Curso de

Estadística Inferencial para Data Science e Inteligencia Artificial

Sílvia Ariza Sentís

🔍 ¿Para quién es este curso?

📊 Estudiantes que saben de análisis descriptivo en Python y quieren explorar el ámbito inferencial.

Estadística descriptiva vs inferencial

Estadística descriptiva vs inferencial

DESCRIPTIVA

Parte de la estadística que arregla los datos de forma que puedan ser analizados e interpretados.

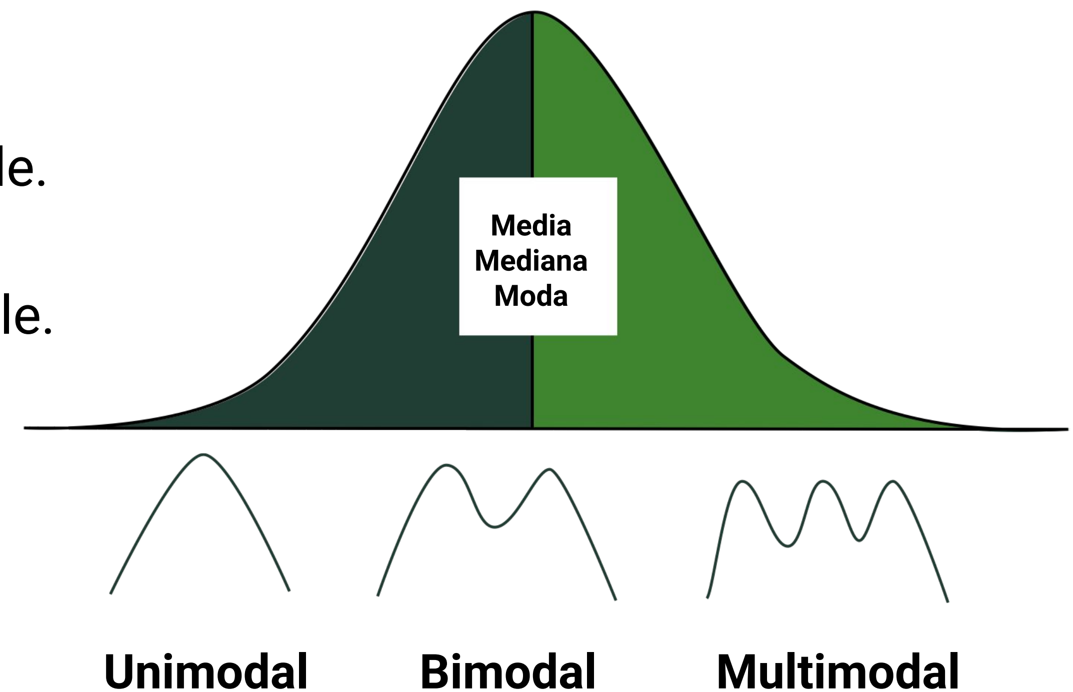
INFERENCIAL

Parte de la estadística que busca predecir o deducir características o resultados esperados de una población, basados en los datos obtenidos de una muestra de esa población.

Estadística descriptiva

Nos ayuda a determinar:

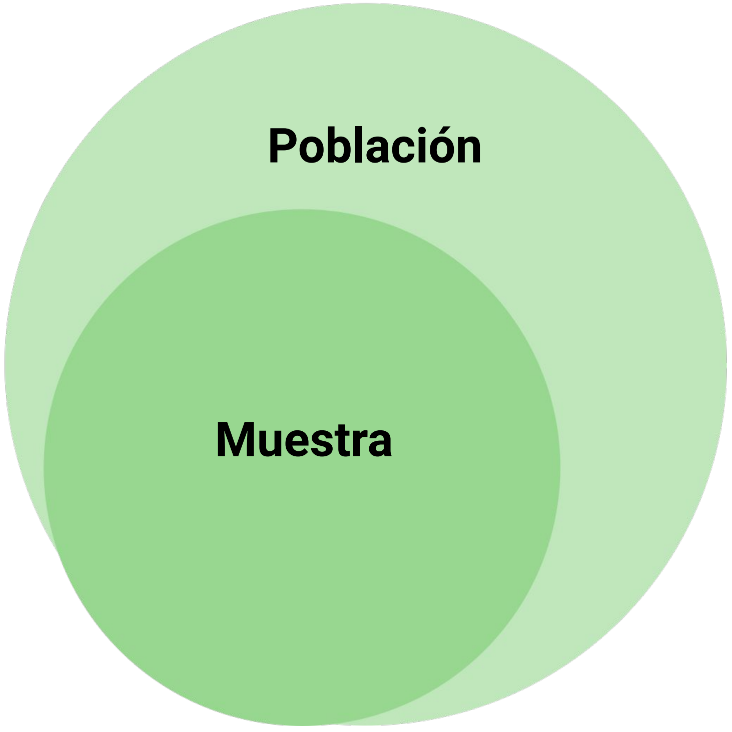
- La **tendencia** central de una variable.
- La **variabilidad** de una variable.
- La **distribución** de una variable.



Estadística inferencial

Nos ayuda a determinar:

- Muestreo
- Intervalos de confianza
- Validación de hipótesis
- Evitar sesgos



Inferencia estadística

- Conclusiones que se obtienen sobre los parámetros de la población de datos.

- Estudio del grado de fiabilidad de los resultados extraídos del estudio.

Uso en data science y machine learning

Tanto en un análisis como en un modelo predictivo, la estadística inferencial servirá para:

- Entender la **distribución** de nuestros datos.
- Crear y validar **hipótesis**.
- Hacer **experimentos**.
- Elegir los **modelos predictivos** adecuados según los datos.

Estadísticos principales

Experimento

Procedimiento que puede repetirse infinitamente y tiene un conjunto bien definido de resultados posibles, conocido como espacio muestral.

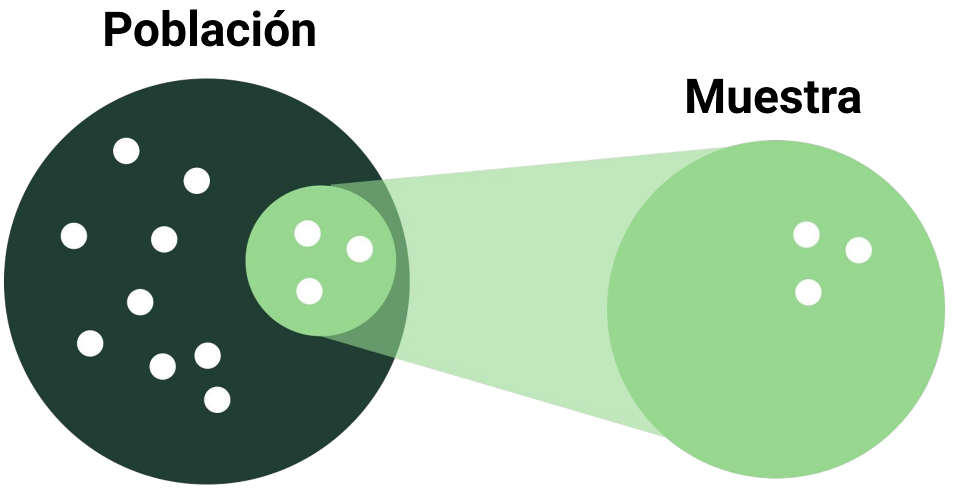
- **Aleatorio**: si tiene más de un resultado posible.
- **Determinista**: si solo tiene un resultado posible.

Población y muestra

Muestra: subconjunto de datos perteneciente a una población.

Condiciones:

- **Número** suficiente de registros para ser estadísticamente significativo.
- Representación **no sesgada** de la información total.



Evento

Cada uno de los posibles resultados de un experimento.

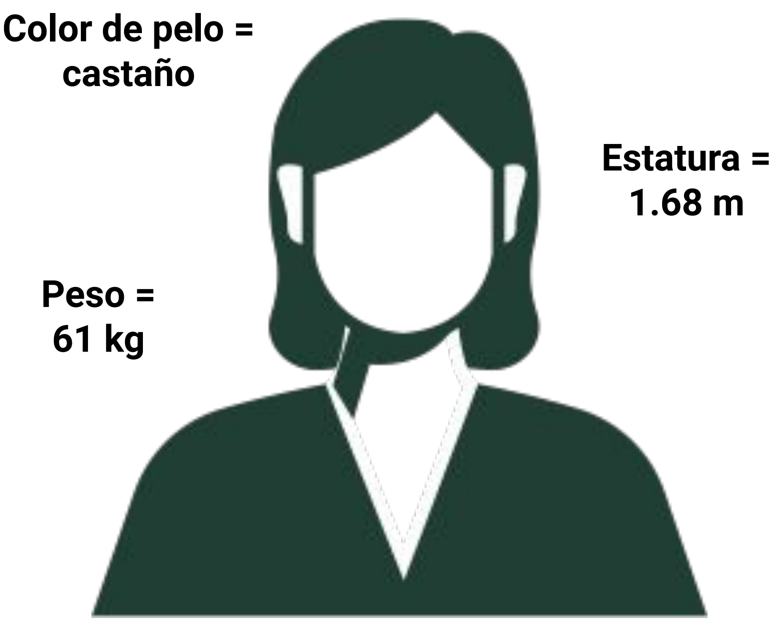
Variable

Es una característica que puede obtener diferentes valores.

Tipos:

- Cualitativas: atributos (no medibles).

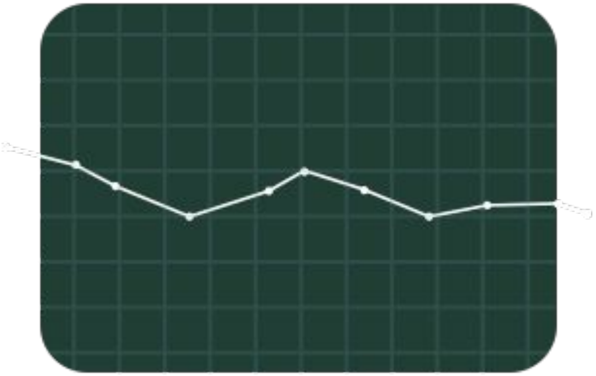
- Cuantitativas: números (medibles).
    - Discretas
    - Continuas

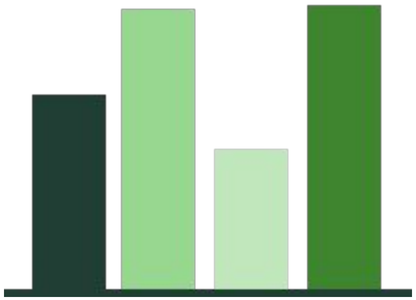


Probabilidad

Mide **qué tan posible** es que ocurra un evento determinado.

El análisis de los eventos probabilísticos se denomina **estadística**.

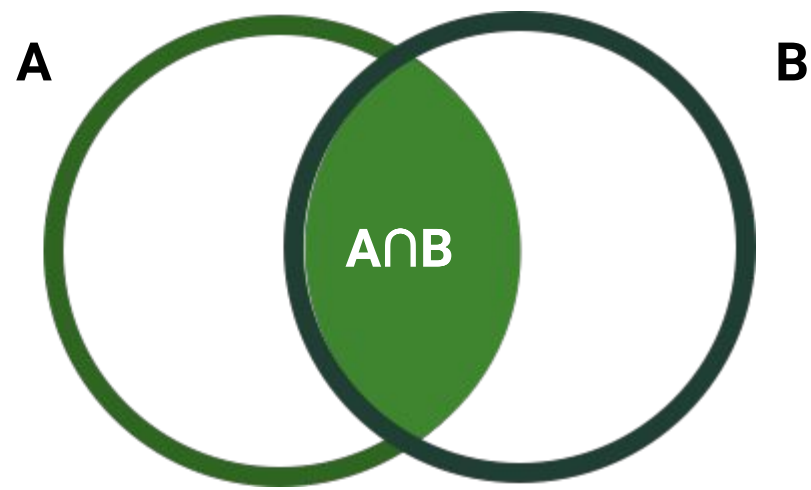




Probabilidad condicionada

Posibilidad de que ocurra un evento como consecuencia de que **otro evento** haya sucedido.

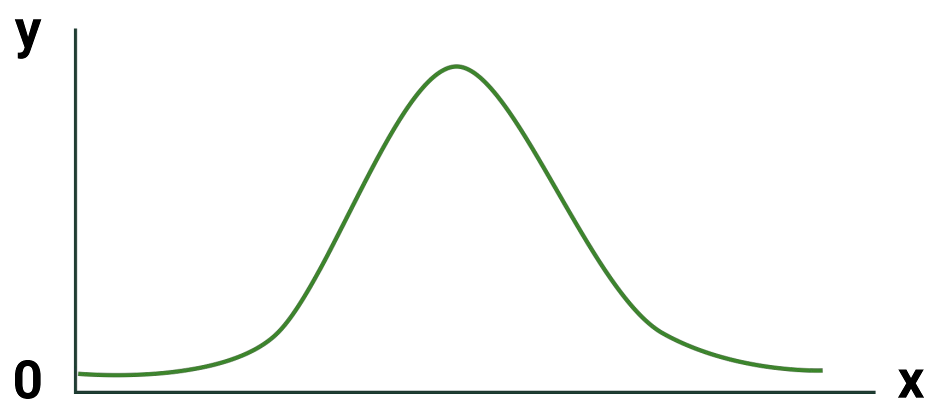
$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$



Poblaciones normales

Distribución normal

- Distribución normal = Distribución de Gauss.
- Su moda = su media = su mediana.
- Es simétrica.
- Tiene forma de campana.



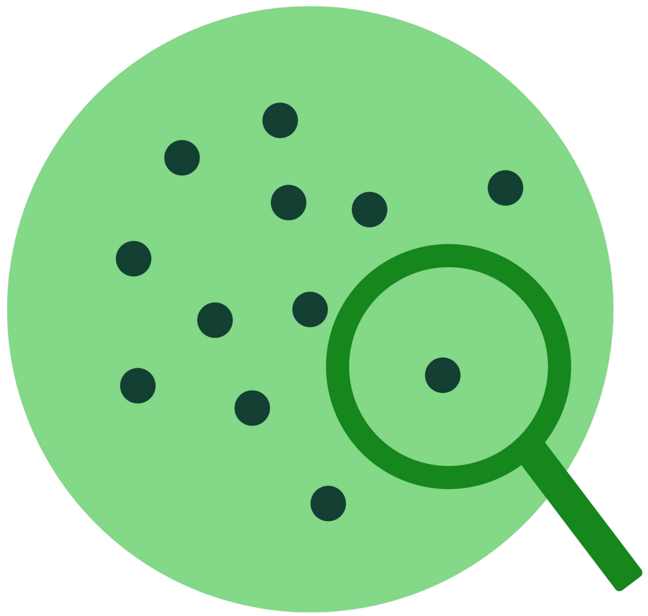
Ejemplos de población normal

- Calorías ingeridas y peso.
- Presión sanguínea.
- Tamaño de los coches producidos por una máquina.

Introducción al muestreo y teorema del límite central

Muestreo

- Técnica para la selección de una muestra.
- Se obtiene a partir de una población estadística.
- La selección tiene que ser aleatoria y se espera que sus propiedades sean extrapolables a la población.

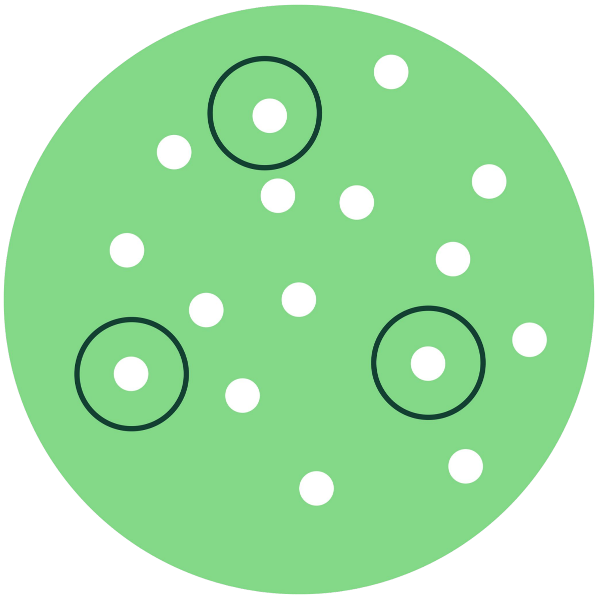


Tipos de muestreo

Aleatorio simple

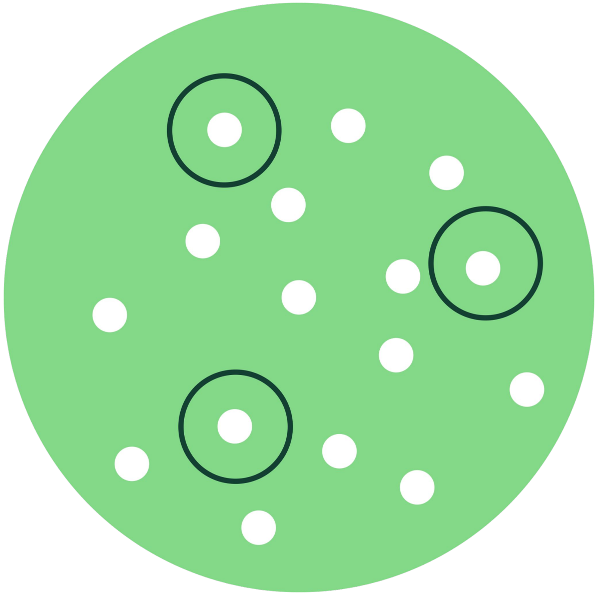
Método de selección de ciertas unidades sacadas de una población de manera que cada una de las muestras tiene la **misma probabilidad** de ser elegida.

Ejemplo: lotería.



Sistemático

Método de selección de ciertas unidades al **azar** y, a continuación, se eligen el resto siguiendo **intervalos regulares**.

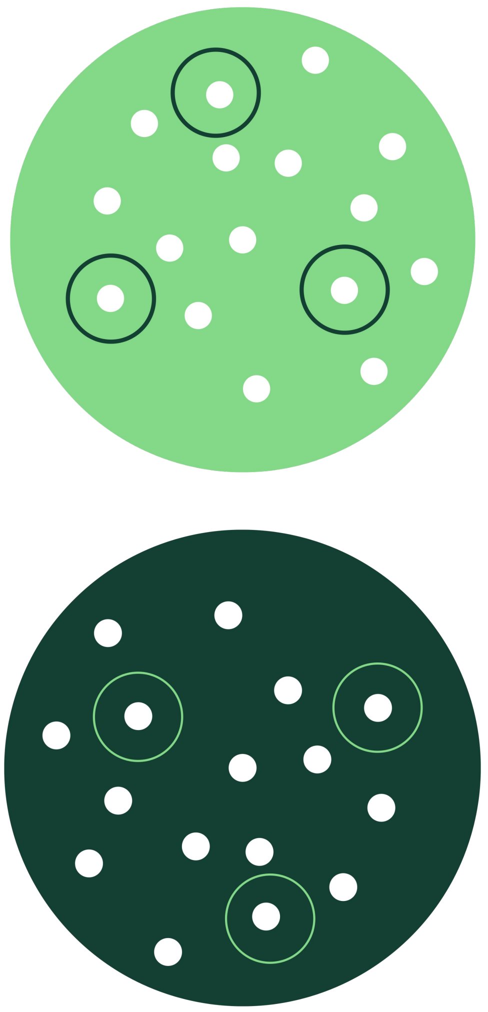


Ejemplo: dar un premio cada cien personas que hagan una inscripción hasta llegar a un total de mil inscritos.

Estratificado

Método de selección de ciertas unidades por segmentos exclusivos y homogéneos y, a continuación, se elige una muestra aleatoria simple de cada segmento.

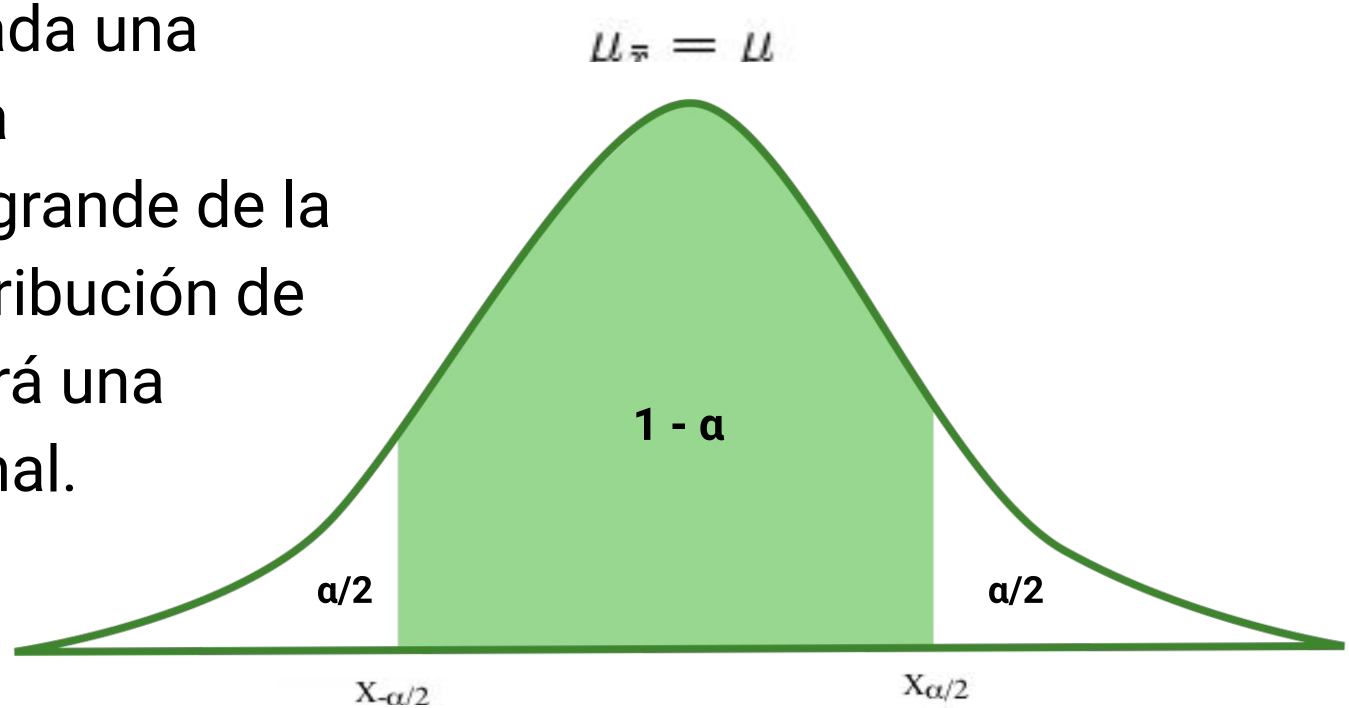
Ejemplo: división por edades.



Teorema del límite central

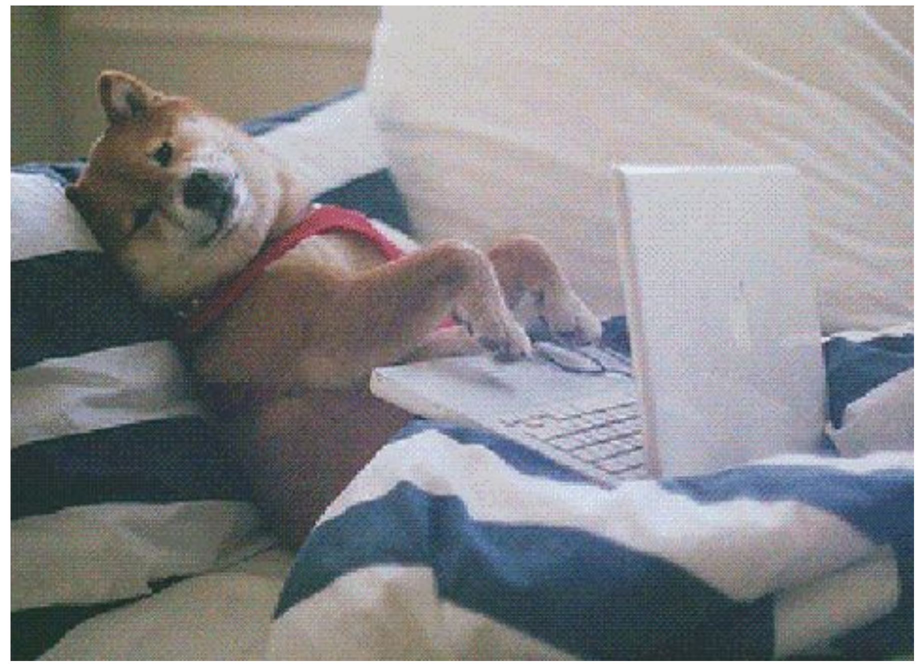
Teorema del límite central

Teoría estadística que establece que, dada una muestra aleatoria suficientemente grande de la población, la distribución de las medias seguirá una distribución normal.



Funciones de muestreo en Python

jupyter



Muestreo estratificado en Python

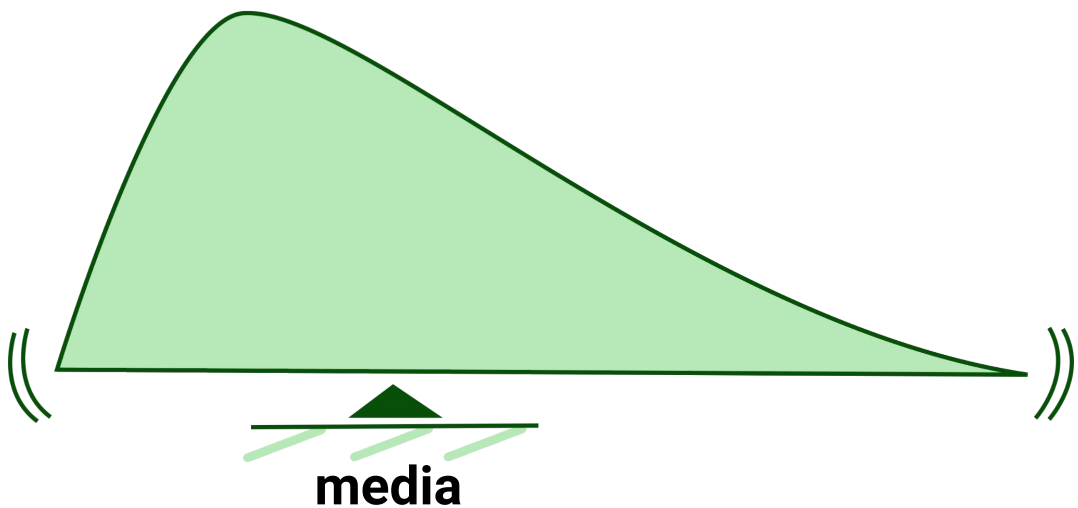
jupyter

La media muestral

Media, moda y mediana

Media

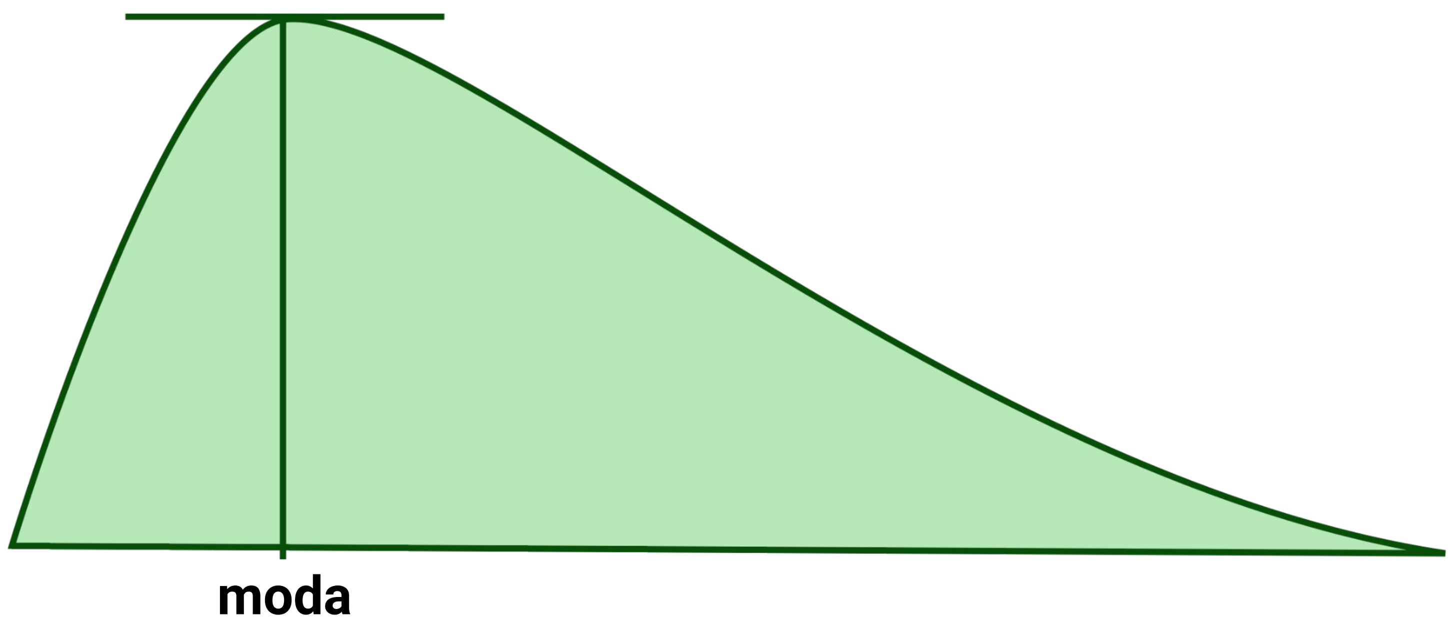
Suma de los datos dividida entre la cantidad de datos.



$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{N}$$

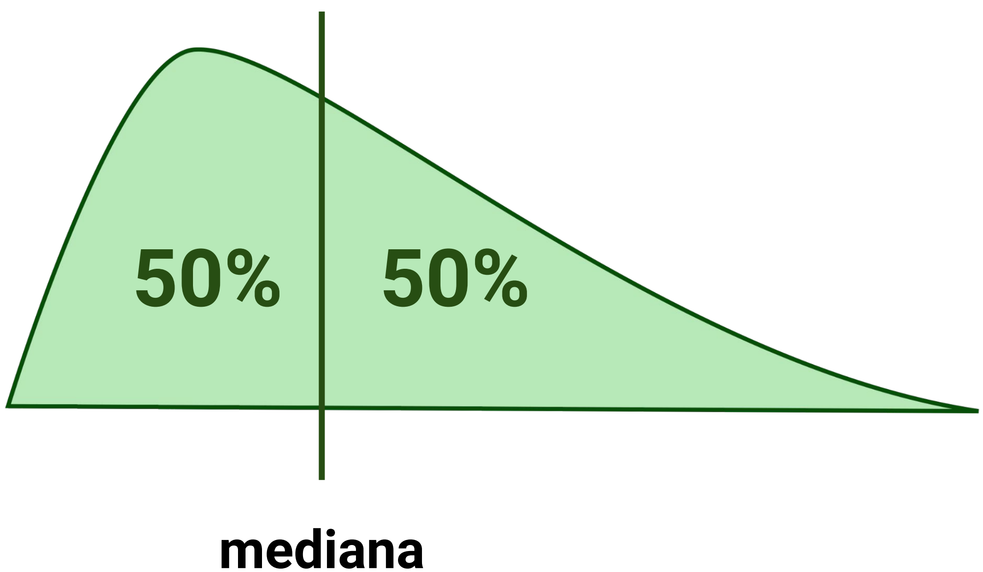
Moda

El dato que más se repite.



Mediana

Es el dato que está en el centro de todos.



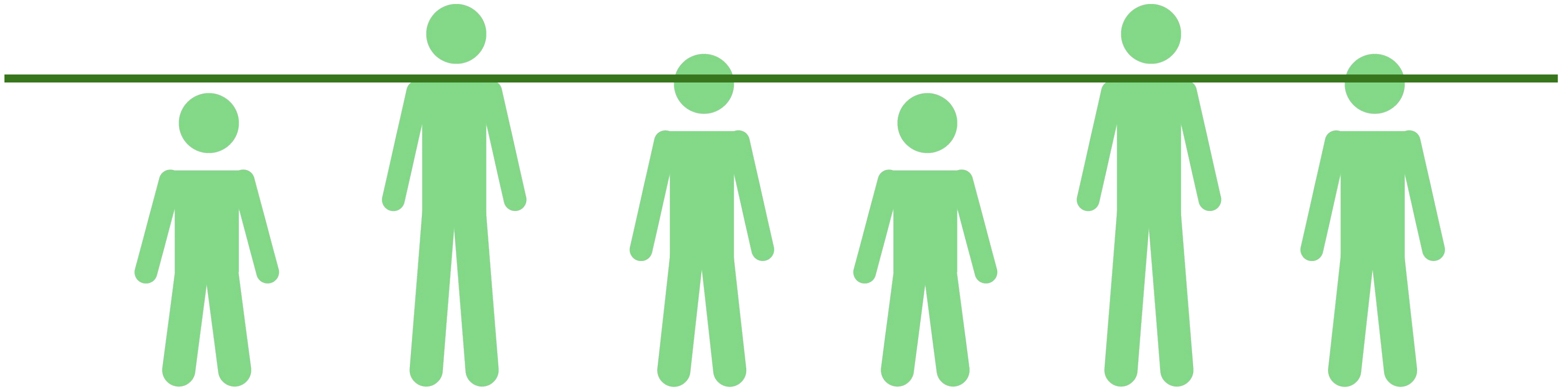
$$x_{\frac{N+1}{2}} \text{ si } N \text{ impar}$$

$$\frac{1}{2} \cdot \left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1} \right) \text{ si } N \text{ par}$$

Media muestral

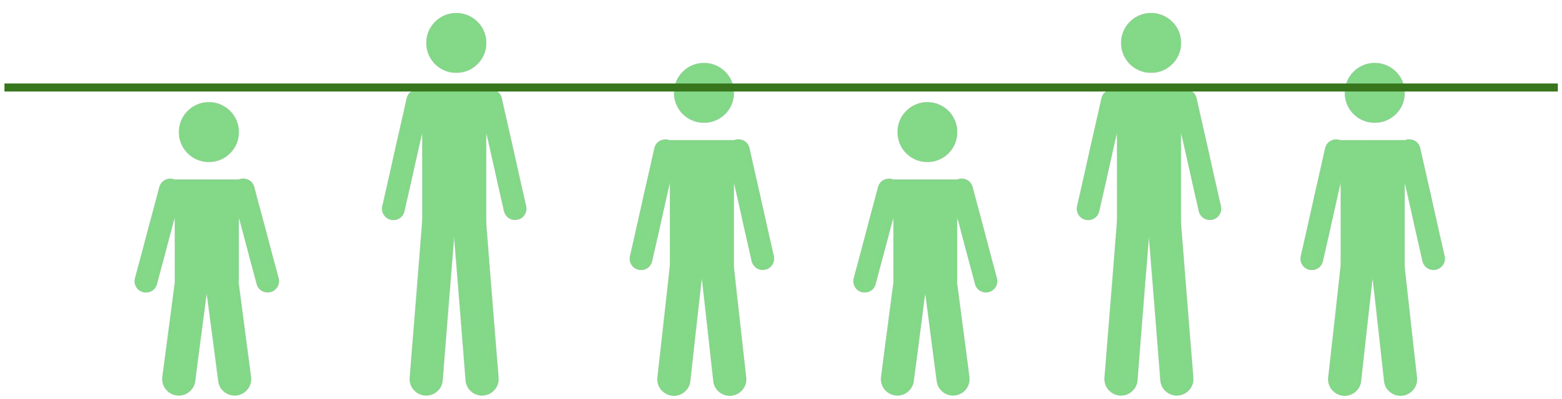
Media muestral

- Media aritmética = promedio = media.
- Valor que se obtiene de sumar un conjunto de valores cuantitativos y dividirlo por el número total de sumados.



Media muestral

- Media muestral $\overline{X}$ $\neq$ media poblacional μ
- **Ejemplo:** estimación puntual de la edad promedio de una población.



Cálculo

$$\bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{N}$$

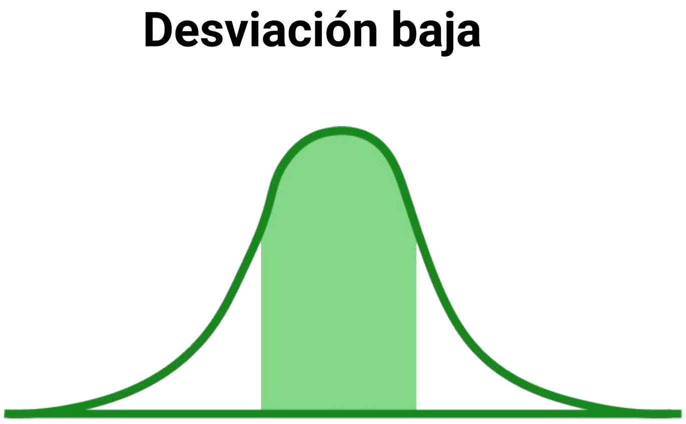
Calcula la edad promedio de una clase donde los estudiantes tienen las siguientes edades: 28, 24, 25, 23, 38, 52.

$$\bar{X} = \frac{28+24+25+23+38+52}{6} = 31.7 \text{ años}$$

Varianza y desviación estándar muestral

Varianza y desviación estándar

- Indica qué tan dispersos están los datos respecto a la media.
- **La desviación estándar es la raíz cuadrada de la varianza.**
- Ejemplo: edades de la población de una ciudad.



Cálculo de la varianza y desviación estándar

Muestral

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

Poblacional

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Desviación estándar de una clase donde los estudiantes tienen las siguientes edades: 28, 24, 25, 23, 38, 52.

Sabemos que la edad promedio es 31.7 años.

Cálculo de la varianza y desviación estándar

Varianza muestral =

$$\frac{(28-31.7)^2+(24-31.7)^2+(28-31.7)^2+(24-31.7)^2+(28-31.7)^2+(24-31.7)^2}{5} = 43.8$$

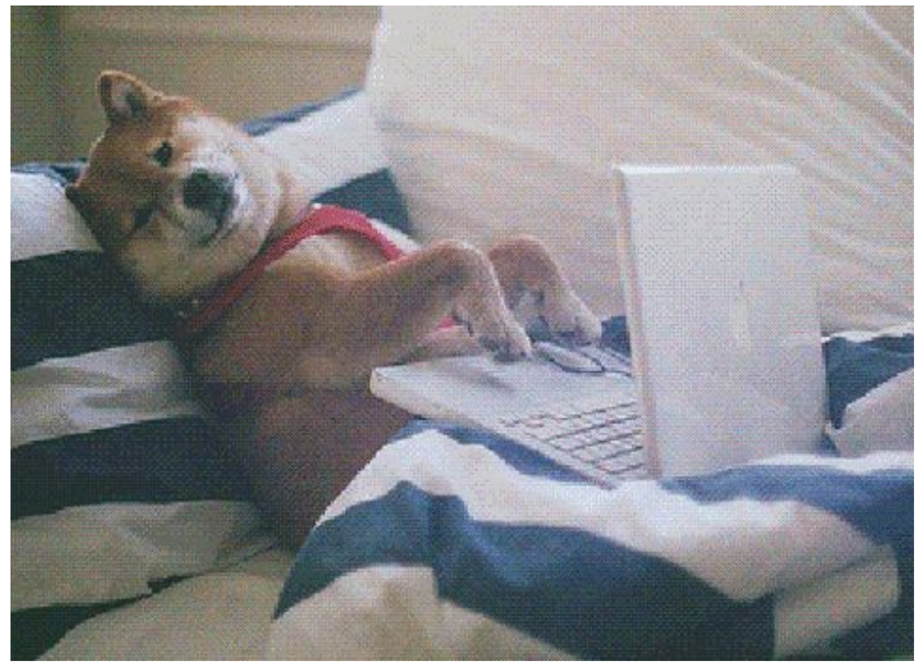
Desviación estándar muestral = $\sqrt{43.8} = 6.62$

Resumen de fórmulas

	Varianza	Desviación estándar	Media
Población	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
Muestra	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Varianza y desviación estándar muestral en Python

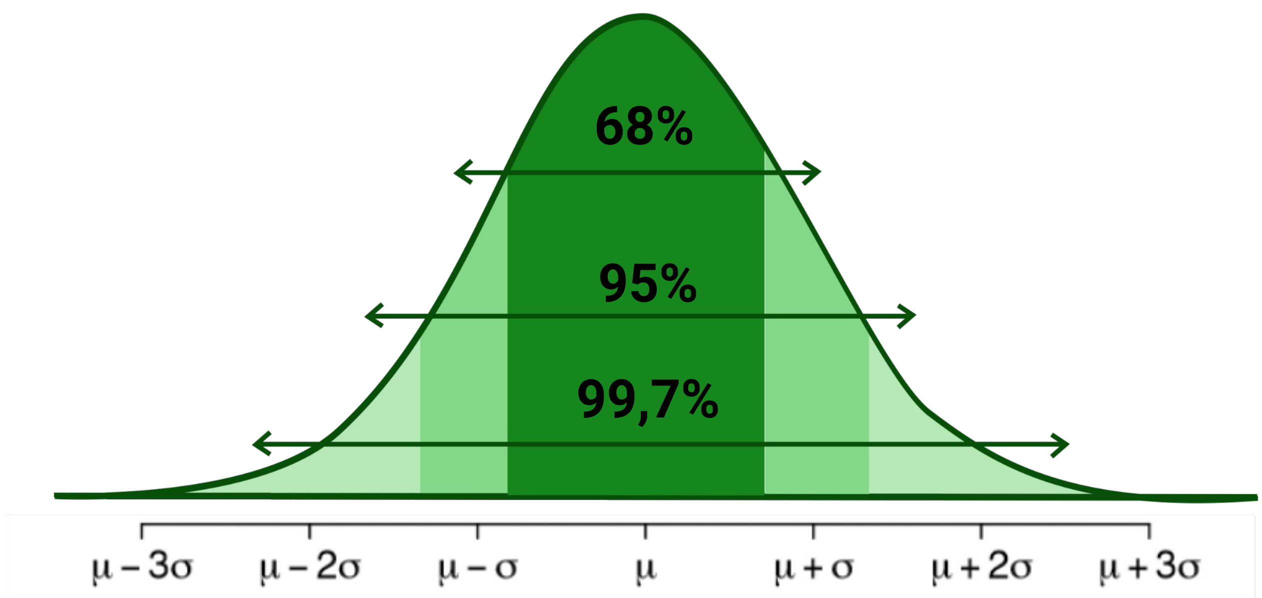
jupyter



Intervalos de confianza

Intervalos de confianza

- Un par o varios pares de números entre los cuales se estima que estará cierto valor desconocido respecto de un parámetro poblacional con un determinado nivel de confianza.
- Son simétricos respecto a la media.



Nivel de significación

- El nivel de significación o alfa es el nivel límite para juzgar si un resultado es o no es estadísticamente significativo.
- Si el valor de significación es menor que el nivel de significación, el resultado es estadísticamente significativo.

Interpretación de un resultado

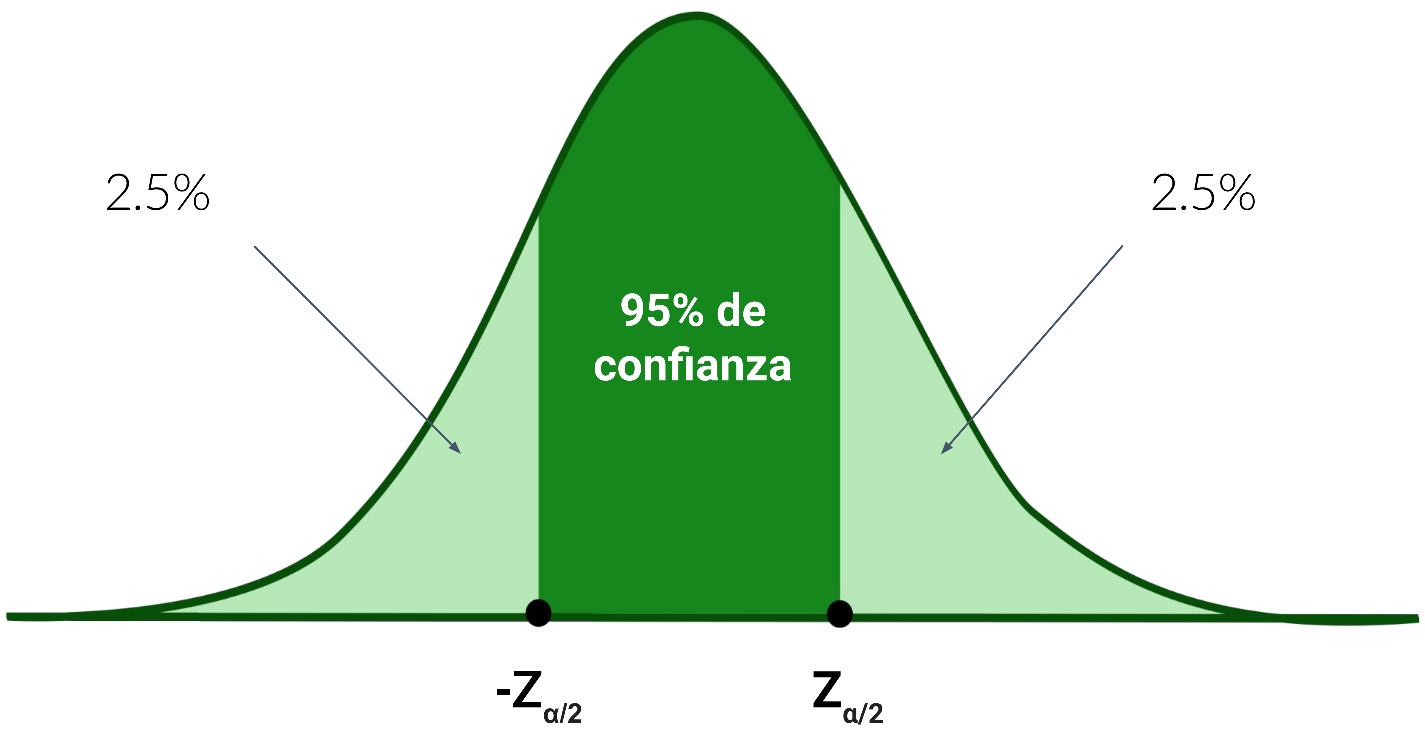
Intervalo de confianza del 95%:

Sabemos que con un 95% de certeza las edades de las personas que esquían están entre dos valores.

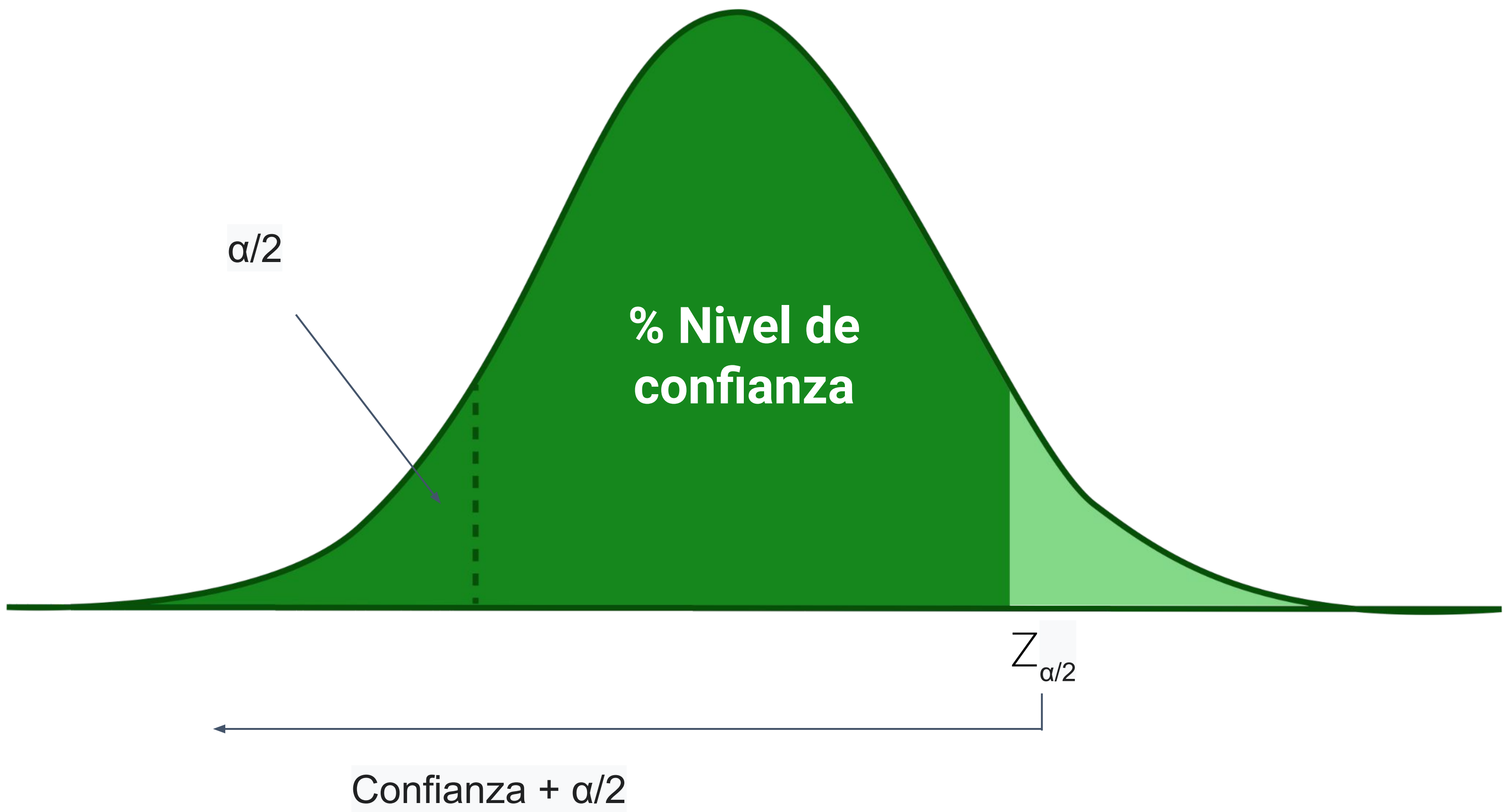
Cálculo de intervalos de confianza

Ejercicio I: nivel de significación

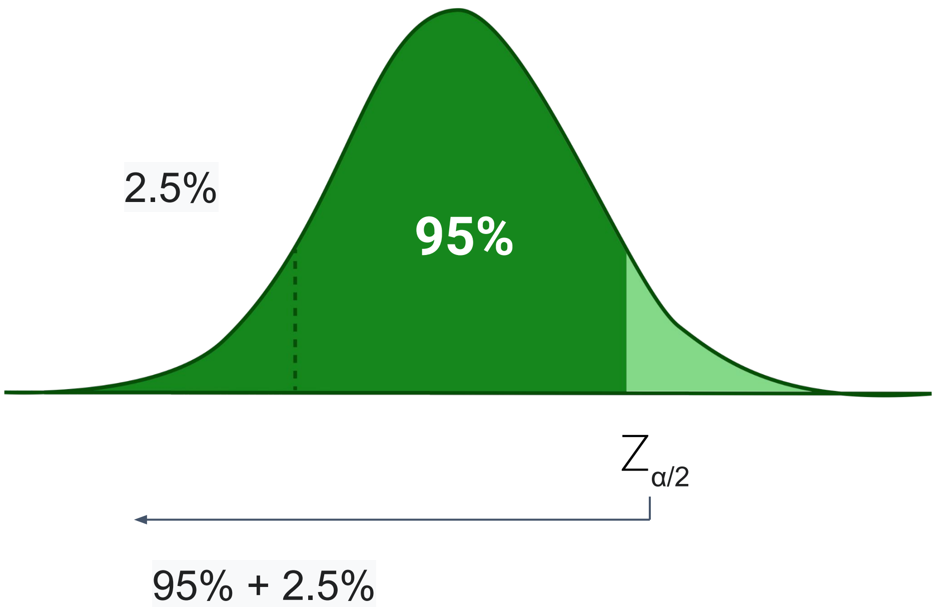
El valor de alfa es del 5%.

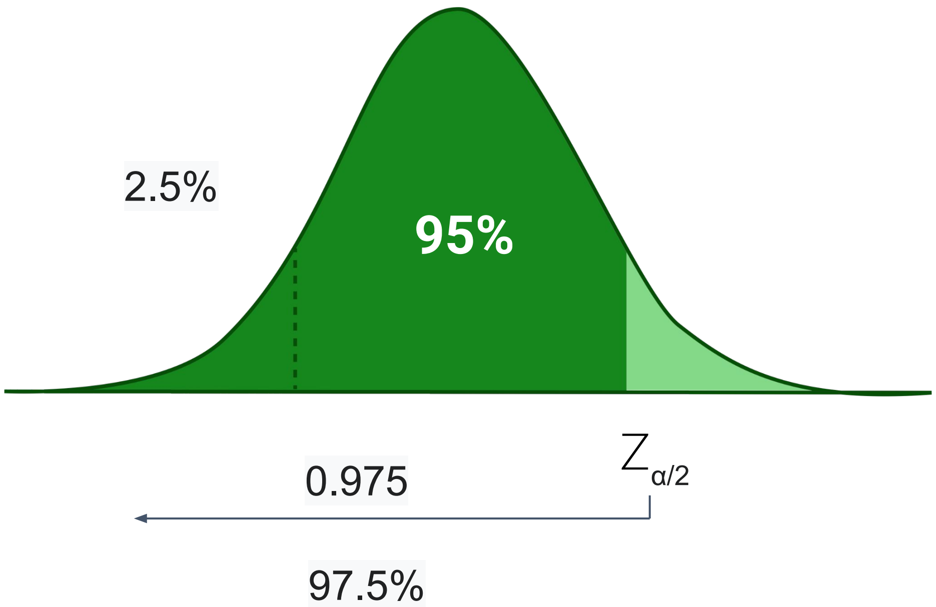


Consideraciones al buscar en la tabla



Ejemplo al 95% de confianza

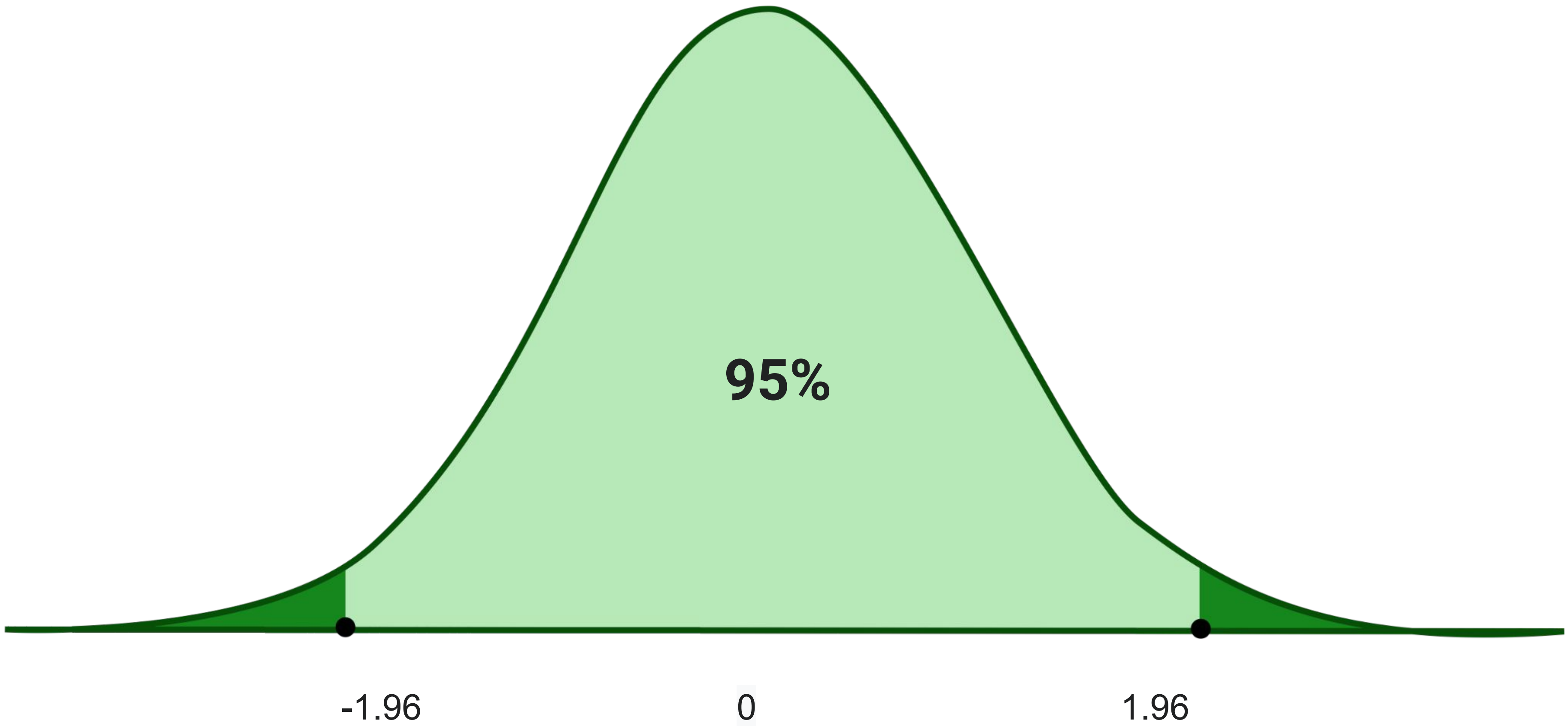




Búsqueda en la tabla

z	0	0.01	0.02	0.03	0.04	0.05	$\boxed{0.06}$	0.07	0.08	0.09
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
+1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
+1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
$\boxed{+1.9}$	.97128	.97193	.97257	.97320	.97381	.97441	$\boxed{.97500}$	.97558	.97615	.97670
+2	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
+2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
+2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899

Resultado



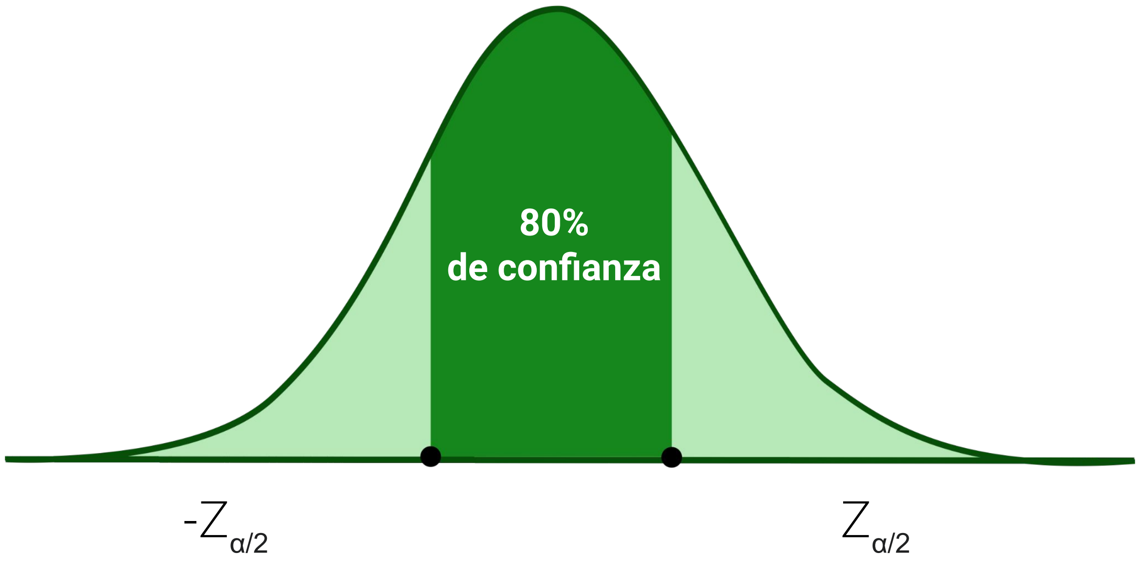
$IC_{95\%} = (-1.96 ; 1.96)$

Ejercicio II

La duración en días de un cepillo de dientes se ajusta a la distribución normal (28, 4). ¿Cuál es el intervalo de confianza al 80%?

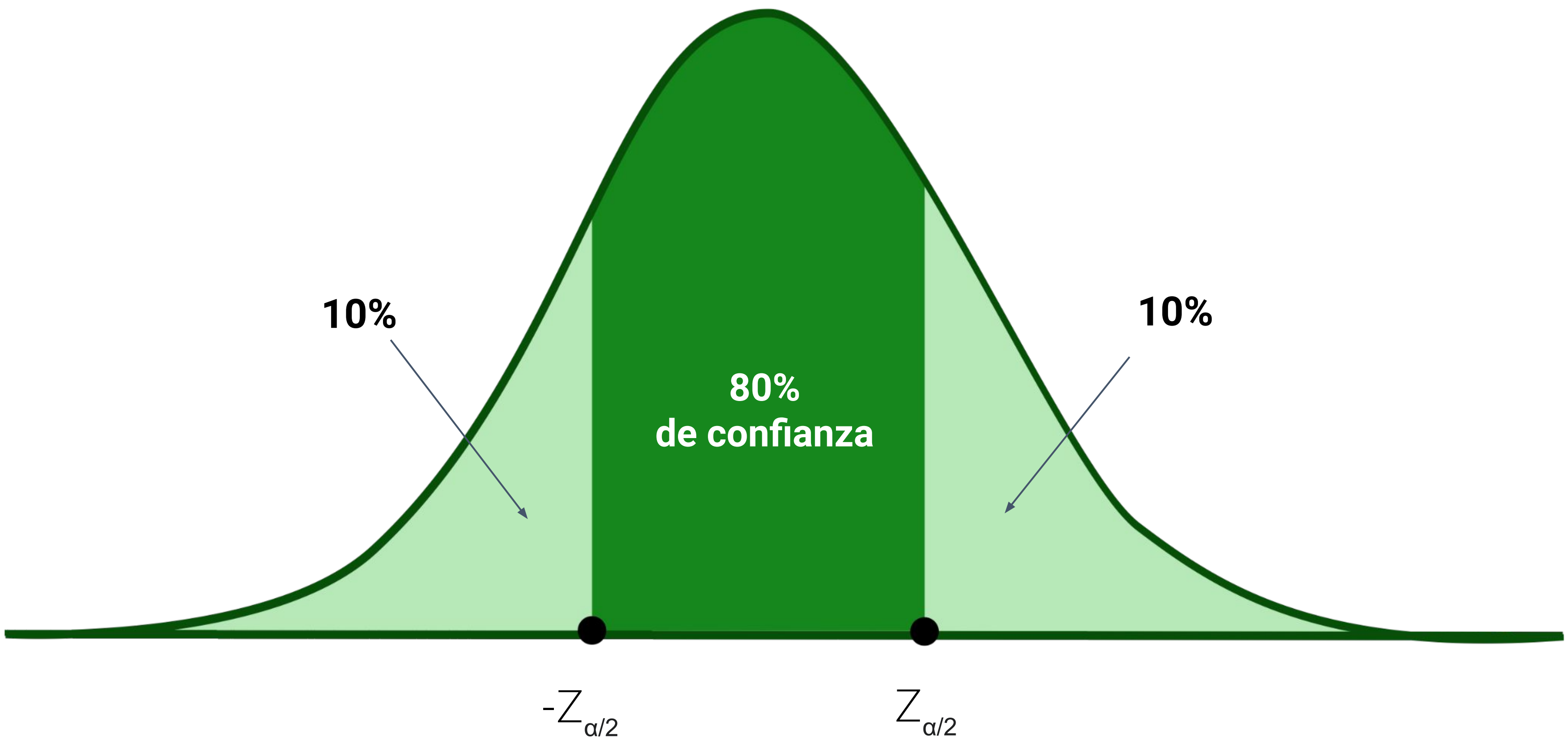
$\mu = 28$

$\sigma = 4$

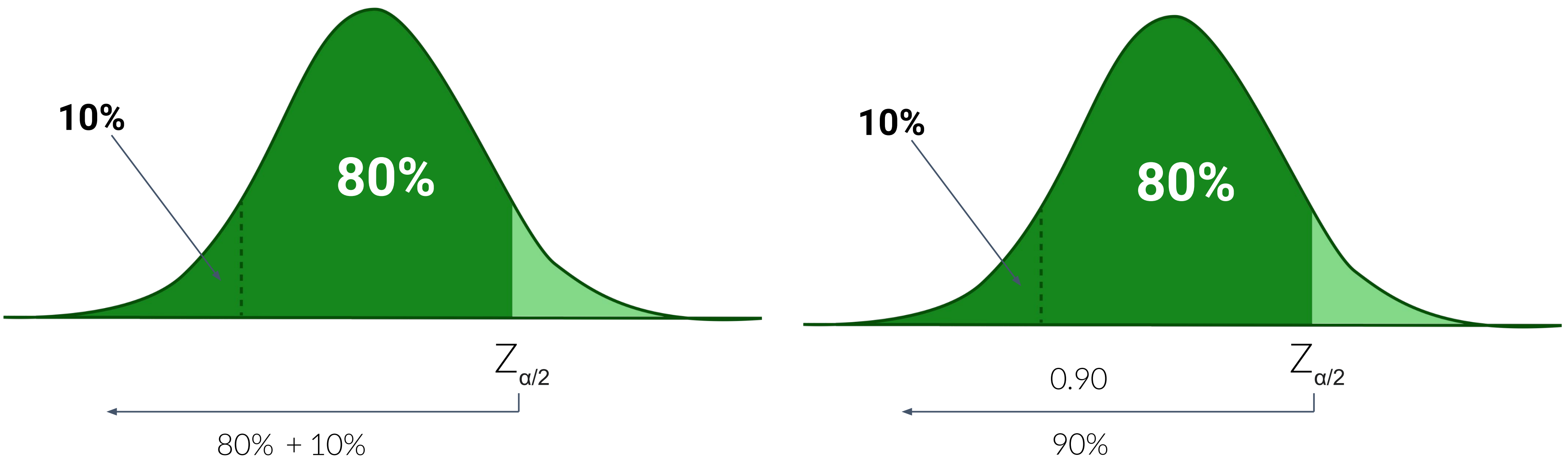


Nivel de significación

El valor de alfa es del 20%.



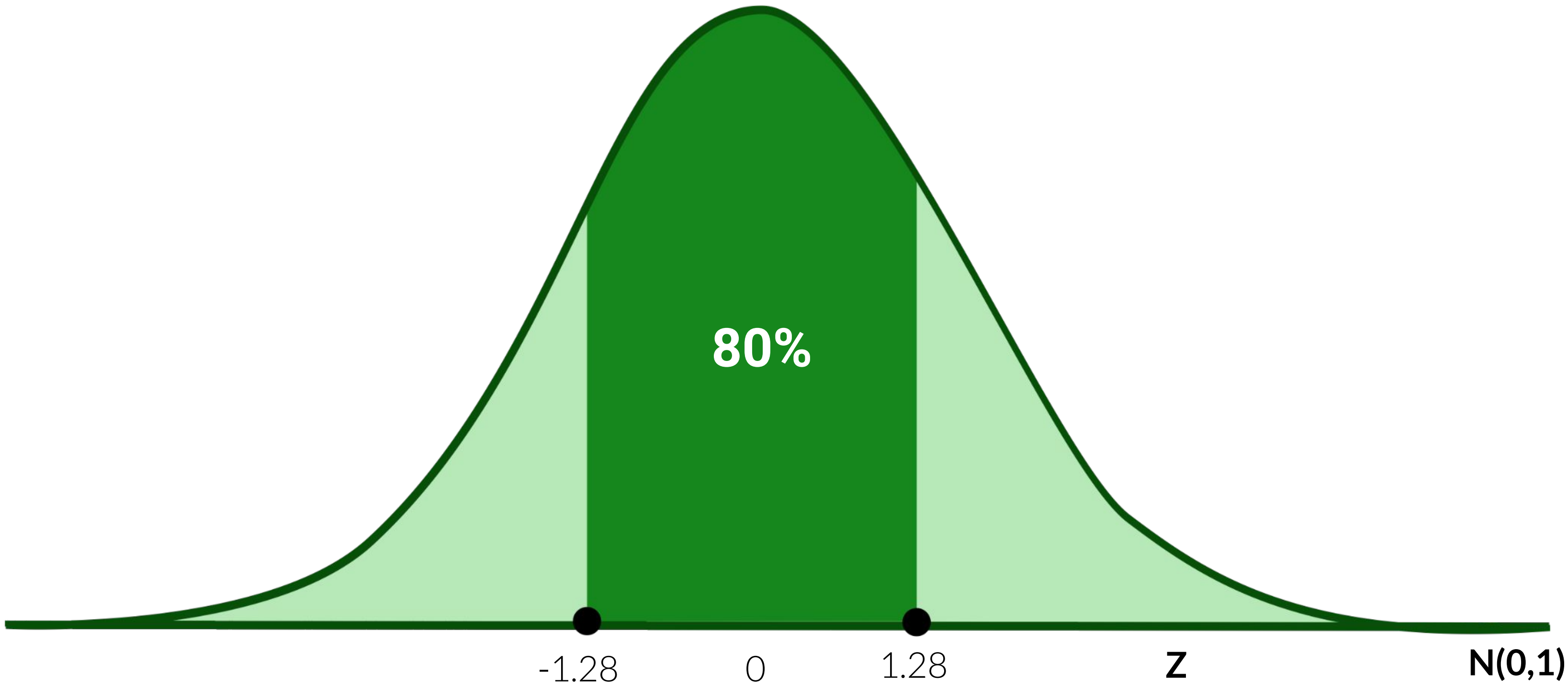
Consideraciones al buscar en la tabla



Búsqueda en la tabla

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
+1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
+1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
+1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
+2	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
+2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
+2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899

Resultado



Este es el caso de media 0, ahora tenemos que convertirlo a la media 28 del ejercicio con la fórmula: $Z = \frac{x - \mu}{\sigma}$

Conversión

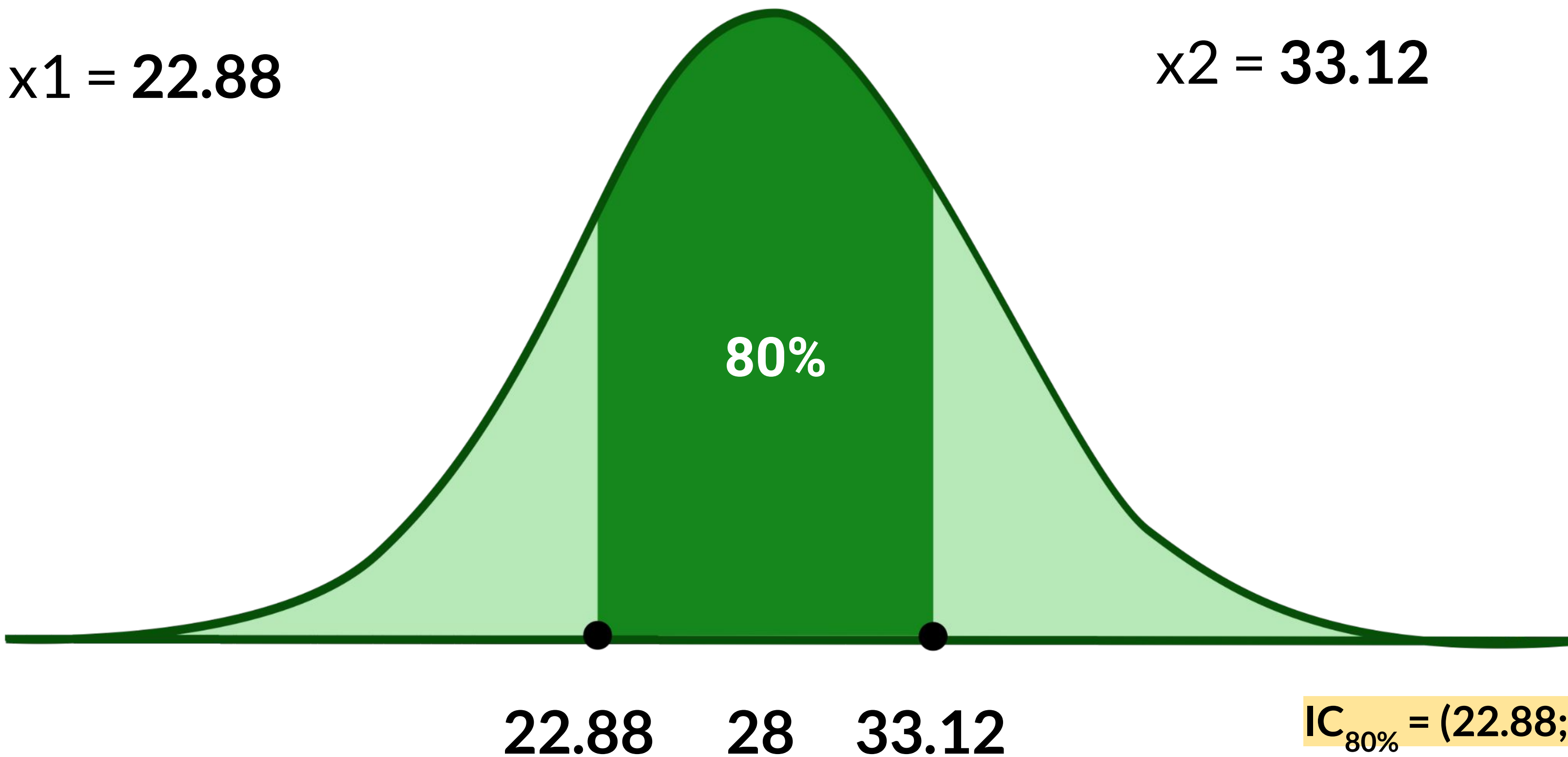
$$Z = \frac{x - \mu}{\sigma}$$

$-1.28 = (x1-28)/4$

$x1 =$ **22.88**

$1.28 = (x2-28)/4$

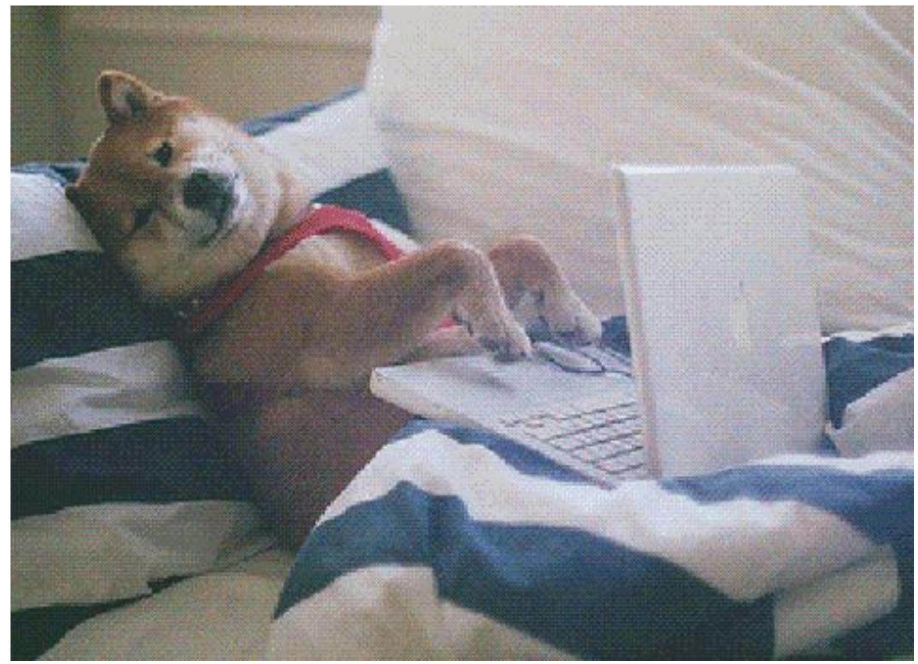
$x2 =$ **33.12**



$IC_{80\%} = (22.88; 33.12)$

Cálculo de intervalo de confianza en Python

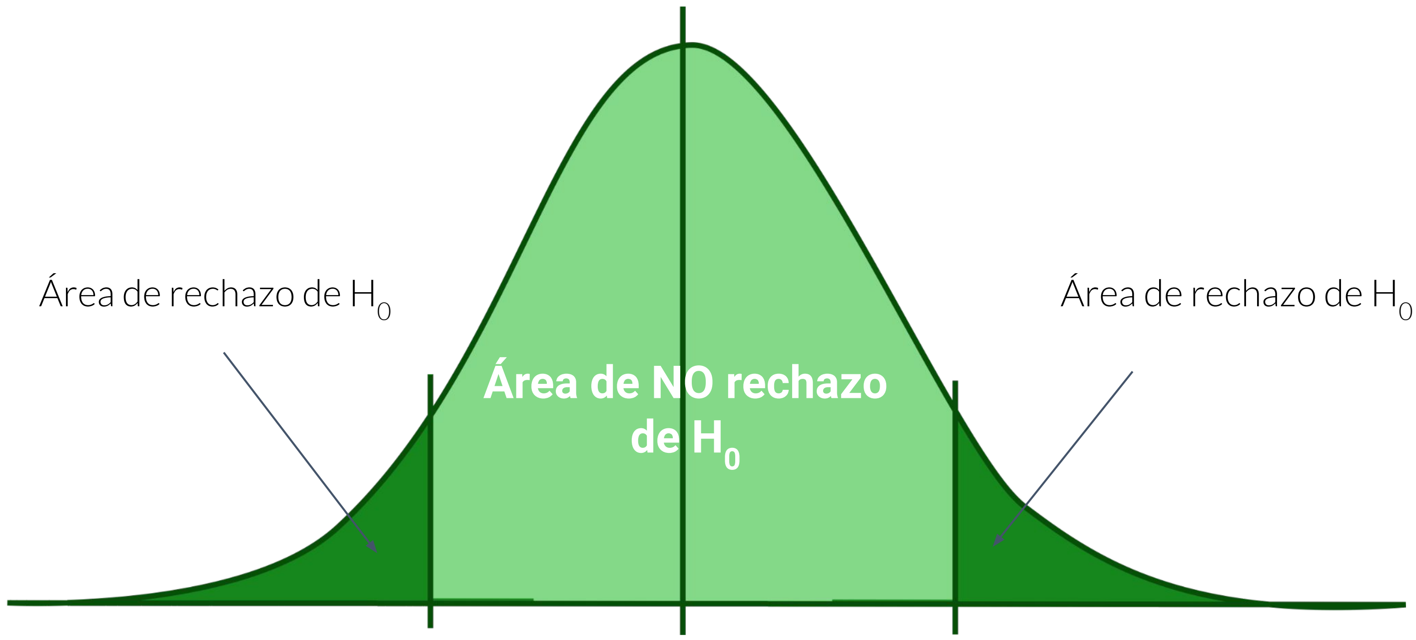
jupyter



Pruebas de hipótesis

Prueba de hipótesis

La prueba de hipótesis o prueba de significación ayuda a juzgar si existe una diferencia significativa entre el tamaño de la muestra y el parámetro general.



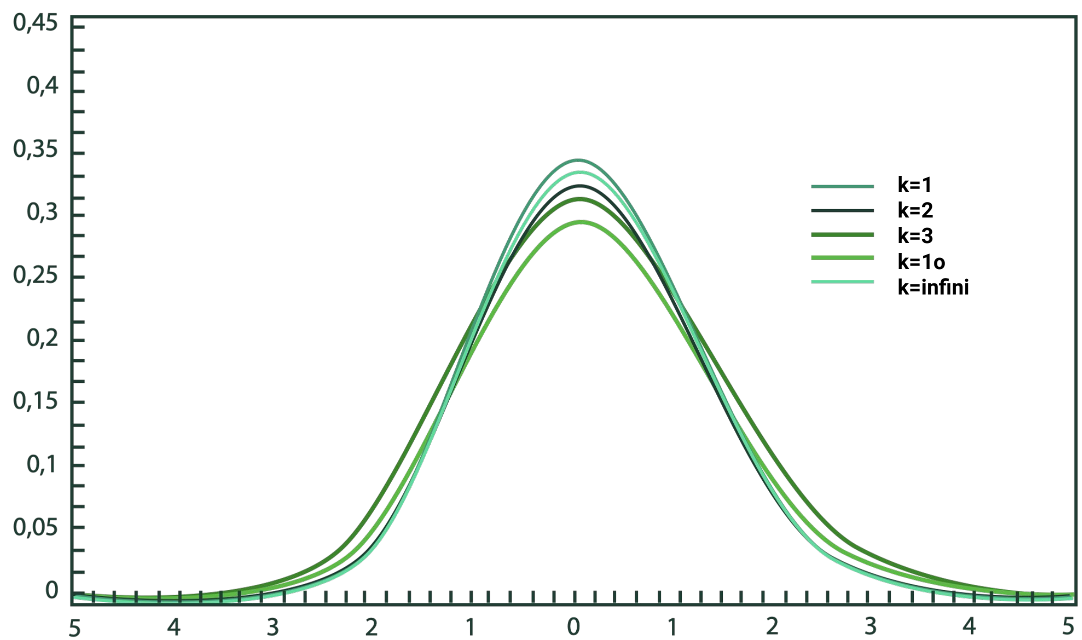
Pasos a seguir

1) Establecer una **hipótesis nula** (H_0) y una **hipótesis alternativa** (H_1).
2) Seleccionar el **nivel de significancia**.
3) Seleccionar el **estadístico de prueba**.
4) Formular la regla de decisión.
5) Interpretar los resultados y tomar una decisión.

Tipos de pruebas de hipótesis

Distribución t de Student

Se usa para estimar una **media de población** normalmente distribuida a partir de una muestra pequeña que sigue una distribución normal y de la que desconocemos la desviación estándar.



$$t = \frac{(x_1 - x_2)}{\sqrt{\dfrac{(s_1)^2}{n_1} + \dfrac{(s_2)^2}{n_2}}}$$

Coeficiente de Pearson

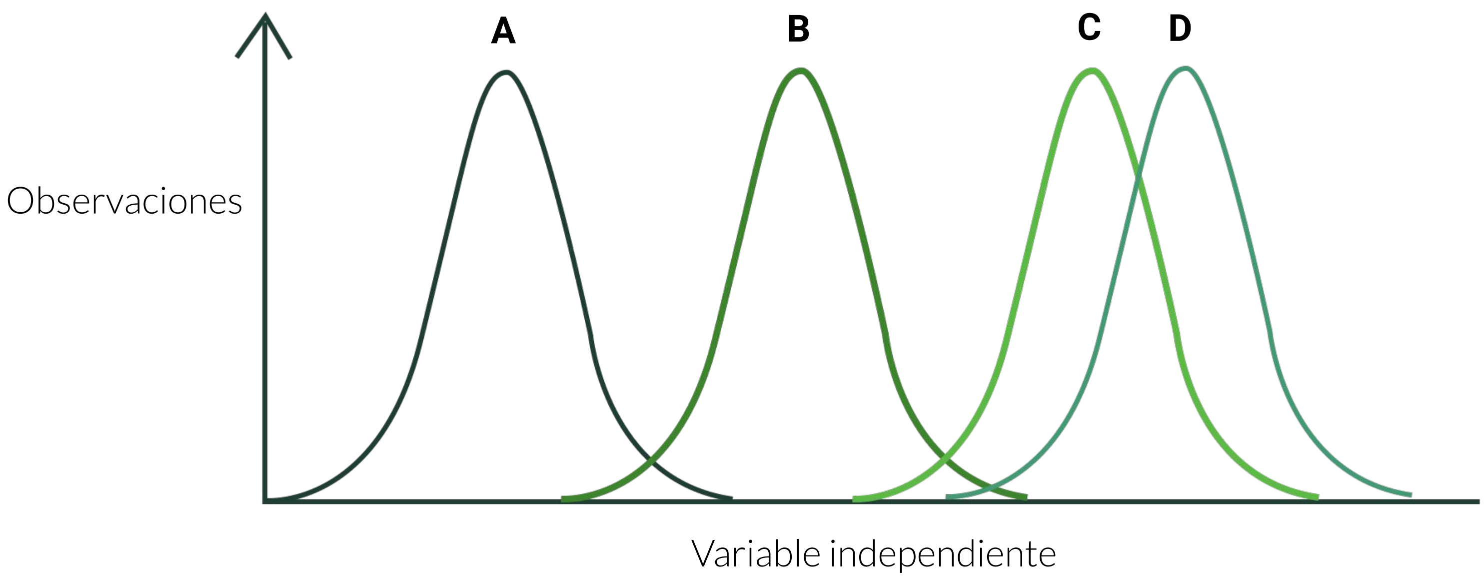
Se usa para medir la dependencia lineal (correlación) entre dos variables aleatorias cuantitativas.

Tabla de interpretación de resultados:

r = 1	correlación perfecta.
0´8 < r < 1	correlación muy alta
0´6< r < 0´8	correlación alta
0´4< r < 0´6	correlación moderada
0´2< r < 0´4	correlación baja
0< r < 0´2	correlación muy baja
r = 0	correlación nula

Análisis de la varianza (ANOVA)

Se usa para comparar las varianzas entre las medias (o el promedio) de diferentes grupos.



Tipos de errores

Contexto

Las conclusiones a las que llegamos se basan en una muestra, por lo que podemos equivocarnos.

Decisiones **correctas**:

1) Rechazar H_0 cuando es falsa.
2) No rechazar H_0 cuando es verdadera.

Decisiones **incorrectas**:

1) Rechazar H_0 cuando es verdadera.
2) No rechazar H_0 cuando es falsa.

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

Ejemplo

HIPÓTESIS:

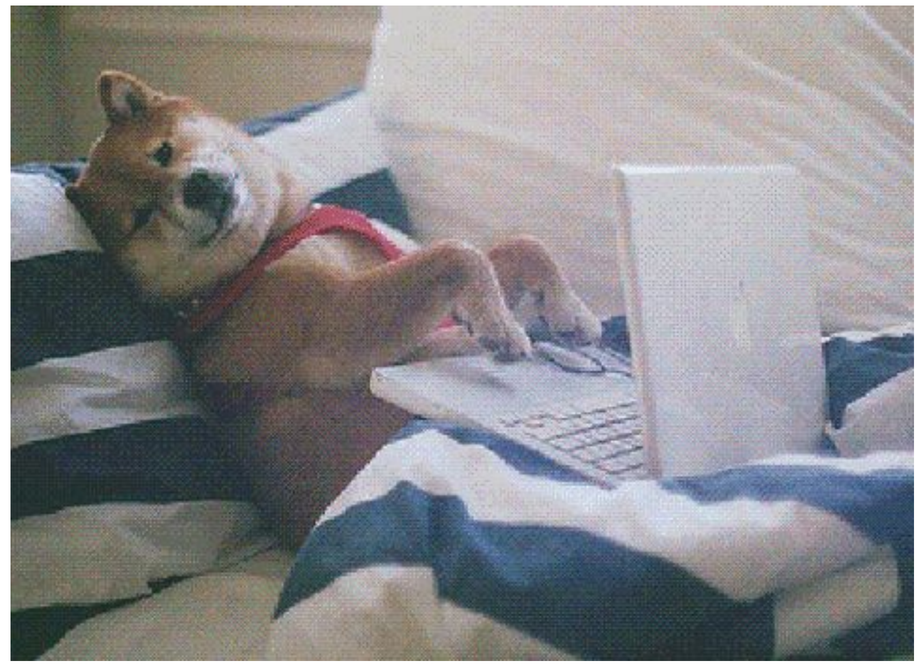
- Hipótesis nula (H_0): $\mu_1 = \mu_2$
  Los dos medicamentos tienen la misma eficacia.
- Hipótesis alternativa (H_1): $\mu_1 \neq \mu_2$
  Los dos medicamentos no tienen la misma eficacia.

ERROR TIPO I: concluir que los dos medicamentos son muy diferentes cuando no lo son.

ERROR TIPO II: concluir que no hay una diferencia significativa entre ambos medicamentos. Muy peligroso.

Pruebas de hipótesis en Python

jupyter

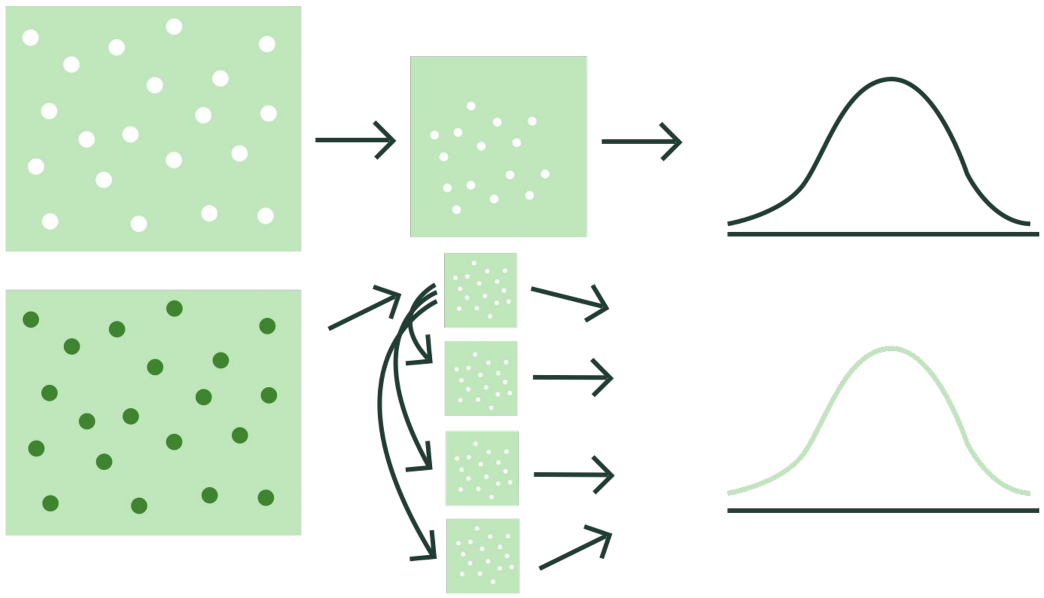


Bootstrapping

Bootstrapping

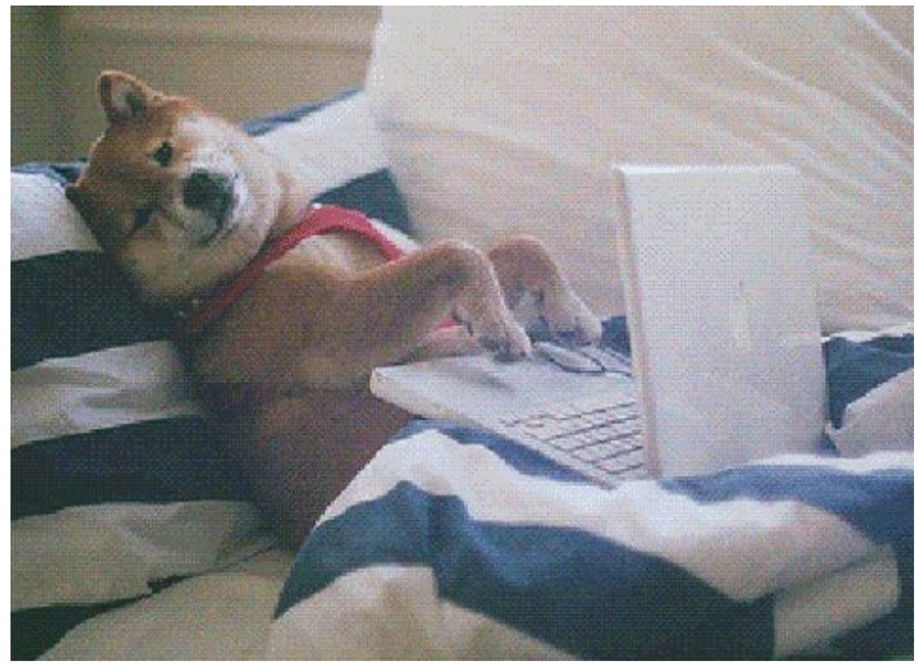
- Método de remuestreo de datos dentro de una muestra aleatoria. Se usa para hallar una aproximación a la distribución de la variable analizada.

- Muy útil en muestras pequeñas o en distribuciones muy sesgadas.



Bootstrapping en Python

jupyter



Validación cruzada

Validación cruzada

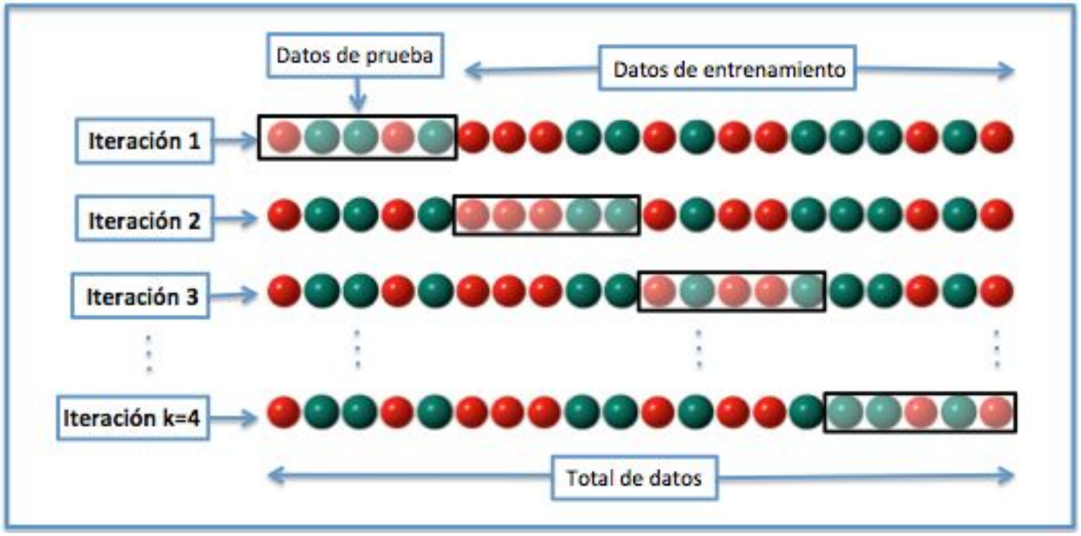
Técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.

Procedimiento

1) División de los datos de forma aleatoria en k grupos de un tamaño similar.

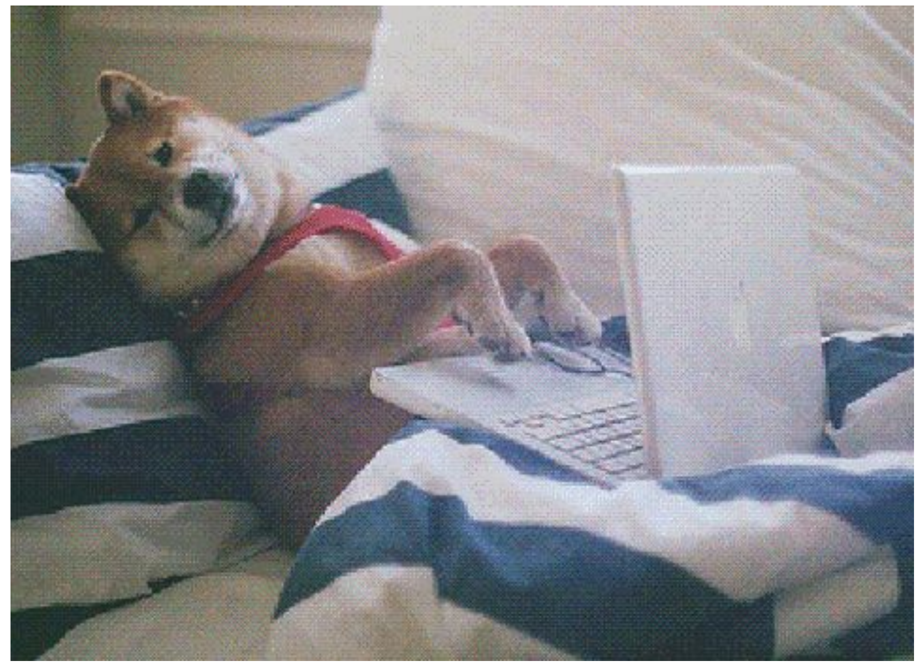
2) Se usan k-1 grupos para entrenar el modelo y uno de ellos se usa para validarlo.

3) El proceso se repite k veces usando un grupo distinto como validación en cada iteración.



Validación cruzada en Python

jupyter



Conclusiones

¡Muchas felicidades!

- Completar los ejercicios y retos.
- Aprobar el examen.
- Compartir qué te pareció el curso en tu reseña.