Máster Data Science - Ed 26

Estadística con Python

Irene Torres Valle

¿Quién soy?

KSCHOOL

2013

CEPSA

Analista de negocio (BI - Marketing)

2016

El Corte Inglés

Data Scientist Estrategia de Clientes

2018

REALE SEGUROS

Analista de Inteligencia de Clientes

- Licenciada en Matemáticas
- Licenciada en CC y TT Estadísticas
UEX

- Programa Óptimus Executive
UFV

- Máster Educación
UI1

- Máster Data Science
UNED

\+

Formación continua *en aspectos técnicos relacionados con la ciencia del dato (machine learning, deep learning) o lenguajes de programación, así como en conocimiento de negocio y habilidades comunicativas, story telling, etc.*
Plataformas online y autodidacta

Contenido

1. Introducción y definiciones:
   estadística, población-muestra
2. Estadística descriptiva.
3. Distribuciones de probabilidad.
4. Estimación puntual y por intervalos de confianza.
5. Contrastes de hipótesis paramétricos.
6. Contrastes de hipótesis no paramétricos.

Introducción y definiciones

Estadística

Ciencia que utiliza un conjunto de datos para obtener, a partir de ellos, análisis descriptivos e inferencias basadas en el cálculo de probabilidades.

Población

Conjunto de elementos sobre los que se quiere estudiar alguna característica o realizar un análisis.

Muestra

Parte de la población, será estudiada y a partir de la cual se generalizan los resultados a la población.

Introducción y definiciones

muestra

Población vs Muestra

Introducción y definiciones

representatividad de la muestra

Una **muestra representativa** es una versión simplificada de la población, y reproduce, de algún modo, el mismo comportamiento y características ante la variable objeto de estudio que ésta pero a pequeña escala, y tal que su estudio sea viable.

Tipos de errores

- Errores muestrales
- Sesgos

AZAR

* excepciones

Introducción y definiciones

tipos de muestreo

Probabilístico (aleatorio)

Todos los individuos de la población tienen la misma probabilidad de formar parte de la muestra.

Aseguran la representatividad de la muestra.

<u>Son los recomendables</u>.

No probabilístico (no aleatorio)

No todos los elementos de la población tienen igual probabilidad de formar parte de la muestra.

Condicionada por la persona que selecciona la muestra o atendiendo a razones de comodidad.

<u>No suele ser un tipo de muestreo riguroso ni científico</u>.

Introducción y definiciones

tipos de muestreo Probabilístico

Muestreo Aleatorio Simple (mas)

Muy sencillo, simple azar.

- ➢ con reemplazamiento
- ➢ sin reemplazamiento (más fiable en muestras pequeñas)

Introducción y definiciones

tipos de muestreo Probabilístico

KSCHOOL

Muestreo Aleatorio Simple (mas)

Muy sencillo, simple azar.

- ➢ con reemplazamiento
- ➢ sin reemplazamiento (más fiable en muestras pequeñas)

1. Extraigo la primera bola: 15073
2. Antes de extraer la siguiente bola:
   - Con reemplazamiento: vuelvo a meter la bola en el bombo
   - Sin reemplazamiento: dejo esa bola fuera y saco otra del bombo
3. Así sucesivamente hasta extraer las n bolas

Introducción y definiciones

tipos de muestreo Probabilístico

Muestreo Sistemático

Poblaciones ordenadas según alguna característica relacionada con la vble en estudio. Se escoge un primer individuo al azar y se van seleccionando los restantes de forma periódica.

Introducción y definiciones

tipos de muestreo Probabilístico

Considero una población de N individuos:
1,2,3,4,5,6,.............., N-2, N-1, N

Buscamos una muestra de tamaño n (n<N)

Partimos de h=N/n, ***coeficiente de elevación***

Se toma un nº al azar $\alpha \in [1,h]$, ***arranque u origen***

Muestra obtenida:
$\alpha, \alpha+h, \alpha+2h, \alpha+3h, \ldots\ldots, \alpha + (n-1)h$

Muestreo Sistemático
Poblaciones ordenadas según alguna característica relacionada con la vble en estudio. Se escoge un primer individuo al azar y se van seleccionando los restantes de forma periódica.

Introducción y definiciones

tipos de muestreo Probabilístico

Considero una población de N individuos:
1,2,3,4,5,6,.............., N-2, N-1, N

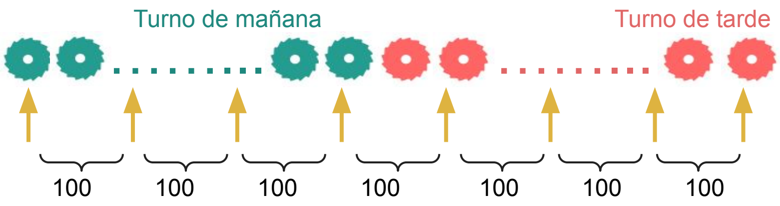
Buscamos una muestra de tamaño n (n<N)

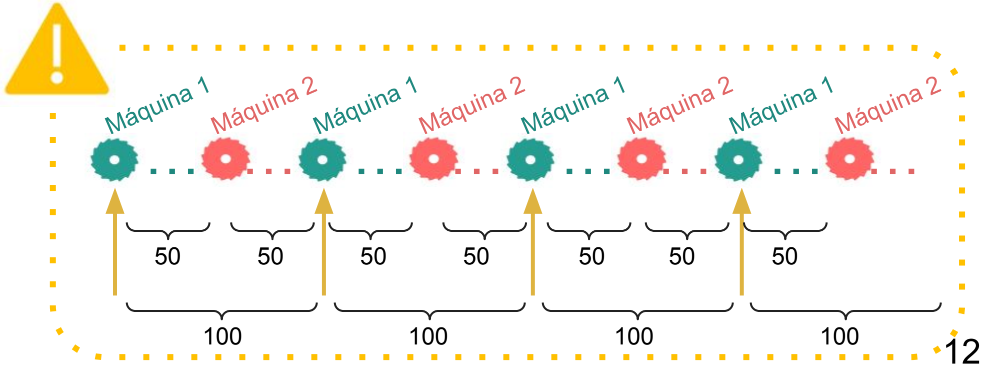
Partimos de h=N/n, ***coeficiente de elevación***

Se toma un nº al azar α ϵ [1,h], ***arranque u origen***

Muestra obtenida:
α, α+h, α+2h, α+3h, …… , α + (n-1)h

Muestreo Sistemático
Poblaciones ordenadas según alguna característica relacionada con la vble en estudio. Se escoge un primer individuo al azar y se van seleccionando los restantes de forma periódica.





Introducción y definiciones

tipos de muestreo Probabilístico

Muestreo estratificado

Dividir la población en estratos homogéneos y extraer una muestra global a partir de la unión de una muestra de cada estrato.

Afijación:

- igual
- proporcional
- óptima (según desviación típica)

Introducción y definiciones

tipos de muestreo Probabilístico

Estudio acerca del consumo de alcohol en la población española mayor de 15 años.

La población se encuentra distribuida por tramos de edad:

Tramos de edad	#	%		D.T.
15-24 años	4.758.009	12,2%	⟶	4,25
25-54 años	21.971.249	56,4%	⟶	6,43
>= 55 años	12.248.573	31,4%	⟶	8,59

Muestreo estratificado

Dividir la población en estratos homogéneos y extraer una muestra global a partir de la unión de una muestra de cada estrato.

Afijación:

- igual
- proporcional
- óptima (según desviación típica)

Extraigamos una muestra de tamaño n=1.000.000.

- Afijación normal:

n1 = n2 = n3 = 1.000.000/3 = 333.333.

- Afijación proporcional:

n1= 12,2% x 1.000.000 = 122.000

n2 = 56,4% x 1.000.000 = 564.000

n3 = 31,4% x 1.000.000 = 314.000

- Afijación óptima (según desviación típica):

$$n1 = \frac{4{,}25 \; x \; 4.758.009 \; x \; 1.000.000}{4{,}25 \; x \; 4.758.009 + 6{,}43 \; x \; 21.971.249 + 8{,}59 \; x \; 12.248.573}$$

= 75.818 individuos

$$n2 = \frac{6{,}43 \; x \; 21.971.249 \; x \; 1.000.000}{4{,}25 \; x \; 4.758.009 + 6{,}43 \; x \; 21.971.249 + 8{,}59 \; x \; 12.248.573}$$

= 529.692 individuos

$$n3 = \frac{8{,}59 \; x \; 12.248.573 \; x \; 1.000.000}{4{,}25 \; x \; 4.758.009 + 6{,}43 \; x \; 21.971.249 + 8{,}59 \; x \; 12.248.573}$$

= 394.940

Introducción y definiciones

tipos de muestreo Probabilístico

. Muestreo por conglomerados

Población dividida en conglomerados heterogéneos, con una variación similar a la del total de la población.

Los conglomerados seleccionados deben ser representativos.

Introducción y definiciones

tipos de muestreo Probabilístico

Muestreo Aleatorio Simple (mas)

Muy sencillo, simple azar.

- ➢ con reemplazamiento
- ➢ sin reemplazamiento (más fiable en muestras pequeñas)

Muestreo Sistemático

Poblaciones ordenadas según alguna característica relacionada con la vble en estudio. Se escoge un primer individuo al azar y se van seleccionando los restantes de forma periódica.

Muestreo estratificado

Dividir la población en estratos homogéneos y extraer una muestra global a partir de la unión de una muestra de cada estrato.

Afijación:

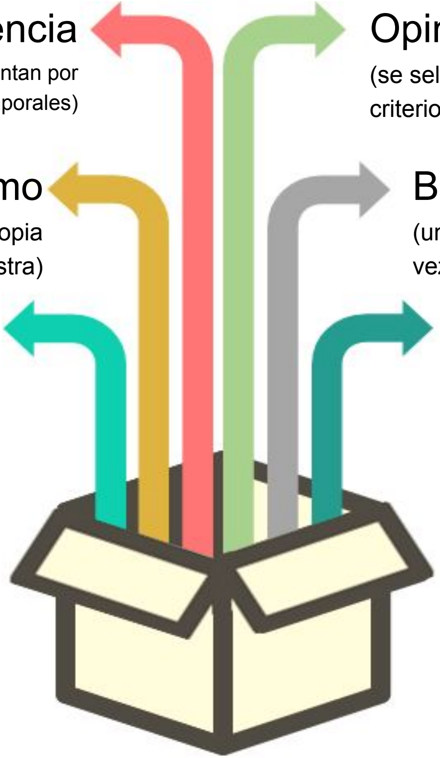
- ➢igual
- ➢proporcional
- ➢óptima (según desviación típica)

Muestreo por conglomerados

Población dividida en conglomerados heterogéneos, con una variación similar a la del total de la población.

Los conglomerados seleccionados deben ser representativos.

Introducción y definiciones
tipos de muestreo No Probabilístico



De conveniencia
(selecciono los individuos que menos esfuerzo presentan por cuestiones económicas, físicas o temporales)

De voluntarismo
(selecciono a los individuos que deciden por iniciativa propia pertenecer a la muestra)

Por cuotas
(se fijan cuotas para cada estrato de la población)

Opinático
(se seleccionan los elementos muestrales según el criterio del investigador)

Bola de nieve
(unos individuos conducen a otros y estos a su vez a otros, hasta formar la muestra)

Otros

KSCHOOL

Introducción y definiciones

Estadística descriptiva

(conjunto de métodos estadísticos que describen un conjunto de datos)

Estadística inferencial

(busca sacar conclusiones generales más allá de los datos analizados)

Introducción y definiciones

Investigación estadística: etapas

Estadística descriptiva

Variable aleatoria: definición

Se define una ***variable aleatoria*** como cada una de las propiedades, rasgos o cualidades que poseen los elementos de una población y que son objeto de estudio.

Estadística descriptiva

Variable aleatoria: clasificación

- **Variable Aleatoria**
  - **Cuantitativa** — aquellas que son medibles
    - **Discreta** — valores de un conjunto finito y numerable
      Nº de hijos
    - **Continua** — cualquier valor real de un intervalo
      m2 de una casa
  - **Cualitativa** — aquellas que no son medibles
    - **Nominal** — establece categorías
      sexo, profesión
    - **Ordinal** — establece un orden
      grado de satisfacción
    - **Binaria o dicotómica** — 2 valores: sí y no

Estadística descriptiva

Tipos de datos en Python

Librería pandas.

float64	números con decimales
int64	números enteros
datetime64	fechas
bool	booleanos: solo 2 valores Verdadero o Falso
timedelta64	periodos de tiempo, diferencias entre fechas
object	texto, vbles cualitativas

Librería datetime

Estadística descriptiva
Principales objetivos

1. Organizar la información recogida a través de tablas de frecuencias.

2. Representar gráficamente la información a través de diagramas de barras, histogramas, diagramas de sectores, polígonos de frecuencias, pictogramas, etc.

3. Resumir adecuadamente la información a través de sus parámetros estadísticos principales: medidas de posición, centrales y no centrales, de dispersión, de forma etc.

4. Detección de valores atípicos o fuera de rango.

Estadística descriptiva

Tablas de frecuencias

k categorías para una determinada vble.

Frecuencia absoluta: n_i (i = 1,...k)

Frecuencia relativa: $f_i = \frac{n_i}{n}$ (n= nº total de datos)

Frecuencia porcentual: $f_i \; x \; 100$

Frecuencia absoluta acumulada: $N_i = \sum_{j \leq i} n_i$

Frecuencia relativa acumulada: $F_i = \sum_{j \leq i} f_i$

Ejemplo 1

Satisfacción	n_i	f_i	%	N_i	F_i
Satisfecho	15	0.3	30%	15	0.3
Indiferente	25	0.5	50%	40	0.8
Insatisfecho	10	0.2	20%	50	1

Ejemplo 2

Hijos	n_i	f_i	%	N_i	F_i
0	30	0.6	60%	30	0.6
[1,3)	15	0.3	30%	45	0.9
[3,+∞)	5	0.1	10%	50	1

Estadística descriptiva

Parámetros estadísticos

KSCHOOL

Parámetros estadísticos

número que resume la gran cantidad de datos de una variable estadística

- **Medidas de posición**

  valores que se caracterizan por su posición en la distribución de la vble

  - **de centralización**

    media aritmética, mediana, moda

  - **de posición no central o cuantiles**

    cuartiles, deciles, percentiles

- **Medidas de dispersión**

  rango, varianza, desviación típica, coeficiente de variación de Pearson

  indican si los datos están concentrados o no alrededor de valores centrales

- **Medidas de forma**

  coeficiente de asimetría de Fisher, coeficiente de Curtosis

  caracterizan la forma de la gráfica

Estadística descriptiva

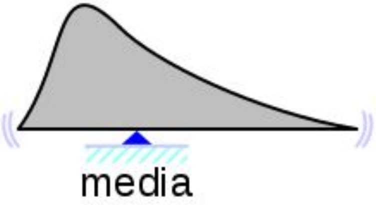
KSCHOOL

Medidas de posición centrales

Valores que se caracterizan por la posición que ocupan. Suelen situarse cerca del centro de la distribución.

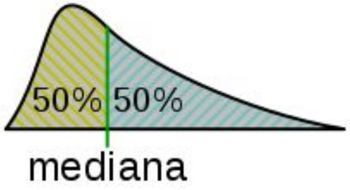
Media aritmética:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k} x_i f_i$$



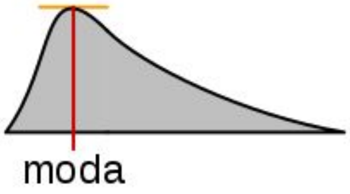
Mediana:

$$\overbrace{\ldots\ldots\ldots\ldots}^{50\%}Me\overbrace{\ldots\ldots\ldots\ldots}^{50\%}$$



Moda:

valor de la vble que más se repite

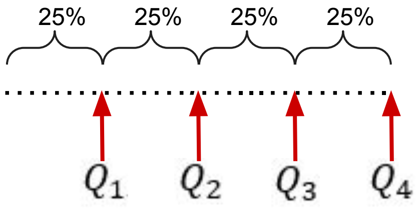


Estadística descriptiva

Medidas de posición no centrales

Valores que se caracterizan por la posición que ocupan. Dividen a la distribución en varias partes iguales:

Cuartiles



Rango Intercuartílico = $Q_3 - Q_1$

Percentiles

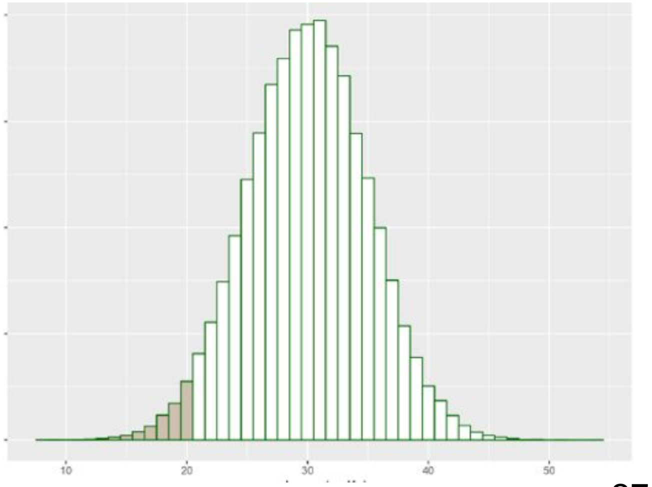
dividen a la distribución de datos en 100 partes iguales

P_{20} : valor de la variable bajo el cual se encuentra el 20% de las observaciones y el 80% restante será mayor

$P_{50} = \text{Me}$

$$P_{25} = Q_1, P_{50} = Q_2, P_{75} = Q_3 \text{ y } P_{100} = Q_4 = \max\{x_i\}$$

Percentil 20



Estadística descriptiva

Medidas de dispersión

Valores que se caracterizan por la posición que ocupan. Suelen situarse cerca del centro de la distribución.

Rango:

$$R = \max\{x_i\} - \min\{x_i\}$$

Varianza:

$$\sigma^2 = Var(X) = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{n}$$

- ★ σ^2 >= 0 (σ^2 = 0 si las mediciones sean todas iguales a la media)
- ★ si a todos los valores se les suma un nº, la varianza continúa igual
- ★ si a todos los valores se les multiplica por un nº, la varianza queda multiplicada por el cuadrado de dicho nº
- ★ la varianza no viene expresada en las mismas unidades que los datos, pues las desviaciones están al cuadrado

Desviación típica:

$$\sigma = +\sqrt{\sigma^2} = +\sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{n}}$$

- ★ σ >= 0 (σ = 0 si las mediciones sean todas iguales a la media)
- ★ Cuanto más pequeña sea σ mayor será la concentración de datos entorno a la media

Estadística descriptiva
Medidas de dispersión

Valores que se caracterizan por la posición que ocupan. Suelen situarse cerca del centro de la distribución.

Coeficiente de variación de Pearson: $CV = \frac{\sigma}{\bar{x}}$

- ★ Permite comparar la dispersión de 2 vbles
- ★ Se calculan de forma indep para cada vble y se comparan los valores obtenidos
- ★ Mayor coeficiente implica mayor variación

Ejemplos:

- ➢ Se toman dos muestras de la misma población, la primera tiene $\overline{x} = 140,\ \sigma_x = 28{,}28$ y la segunda $\overline{w} = 150,\ \sigma_w = 24$ ¿Cuál de las dos muestras presenta menor dispersión de los datos?

- ➢ En marzo del año pasado, los datos de préstamos personales de un Banco mostraron un promedio de $6.500.000 y una desviación estándar de $3.000.000 . Recientemente se calculó la media y la desviación estándar correspondiente a los préstamos personales de marzo del presente año resultando las mismas 9.000.000 y 3.500.000 respectivamente. ¿En cuál de los dos años los préstamos personales presentaron menor dispersión

Estadística descriptiva

Medidas de dispersión

Valores que se caracterizan por la posición que ocupan. Suelen situarse cerca del centro de la distribución.

Coeficiente de variación de Pearson: $CV = \frac{\sigma}{\bar{x}}$

- ★ Permite comparar la dispersión de 2 vbles
- ★ Se calculan de forma indep para cada vble y se comparan los valores obtenidos
- ★ Mayor coeficiente implica mayor variación

Ejemplos:

➢ Se toman dos muestras de la misma población, la primera tiene $\overline{x} = 140,\ \sigma_x = 28{,}28$ y la segunda $\overline{w} = 150,\ \sigma_w = 24$ ¿Cuál de las dos muestras presenta menor dispersión de los datos?

$$CV_x = \frac{28{,}28}{140} = 0{,}202 \text{ y } \boxed{CV_w = \frac{24}{150} = 0{,}16}$$

➢ En marzo del año pasado, los datos de préstamos personales de un Banco mostraron un promedio de \$6.500.000 y una desviación estándar de \$3.000.000 . Recientemente se calculó la media y la desviación estándar correspondiente a los préstamos personales de marzo del presente año resultando las mismas 9.000.000 y 3.500.000 respectivamente. ¿En cuál de los dos años los préstamos personales presentaron menor dispersión

$$CV_x = \frac{3}{6{,}5} = 0{,}462 \text{ y } \boxed{CV_w = \frac{3{,}5}{9} = 0{,}389}$$

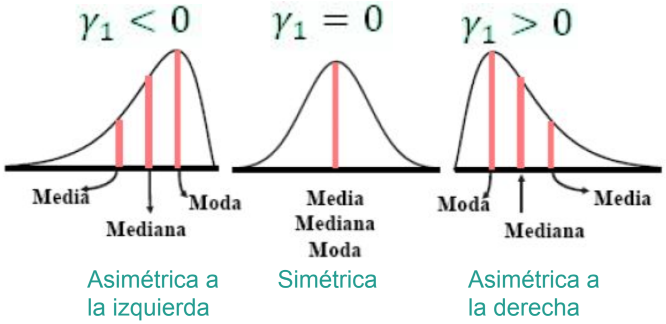
Estadística descriptiva

Medidas de forma

Valores que caracterizan la forma de la gráfica de una distribución de datos.

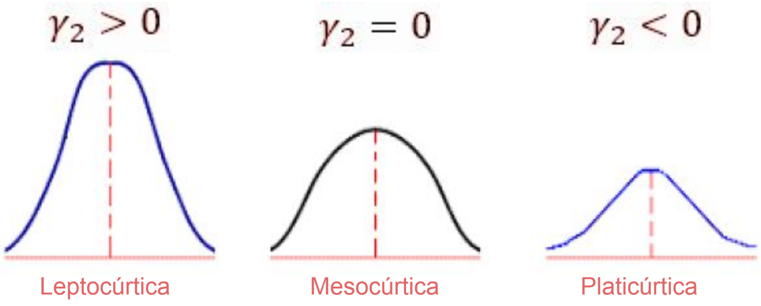
Coeficiente de asimetría de Fisher:

$$\gamma_1 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^3}{n\sigma^3}$$



Coeficiente de Curtosis:

$$\gamma_2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^4}{n\sigma^4} - 3$$



Estadística descriptiva

Relación entre variables

El coeficiente de Correlación de Pearson nos indica si entre dos variables cualesquiera x, y existe relación.

Covarianza poblacional:

$$\sigma(x,y) = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$$

Covarianza muestral:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y})$$

Correlación entre dos variables:

$$r = corr(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

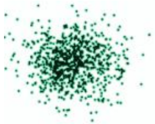
r ϵ [-1,1]

Gráfico de dispersión

- ➢ r = 1, correlación positiva perfecta.

- ➢ 0 < r < 1, existe una correlación positiva.

- ➢ r = 0, no existe relación lineal.



- ➢ -1 < r < 0, existe una correlación negativa.

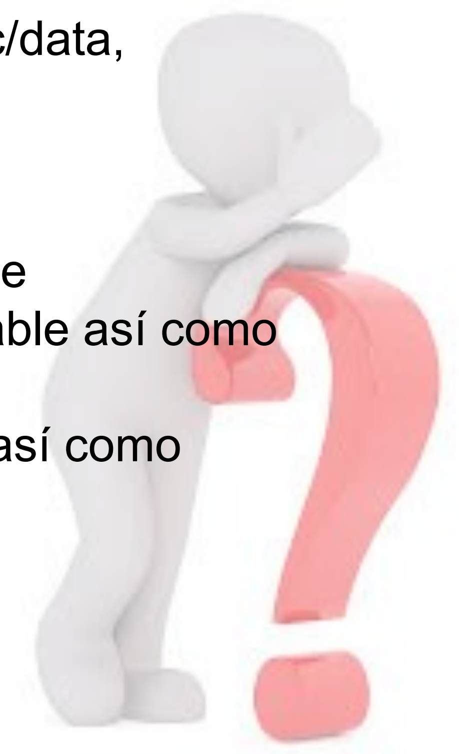
- ➢ r = -1, existe una correlación negativa perfecta.

Estadística descriptiva

Actividad a desarrollar - python

A partir del conjunto de datos de https://www.kaggle.com/c/titanic/data,

1. Analiza tipo de variables que contiene
2. Genera tablas de frecuencias para cada variable
3. Obtén las medidas de posición principales para cada variable
4. Obtén las medidas de dispersión principales para cada variable así como representación de las mismas.
5. Obtén las medidas de forma principales para cada variable así como representación de las mismas.
6. Analiza las posibles correlaciones entre variables.



Distribuciones de probabilidad

Definición

La ***Función de probabilidad*** de una variable aleatoria, es una función que asigna a cada suceso la probabilidad de que dicho suceso ocurra.

P[X=x]

Por ejemplo:
Definimos la v.a. X como resultado de lanzar un dado.
Esta vble puede tomar los valores de 1 a 6
(v.a. discreta)

x	P[X=x]
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Distribuciones de probabilidad

Definición

La ***Función de distribución*** de una variable aleatoria, es una función definida sobre $\mathbb{R}$ cuyo valor en cada x es la probabilidad de que la v.a. sea menor o igual que x.

$$\boxed{F(x) = p(X \le x)} = \begin{cases} \sum_{x_i \le x} f(x_i), & si\ es\ una\ v.a.\ discreta \\ \int_{-\infty}^{x} f(t)dt, & si\ es\ una\ v.a.\ continua \end{cases}$$

Por ejemplo:
Definimos la v.a. X como resultado de lanzar un dado.
Esta vble puede tomar los valores de 1 a 6
(v.a. discreta)

x	P[X=x]	P[X<=x]
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6=1

Distribuciones de probabilidad

Principales

Distribution	Probability Function	Moment-Generating Function	Mean	Variance
Discrete uniform	$p(x) = \frac{1}{n}$ $x = 1, 2, \ldots, n$		$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Hyper-geometric	$\frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}}$ $\text{Max}[0, n-(N-N_1)]$ $\leq x \leq \text{Min}(n, N_1)$		$\mu = n\theta$ $\theta = \frac{N_1}{N}$	$\sigma^2 = \frac{N-n}{N-1} n\theta(1-\theta)$ $\theta = \frac{N_1}{N}$
Bernoulli	$\theta^x(1-\theta)^{1-x}$ $x = 0, 1 \quad 0 \leq \theta \leq 1$	$\theta e^t + (1-\theta)$	θ	$\theta(1-\theta)$

Distribuciones de probabilidad

Principales

Distribution	Probability Function	Moment-Generating Function	Mean	Variance
Binomial	$\binom{n}{x}\theta^x 1-\theta^{n-x}$ $x=0,1,\ldots,n;\ 0\leq\theta\leq 1$	$(\theta e^t+(1-\theta))^n$	$n\theta$	$n\theta(1-\theta)$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$ $x=0,1,\ldots;\ \lambda>0$	$e^{\lambda(e^t-1)}$	λ	λ
Uniform	$f(x)=\frac{1}{b-a}$ $a\leq x\leq b$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$ $-\infty<x<\infty, -\infty<\mu<\infty,$ $\sigma>0$	$e^{\mu t+(\sigma^2t^2/2)}$	μ	σ^2
Chi-square	$\frac{1}{2^{n/2}\Gamma(n/2)}w^{n/2-1}e^{-w/2}$ $w\geq 0, n>0$	$(1-2t)^{-n/2}$	n	$2n$
Student-t	$f(t)=\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}$ $\left(1+\frac{t^2}{n}\right)^{(n+1)/2}$ $-\infty<t<\infty$		0	$\frac{n}{n-2}$

Distribuciones de probabilidad

Distribución normal

- O distribución gaussiana, curva de Gauss o campana de Gauss.
- Variables aleatorias continuas.
- Teorema Central del Límite: bajo ciertas condiciones (como pueden ser independientes e idénticamente distribuidas con varianza finita), la suma de un gran número de variables aleatorias se distribuye aproximadamente como una normal.

> Sean $X_1, X_2,..., X_n$ un conjunto de variables aleatorias, independientes e idénticamente distribuidas con media μ y varianza σ^2 (donde $\sigma^2 \in (0, \infty)$).
> Sea $S_n = X_1 + \cdots + X_n$, entonces
> $$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z)$$
> donde $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ de media 0 y desviación 1, es decir las Z_n convergen en distribución a la normal estándar N(0,1) y $\Phi(z)$ función de distribución de N(0,1).

- Es muy cómoda y fácil de manipular matemáticamente
- Satisface la propiedad de reproductividad
- Depende de dos parámetros, la media y la varianza.
- No debemos abusar de sus beneficios

Distribuciones de probabilidad
Distribución normal

KSCHOOL

- Distribución más utilizada en estadística, por su aplicación.
- Permite modelar numerosos fenómenos naturales, sociales, psicológicos, etc.
- Ejemplos:
    - características morfológicas de individuos, como la estatura;
    - características sociológicas, como el consumo de cierto producto por un mismo grupo de individuos;
    - características psicológicas, como el cociente intelectual;
    - nivel de ruido en telecomunicaciones;
    - errores cometidos al medir ciertas magnitudes;
    - etc.

Distribuciones de probabilidad

Distribución normal

$$X \sim N(\mu, \sigma)$$

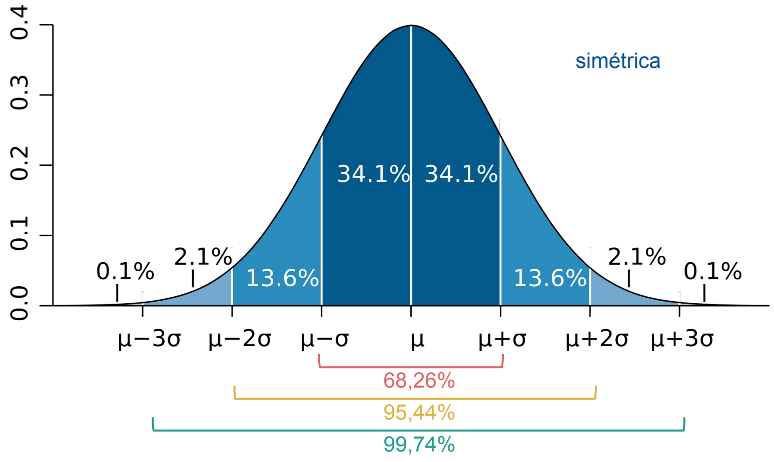
Función de densidad:

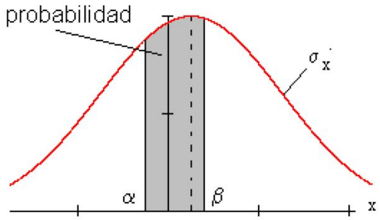
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

Función de distribución:

$$P(\alpha \leq x \leq \beta) = \int_{\alpha}^{\beta} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Distribución de probabilidad alrededor de la media:





Distribuciones de probabilidad

Distribución normal

$$X \sim N(\mu, \sigma)$$

Definiendo:

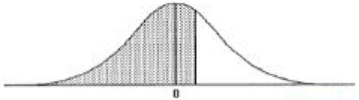
$$Z = \frac{X - \mu}{\sigma}$$

se tiene que Z ~ N(0,1).

Y para esta distribución existen tablas para obtener cualquier P(Z<z):

TABLA-T3: DISTRIBUCIÓN NORMAL ESTANDAR

$$Z \approx N(\mu = 0; \sigma^2 = 1)$$



$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t/2}\, dt$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327

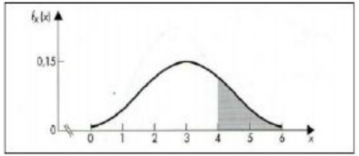
$$P(Z \leq 0{,}34) = 0{,}63307$$

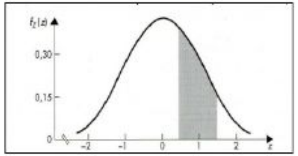
$$P(Z \leq 1{,}36) = 0{,}91308$$

Ejemplo:

Si X ~ N(3,2). Calcular la probabilidad de que tome un valor entre 4 y 6.

$$X \to N(3, 2) \qquad Z \to N(0, 1)$$





$$P(4 < X < 6) = P\left(\frac{4-3}{2} < \frac{X-\mu}{\sigma} < \frac{6-3}{2}\right) = P(0{,}5 < Z < 1{,}5) =$$

$$= P(Z < 1{,}5) - P(Z < 0{,}5) = 0{,}9332 - 0{,}6915 = 0{,}2417$$

Pruebas de normalidad

KSCHOOL

- ★ Analizando el histograma de frecuencias y las medidas de forma (asimetría y curtosis).
- ★ Analizando el gráfico box-plot.
- ★ Analizando las gráficas Q-Q-plot (Quantile-Quantile-Plot).
- ★ Contrastes de normalidad: Kolmogorov-Smirnov, Shapiro-Wilk, Chi-Cuadrado.

Pruebas de normalidad

Quantile-Quantile-Plot

Técnica gráfica para determinar si dos conjuntos provienen de poblaciones con distribución común.

Un ejemplo es la prueba de no-normalidad de la distribución de una variable.

Gráfica de los cuantiles del primer conjunto de datos contra los cuantiles del segundo.

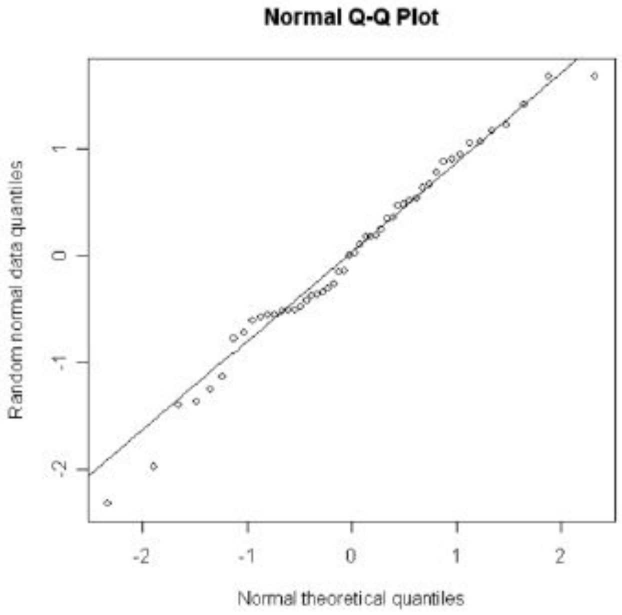
Línea de referencia de 45 grados.
Misma distribución, entonces se aproximan a la recta.

Ubicación, escala, simetría y presencia de valores atípicos detectados desde el qqplot.

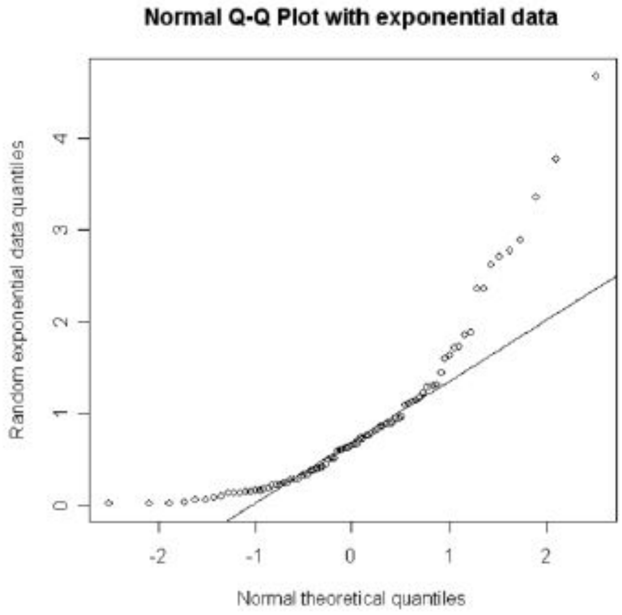
Pruebas de normalidad

Quantile-Quantile-Plot

Un gráfico Q-Q normal de datos N(0,1) generados aleatoriamente.



Un gráfico Q-Q normal de datos exp(1) generados aleatoriamente.



Estimación

Puntual vs Intervalos de confianza

KSCHOOL

Una estimación es **puntual** cuando se usa un solo valor extraído de la muestra para estimar el parámetro desconocido de la población. Al valor usado se le llama estimador.

- ★ La media de la población se puede estimar puntualmente mediante la media de la muestra: $\bar{x} = \mu$
- ★ La proporción de la población se puede estimar puntualmente mediante la proporción de la muestra: $\hat{p} = p$
- ★ La desviación típica de la población se puede estimar puntualmente mediante la desviación típica de la muestra, aunque hay mejores estimadores: $s = \sigma$

A veces es conveniente obtener unos límites entre los cuales se encuentre el parámetro con un cierto **nivel de confianza**, en este caso hablamos de estimación **por intervalos**.

El **nivel de confianza** ($1 - \alpha$) es la probabilidad de que el intervalo de confianza contenga el parámetro estimado.
** De cada 100 intervalos construidos a partir de 100 muestras, 100*(1−α)% deberían contener al verdadero valor del parámetro.*

Estimación. Distribución Normal

Estimación Puntual

★ Como vimos, la distribución normal depende de dos parámetros, la media μ y la desviación estandar σ.

En este caso, los estimadores, por el método de máxima verosimilitud (EMV), son

$$\mu = \bar{x} = \frac{\sum x_i}{n}, \quad \sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Para cada tipo de distribución existen diferentes estimadores.

Estimación. Distribución Normal

Estimación por intervalos de confianza

KSCHOOL

★ Un intervalo de confianza es un intervalo de números que contiene los valores más plausibles para nuestro parámetro de población.

★ Proporcionan el valor de un estadístico mediante un intervalo, bajo una confianza: $(\bar{X} - Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + Z\alpha_{/2} \cdot \frac{\sigma}{\sqrt{n}})$

Nivel de confianza ($1-\alpha$): 95%

Calculamos la probabilidad:

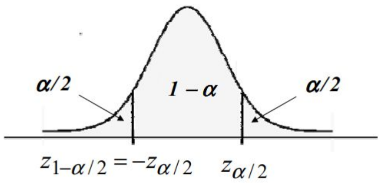
$$p = \frac{1+NC}{2} = \frac{1+0.95}{2} = 0.975$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808

$Z\alpha_{/2} = 1.96$

IC al 95%:

$(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}})$



Estimación. Distribución Normal

Pruebas de Hipótesis

★ Afirmación acerca del valor que puede tomar el parámetro de la población bajo estudio.

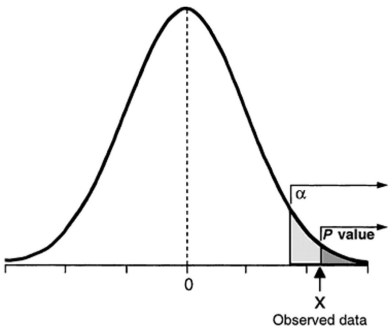
★ Basada en alguna creencia o experiencia pasada.

★ Contrastada con la evidencia de la muestra.α

★ Elementos:

- H_0 : hipótesis nula, supuesta cierta de partida.
- H_1 : hipótesis alternativa.
- Significatividad.
- p-valor.
- estadístico de prueba.
- Objetivo: aceptación o no de H_0

Máxima probabilidad de equivocarnos que estamos dispuestos a asumir en caso de rechazar H0
Suele ser 0.05.



Probabilidad de error en que incurriríamos en caso de rechazar H_0 con los datos de que disponemos. Cuantifica el riesgo que hay que asumir si queremos rechazar H_0 .

$$P - valor < \alpha \Rightarrow \ Rechazamos\ H_0$$

$$P - valor > \alpha \Rightarrow \ Aceptamos\ H_0$$

Estimación. Distribución Normal

Pruebas de Hipótesis

Tipos de errores:

	H_0 Verdadera	H_0 Falsa
Rechazamos H_0	Error Tipo I P(error Tipo I) = α	Decisión Correcta
No Rechazamos H_0	Decisión Correcta	Error Tipo II P(error Tipo II) = β

Documentación

- Determinación del tamaño muestral. Pita Fernández, S.Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Universitario de A Coruña. Recuperado de: https://www.fisterra.com/mbe/investiga/9muestras/9muestras2.asp
- Funciones Python vs R estadística. Recuperado de: https://rpubs.com/rparra/438555
- Documentación Pandas. Recuperado de: https://pandas.pydata.org/pandas-docs/stable/reference/frame.html
- Variables aleatorias y sus momentos. Recuperado de: http://halweb.uc3m.es/esp/Personal/personas/mwiper/docencia/Spanish/Teoria_Est_El/tema5_orig.pdf
- https://bookdown.org/aquintela/EBE/inferencia-estadistica.html
- Documentación scipy. Recuperado de https://pybonacci.org/2012/04/21/estadistica-en-python-con-scipy/
- Distribución normal. Recuperado de: https://www.uv.es/ceaces/pdf/normal.pdf
- Distribuciones de probabilidad en python. Recuperado de https://www.math.purdue.edu/~lin491/ME597/lec_03.pdf
- Tabla valores distribución normal. Recuperado de: https://ematecs.com/tabla-de-probabilidades-de-la-distribucion-normal/ o https://www.um.es/documents/877924/4630870/Mayores2018+Mat+Apli+CCSS+-+tabla_de_la_distribucion_normal3-1.pdf/fdcdf99d-b6d6-49c8-82af-eded690dbf4f
- Ejemplos funciones normales python. Recuperado de: https://www.programcreek.com/python/example/103629/scipy.stats.norm.ppf
- Distribuciones normales de probabilidad. Recuperado de: https://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm
- Estimación. Recuperado de: http://www4.ujaen.es/~dmontoro/Metodos/Temas/Tema7.pdf
- Estimación puntual y por intervalos de confianza. http://matematicas.unex.es/~mota/ciencias_ambientales/tema5.pdf
- Aplicación EMV. https://tereom.github.io/est-computacional-2018/maxima-verosimilitud.html
- EMV. http://benasque.org/benasque/2005tae/2005tae-talks/233s6.pdf
- Contraste de hipótesis con ejemplos. Recuperado de https://www.ucm.es/data/cont/docs/518-2013-11-13-tests.pdf


FEEDBACK

¡Gracias!

irenetorresvalle@gmail.com