Nominal</i>	<i>Sexo (0 = Hombre 1 = Mujer)</i>
<i>Edad</i>	<i>Cuantitativa Discreta</i>	<i>Edad del Paciente</i>
<i>Peso</i>	<i>Cuantitativa Continua</i>	<i>Peso del paciente en kg</i>
<i>Altura</i>	<i>Cuantitativa Continua</i>	<i>Altura en metros</i>
<i>LC</i>	<i>Cualitativa Nominal</i>	<i>Tipo de Lumbociática (0 = NO Lumbociática, 1 = Lumbociática Derecha y 2 = Lumbociática Izquierda)</i>
<i>PC</i>	<i>Cualitativa Nominal</i>	<i>Pierna corta aparente (0 = No Pierna Corta, 1 = Pierna Corta Derecha, 2 = Pierna Corta Izquierda)</i>
<i>NHD</i>	<i>Cualitativa Ordinal</i>	<i>Número hernia discal (0 = Ninguna, 1 = Una, 2 = Dos)</i>
<i>ODIMes0</i>	<i>Cuantitativa Continua</i>	<i>Índice de Discapacidad de Oswestry Mes0 (Porcentaje Antes Tratamiento)</i>
<i>ODIMes1</i>	<i>Cuantitativa Continua</i>	<i>Índice de Discapacidad de Oswestry Mes1 (Porcentaje Después Tratamiento)</i>
<i>Grupo</i>	<i>Cualitativa Nominal</i>	<i>Tipo de Tratamiento (0 = Convencional, 1 = Investigación)</i>
<i>LF_Mes0</i>	<i>Cualitativa Ordinal</i>	<i>Límite Funcional Mes0 (0 = Mínima, 1 = Moderada, 2 = Intensa, 3 = Discapacidad, 4 = Máxima)</i>
<i>LF_Mes1</i>	<i>Cualitativa Ordinal</i>	<i>Límite Funcional Mes1 (0 = Mínima, 1 = Moderada, 2 = Intensa, 3 = Discapacidad, 4 = Máxima)</i>
<i>IMC_Grupos</i>	<i>Cualitativa Ordinal</i>	<i>Estados Nutricional (0 = desnutrido, 1 = Normal, 2 = Sobrepeso, 3 = Obesidad, 4 = Obesidad Mórbida)</i>

	A	B	C	D	E	F	G	H	I	J	K
1	Id Paciente	Sexo	Edad	Peso	Altura	LC	PC	NHD	ODI Mes0	ODI Mes1	Grupo
2	1	0	44	64.5	1.63	0	0	0	44.3	35.2	0
3	2	0	33	67.52	2.02	0	0	0	38.9	29.4	0
4	3	0	47	72.72	1.6	0	0	0	39.3	31.7	0
5	4	0	34	71.83	1.85	0	0	0	33.4	23.6	0
6	5	0	41	52.08	1.67	0	0	0	42.7	34.6	0
7	6	0	38	72.24	1.61	1	2	1	69.4	55.1	0
8	7	0	47	77.98	1.57	0	0	0	31.3	22.4	0
9	8	0	60	60.54	1.71	1	2	1	61.2	47.9	0
10	9	0	75	89.12	1.98	1	2	1	65	52.3	0
11	10	0	40	68.88	1.54	1	2	1	68.6	54.1	0
12	11	0	71	86.48	1.78	2	1	0	63.9	55.2	0
13	12	0	51	60.13	1.65	0	0	0	34	26.8	0
14	13	0	46	74.17	1.75	0	0	0	36.3	27.6	0
15	14	0	57	69.93	1.82	1	2	0	71.8	56.9	0
16	15	0	56	89.46	1.89	0	0	0	74.6	66	0
17	16	0	72	109.2	2.08	1	2	0	65.3	53	0
18	17	0	64	66.33	1.62	1	2	0	65.5	58	0
19	18	0	71	65.79	1.67	2	1	0	60.2	54.7	0
20	19	0	56	72.21	2.02	0	0	0	68.3	62.8	0
21	20	0	35	61.85	1.59	0	0	0	58	50.1	0
22	21	0	35	63.42	1.78	0	0	0	25.2	20.2	0
23	22	0	73	72.27	1.69	2	1	0	66.6	57.7	0
24	23	0	46	63.3	1.38	0	0	0	48.2	38.9	0
25	24	0	62	64.33	1.78	1	2	0	67.1	61.5	0
26	25	0	79	73.55	1.71	2	1	1	72.6	49.6	0
27	26	0	63	65.29	1.66	1	2	1	62.3	54.6	0
28	27	0	64	66.55	1.31	1	2	1	64.8	58.7	0
29	28	0	63	69.34	1.88	2	1	1	67.6	43.1	0
30	29	0	31	66.01	1.68	2	1	1	71.7	53.5	0
31	30	0	55	71.57	1.8	2	1	1	67.7	46.2	0
32	31	0	36	64.81	1.78	0	0	0	36.8	27.1	0
33	32	0	73	60.42	1.64	1	2	1	73.3	61.9	0
34	33	0	77	88.9	1.87	1	2	1	68.9	62.1	0
35	34	0	66	60.15	1.44	1	2	1	72.7	62.2	0

Los 4 pasos fundamentales en el Análisis de Datos

- **Objetivo:** definir qué queremos obtener con los datos

OBJETIVO

Comparar dos tratamientos

Control vs investigación

La diferencia entre el ODI en el mes 0 y el mes 1 y analizar qué tratamiento mejora más a los pacientes

	A	B	C	D	E	F	G	H	I	J	K
1	Id Paciente	Sexo	Edad	Peso	Altura	LC	PC	NHD	ODI Mes0	ODI Mes1	Grupo
2	1	0	44	64.5	1.63	0	0	0	44.3	35.2	0
3	2	0	33	67.52	2.02	0	0	0	38.9	29.4	0
4	3	0	47	72.72	1.6	0	0	0	39.3	31.7	0
5	4	0	34	71.83	1.85	0	0	0	33.4	23.6	0
6	5	0	41	52.08	1.67	0	0	0	42.7	34.6	0
7	6	0	38	72.24	1.61	1	2	1	69.4	55.1	0
8	7	0	47	77.98	1.57	0	0	0	31.3	22.4	0
9	8	0	60	60.54	1.71	1	2	1	61.2	47.9	0
10	9	0	75	89.12	1.98	1	2	1	65	52.3	0
11	10	0	40	68.88	1.54	1	2	1	68.6	54.1	0
12	11	0	71	86.48	1.78	2	1	0	63.9	55.2	0
13	12	0	51	60.13	1.65	0	0	0	34	26.8	0
14	13	0	46	74.17	1.75	0	0	0	36.3	27.6	0
15	14	0	57	69.93	1.82	1	2	0	71.8	56.9	0
16	15	0	56	89.46	1.89	0	0	0	74.6	66	0
17	16	0	72	109.2	2.08	1	2	0	65.3	53	0
18	17	0	64	66.33	1.62	1	2	0	65.5	58	0
19	18	0	71	65.79	1.67	2	1	0	60.2	54.7	0
20	19	0	56	72.21	2.02	0	0	0	68.3	62.8	0
21	20	0	35	61.85	1.59	0	0	0	58	50.1	0
22	21	0	35	63.42	1.78	0	0	0	25.2	20.2	0
23	22	0	73	72.27	1.69	2	1	0	66.6	57.7	0
24	23	0	46	63.3	1.38	0	0	0	48.2	38.9	0
25	24	0	62	64.33	1.78	1	2	0	67.1	61.5	0
26	25	0	79	73.55	1.71	2	1	1	72.6	49.6	0
27	26	0	63	65.29	1.66	1	2	1	62.3	54.6	0
28	27	0	64	66.55	1.31	1	2	1	64.8	58.7	0
29	28	0	63	69.34	1.88	2	1	1	67.6	43.1	0
30	29	0	31	66.01	1.68	2	1	1	71.7	53.5	0
31	30	0	55	71.57	1.8	2	1	1	67.7	46.2	0
32	31	0	36	64.81	1.78	0	0	0	36.8	27.1	0
33	32	0	73	60.42	1.64	1	2	1	73.3	61.9	0
34	33	0	77	88.9	1.87	1	2	1	68.9	62.1	0
35	34	0	66	60.15	1.44	1	2	1	73.7	63.3	0

Los 4 pasos fundamentales en el Análisis de Datos

- **Exploración/Descripción:** interpretar los datos

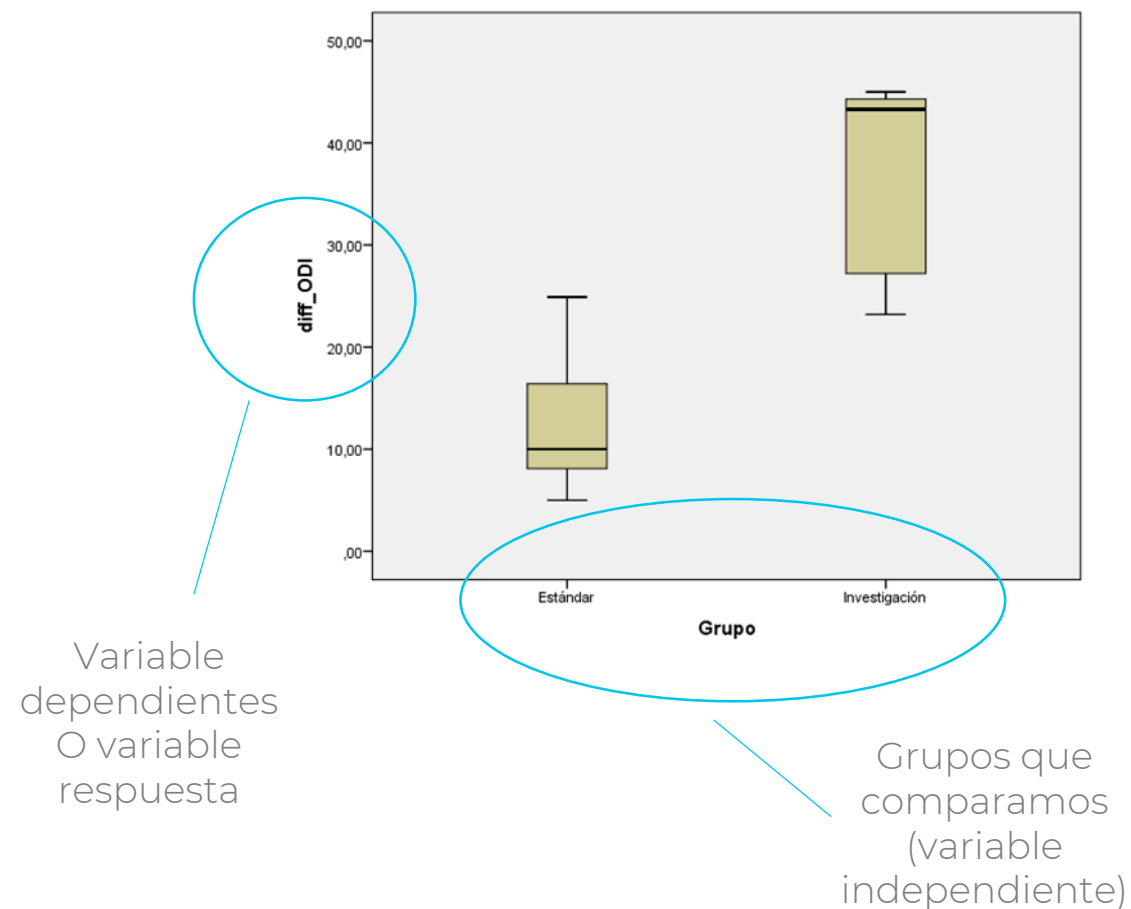
OBJETIVO

Comparar dos tratamientos

Control vs investigación

La diferencia entre el ODI en el mes 0 y el mes 1 y analizar qué tratamiento mejora más a los pacientes

Comparamos las distribuciones de la diferencia de ODI – limitación funcional



Los 4 pasos fundamentales en el Análisis de Datos

- **Análisis y conclusión:** aplicar técnicas de análisis

OBJETIVO

Comparar dos tratamientos

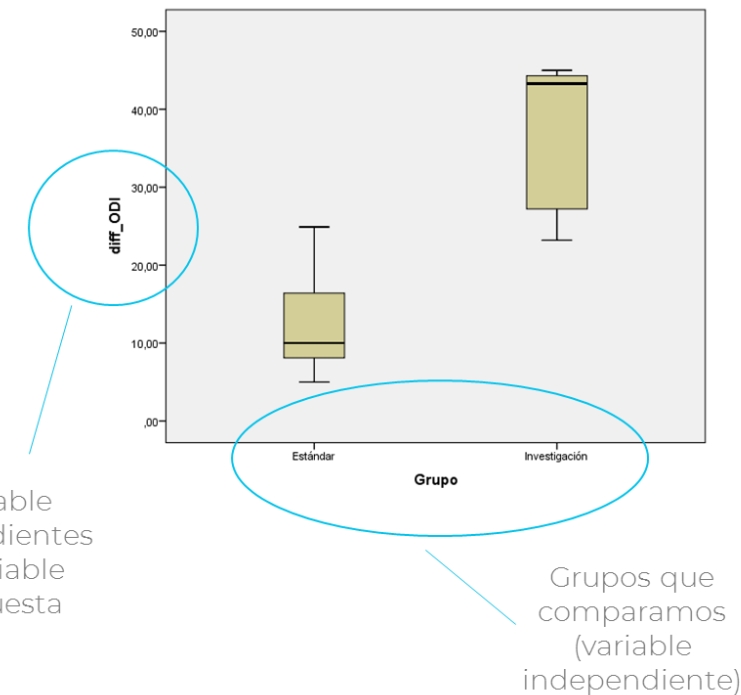
Control vs investigación

La diferencia entre el ODI en el mes 0 y el mes 1 y analizar qué tratamiento mejora más a los pacientes

Comparación de medias

Estadísticos de prueba ^a	
	diff_ODI
U de Mann-Whitney	27,000
W de Wilcoxon	5178,000
Z	-12,151
Sig. asintótica (bilateral)	,000

Detalles:
Test NO paramétrico
Grupos independientes



Hay diferencias significativas en el efecto sobre la limitación funcional entre el tratamiento control y el de investigación

Take away

El resumen de la lección

Lo más importante de la lección

- 7 etapas fundamentales en un proyecto con datos
 - Problema
 - Recolección
 - Limpieza
 - Objetivo
 - Exploración
 - Análisis
 - Conclusión
- Una vez tengamos la tabla de datos. Pasamos por el: **objetivo, descripción, análisis y resultados**

Tú turno

Un pequeño ejercicio para empezar

Tú turno

- Aplica lo que has aprendido en tu proyecto
- Tienes esta presentación para inspirarte