



¿Cómo se realiza un análisis de los resultados de un trabajo científico?

Parte 1



Dr. Diego Stalder (FIUNA)
dstalder@ing.una.py

Contenido: Análisis estadístico

Día 1

- Revisión de algunos conceptos
- Medición/Muestreo
- Estadística Descriptiva
- Probabilidades
 - Distribuciones de Discretas y Continuas

Día 2

- Estimación de Parámetros
 - Regresión Lineal
 - Regresión Logística
 - Bondad de ajuste

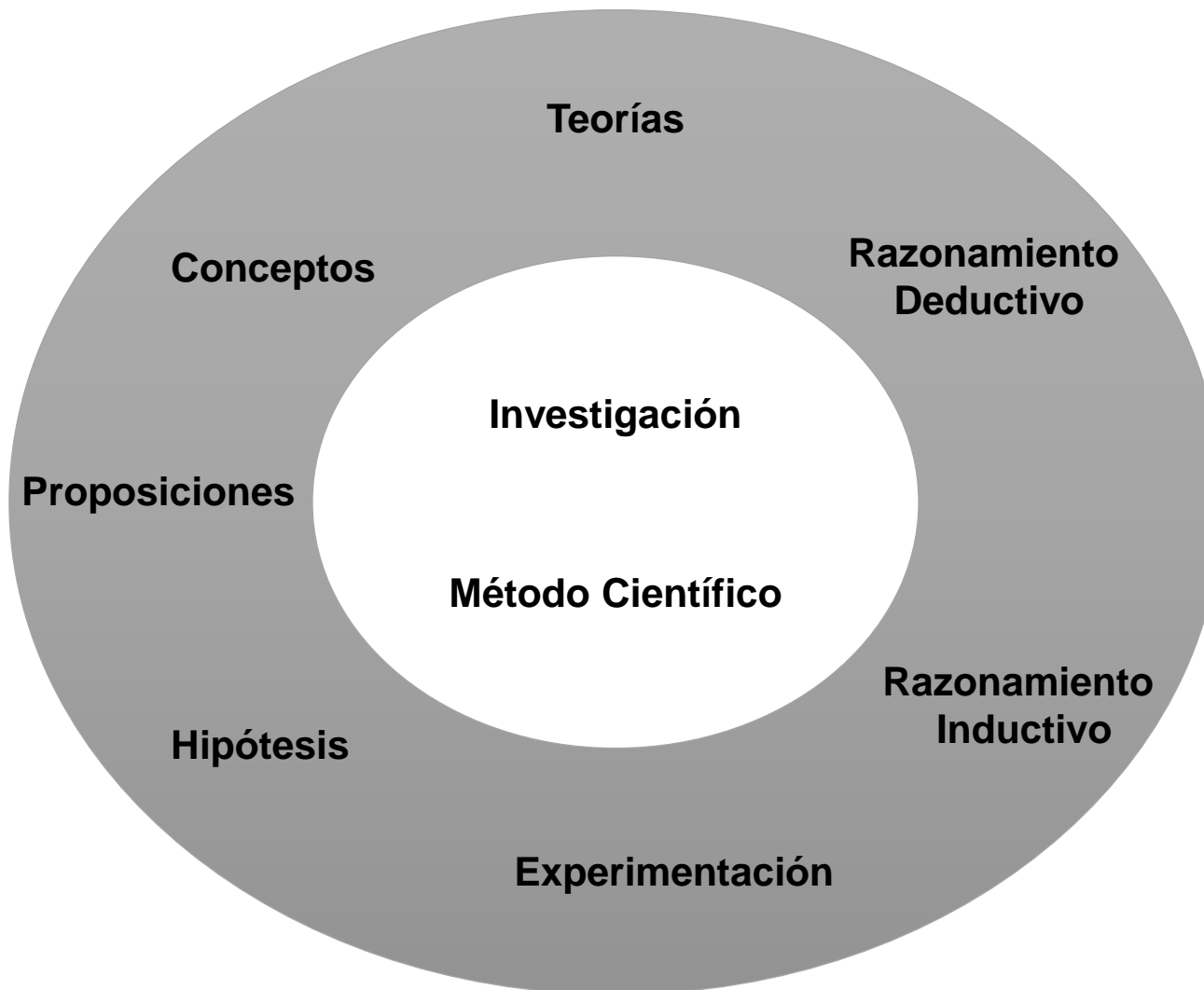
Tipos de Investigación



La investigación se realiza en diferentes areas de la ciencia, por ejemplo:

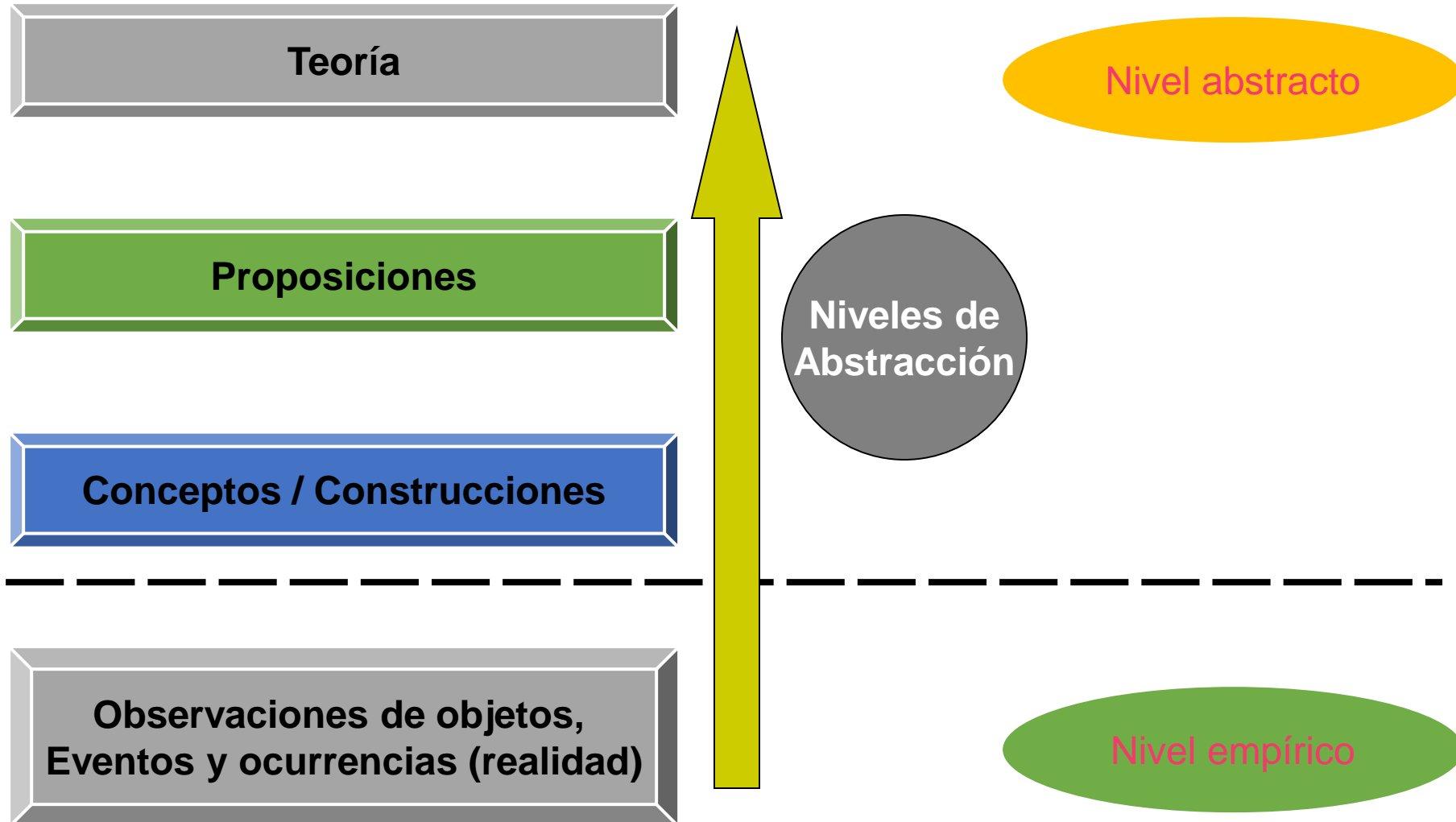
- ☐ Ciencias de la Salud
- ☐ Ciencias Sociales
- ☐ Ingeniería
- ☐ Matemáticas
- ☐ Física
- ☐ Computación.

Dimensiones de la Investigación



En el marco del curso "De la mano de científicos:
¿Cómo se realiza un análisis de los resultados de un trabajo científico?"

La escalera de la abstracción



En el marco del curso "De la mano de científicos:
¿Cómo se realiza un análisis de los resultados de un trabajo científico?"

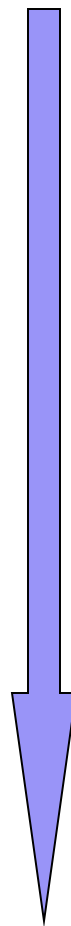
Razonamiento deductivo

Teoría

Hipótesis

Observación

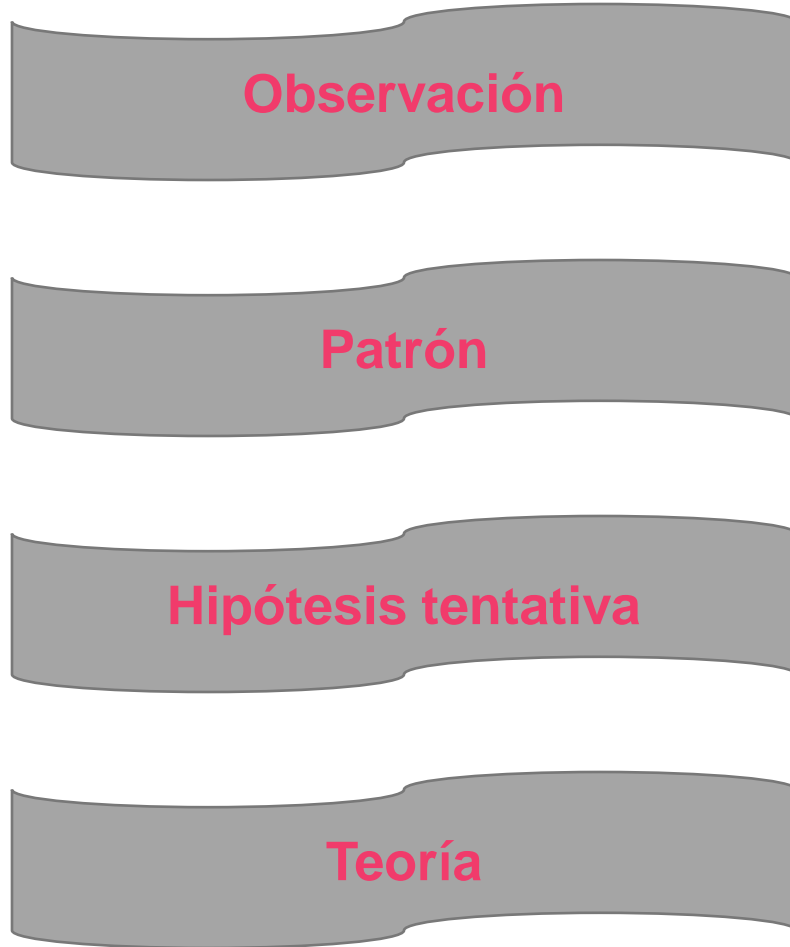
Confirmación



Usando el razonamiento deductivo, uno comienza con una teoría dada como el base por la cual desarrollamos hipótesis y luego confirmar estos con datos adquiridos mediante observación o experimentación

(¿Nuestra teoría es válida o no?)

Razonamiento inductivo



Usando el razonamiento inductivo, uno comienza con una observación específica como base para la cual desarrollamos un patrón general y tentativo hipótesis como fundamento de una teoría

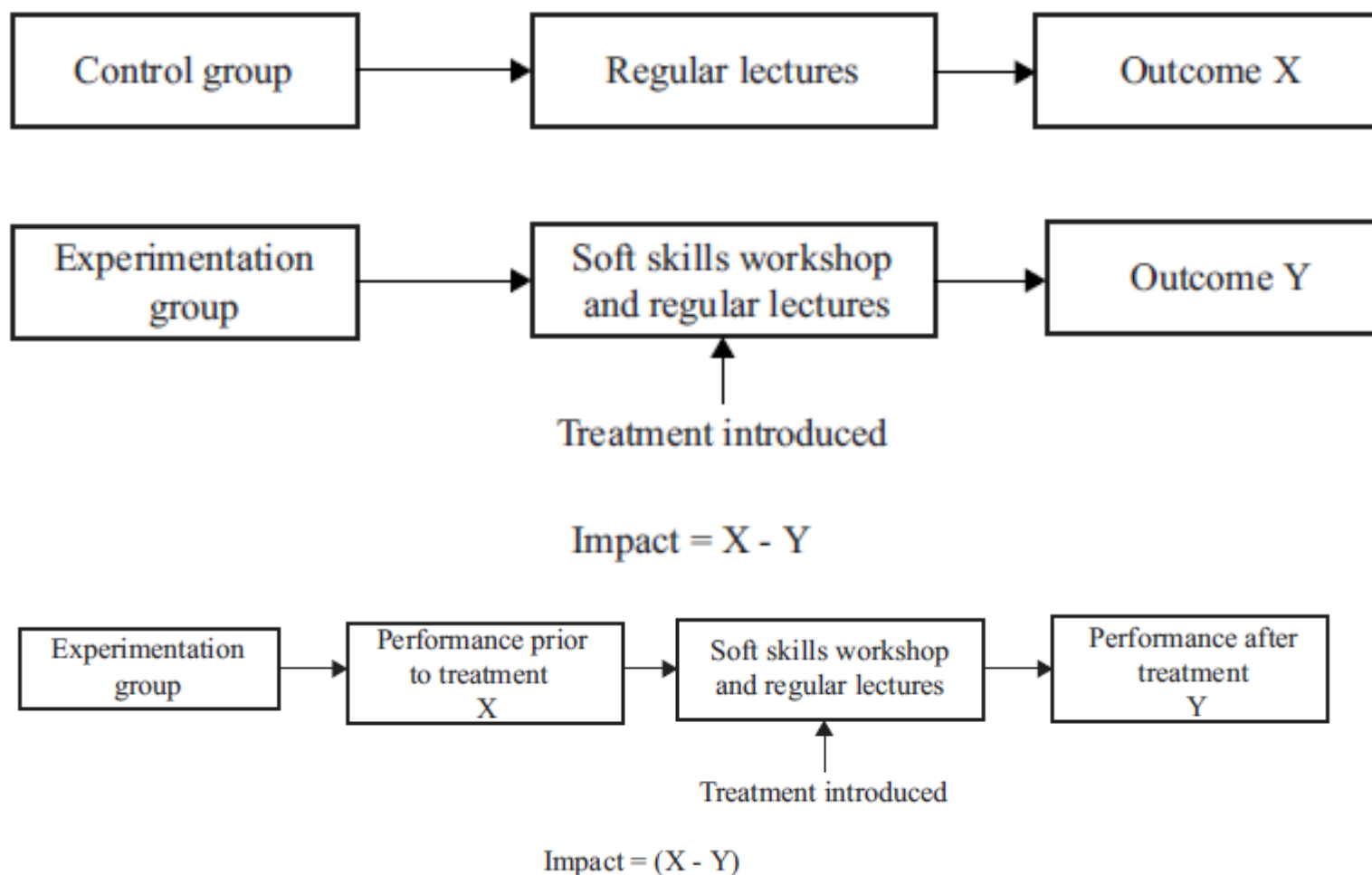
Abordaje Experimental

- **Población:** es el conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencia).
 - Normalmente es demasiado grande para poder abarcarlo.
- **Muestra:** es un subconjunto de la población al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)
 - Debería ser “representativo”
 - Esta formado por miembros “seleccionados” de la población (individuos, unidades experimentales).

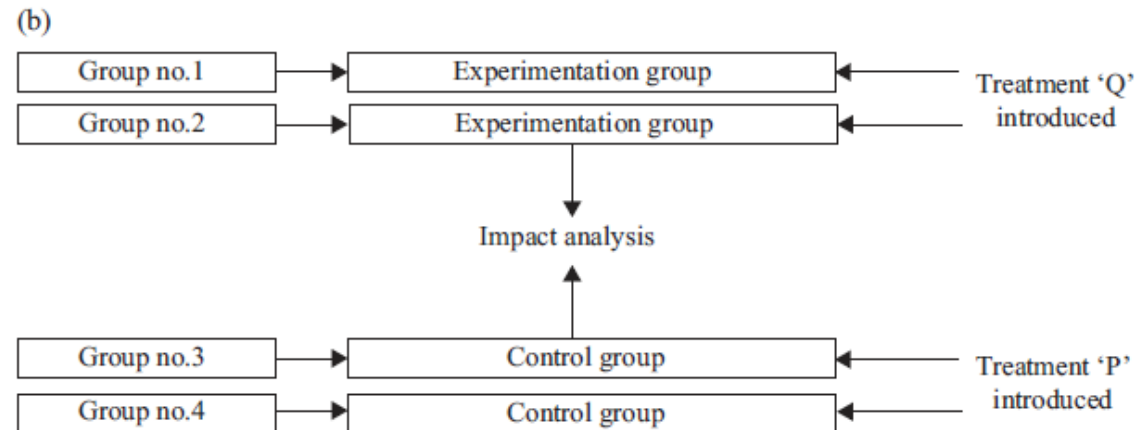
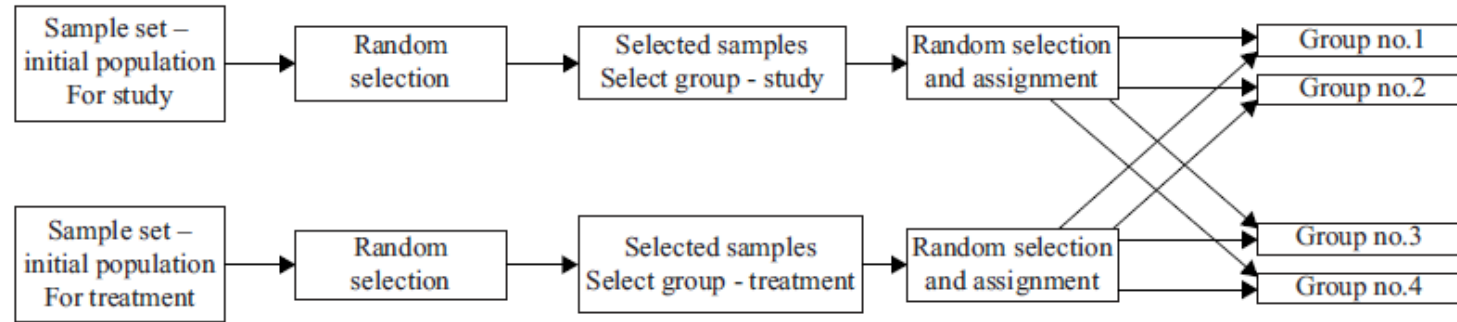


En el marco del curso "De la mano de científicos: ¿Cómo hacer investigación?"
¿Cómo se realiza un análisis de los resultados de un trabajo científico?

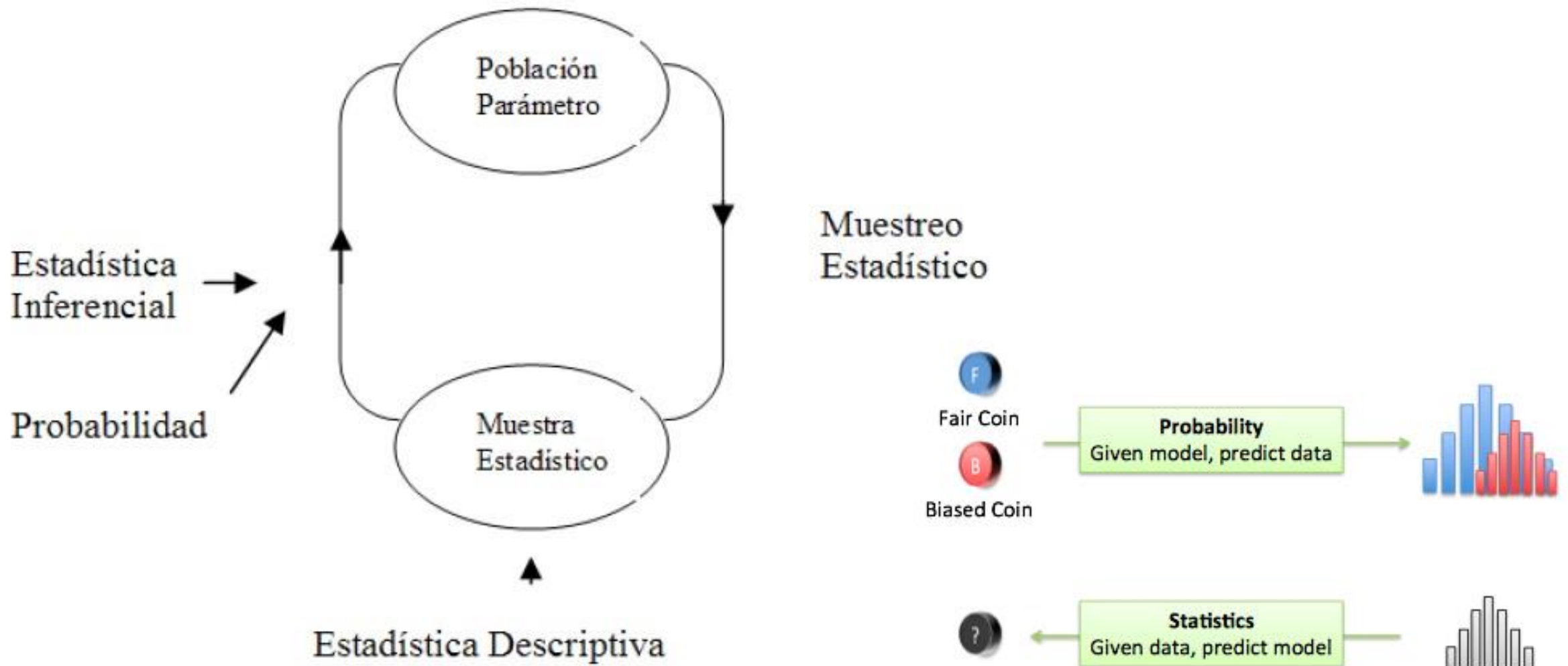
Diseño Experimental: Pre-Test Post-Test



Diseño Experimental: Aleatorización



Análisis Estadístico



Contenido: Análisis estadístico

Día 1

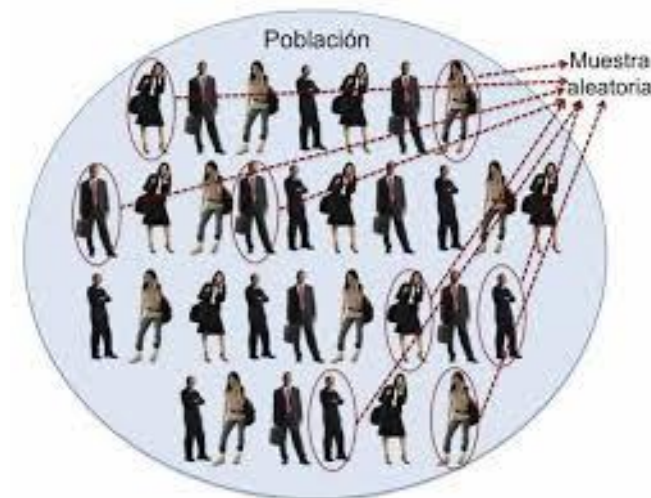
- Revisión de algunos conceptos
- **Medición/Muestreo**
- Estadística Descriptiva
- Probabilidades
 - Distribuciones de Probabilidad Discretas y Cont[inuas]

Experimento

<https://forms.gle/kXggkbiTsik7eYFr7>

Principios del Muestreo

- (i) Principio de regularidad estadística: Este principio se basa en la teoría de la probabilidad. Afirma que un número suficiente de las muestras seleccionadas al azar de la población objetivo de estudio poseen las características requeridas de la población.
 - (ii) Principio de inercia de grandes números: Este principio establece que cuanto mayor sea el número de muestras, más precisos serán los resultados sería.
- **Muestra Aleatoria:** es una muestra bien representativa de la población. Se considera que cada elemento de la población ha tenido la misma oportunidad de formar parte de la muestra. Las conclusiones basadas en una muestra aleatoria son confiables.



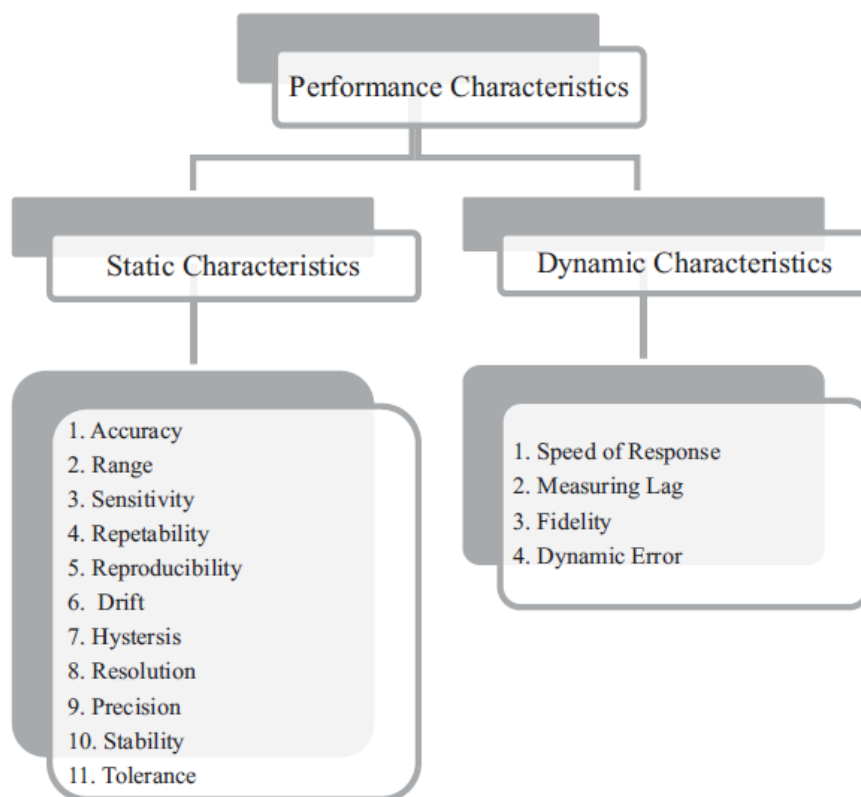
Métodos de muestreo

Aleatorio
Sistemático
Estratificado
Por grupos
Multi-etapas

No probabilísticos
Por cuotas
Conveniencia
Bola de Nieve

Para Machine learning:
Por Percentiles
Validación Cruzada
Hold Out
KFold

Proceso de Medición



RESEARCH METHODOLOGY

A Practical and Scientific Approach



EDITED BY

Vinayak Bairagi

Mousami V. Munot

Experimento

Suponga que desea saber el tamaño promedio de la familia en los Paraguay y tenemos una muestra de conveniencia

Podríamos preguntar, "¿Cuántos hijos tienes?" Pero los encuestados si son jóvenes.

Cuántos hijos tiene tu mama?



Encuesta

Intento de estimar el promedio de hijos en Paraguay, preguntando cuantos hijos tienen sus padres

*Obligatorio

Cuantos hijos (nacidos vivos) tiene tu madre ? *

Tu respuesta

Enviar

Ahora analizaremos los datos

<https://github.com/diegostaPy/cursoSociedadCientifica>



diegostaPy Update README.md 65f7844

- LICENSE Initial commit
- README.md Update README.md
- paradojaInspeccion.ipynb Create paradojaInspeccion.ipynb

README.md

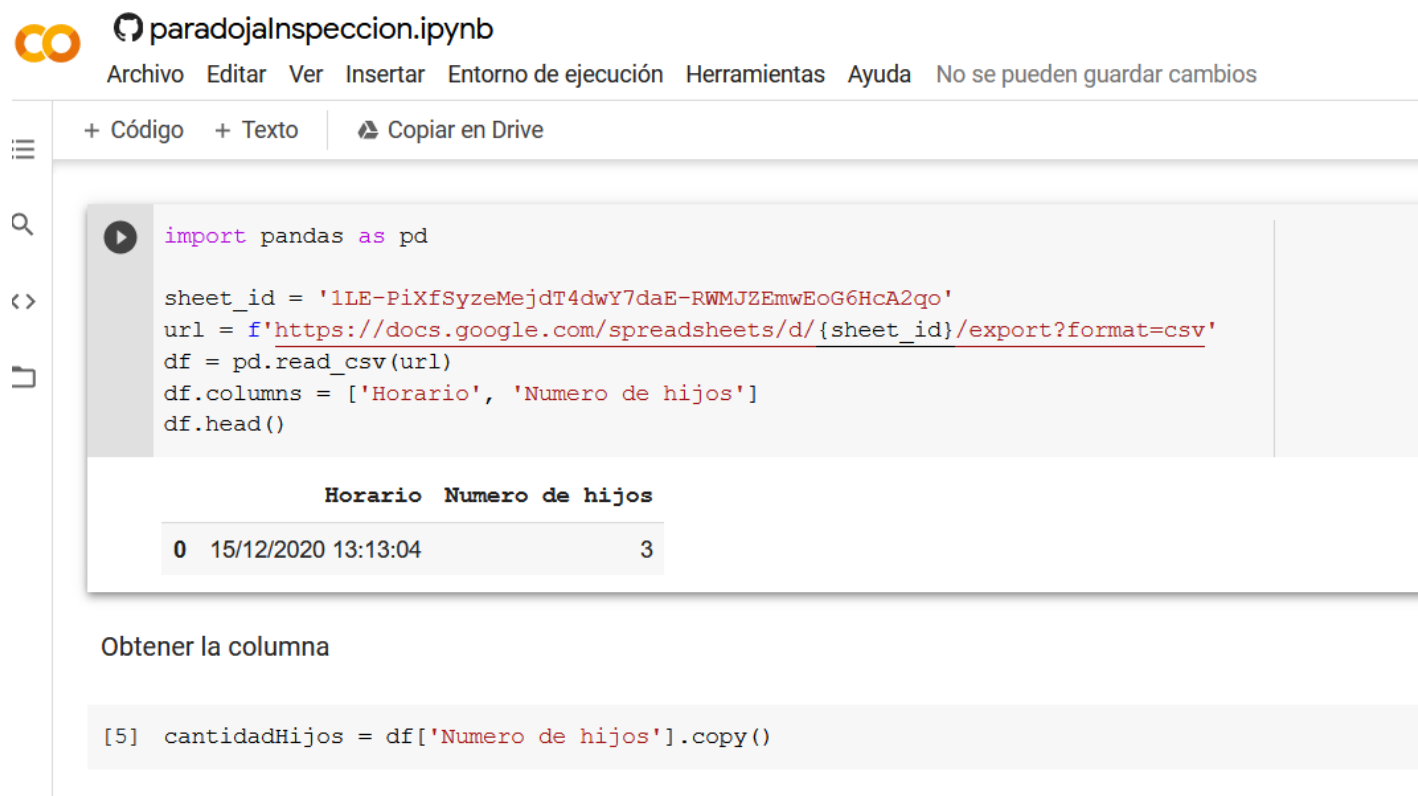
cursoSociedadCientifica

¿Cómo se realiza un análisis de los resultados de un trabajo científico?

La paradoja de la inspección

Experimento de clase:

1. [Haga Click aqui para responder al siguiente cuestionario](#)
2. [Haga un click aqui para analizar los datos](#)



paradojaInspeccion.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda No se pueden guardar cambios

+ Código + Texto Copiar en Drive

```
import pandas as pd

sheet_id = '1LE-PiXfSyzeMejdT4dwY7daE-RWMJZEmwEoG6HcA2qo'
url = f'https://docs.google.com/spreadsheets/d/{sheet_id}/export?format=csv'
df = pd.read_csv(url)
df.columns = ['Horario', 'Numero de hijos']
df.head()
```

	Horario	Numero de hijos
0	15/12/2020 13:13:04	3

Obtener la columna

```
[5] cantidadHijos = df['Numero de hijos'].copy()
```

Cuáles son los posibles problemas en la muestra?

¿Por qué la muestra podría provenir de familias numerosas?

Nivel de Educación, Ingreso Económico, Edad, etc.

La muestra conveniente puede tener muchas fuentes de bias o tendencia.???

Las familias sin hijos no están representadas

Las familias con muchos hijos están sobre representadas

En general, las familias con x hijos están sobrerrepresentadas por un factor de x .

Los procesos de muestreo sutilmente diferentes producen resultados sorprendentemente diferentes.

La paradoja del
muestreo

Cantidad de alumnos promedio

Pregunte a los profesores. Promedio = 31

Pregunte a los estudiantes. Promedio = 56

¿Quién miente?

138 clases con 1 alumno = 138 alumnos
333 clases con más de 100 estudiantes =
33,300+ Las clases grandes se
sobremuestran.
El tamaño de la clase x es
sobremuestreado por x .



PURDUE UNIVERSITY Data Digest 2013-14

Home Fast Facts Students Instruction and Student Life Faculty and Staff Diversity Finance Facilities Research

Strategic Plan Peer University Comparisons Additional Facts and Figures Regional Campuses System-wide Definitions

Distribution of Undergraduate¹ Classes by Course Level and Class Size

(for Fall 2012)

Download a PDF of this page ([Adobe Acrobat Reader](#) Required).

Course Level	1	2-9	10-19	20-29	30-39	40-49	50-99	100+	Total
000-199	38	164	659	917	241	70	99	123	2,311
200-299	82	108	370	486	307	84	109	134	1,680
Lower Level	120	272	1,029	1,403	548	154	208	257	3,991
Percent of Lower Level Total	3.0%	6.8%	25.8%	35.2%	13.7%	3.9%	5.2%	6.4%	100.0%
300-399	4	148	387	314	115	96	186	53	1,303
400-499	14	132	256	190	83	67	64	17	823
Upper Level	18	280	643	504	198	163	250	70	2,126
Percent of Upper Level Total	0.8%	13.2%	30.2%	23.7%	9.3%	7.7%	11.8%	3.3%	100.0%
500-599	0	79	102	67	43	29	23	2	345
600-699	0	4	14	5	7	8	6	4	48
800-899	0	0	0	0	0	0	0	0	0
Dual Level	0	83	116	72	50	37	29	6	393
Percent of Dual Level Total	0.0%	21.1%	29.5%	18.3%	12.7%	9.4%	7.4%	1.5%	100.0%
Total All Classes	138	635	1,788	1,979	796	354	487	333	6,510
Percent of All Classes	2.1%	9.8%	27.5%	30.4%	12.2%	5.4%	7.5%	5.1%	100.0%

¹Undergraduate¹ Classes refers to organized classes with one or more undergraduate students enrolled.



paradojalnspeccion.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda No se pueden guardar cambios

+ Código + Texto Copiar en Drive

```
import pandas as pd

sheet_id = '1LE-PiXfSyzeMejdT4dwY7daE-RWMJZEEmwEoG6HcA2qo'
url = f'https://docs.google.com/spreadsheets/d/{sheet_id}/export?format=csv'
df = pd.read_csv(url)
df.columns = ['Horario', 'Numero de hijos']
df.head()
```

Horario	Numero de hijos
0 15/12/2020 13:13:04	3

Obtener la columna

```
[5] cantidadHijos = df['Numero de hijos'].copy()
```

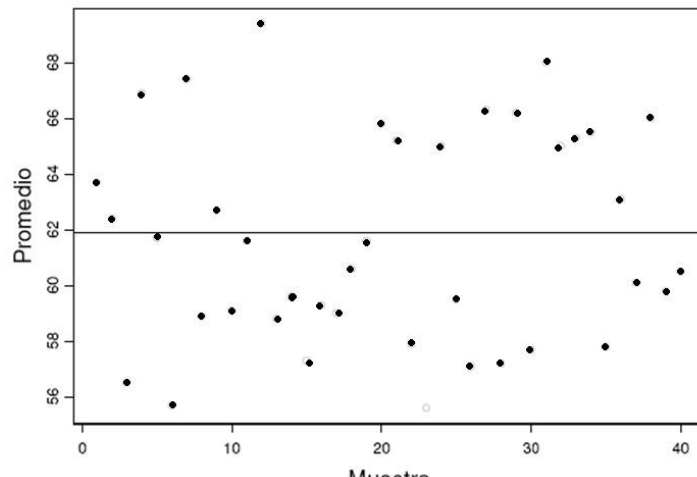
Variabilidad debido a la aleatoriedad

64	66	46	71	65	73	61
75	58	90	73	85	75	44
64	76	73	50	59	54	74
84	65	41	73	57	73	69
73	59	63	66	48	60	55
79	75	93	45	72	60	78
63	73	75	49	61	41	70
71	42	45	71	62	38	79
76	44	72	65	64	49	60
51	50	73	78	58	76	53
49	63	68	62	71	67	60
51	63	59	67	33	62	61
65	38	40	80	63	57	67
68	76	81	65	50	79	42
49	63	72	62	62	53	86
84	59	40	57	67	48	54
60	67	70	44	52	68	76
68	47	59	73	63	61	59
63	63	72	95	61	61	86
33	52	63	69	51	53	54

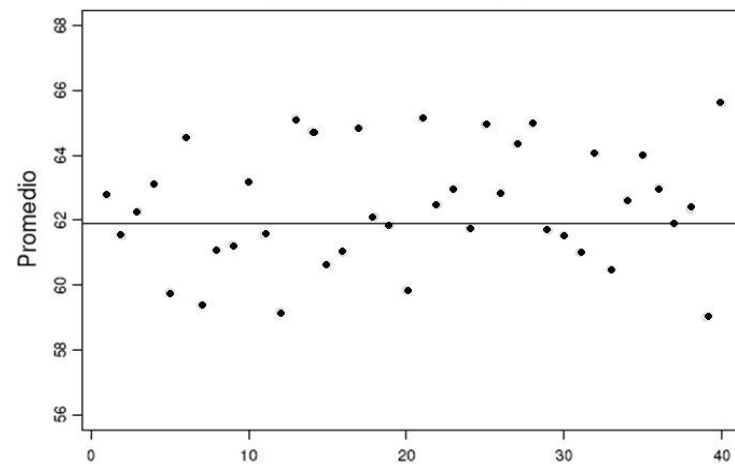
- La media de estos 350 datos es 61.9, lo que corresponde a la media poblacional
- Si calculamos el promedio de la muestra de tamaño 10, obtenemos 63.7
- Al repetir 40 veces el experimento se obtienen los siguientes resultados:

63.7	62.4	56.5	66.9	61.7	55.7	67.4	58.9	62.7	59.1
61.6	70.1	58.8	59.6	57.3	59.3	59.0	60.6	61.6	65.8
65.2	57.9	53.6	65.0	59.5	57.1	66.3	57.2	66.2	57.7
68.0	65.0	65.3	65.5	57.8	63.1	60.1	66.0	59.8	60.5

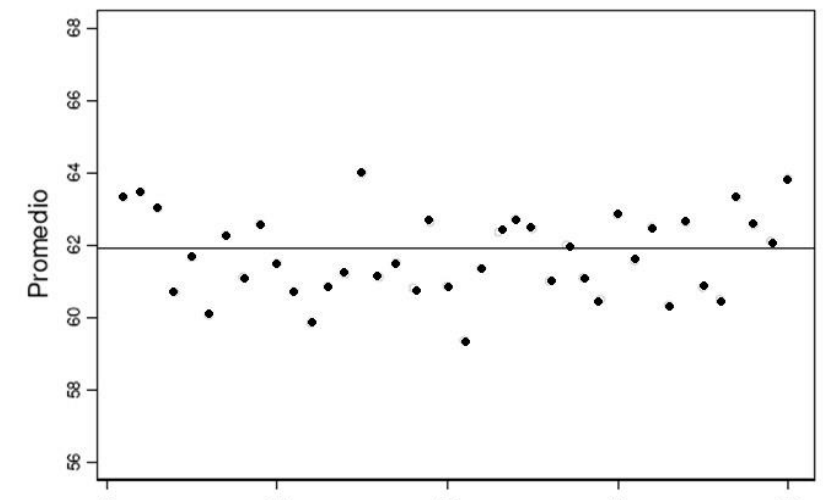
40 muestras de tamaño 10



40 muestras de tamaño 30



40 muestras de tamaño 100



En el marco del curso "De la mano de científicos:

¿Cómo se realiza un análisis de los resultados de un trabajo científico?

Pero muchas veces se hace mucho énfasis a las pruebas estadísticas en relación a la variabilidad aleatoria



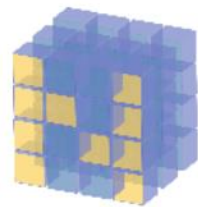
Primero es necesario medir el tamaño del efecto!!

Allen Downey - Computational Statistics - PyCon 2016

Contenido: Análisis estadístico

Día 1

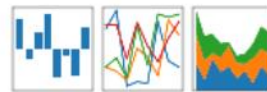
- Revisión de algunos conceptos
- Medición/Muestreo
- **Estadística Descriptiva: Analisis Exploratorio**
- Probabilidades
 - Distribuciones y variables aleatorias



NumPy

pandas

$y_t = \beta'x_{it} + \mu_i + \epsilon_{it}$



SciPy

matplotlib

Ejemplo: Tipo de variables cont.

En un programa para la detección de hipertensión en una muestra de 30 hombres en edades entre 30 y 40 años, la distribución de la presión diastólica (mínima) en mm Hg fue la siguiente:

70	85	85	75	65	90	110	95	90	70
60	75	80	120	85	95	90	70	100	65
80	90	95	90	95	110	100	85	80	75

La variable en estudio es :

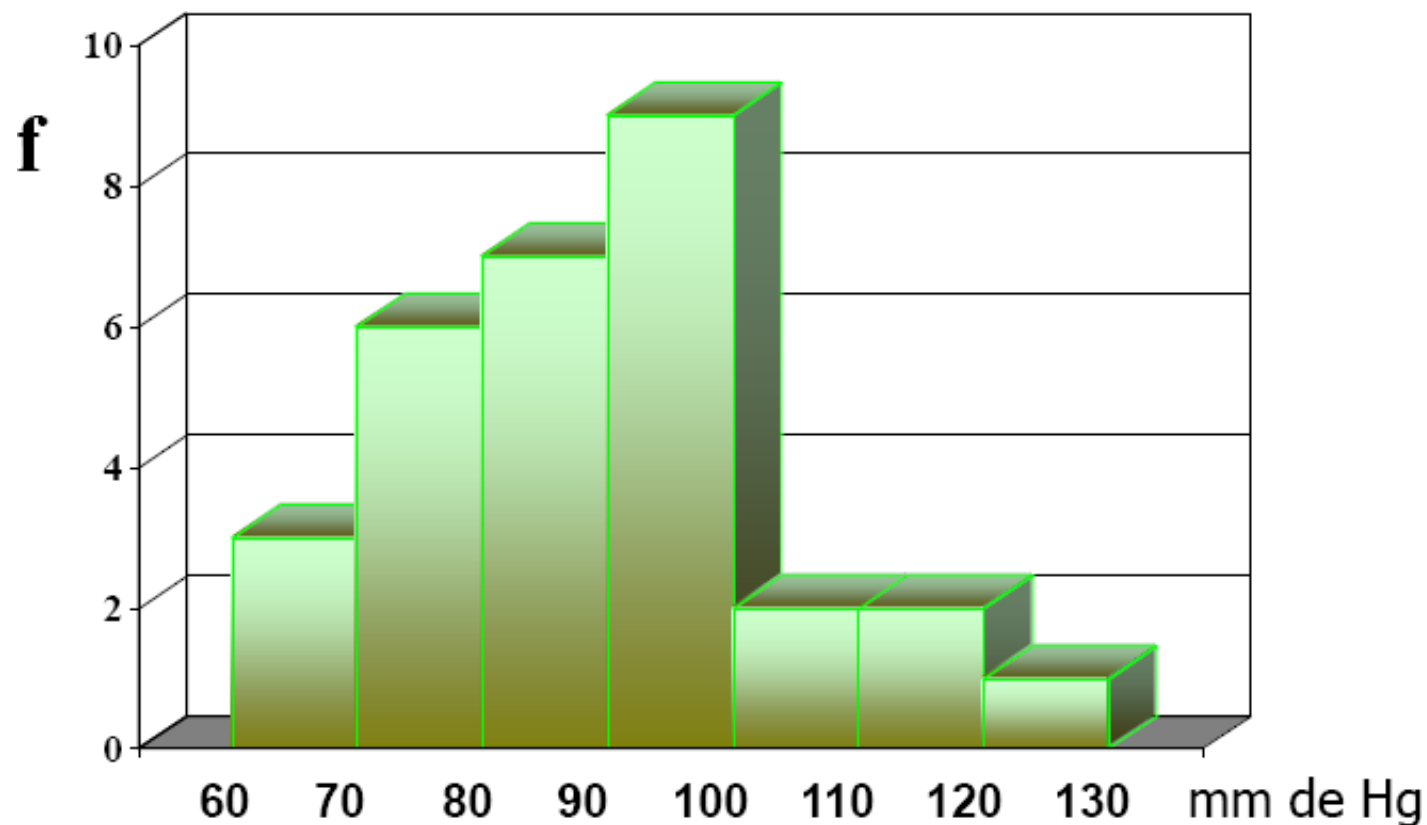
Presión diastólica (medida en mm de Hg)

una variable numérica continua.

Tabla de Frecuencias cont.

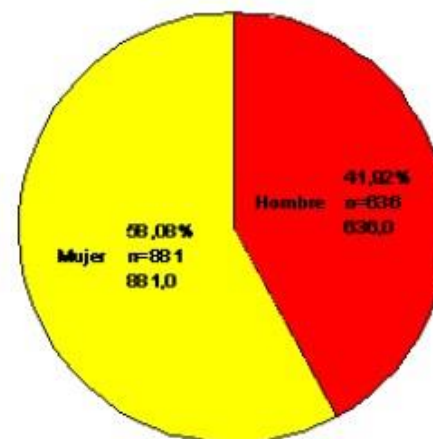
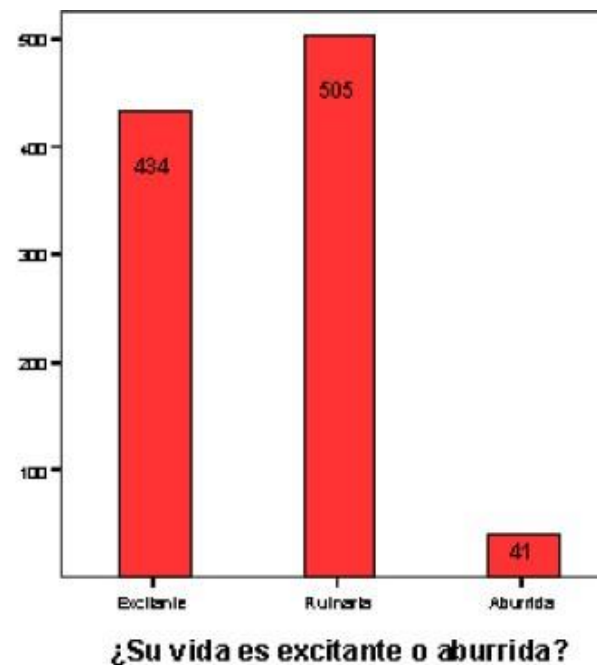
Variable	Frecuencia Absoluta	Frecuencia Absoluta Acumulada	Frecuencia Relativa	Frecuencia Relativa Acumulada
X_i	n_i	N_i	h_i	H_i
60 - 70	3	3	0.1	0.1
70 - 80	6	9	0.2	0.3
80 - 90	7	16	0.23	0.53
90 - 100	9	25	0.3	0.83
100 - 110	2	27	0.07	0.90
110 - 120	2	29	0.07	0.97
120 - 130	1	30	0.03	1.00
total	30		1.0	

Histograma de la distribución de presión diastólica en mm de Hg según las frecuencias absolutas:



Gráficos para variables cualitativas

- **Diagramas de barras**
 - Alturas proporcionales a las frecuencias (abs. o rel.)
 - Se pueden aplicar también a variables discretas
- **Diagramas de sectores (tartas, polares)**
 - El área de cada sector es proporcional a su frecuencia (abs. o rel.)

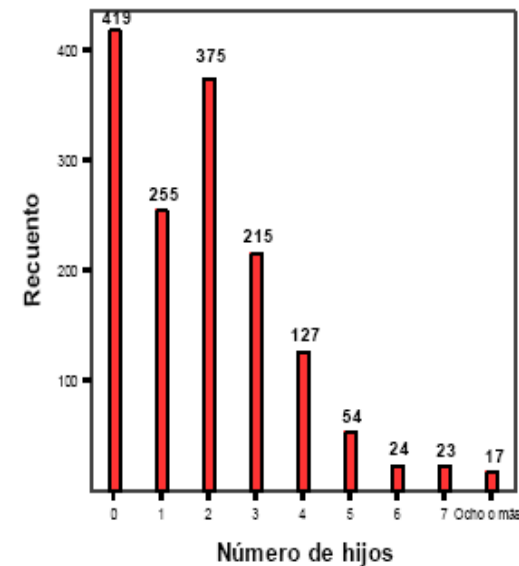


Gráficos diferenciales para variables numéricas

Son diferentes en función de que las variables sean **discretas** o **continuas**.
Valen con frec. absolutas o relativas.

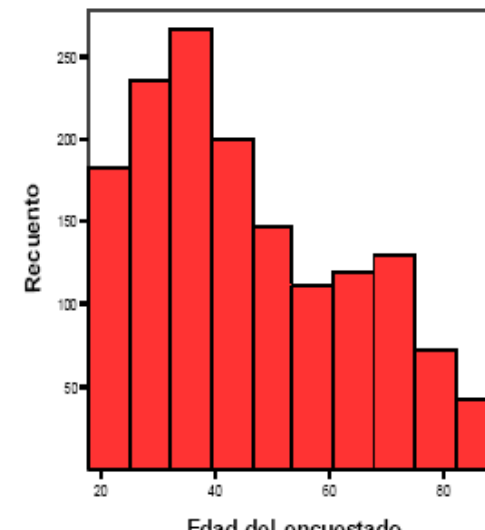
– Diagramas barras para variables discretas

- Se deja un espacio entre barras para indicar los valores que no son posibles

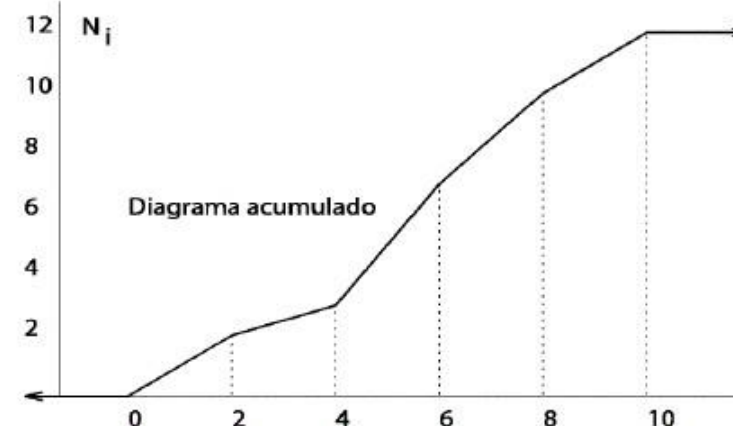
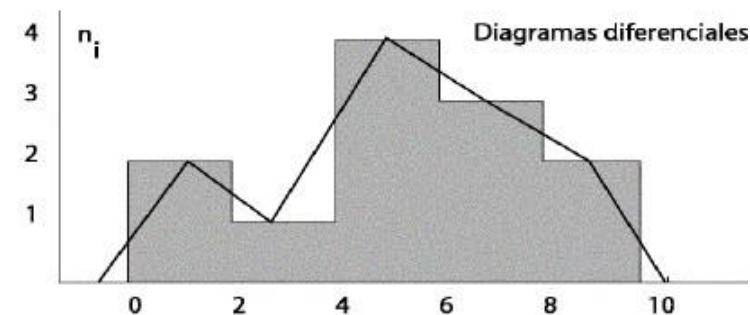
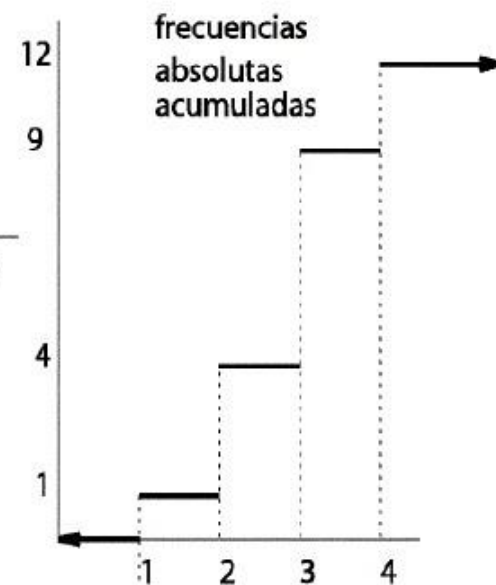
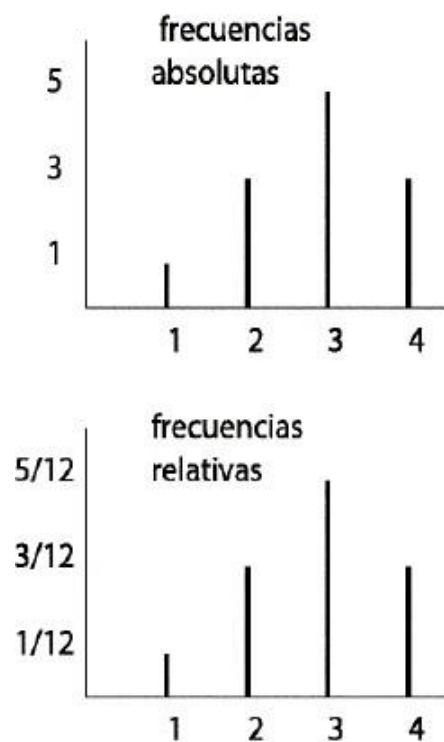


– Histogramas para v. continuas

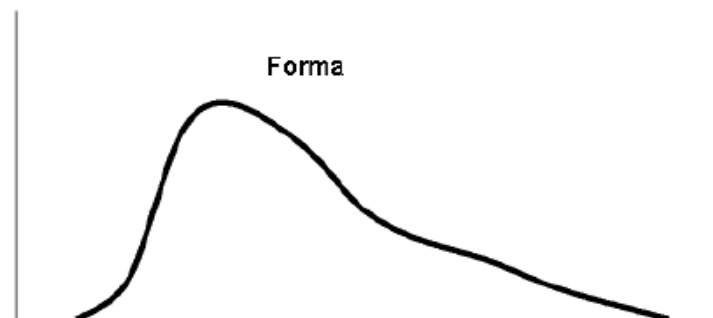
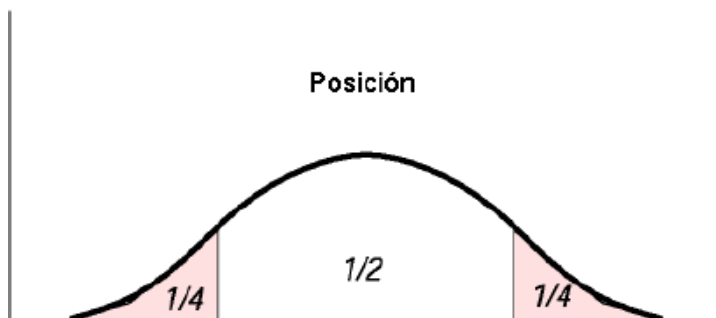
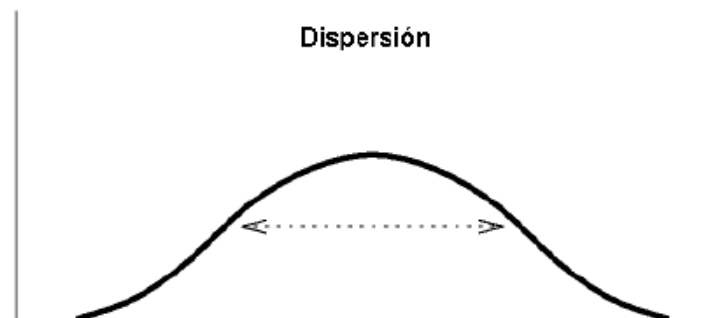
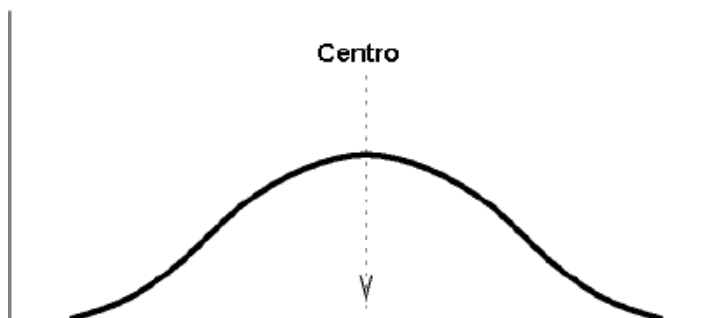
- El área que hay bajo del histograma entre dos puntos cualesquiera indica la cantidad (porcentaje o frecuencia) de individuos en el intervalo.



- Cada uno de los anteriores diagramas tiene su correspondiente **diagrama integral**. Se realizan a partir de las **frecuencias acumuladas**. Indican, para cada valor de la variable, **la cantidad (frecuencia) de individuos que poseen un valor inferior o igual al mismo**.



Estadísticos de forma intuitiva



Estadísticos

- **Posición (Basados en el orden)**
 - Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
 - Cuantiles, percentiles, cuartiles, deciles,...
- **Centralización**
 - Indican valores con respecto a los que los datos parecen agruparse.
 - Media, mediana y moda
- **Dispersión**
 - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
 - Desviación estándar, coeficiente de variación, rango, varianza
- **Forma**
 - Asimetría
 - Apuntamiento o curtosis

Centralización

- Añaden unos cuantos casos particulares a las medidas de posición. Son medidas que buscan posiciones (valores) con respecto a los que los datos muestran tendencia a agruparse.
- **Media:** es la media aritmética (promedio) de los valores de una variable. Suma de los valores dividido por el tamaño muestral.
 - Media de $\{2, 2, 3, 7\}$ es $(2+2+3+7)/4 = 3.5$
 - Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos.
 - Centro de gravedad de los datos.



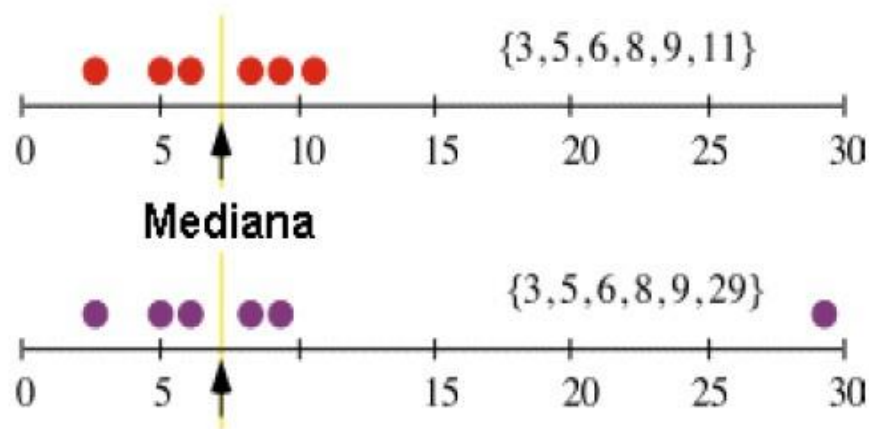
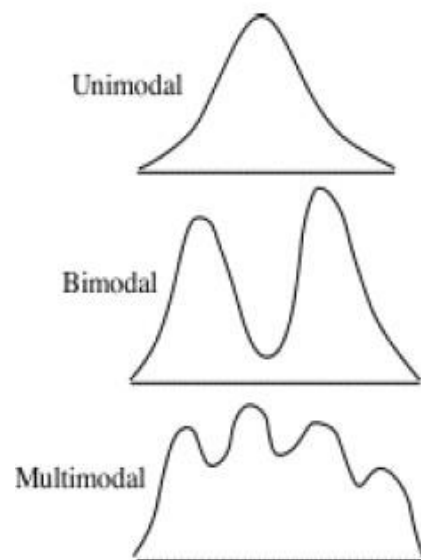
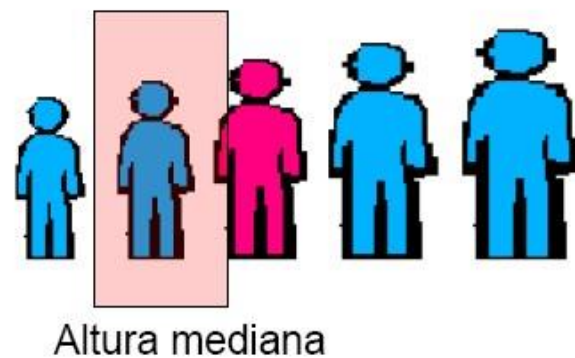
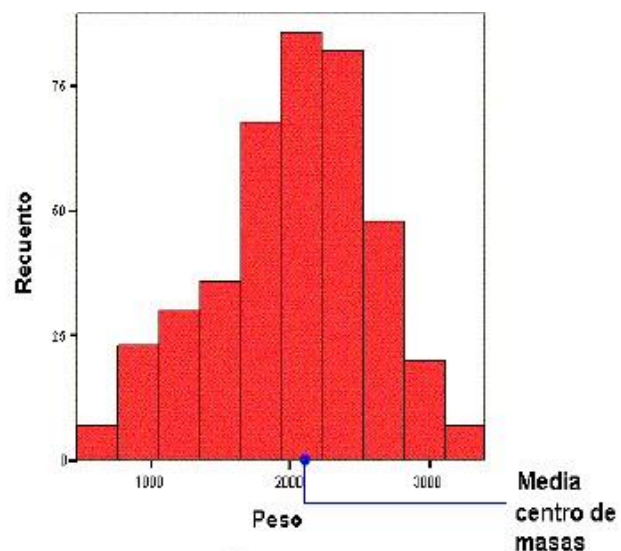
Centralización

- **Mediana:** es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
 - Mediana de 1, 2, 4, **5**, 6, 6, 8 es 5
 - Mediana de 1, 2, 4, **5**, **6**, 6, 8, 9 es $(5+6)/2 = 5.5$
 - Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos.
 - Mediana de 1, 2, 4, **5**, 6, 6, 800 es 5. ¡La media es 117.7!
- **Moda:** es el/los valor/es donde la distribución de frecuencia alcanza un máximo.



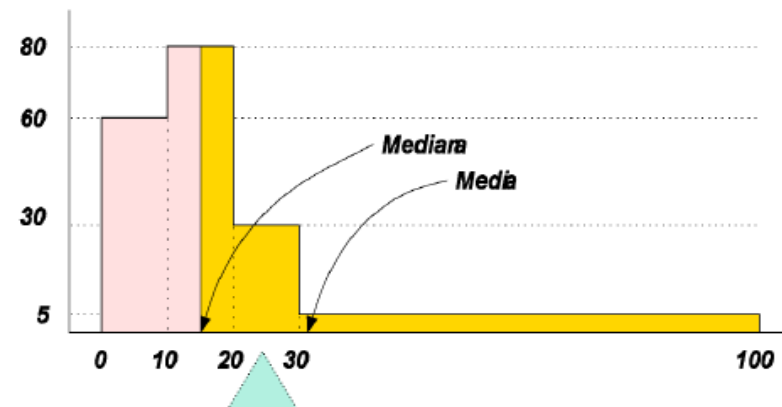
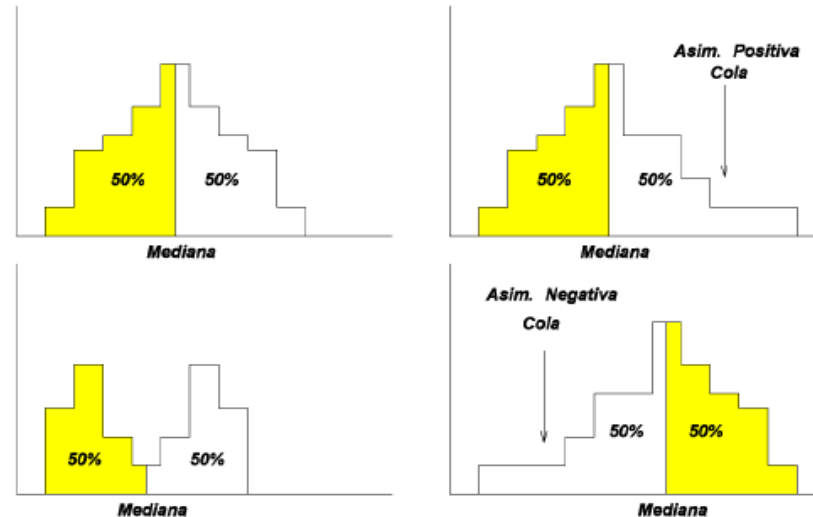
En el marco del curso "De la mano de científicos:

¿Cómo se realiza un análisis de los resultados de un trabajo científico?



Asimetría o sesgo

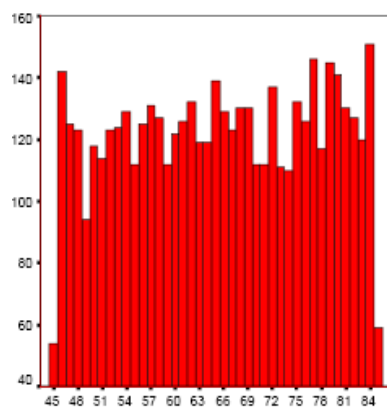
- Una distribución es simétrica si la mitad izquierda de su distribución es la imagen especular de su mitad derecha.
- En las distribuciones simétricas media y mediana coinciden. Si sólo hay una moda también coincide.
- La asimetría es positiva o negativa en función de a qué lado se encuentra la cola de la distribución.
- La media tiende a desplazarse hacia los valores extremos (colas).
- Las discrepancias entre las medidas de centralización son indicación de asimetría.



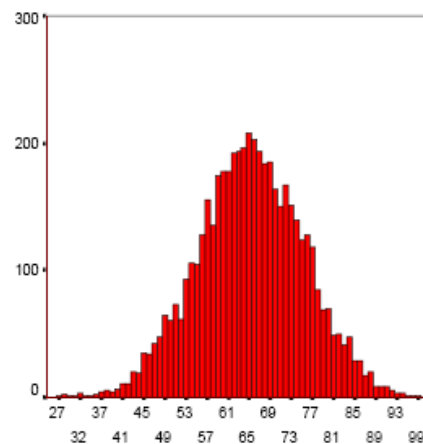
Apuntamiento o curtosis (kurtosis)

- La **curtosis** nos indica el grado de apuntamiento (aplastamiento) de una distribución con respecto a la distribución normal o gaussiana. Es adimensional.
- **Platicúrtica**: $\text{curtosis} < 0$
- **Mesocúrtica**: $\text{curtosis} = 0$
- **Leptocúrtica**: $\text{curtosis} > 0$

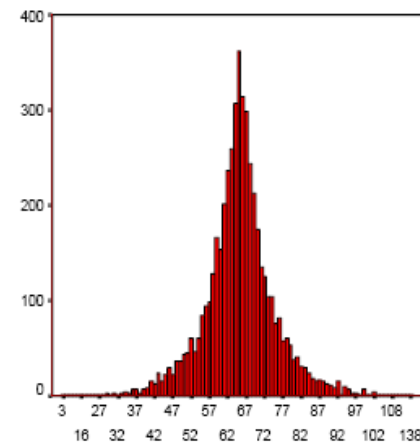
Los gráficos poseen la misma media y desviación típica, pero diferente grado de apuntamiento o curtosis.



Platicúrtica



Mesocúrtica



Leptocúrtica

¿Cómo se realiza un análisis de los resultados de un trabajo científico?

Medidas de dispersión

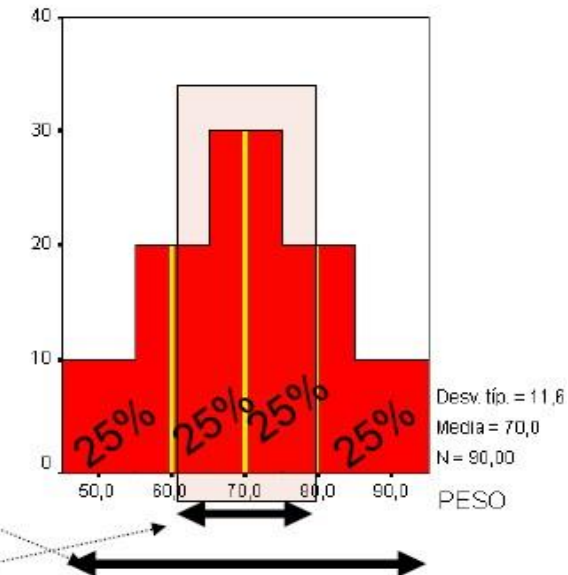
- Miden el grado de dispersión (variabilidad) de los datos, independientemente de su causa.

- **Amplitud o Rango** ('range'): La diferencia entre las observaciones extremas.

- 2,1,4,3,8,4. El rango es $8-1=7$
- Es muy sensible a los valores extremos.

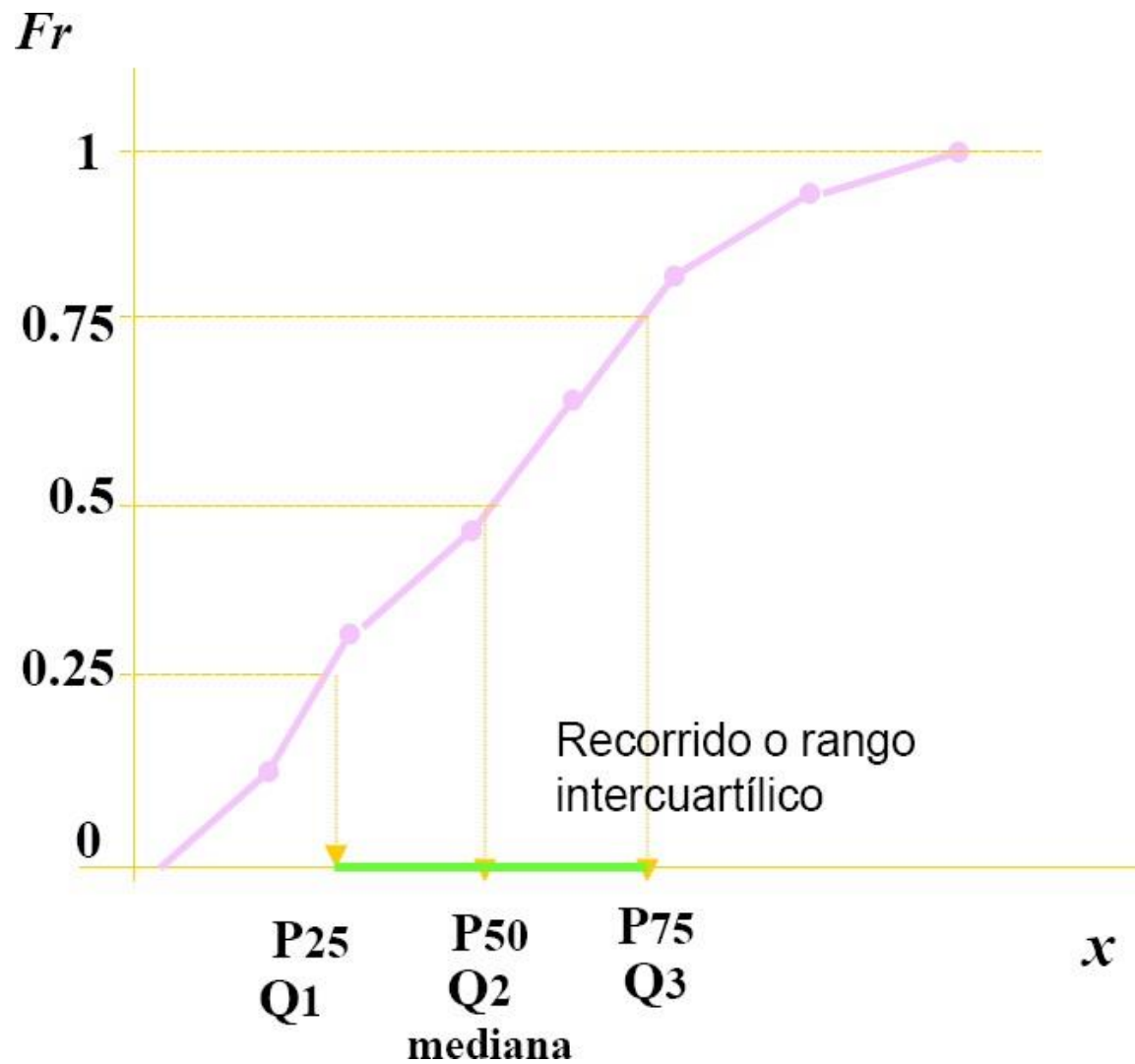
- **Rango intercuartílico** ('interquartile range'):

- Es la distancia entre el primer y tercer cuartil.
 - Rango intercuartílico = $P_{75} - P_{25}$
- Parecida al rango, pero eliminando las observaciones más extremas inferiores y superiores.
- No es tan sensible a valores extremos.



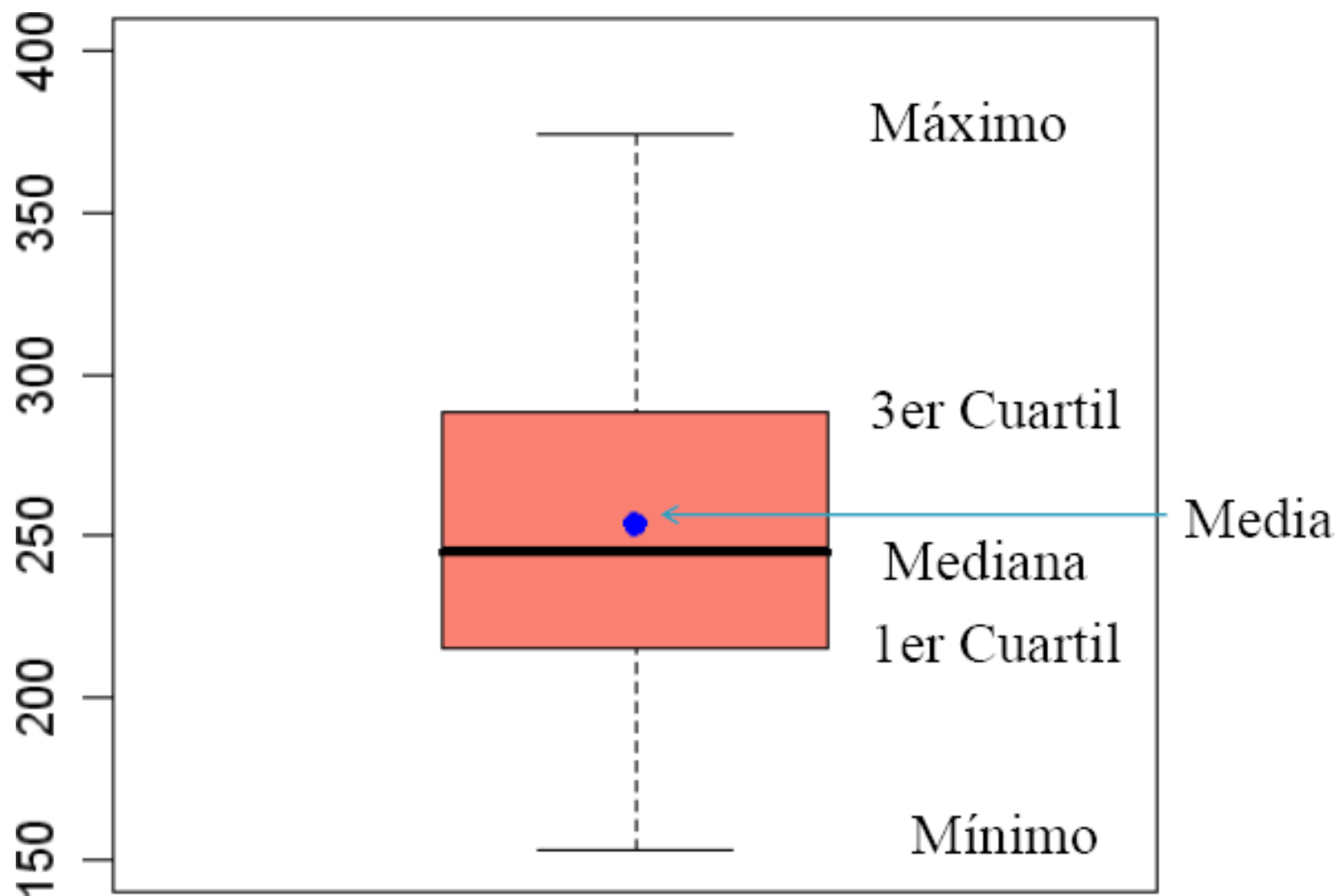
En el marco del curso "De la mano de científicos:

¿Cómo se realiza un análisis de los resultados de un trabajo científico?



En el marco del curso "De la mano de científicos:
¿Cómo se realiza un análisis de los resultados de un trabajo científico?

Box-plot



Análisis Exploratorio

<https://github.com/sborquez/Python-LEC/>

- Antes de Comenzar
- Análisis Exploratorio de Datos
 - ¿Por qué es importante?
 - Manipulación de datos con Pandas
 - ¿Qué Gráfico debería usar?
- Caso de estudio: Migraciones en Chile
 - Carga de Datos
 - Conociendo el Dataframe
 - Realizar Consultas
 - Operaciones sobre el DataFrame
 - Agrupar datos
 - Visualizaciones Básicas
- Caso de estudio: Pokemon Dataset
 - Estadísticas Básicas
 - Operaciones y Comparaciones entre Columnas
 - Visualización Estadística de Datos
- Caso de estudio: SARS-CoV-2 Total Cases Dataset Pronto...
 - How to combine data from multiple tables?
 - How to handle time series data with ease?
 - Log Scales
 - Gráficos Interactivos con Plotly

Análisis Exploratorio

visas.head(5)

	SEXO	NACIMIENTO	ACTIVIDAD	PROFESION	PAIS	ESTUDIOS	COMUNA	PROVINCIA	REGION	TIT_DEP	AÑO	BENEFICIO
0	Femenino	1974-10-05	EMPLEADO	MATRONA	PERÚ	no indica	SANTIAGO	SANTIAGO	METROPOLITANA	T	2006	PERMANENC DEFINITIVA
1	Masculino	1949-09-13	EMPLEADO	INGENIERO	ECUADOR	no indica	PROVIDENCIA	SANTIAGO	METROPOLITANA	T	2007	PERMANENC DEFINITIVA
2	Femenino	1949-12-07	EMPLEADO	ASESORA DEL HOGAR	BOLIVIA	BASICO	ARICA	ARICA	ARICA Y PARINACOTA	T	2007	PERMANENC DEFINITIVA
3	Femenino	1966-09-20	DUEÑA DE CASA	DUEÑA DE CASA	BOLIVIA	MEDIO	ARICA	ARICA	ARICA Y PARINACOTA	T	2006	PERMANENC DEFINITIVA
4	Masculino	1981-08-15	EMPRESARIO O PATRON	COMERCIANTE	BRASIL	no indica	LAS CONDES	SANTIAGO	METROPOLITANA	T	2008	PERMANENC DEFINITIVA

Tipos de datos:

- Categóricos: Sexo, Actividad, Profesión, Estudios ...
- Discretos: Año, Nacimiento(Fecha, dato compuesto)
- Continuos: No hay



Valores por columna

A continuación, se hará una exploración de las columnas.

```
# Al iterar un df, se obtienen los nombres de las columnas
for columna in visas:

    # Una forma de accesos a las columnas es como usarlo como un diccionario
    datos_columna = visas[columna]

    # Cantidad de valores unicos
    distintos = datos_columna.nunique()

    print(f"La columna {columna} tiene {distintos} valores diferentes.")
```

```
La columna SEXO tiene 2 valores diferentes.
La columna NACIMIENTO tiene 27688 valores diferentes.
La columna ACTIVIDAD tiene 15 valores diferentes.
La columna PROFESION tiene 599 valores diferentes.
La columna PAIS tiene 164 valores diferentes.
La columna ESTUDIOS tiene 8 valores diferentes.
La columna COMUNA tiene 351 valores diferentes.
La columna PROVINCIA tiene 55 valores diferentes.
La columna REGION tiene 15 valores diferentes.
La columna TIT DEP tiene 2 valores diferentes.
La columna AÑO tiene 12 valores diferentes.
La columna BENEFICIO tiene 2 valores diferentes.
```


Análisis Exploratorio

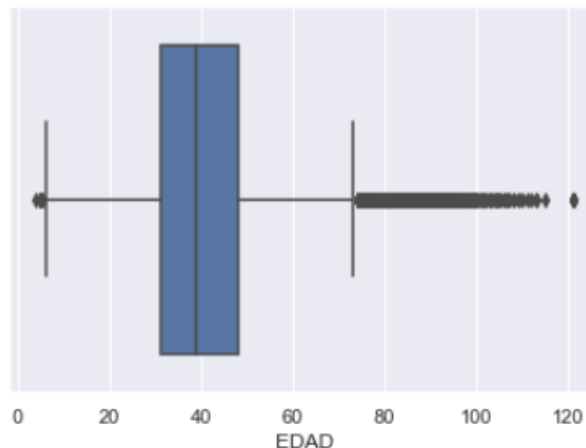
Variable Numérica Discreta

	AÑO	EDAD
count	324932.000000	324931.000000
mean	2011.956988	39.448908
std	3.314549	14.184114
min	2005.000000	4.000000
25%	2009.000000	31.000000
50%	2013.000000	39.000000
75%	2015.000000	48.000000
max	2016.000000	121.000000

Columna Calculada a partir de la fecha de nacimiento

```
import seaborn as sns
```

```
ax = sns.boxplot(x=visas["EDAD"])
```

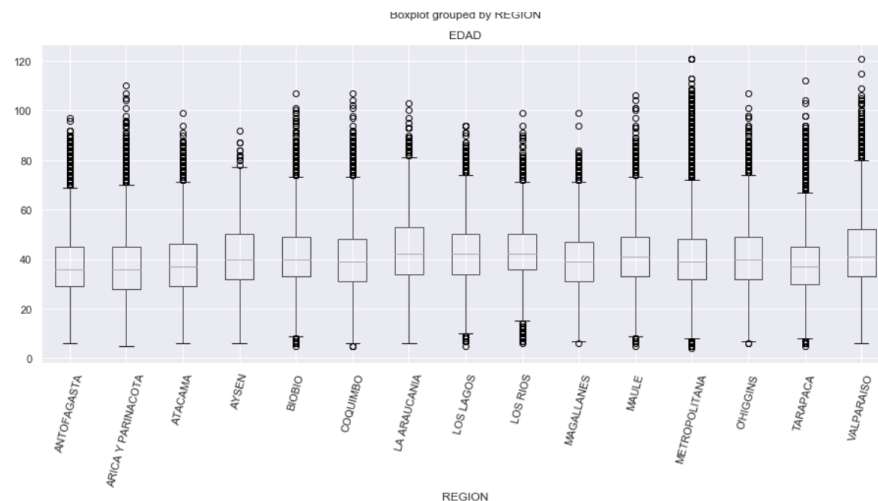
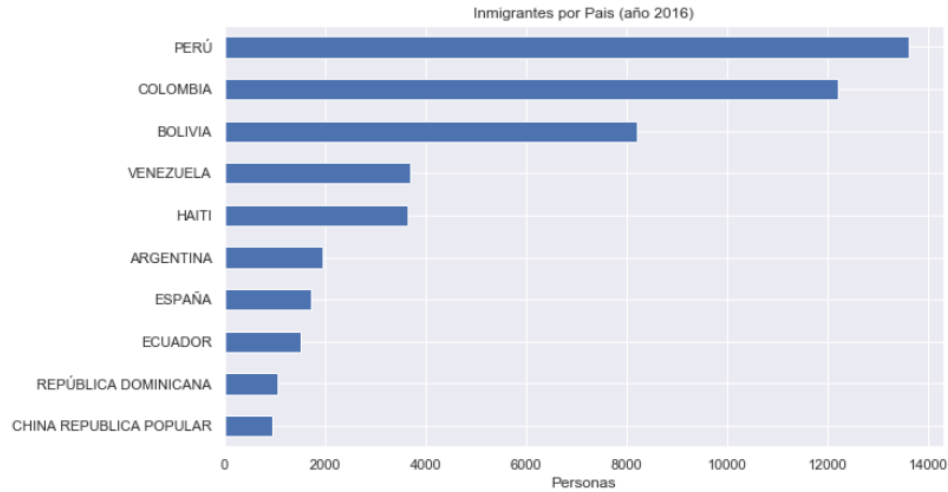


En el marco del curso "De la mano de científicos:

¿Cómo se realiza un análisis de los resultados de un trabajo científico?

Visualizaciones Básicas

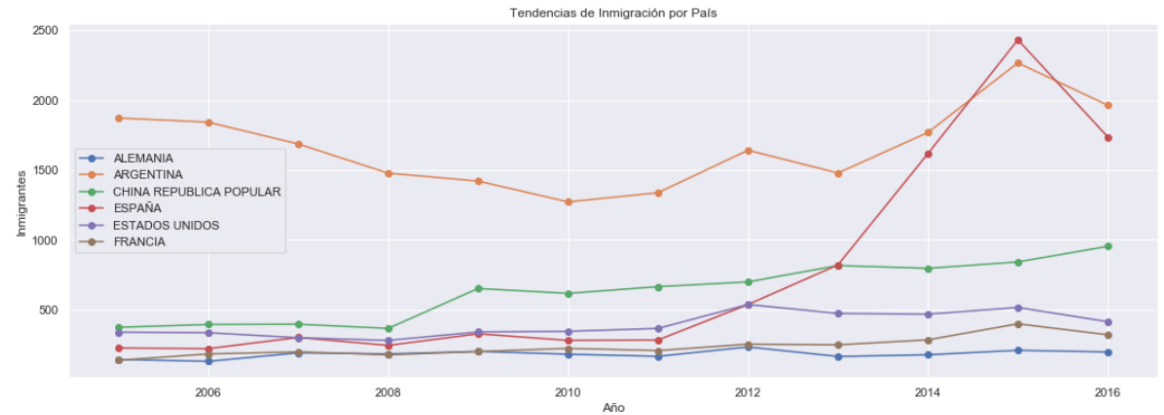
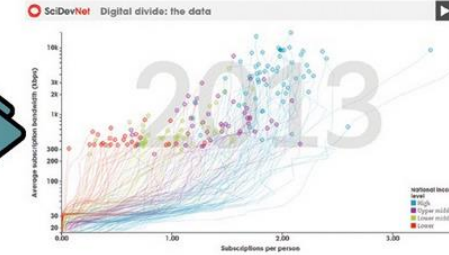
Análisis Exploratorio



Raw unprocessed data

AI	AK	AL	AM	AN	AO	AP	AQ
Income O	OECD me	Fixed UL	1988	1987	1988	1989	1990
Upper mi	non-OECD	Albania	310,126	320,359	330,609	340,991	350,51
Upper mi	non-OECD	Algeria	164,270	183,302	200,131	223,242	251,91
High inco	non-OECD	Andorra	415,111	515,184	561,080	578,344	604,24
Upper mi	non-OECD	Angola	87,356	107,357	128,502	156,822	190,84
High inco	non-OECD	Antigua and Barb	525,647	606,130	678,664	741,901	854,15
Upper mi	non-OECD	Argentina	746,074	788,347	869,615	978,111	1,088,000
Lower mi	non-OECD	Australia	564,449	621,981	684,924	747,534	817,71
High inco	non-OECD	Azerbaijan	115,784	130,314	136,818	145,805	154,00
Upper mi	non-OECD	Bahamas	115,784	130,314	136,818	145,805	154,00
High inco	non-OECD	Bahrain	115,784	130,314	136,818	145,805	154,00
Upper mi	non-OECD	Bangladesh	115,784	130,314	136,818	145,805	154,00
High inco	non-OECD	Barbados	115,784	130,314	136,818	145,805	154,00
Upper mi	non-OECD	Belarus	115,784	130,314	136,818	145,805	154,00
High inco	non-OECD	Belgium	115,784	130,314	136,818	145,805	154,00
Upper mi	non-OECD	Belize	115,784	130,314	136,818	145,805	154,00
High inco	non-OECD	Benin	115,784	130,314	136,818	145,805	154,00

Information (trends and patterns within the data)



Contenido: Análisis estadístico

Día 1

- Revisión de algunos conceptos
- Medición/Muestreo
- Estadística Descriptiva: Analisis Exploratorio
- **Probabilidades**
 - **Distribuciones y variables aleatorias**













Variable Aleatoria Discreta

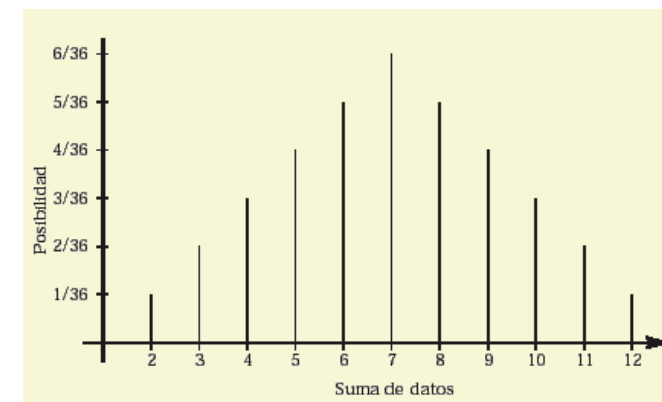
Lanzar dos dados

Variable aleatoria Valores posibles Eventos aleatorios

$$X = \begin{cases} 0 \\ 1 \end{cases}$$

Lanzar una moneda

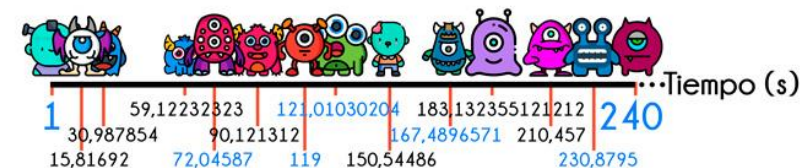
						
	(1 1)	(1 2)	(1 3)	(1 4)	(1 5)	(1 6)
	(2 1)	(2 2)	(2 3)	(2 4)	(2 5)	(2 6)
	(3 1)	(3 2)	(3 3)	(3 4)	(3 5)	(3 6)
	(4 1)	(4 2)	(4 3)	(4 4)	(4 5)	(4 6)
	(5 1)	(5 2)	(5 3)	(5 4)	(5 5)	(5 6)
	(6 1)	(6 2)	(6 3)	(6 4)	(6 5)	(6 6)



Variable aleatoria continua

Una variable aleatoria continua, es aquella que puede asumir un número incontable de valores.

Ejemplo: si vamos a una agencia del banco y registramos los datos de atención a los clientes, podemos definir la variable aleatoria D:

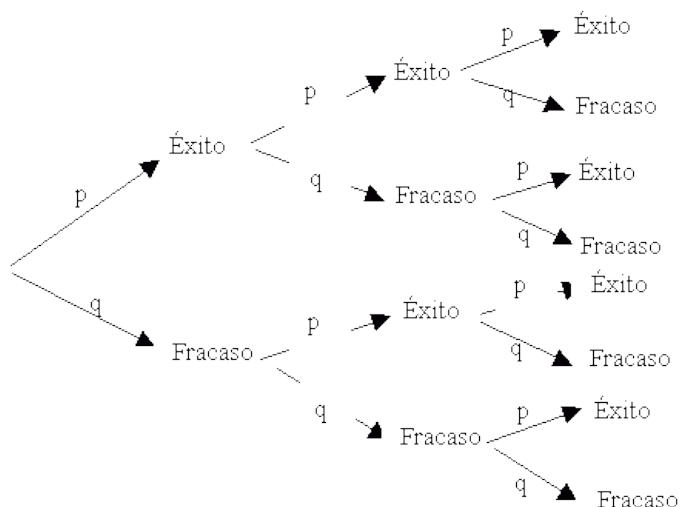


>> D = tiempo de atención en ventanilla (en s).

↪ $R_D : 1 \leq d \leq 240$

Distribuciones de Probabilidades Discretas

Binomial

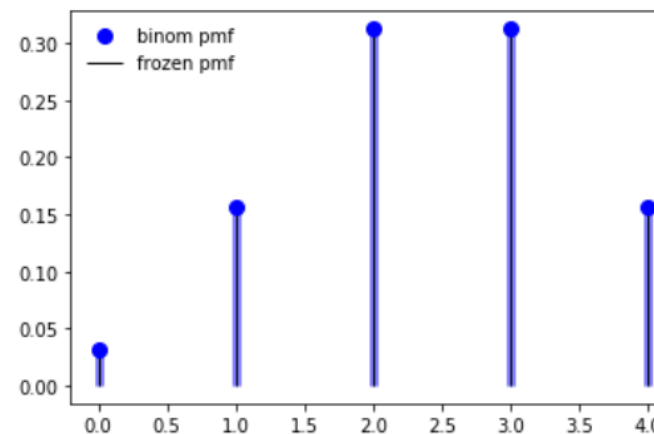


Ejemplo: Lanzar 3 veces una monedas y contar cuantas veces sale cara

Las binomial, para predecir la variable lanzaR 5 veces una moneda

```

|: from scipy.stats import binom
n, p = 5, 0.5
fig, ax = plt.subplots(1, 1)
x = np.arange(binom.ppf(0.01, n, p),
              binom.ppf(0.99, n, p))
ax.plot(x, binom.pmf(x, n, p), 'bo', ms=8, label='binom pmf')
ax.vlines(x, 0, binom.pmf(x, n, p), colors='b', lw=5, alpha=0.5)
rv = binom(n, p)
ax.vlines(x, 0, rv.pmf(x), colors='k', linestyle='--', lw=1,
          label='frozen pmf')
ax.legend(loc='best', frameon=False)
plt.show()
  
```



Distribuciones de Probabilidades Discretas

La distribución binomial tiende a una distribución de Poisson cuando en una distribución binomial se realiza el experimento muchas veces, la muestra n es grande y la probabilidad de éxito p en cada ensayo es baja, es aquí donde aplica el modelo de distribución de Poisson. Se tiene que cumplir que: $p < 0.10$ $p * n < 10$

La probabilidad de que haya un accidente en una compañía de manufactura es de 0.02 por cada día de trabajo. Si se trabajan 300 días al año, ¿cuál es la probabilidad de tener 3 accidentes? Como la probabilidad p es menor que 0.1, y el producto $n * p$ es menor que 10 ($300 * 0.02 = 6$), entonces, aplicamos el modelo de distribución de Poisson:

- $\text{Poisson}(\lambda = 6)$: `stats.poisson(mu=6)`

```

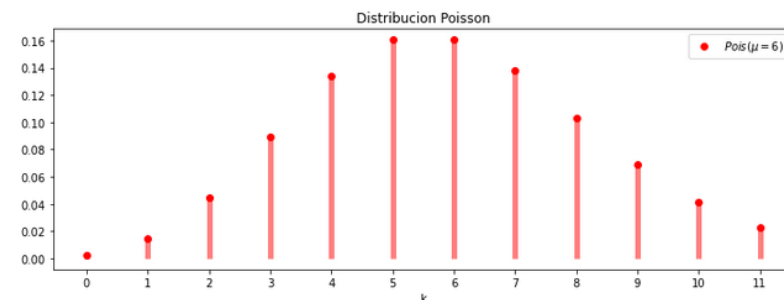
# Instanciar Distribuciones
Pois = stats.poisson(mu=6)

# Generar figura
plt.figure(figsize=(12,4))

# Generar puntos
k = np.arange(12)

# Generar probabilidades para Poisson
plt.plot(k, Pois.pmf(k), "ro", label="$Pois(\mu=6)$")
plt.vlines(k, 0, Pois.pmf(k), colors='r', lw=5, alpha=0.5)

# Agregar estilo
plt.title("Distribucion Poisson")
plt.xlabel("k")
plt.xticks(k)
plt.legend()
plt.show();
  
```



Poisson Distribution Formula

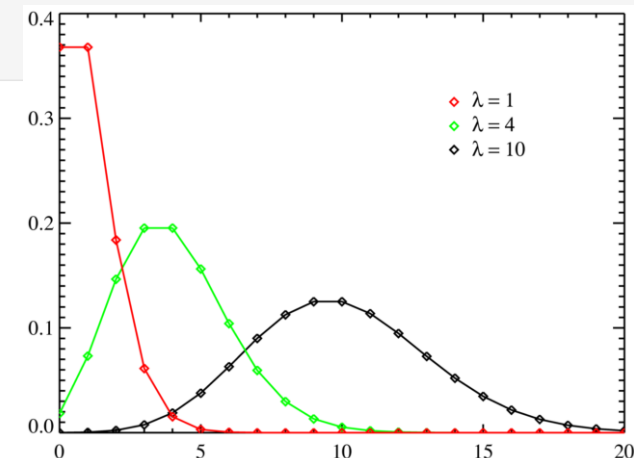
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

$x = 0, 1, 2, 3, \dots$

λ = mean number of occurrences in the interval

e = Euler's constant ≈ 2.71828



Distribución de Probabilidades Continuas

- Normal($\mu = 5, \sigma = 2$): `stats.norm(loc=5, scale=2)`

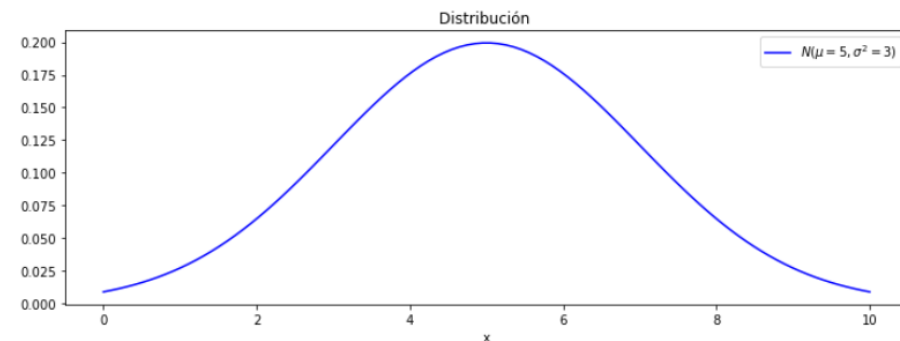
```
# Instanciar Distribucion
N = stats.norm(loc=5, scale=2)

# Generar figura
plt.figure(figsize=(12,4))

# Generar puntos
x = np.linspace(0,10, 100)

# Generar probabilidades para Normal
plt.plot(x, N.pdf(x), "b", label="$N(\mu=5, \sigma^2=3)$")

# Agregar estilo
plt.title("Distribución ")
plt.xlabel("x")
plt.legend()
plt.show()
```



- Gamma($\alpha = 9, \beta = 2$): `stats.gamma(a=9, scale=(1/2))`

```
# Instanciar Distribucion
Ga = stats.gamma(a=9, scale=0.5)

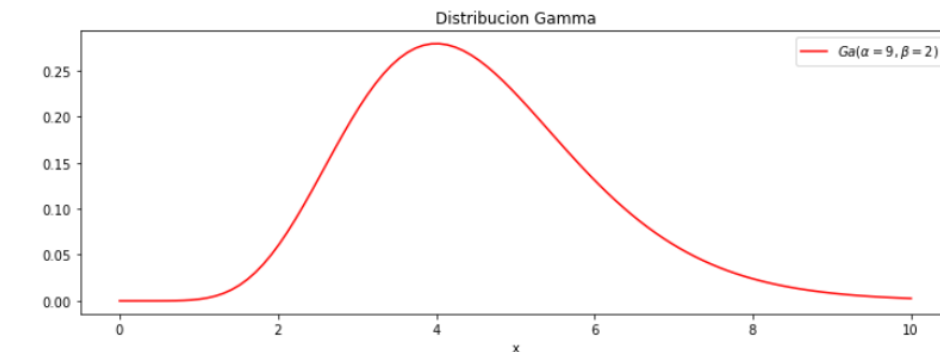
# Generar figura
plt.figure(figsize=(12,4))

# Generar puntos
x = np.linspace(0,10, 100)

# Generar probabilidades para Gamma
plt.plot(x, Ga.pdf(x), "r", label="$Ga(\alpha=9, \beta=2)$")


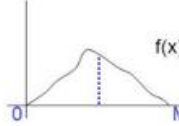
# Agregar estilo
plt.title("Distribucion Gamma")
plt.xlabel("x")
plt.legend()
```

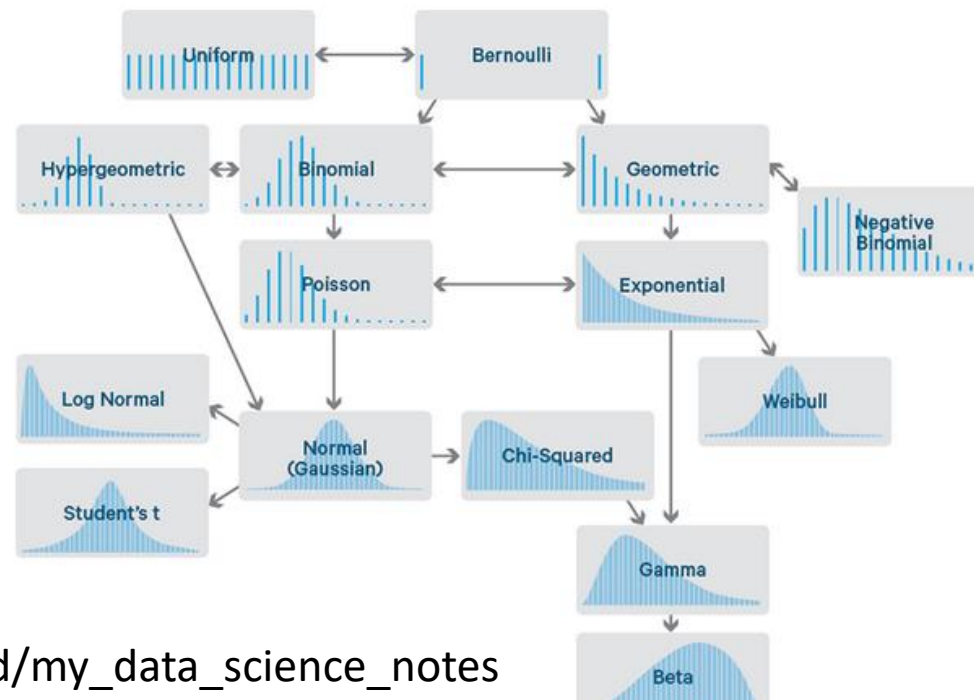
: <matplotlib.legend.Legend at 0x180396dc908>



En el marco del curso "De la mano de científicos:

¿Cómo se realiza un análisis de los resultados de un trabajo científico?

	Definition	Discrete R.V.s	Continuous R.V.s
<i>pmf</i> (Discrete) vs. <i>pdf</i> (Continuous)			
			
			
Mean: μ	$E(X)$	$\sum_i p(x_i)x_i$	$\int_{-\infty}^{\infty} p(x)x dx$
Variance: σ^2	$E((X - \mu)^2)$	$\sum_i p(x_i)(x_i - \mu)^2$	$\int_{-\infty}^{\infty} p(x)(x - \mu)^2 dx$



https://github.com/jirvingphd/my_data_science_notes

En el marco del curso "De la mano de científicos: ¿Cómo hacer investigación?"
¿Cómo se realiza un análisis de los resultados de un trabajo científico?

Gracias por su atención