



¿Cómo se realiza un análisis de los resultados de un trabajo científico? Parte 2



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

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

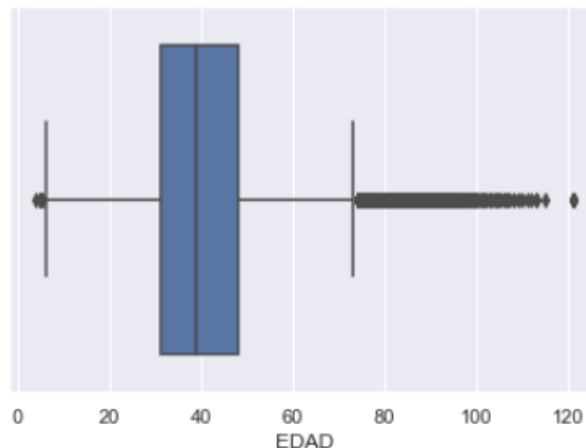
Variable Numérica Discreta

	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"])
```

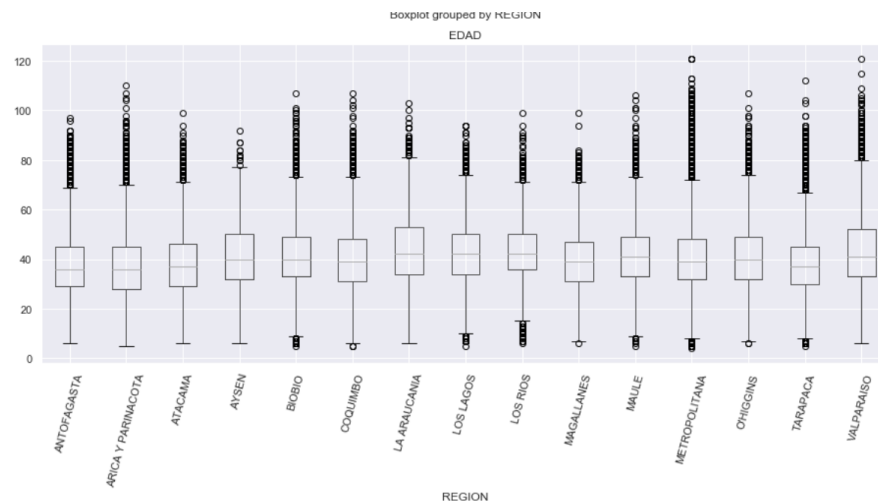
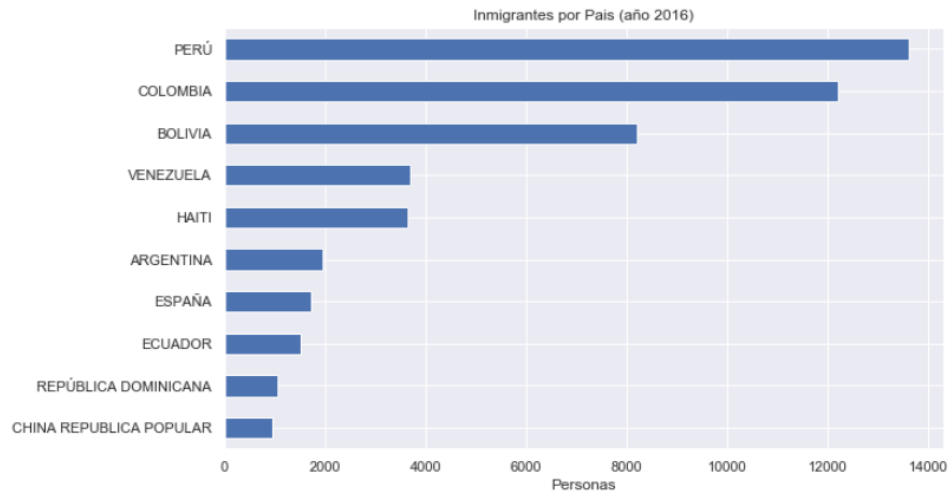


En el marco del curso "De la mano de científicos:

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

Visualizaciones Básicas

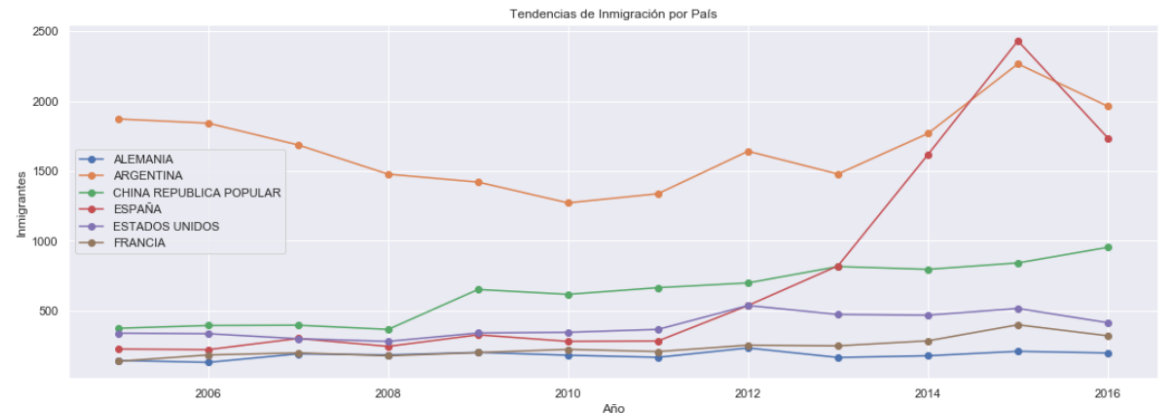
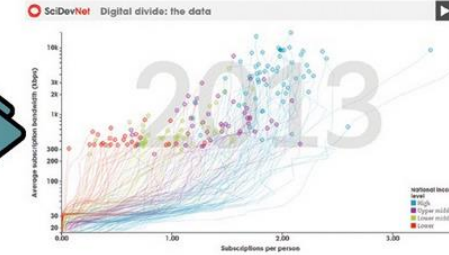
Análisis Exploratorio



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 in	non-OECD	Andorra	415,111	535,184	561,080	578,344	604,24
Upper mi	non-OECD	Angola	87,356	107,357	128,502	156,822	190,84
High in	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	Armenia	564,449	621,981	684,924	747,534	817,71
High in	non-OECD	Australia	82,872	106,449	123,950	171,880	199,34
High in	OECD me	Austria	116,784	130,314	136,818	145,805	154,00
Upper mi	non-OECD	Azerbaijan					
Low in	non-OECD	Bahamas					
High in	non-OECD	Bahrain					
Low in	non-OECD	Bangladesh					
High in	non-OECD	Barbados					
Upper mi	non-OECD	Belarus					
High in	OECD me	Belgium					
Upper mi	non-OECD	Belize					
Low in	non-OECD	Benin					

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

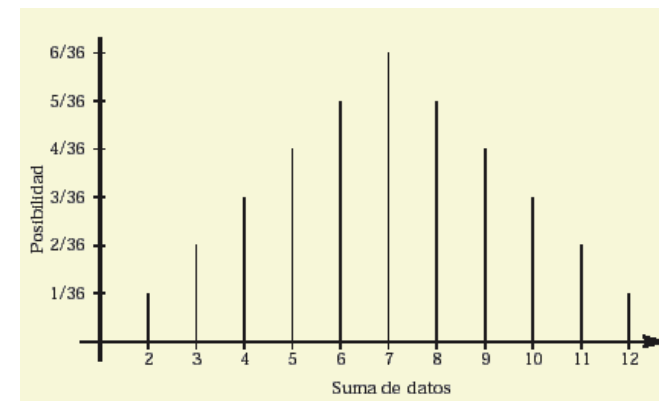
Lanzar dos dados

Variable aleatoria Valores posibles Eventos aleatorios

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

Lanzar una moneda

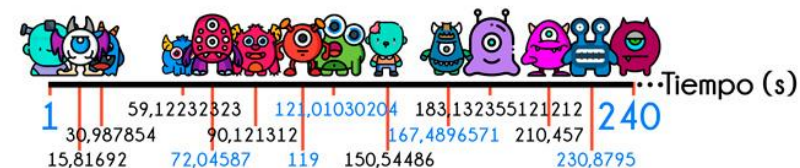
	(1 1)	(1 2)	(1 3)	(1 4)	(1 5)	(1 6)
	(2 1)	(2 2)	(2 3)	(2 4)	(2 5)	(2 6)
	(3 1)	(3 2)	(3 3)	(3 4)	(3 5)	(3 6)
	(4 1)	(4 2)	(4 3)	(4 4)	(4 5)	(4 6)
	(5 1)	(5 2)	(5 3)	(5 4)	(5 5)	(5 6)
	(6 1)	(6 2)	(6 3)	(6 4)	(6 5)	(6 6)



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:

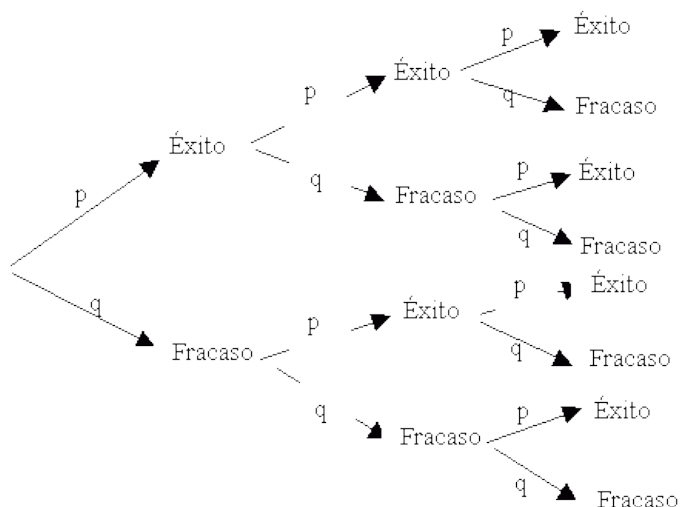


>> D = tiempo de atención en ventanilla (en s).

↪ $R_D : 1 \leq d \leq 240$

Distribuciones de Probabilidades Discretas

Binomial

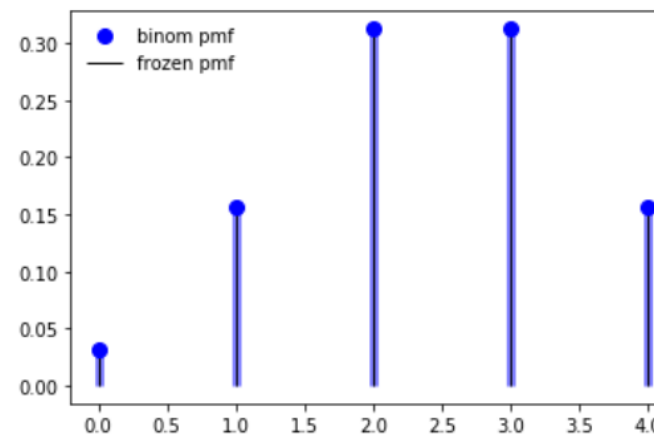


Ejemplo: Lanzar 3 veces una monedas y contar cuantas veces sale cara

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)`

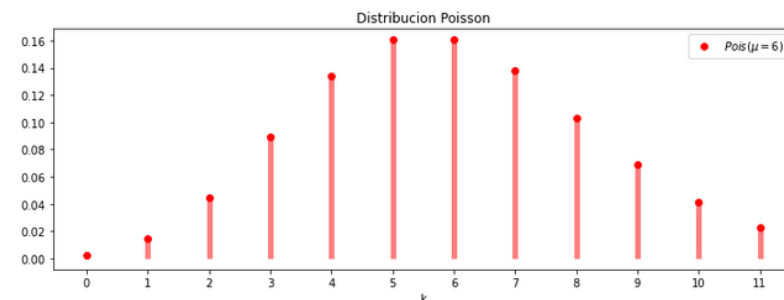
```
# Instanciar Distribuciones
Pois = stats.poisson(mu=6)

# 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

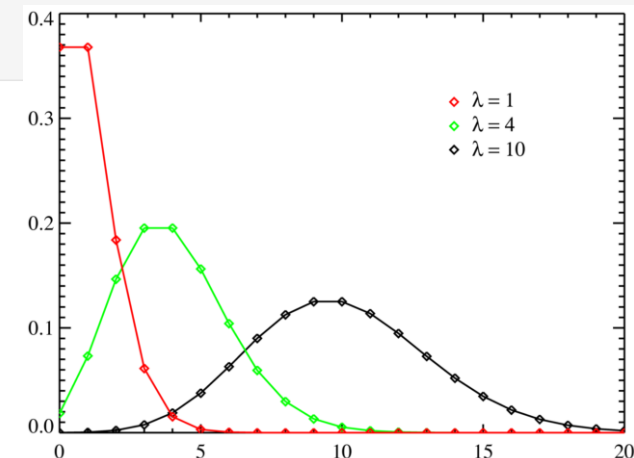
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

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)`

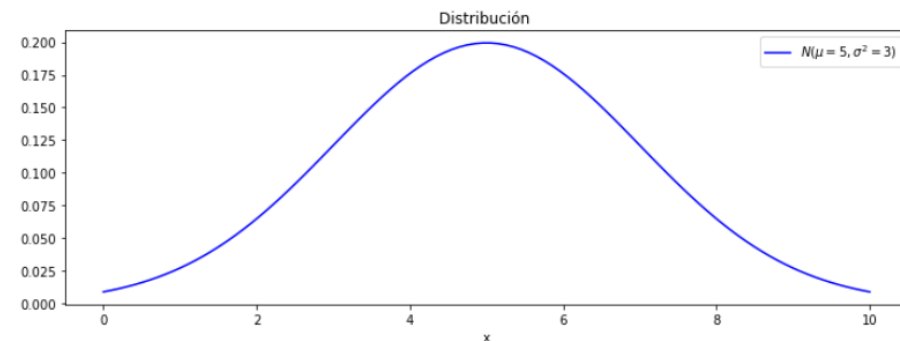
```
# Instanciar Distribucion
N = stats.norm(loc=5, scale=2)

# 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)

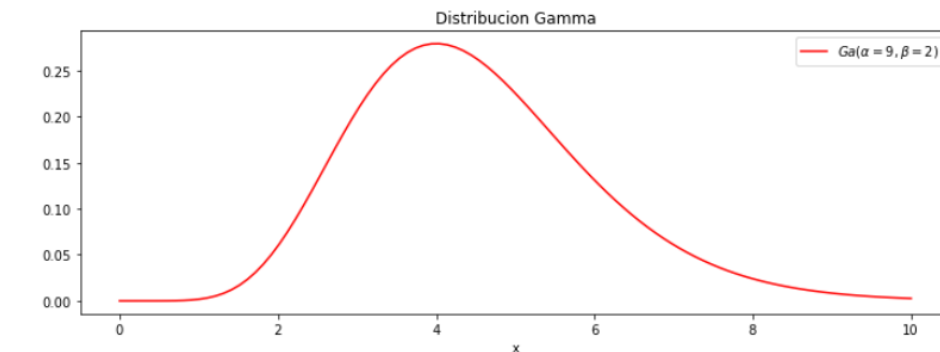
# Generar figura
plt.figure(figsize=(12,4))

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

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


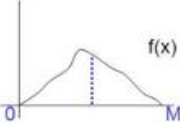
# Agregar estilo
plt.title("Distribucion Gamma")
plt.xlabel("x")
plt.legend()
```

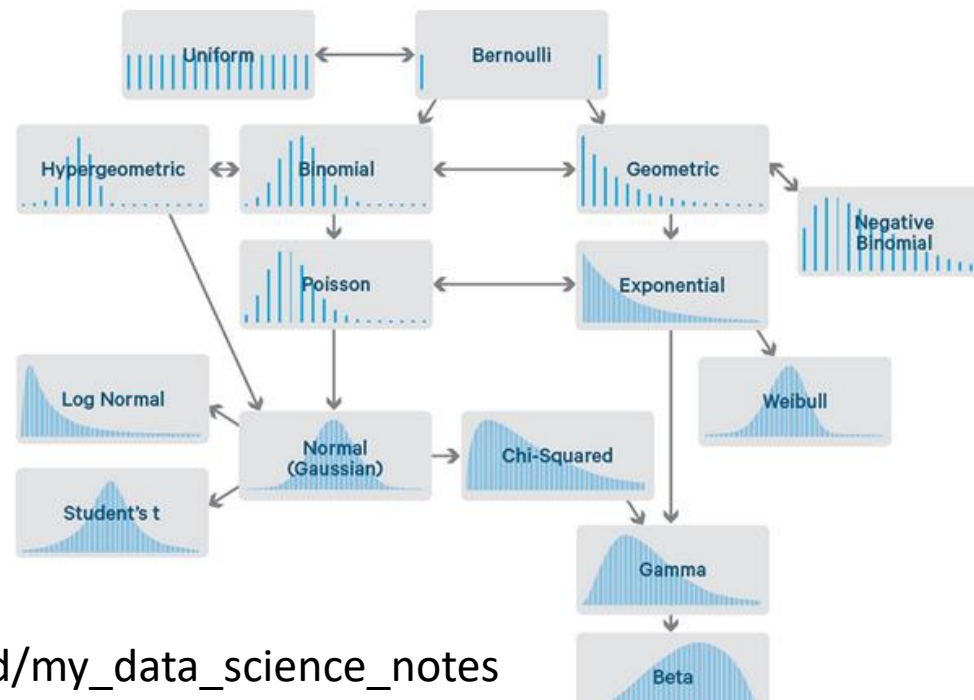
: <matplotlib.legend.Legend at 0x180396dc908>



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

Contenido: Análisis estadístico

Día 2

- Pruebas de Hipótesis
 - Nivel de significancia
 - Tipos de Errores
- Estimación de Parámetros
 - Regresión Lineal
 - Regresión Logística
 - Bondad de ajuste

One-Sample Hypothesis Tests for Discrete Data (Orange)

Select:	Two-tail test	One-tail test	
	Two-tail	Lower/left-tail	Upper/right-tail
	$H_0: p = p_0$	$H_0: p \geq p_0$	$H_0: p \leq p_0$
	$H_a: p \neq p_0$	$H_a: p < p_0$	$H_a: p > p_0$
Choose:	Sample size		
	Must have $np \geq 5$ $n(1-p) \geq 5$ $n \geq 30$		Where $p = X/n$ $X = \text{no. of items of interest in sample}$
Calculate:	Test statistic		
	$z = \frac{p - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n}$		
Identify:	p-value		
	Two-tail $p = 2 \times \text{area past } Z$	Lower/left-tail $p = \text{area left of } Z$	Upper/right-tail $p = \text{area right of } Z$

Two-Sample Hypothesis Tests for Discrete Data (Pink)

Select:	Two-tail test	One-tail test	
	Two-tail	Lower/left-tail	Upper/right-tail
	$H_0: p_1 = p_2$	$H_0: p_1 \geq p_2$	$H_0: p_1 \leq p_2$
	$H_a: p_1 \neq p_2$	$H_a: p_1 < p_2$	$H_a: p_1 > p_2$
Choose:	Sample size		
	Must have $n_1 + n_2 \geq 30$		Where $p_1 = X_1/n_1$ and $p_2 = X_2/n_2$ $X = \text{no. of items of interest in sample}$
Calculate:	Test statistic		
	$Z = \frac{p_1 - p_2}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$ Where $p = \frac{X_1 + X_2}{n_1 + n_2}$		
Identify:	p-value		
	Two-tail $p = 2 \times \text{area past } Z$	Lower/left-tail $p = \text{area left of } Z$	Upper/right-tail $p = \text{area right of } Z$

One-Sample Hypothesis Tests for Continuous Data (Purple)

Select:	Two-tail test	One-tail test	
	Two-tail	Lower/left-tail	Upper/right-tail
	$H_0: \mu = \mu_0$	$H_0: \mu \geq \mu_0$	$H_0: \mu \leq \mu_0$
	$H_a: \mu \neq \mu_0$	$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$
Choose:	Sample size		
	Large $n \geq 30$ (or σ known)	Small $n < 30$ (or σ unknown)	
Calculate:	Test statistic		
	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ Can replace σ with s if known	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ $df = n - 1$	
Identify:	p-value		
	Two-tail $p = 2 \times \text{area past } Z \text{ or } t$	Lower/left-tail $p = \text{area left of } Z \text{ or } t$	Upper/right-tail $p = \text{area right of } Z \text{ or } t$

Two-Sample Hypothesis Tests for Continuous Data (Green)

Select:	Two-tail test	One-tail test	
	Two-tail	Lower/left-tail	Upper/right-tail
	$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 \geq \mu_2$	$H_0: \mu_1 \leq \mu_2$
	$H_a: \mu_1 \neq \mu_2$	$H_a: \mu_1 < \mu_2$	$H_a: \mu_1 > \mu_2$
Choose:	Sample size		
	Large $n_1 + n_2 \geq 30$ (or σ known)	Small $n_1 + n_2 < 30$ (or σ unknown)	
Calculate:	Test statistic		
	$Z = \frac{\bar{X} - 1 - \bar{X} - 2s_1^2n_1 + s_2^2n_2}{2s_1^2n_1 + s_2^2n_2}$	$t = \frac{\bar{X} - 1 - \bar{X} - 2s_1^2n_1 + s_2^2n_2}{df = n_1 + n_2 - 2}$	
Identify:	p-value		
	Two-tail $p = 2 \times \text{area past } Z \text{ or } t$	Lower/left-tail $p = \text{area left of } Z \text{ or } t$	Upper/right-tail $p = \text{area right of } Z \text{ or } t$

Prueba de Hipótesis

Una prueba de hipótesis es una regla que especifica si se puede aceptar o rechazar una afirmación acerca de una población dependiendo de la evidencia proporcionada por una muestra de datos.

- ¿Tienen las estudiantes de pregrado una estatura media diferente de 66 pulgadas?
- ¿Es la desviación estándar de su estatura igual a o menor que 5 pulgadas?
- ¿Es diferente la estatura de las estudiantes y los estudiantes de pregrado en promedio?
- ¿Es la proporción de los estudiantes de pregrado significativamente más alta que la proporción de las estudiantes de pregrado?



Prueba de Hipótesis

Una prueba de hipótesis es una regla que especifica si se puede aceptar o rechazar una afirmación acerca de una población dependiendo de la evidencia proporcionada por una muestra de datos.

- ¿Tienen las estudiantes de pregrado una estatura media diferente de 1.75m?
- ¿Es la desviación estándar de su estatura igual a o menor que 15cm pulgadas?
- ¿Es diferente la estatura de las estudiantes y los estudiantes de pregrado en promedio?
- ¿Es la proporción de los estudiantes de pregrado significativamente más alta que la proporción de las estudiantes de pregrado?



Prueba de Hipótesis

<https://www.youtube.com/watch?v=AJcy4eZMwWM>

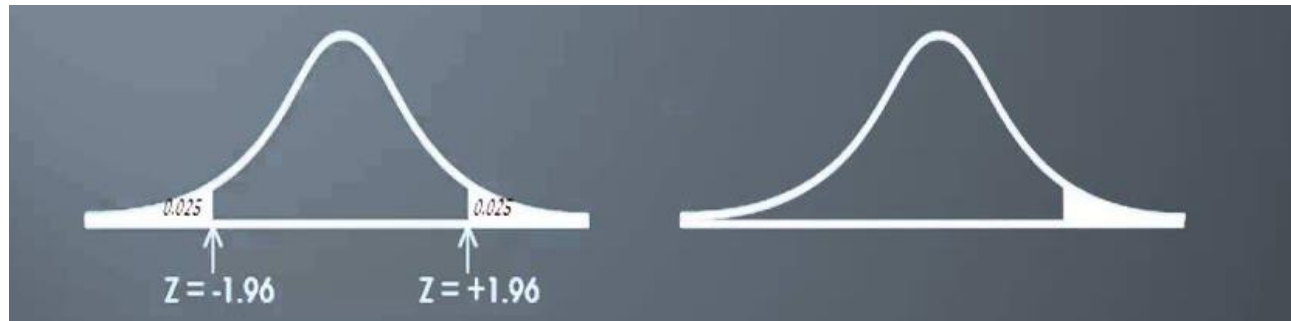
Las hipótesis nula y alternativa son dos enunciados mutuamente excluyentes acerca de una población.

Hipótesis nula (H_0)

La hipótesis nula indica que un parámetro de población (tal como la media, la desviación estándar, etc.) es igual a un valor hipotético

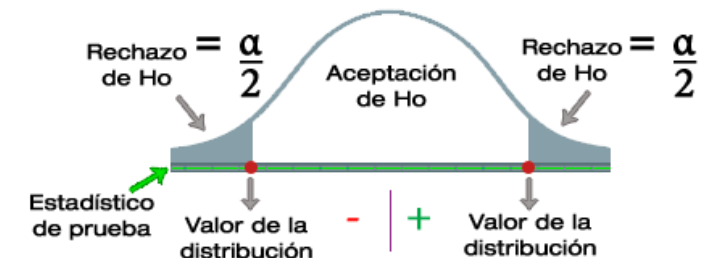
Hipótesis alternativa (H_1)

La hipótesis alternativa indica que un parámetro de población es más pequeño, más grande o diferente del valor hipotético de la hipótesis nula.



. ($H_0: \mu = 850$ vs. $H_1: \mu \neq 850$)

($H_0: \mu = 850$ vs. $H_1: \mu > 850$)



Prueba de Hipótesis

Pasos de la prueba de hipótesis

PASO 1

Planteamiento de la hipótesis nula y alternativa
3 situaciones

$$① \begin{cases} H_0 : P = p \\ H_1 : P \neq p \end{cases}$$

$$② \begin{cases} H_0 : P \leq p \\ H_1 : P > p \end{cases}$$

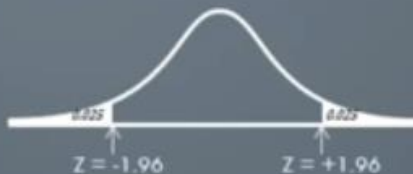
$$③ \begin{cases} H_0 : P \geq p \\ H_1 : P < p \end{cases}$$

PASO 2

Elegir el nivel de significancia ($\alpha = 0.05$ o en su forma 5%.

PASO 3

Determinación de la zona de aceptación y rechazo de la hipótesis nula (H_0)



PASO 4

Determinación de la Función Pivotal

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Cuando $n > 30$

$$t = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (n-1) \text{ GL}$$

Cuando $n \leq 30$

Prueba de Hipótesis

Pasos de la prueba de hipótesis

PASO 5

Cálculo de la función Pivotal

Se reemplaza en la formula correcta la información obtenida y se obtiene un valor.

Por ejemplo si deseamos realizar una prueba de hipótesis para la media poblacional de los estudiantes de la USP y planteamos la hipótesis de interés de que la edad es diferente a 25 años.

Posteriormente cogemos una muestra de 40 alumnos y encontramos que el promedio de su edad es de 22.5 años, con una Desviación estándar de 4.5 años.

La función Pivotal elegida sería:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

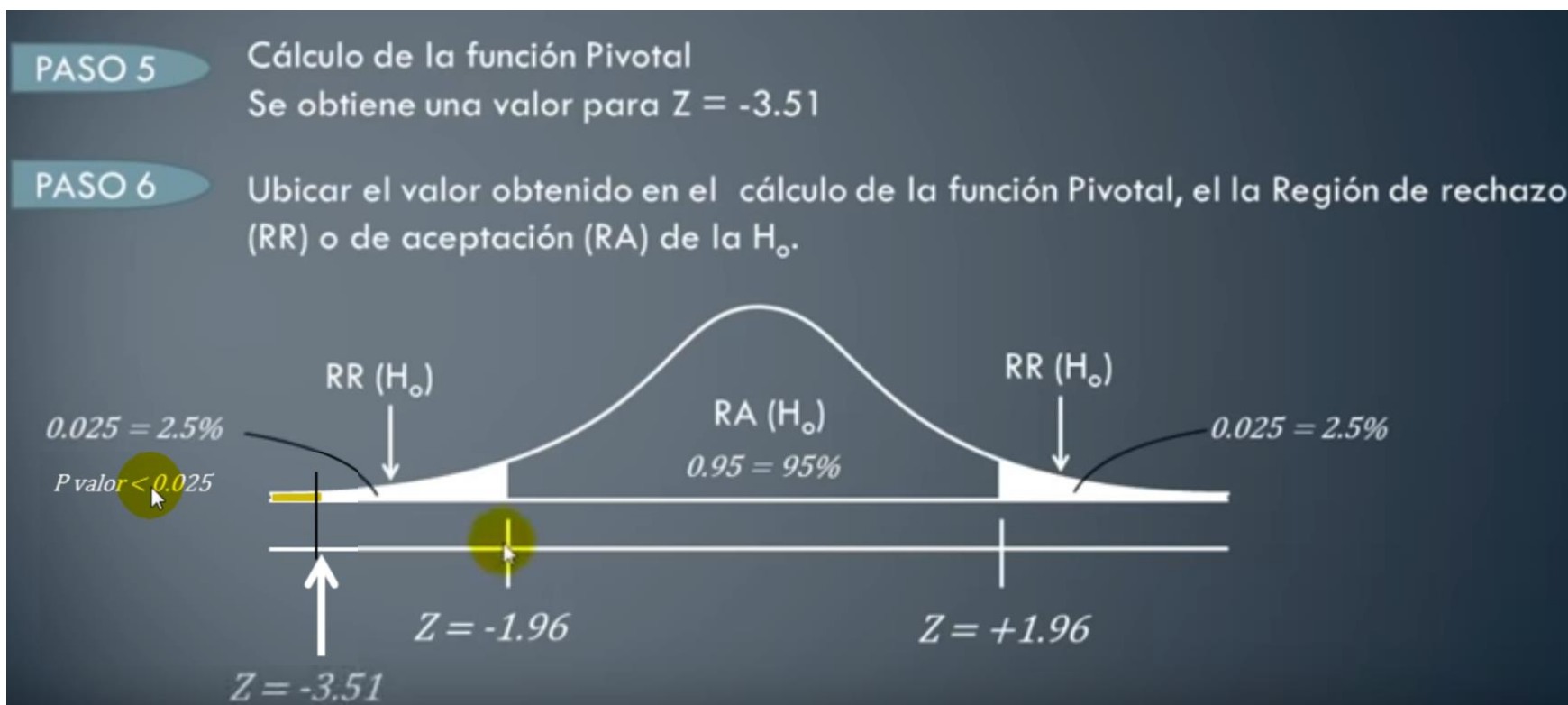
Cuando $n > 30$

Reemplazando los valores

$$Z = \frac{22.5 - 25}{4.5 / \sqrt{40}} = \frac{-2.5}{0.71}$$

Finalmente $Z = -3.51$

Prueba de Hipótesis



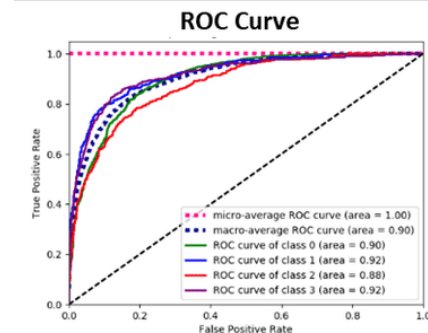
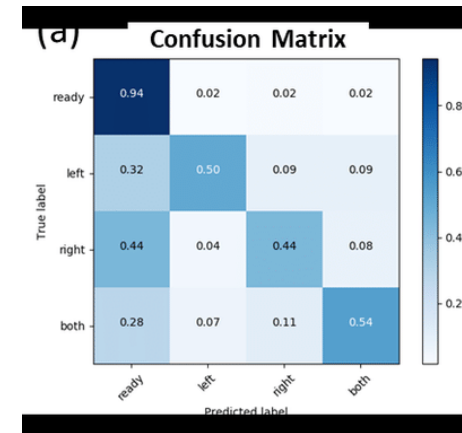
P-value

Tipos de errores

Decisión basada en la muestra	Verdad acerca de la población	
	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta (probabilidad = $1 - \alpha$)	Error tipo II - no rechazar H_0 cuando es falsa (probabilidad = β)
Rechazar H_0	Error tipo I - rechazar H_0 cuando es verdadera (probabilidad = α)	Decisión correcta (probabilidad = $1 - \beta$)

- Un investigador médico desea comparar la efectividad de dos medicamentos. Las hipótesis nula y alternativa son: Hipótesis nula (H_0): $\mu_1 = \mu_2$
- Los dos medicamentos tienen la misma eficacia.
- Hipótesis alternativa (H_1): $\mu_1 \neq \mu_2$
- Los dos medicamentos no tienen la misma eficacia.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)



Tipos de errores

Un investigador médico desea comparar la efectividad de dos medicamentos. Las hipótesis nula y alternativa son:

- **Hipótesis nula (H_0):** $\mu_1 = \mu_2$

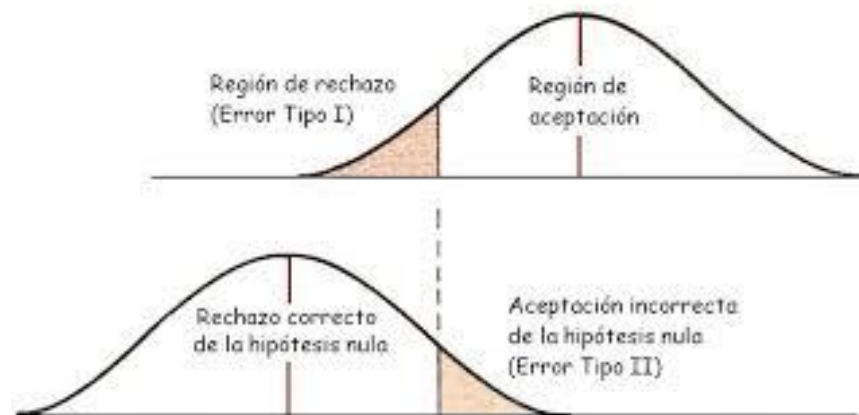
Los dos medicamentos tienen la misma eficacia.

- **Hipótesis alternativa (H_1):** $\mu_1 \neq \mu_2$

Los dos medicamentos no tienen la misma eficacia.

Un error de tipo I se produce si el investigador rechaza la hipótesis nula y concluye que los dos medicamentos son diferentes cuando, en realidad, no lo son.

Un error de tipo II, el investigador no rechaza la hipótesis nula cuando debe rechazarla y concluye que los medicamentos son iguales cuando en realidad son diferentes.



Prueba de Normalidad

Prueba de Kolmogorov-Smirnov

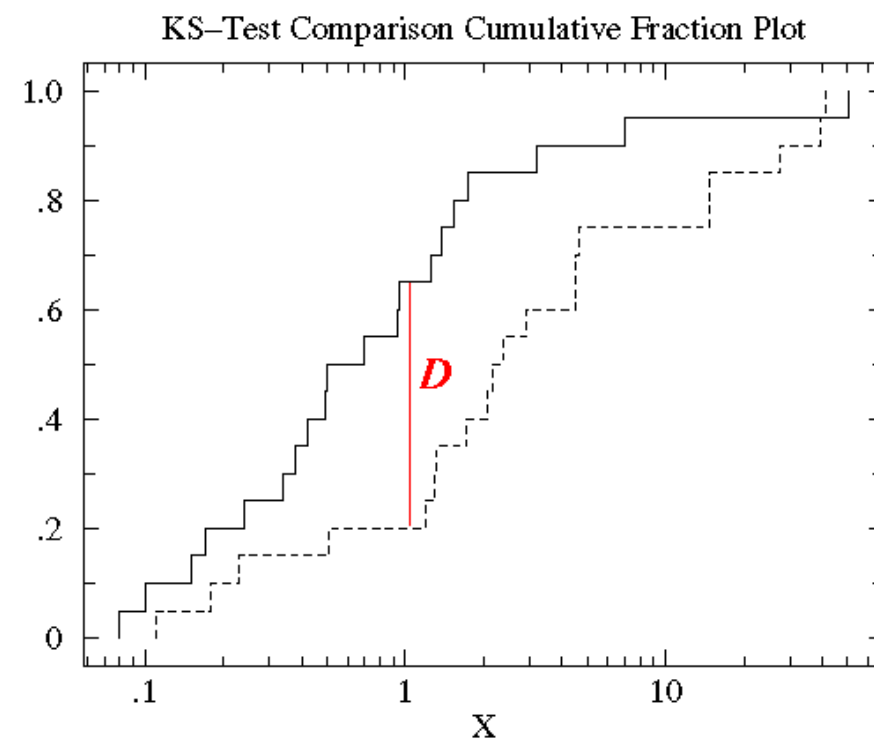
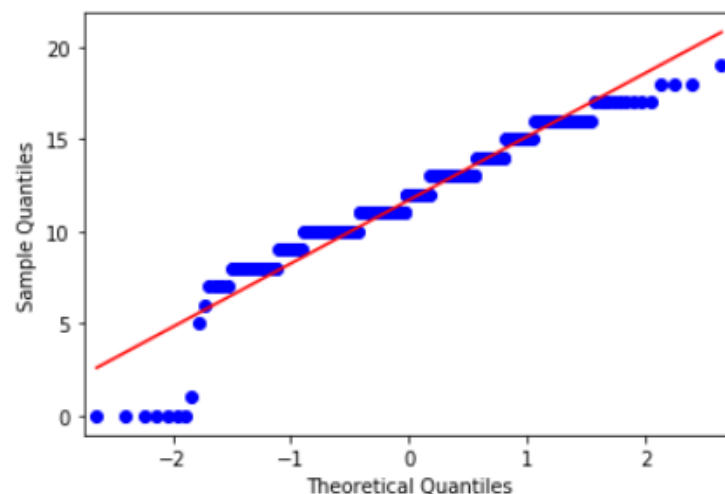
El supuesto de normalidad es importante para muchos modelos:

Se puede examinar con graficos:

- qqplots
- diagramas de caja

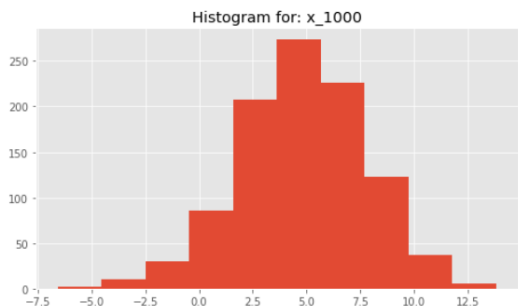
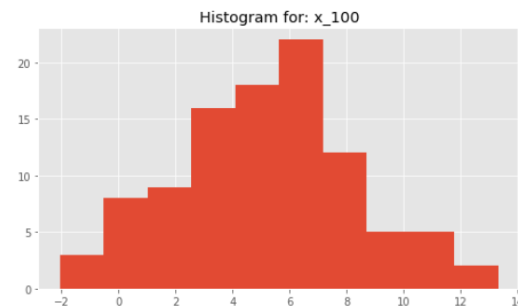
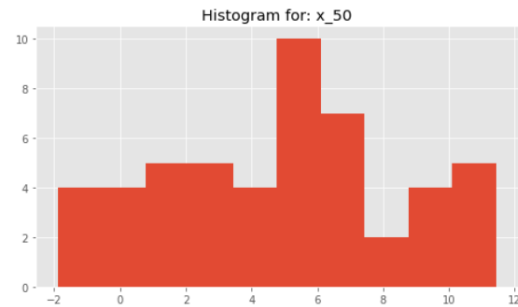
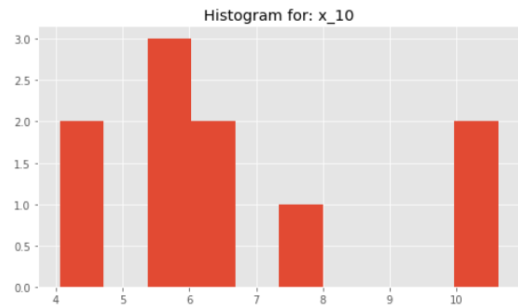
Prueba con estadísticas:

- Shapiro-Wilk;
- Anderson-Darling y;
- Kolmogorov-Smirnov.

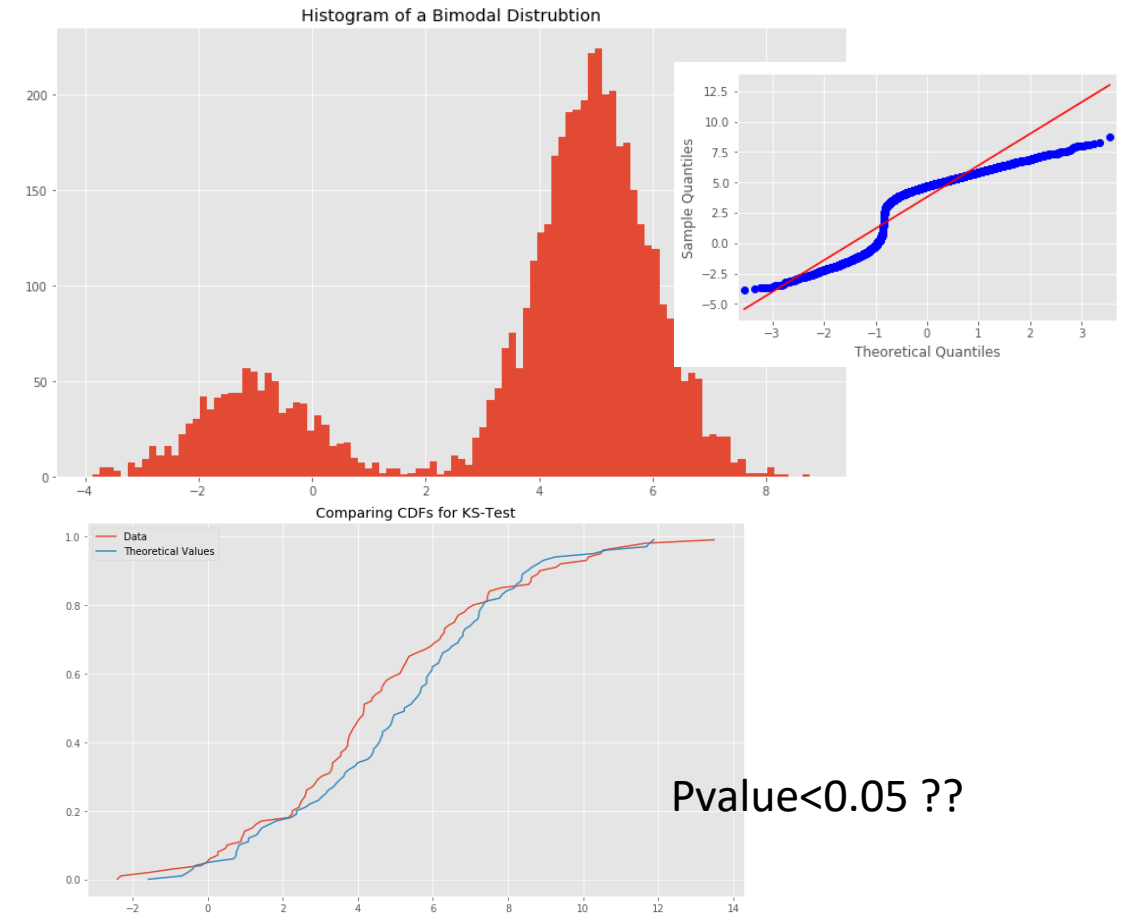


Prueba de Normalidad

<https://github.com/robertdefilippi/normality-applications-python/blob/master/normality-tests.ipynb>



```
(0.8637195825576782, 0.08440198749303818)
(0.9710047245025635, 0.2541665732860565)
(0.9930154085159302, 0.8889862298965454)
(0.9976759552955627, 0.17184072732925415)
For x_10: Don't reject normality hypothesis
None
For x_50: Don't reject normality hypothesis
None
For x_100: Don't reject normality hypothesis
None
For x_1000: Don't reject normality hypothesis
None
```



KstestResult(statistic=0.061047174276107175, pvalue=0.8501205705664947)

Prueba de Hipótesis(dos muestras)

Prueba KS de dos muestras

- comprueba si se han extraído dos muestras ** independientes * dos poblaciones idénticas ($X = Y$).
- compara dos distribuciones de ** muestra ** (en lugar de teóric

$$d = \max[abs[F_{n1}(X) -$$

- n_1 = Observaciones de la primera muestra.
- n_2 = Observaciones del segundo muestreo

Hipótesis nula: Se extraen 2 muestras independientes de la misma distribución continua.

```
from scipy import stats

np.random.seed(12345678) #fix random seed to get the same result
n1 = 200 # size of first sample
n2 = 300 # size of second sample
n3 = 100 # size of second sample

rvs1 = stats.norm.rvs(size=n1, loc=0., scale=1)

rvs2 = stats.norm.rvs(size=n2, loc=0.5, scale=1.5)

rvs3 = stats.norm.rvs(size=n3, loc=0., scale=1)
```

```
stats.ks_2samp(rvs1, rvs2)
```

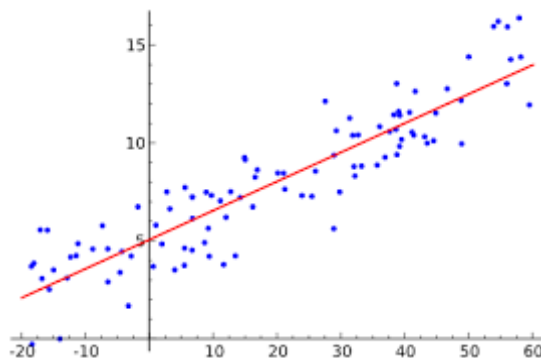
```
Ks_2sampResult(statistic=0.20833333333333334, pvalue=5.129279597815284e-05)
```

```
stats.ks_2samp(rvs1, rvs3 )
```

```
Ks_2sampResult(statistic=0.105, pvalue=0.44546367341695026)
```

Regresión Lineal

La regresión lineal se basa en la estimación de mínimos cuadrados ordinarios (OLS)



Divide by the total number of data points

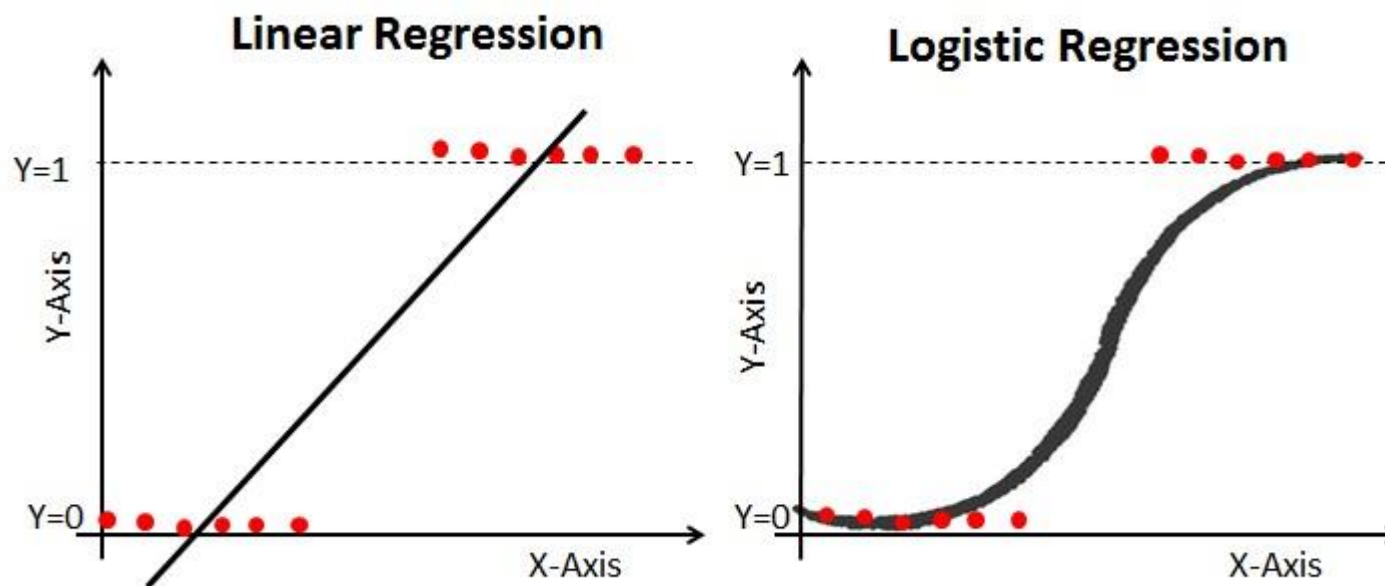
Predicted output value

Actual output value

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

Regresión Lineal vs Regresión Logística

La regresión lineal se utiliza para predecir salidas continuas, mientras que la regresión logística se utiliza para predecir un conjunto discreto de salidas que se asigna a diferentes clases.



Bondad de ajuste

"Prueba de bondad de ajuste chi cuadrado" Esto también se usa para probar si las frecuencias de muestra observadas son consistentes con las frecuencias esperadas.

Supuestos

- El método de muestreo es un muestreo aleatorio simple.
- La variable en estudio es ** categórica **.
- El valor esperado del número de observaciones de muestra en cada nivel de la variable es al menos 5.

Hipótesis

** Hipótesis nula **: los valores observados y esperados son consistentes

** Hipótesis alternativa **: los valores observados y esperados son significativamente diferentes

Estadística de prueba

La estadística de prueba (CV) se calcula como

$$\chi^2 = \sum_i (O_i - E_i)^2 / E_i$$

donde se observan O_i y E_i y se cuentan las frecuencias esperadas.

<https://archive.ics.uci.edu/ml/index.php>

Actividades

- Seleccionar un dataset
(assumir que este representa la población total de estudio)
- Seleccionar dos variables de interés
- Seleccionar una muestra aleatoria de los datos
- Graficar el histograma de las variables (población y de la muestra)
- Construir um gráfico de dispersión
- Graficar un box plot(si es uma variable numérica) o una torta si es categórica
- Comparar los estadísticos de la muestra y de la población

Most Popular Data Sets (hits since 2007):	
3722311:	 Iris
2020349:	 Adult
1562059:	 Wine
1396138:	 Breast Cancer Wisconsin (Diagnostic)
1393114:	 Heart Disease
1389636:	 Wine Quality
1355281:	 UCI Bank Marketing
1293901:	 Car Evaluation
1068686:	 UCI Human Activity Recognition Using Smartphones
1026779:	 Abalone
957401:	 Forest Fires
838077:	 UCI Student Performance

Actividades Opcionales

- Realizar una prueba de hipótesis para determinar si la media de la muestra es consistente con la de la población.
- Definir dos muestras (o dos grupos) y pueden venir de una misma distribución continua o no(KS).
- (Opcional)Definir dos muestras (o dos grupos) y determinar si la diferencia de las medias es significativa o no. <https://github.com/trangel/stats-with-python/blob/master/notebooks//Difference%20between%20means.ipynb>
- Seleccionar dos variables numéricas y realizar un ajuste lineal
- Escribir un informe de 1-2 páginas (se puede enviar el notebook o código si utilizaron python o R)

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