KIBERNUM
IT Academy

Módulo 4 – Distribuciones de Probabilidad
Clase 8-2

Bootstrapping

Especialización en Ciencia de Datos

Objetivos de aprendizaje

- Utiliza los conceptos básicos de estadística Inferencial.
- Describir el concepto de intervalo de confianza.
- Realizar estimaciones de la media de una población utilizando intervalos de confianza implementados mediante bootstrapping.

Contenido:

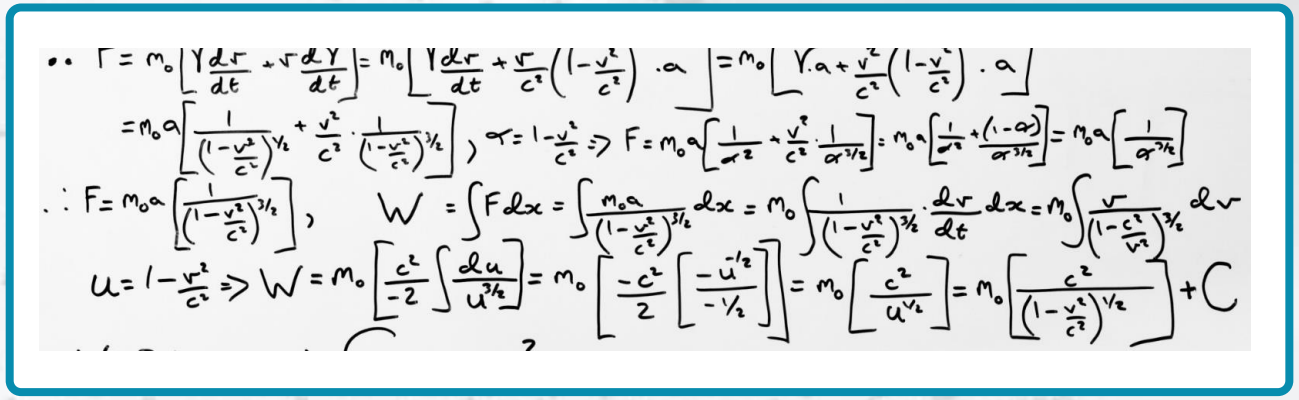
1. ¿Qué es Bootstrapping?
2. Implementación de Bootstrapping en Python
3. Cálculo de intervalos de confianza con Bootstrapping

¿Qué es Bootstrapping?

¿Qué es Bootstrapping?

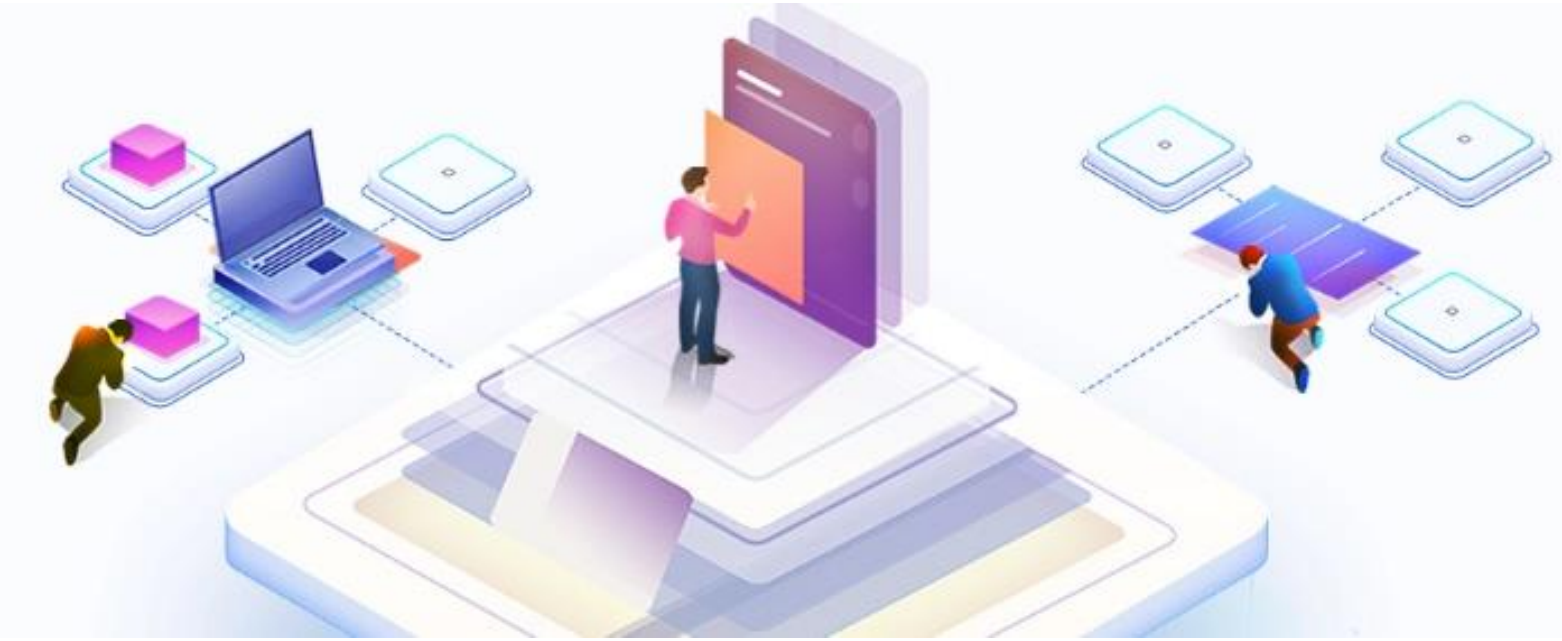
En ciencia de datos, hay situaciones en que es necesario cuantificar la incertidumbre que hay alrededor de un valor, sea una medición, una predicción, una estimación, etc. Por ejemplo, un caso típico es resumir un conjunto de valores que siguen una determinada distribución aleatoria usando un estadístico, como por ejemplo la media.

El enfoque tradicional ha sido utilizar métodos analíticos para cuantificar esta incertidumbre, pero no siempre existe una fórmula analítica para todas las posibles situaciones, además, muchas veces estos métodos deben cumplir con requisitos o supuestos para su aplicación que no siempre se cumplen.



¿Qué es Bootstrapping?

Hoy en día existe la alternativa de utilizar métodos computacionales (algorítmicos) que obtienen resultados equivalentes a los métodos analíticos. Estos métodos tienen muchas menos restricciones y son más fáciles de entender y aplicar. Sin embargo, son más lentos puesto que necesitan realizar un número elevado de operaciones en su procesamiento, aunque esto no es un problema, dada la creciente potencia de cálculo de las máquinas.



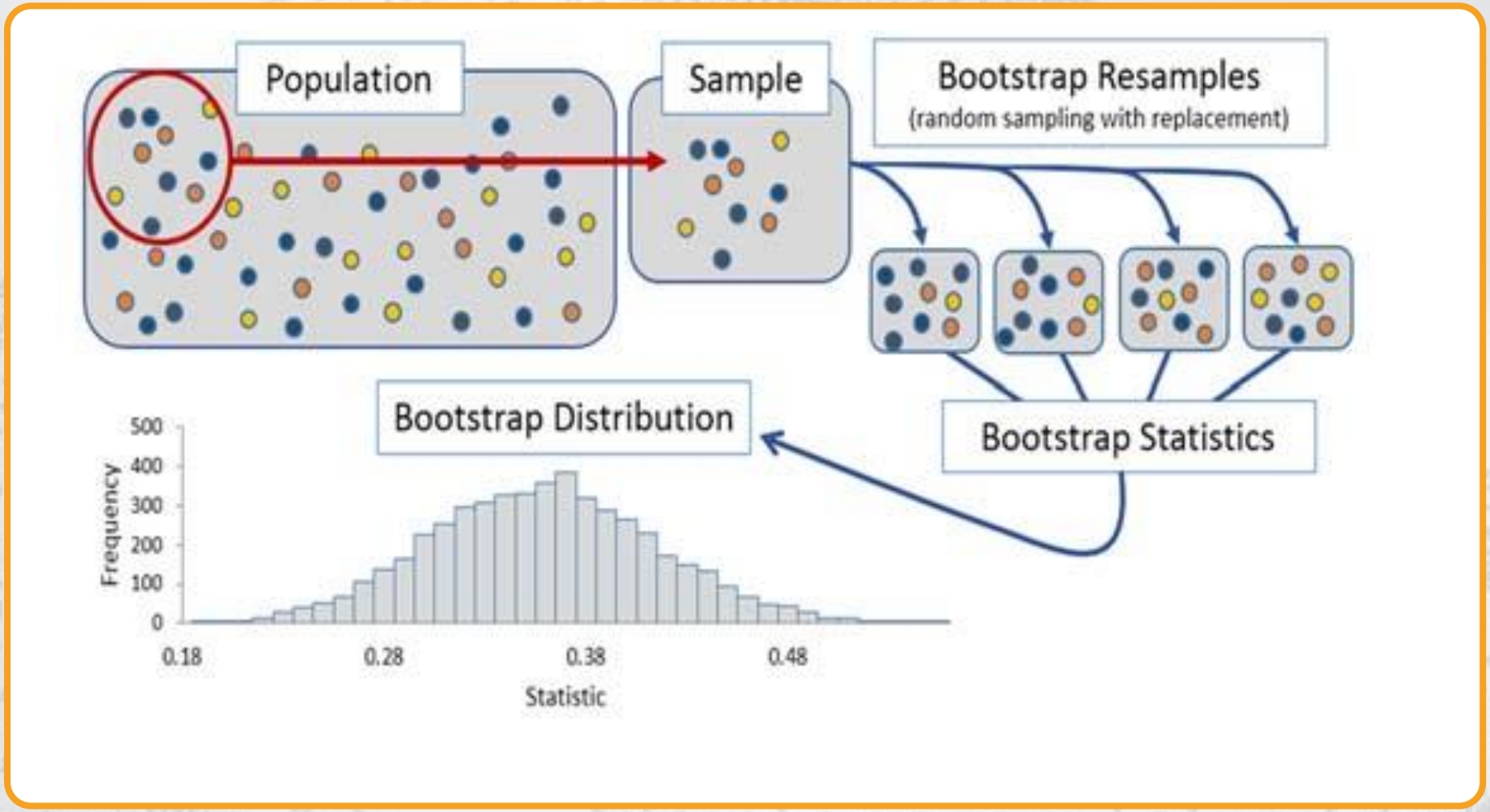
¿Qué es Bootstrapping?

Justamente uno de estos métodos computacionales es el Bootstrapping, y se le llama así a un conjunto de técnicas que utilizan muestreos aleatorios con reemplazo a partir de una distribución de partida para estimar intervalos de confianza, distribuciones muestrales de un parámetro, etc.

La idea es simple, partiendo de la muestra original, extraemos nuevas muestras aleatorias con reemplazo del mismo tamaño y calculamos el estadístico de interés (media, varianza, etc), repitiendo este proceso un número elevado de veces. Esto nos dará una distribución muestral del estadístico que nos interese, sobre la que podremos calcular, por ejemplo, intervalos de confianza. Cuantas más veces lo repitamos, mayor será la distribución muestral generada, y más precisa será la estimación que hagamos (de aquí el coste computacional de este método).

Más info:
https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

¿Qué es Bootstrapping?



¿Qué es Bootstrapping?

Desde el punto de vista teórico, el escenario ideal para realizar inferencia sobre una población es disponer de infinitas (o una gran cantidad) de muestras de dicha población. Si para cada muestra se calcula el estadístico de interés, por ejemplo, la media, se obtiene lo que se conoce como distribución muestral. Esta distribución tiene dos características: su promedio tiende a converger con el valor real del parámetro poblacional, y su dispersión permite conocer el error esperado al estimar el estadístico con una muestra de un tamaño determinado.

En la práctica, no suele ser posible acceder a múltiples muestras. Si solo se dispone de una muestra, y ésta es representativa de la población, cabe esperar que los valores en la muestra aparezcan aproximadamente con la misma frecuencia que en la población.

¿Qué es Bootstrapping?

Así pues, bootstrapping es un proceso de simulación gracias al cual se puede aproximar la distribución muestral de un estadístico empleando únicamente una muestra inicial. Ahora bien, es importante destacar qué información puede y no puede extraerse.

- Bootstrapping no proporciona una mejor estimación del estadístico que la obtenida con la muestra original.

- Bootstrapping simula el proceso de muestreo y con ello la variabilidad generada por este proceso. Gracias a esto, permite estimar la incertidumbre que se puede esperar de un estadístico calculado a partir de una muestra.

¿Qué es Bootstrapping?

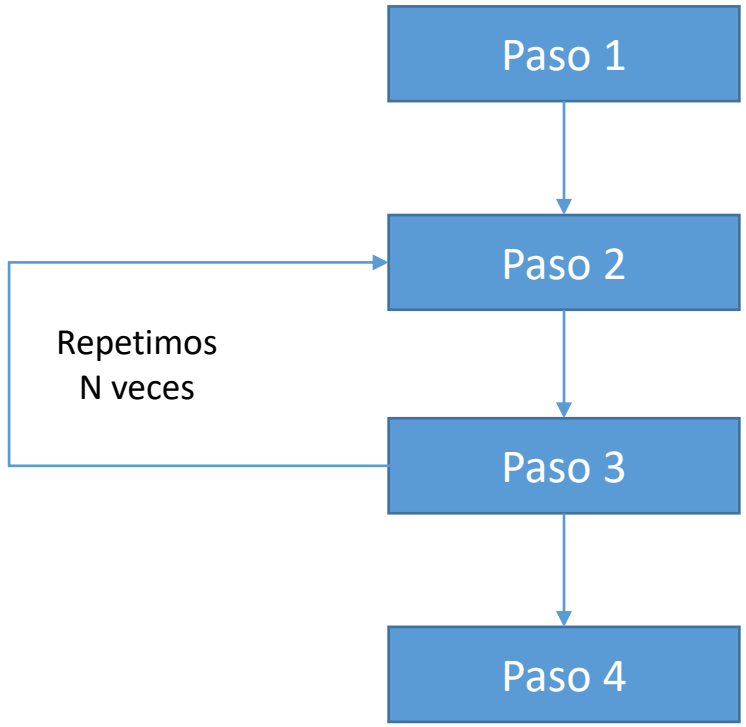
La estrategia de bootstrapping se puede emplear para resolver varios problemas, entre ellos, los siguientes:

- Calcular intervalos de confianza de un parámetro poblacional.
- Calcular la significancia estadística (p-value) de la diferencia entre poblaciones.
- Calcular intervalos de confianza para la diferencia entre poblaciones.

Implementación de Bootstrapping en Python

Implementación en Python

Dependiendo del estadístico que se quiera estimar, va a variar la implementación en código, pero en general, se tiene la siguiente estructura:



Partimos de una muestra aleatoria x

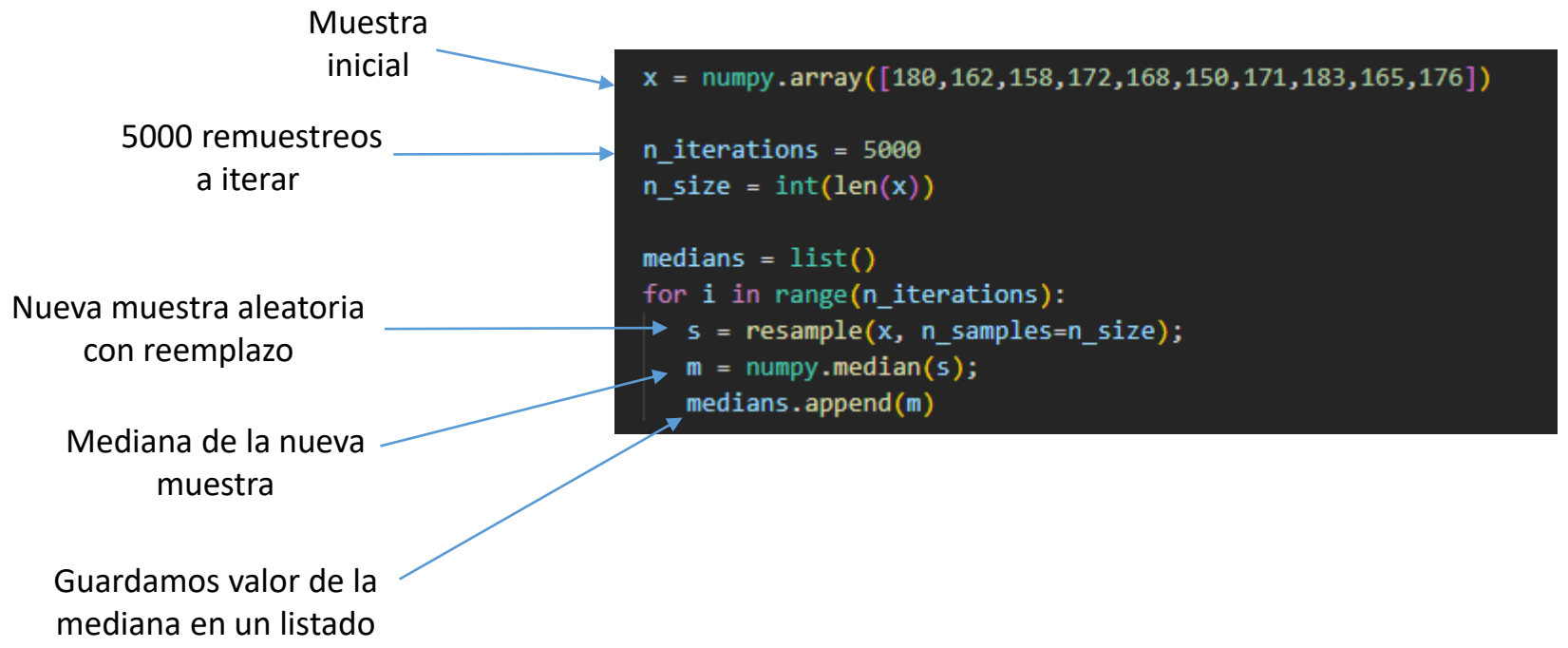
A partir de x obtenemos una nueva muestra aleatoria con reemplazo, del mismo tamaño, llamémosle x'

Calculamos el estadístico de interés sobre x' y lo guardamos

Obtenemos distribución, estimadores, intervalos, etc.

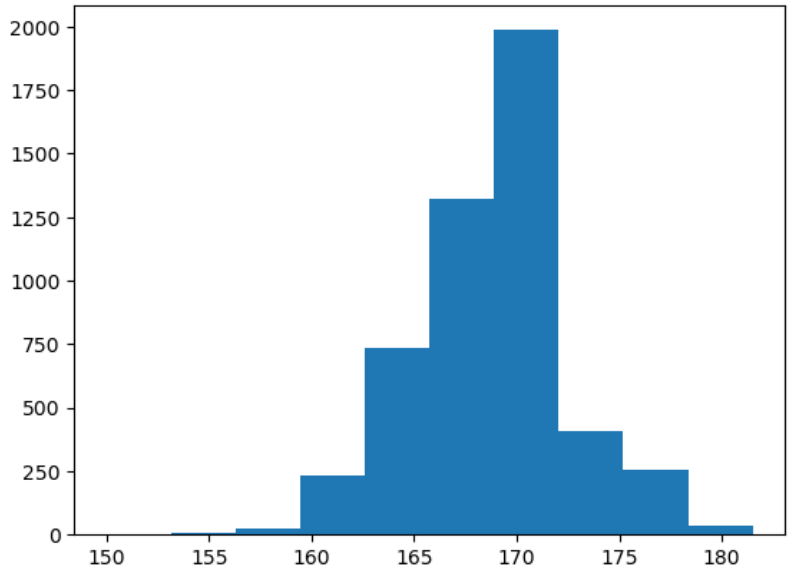
Implementación en Python

Por ejemplo, si queremos calcular la distribución muestral de la media usando bootstrapping, se puede utilizar la librería numpy, el cual es muy eficiente para las operaciones con datos. A continuación, se presenta un ejemplo:



Implementación en Python

```
plt.hist(medians, bins=10)
plt.show()
```




```
print("Mediana muestreo", numpy.median(medians))
print("Desv Est muestreo", numpy.std(medians))
```

```
Mediana muestreo 169.5
Desv Est muestreo 3.851252346964556
```

Bootstrapping para Calcular Intervalos de Confianza

Bootstrapping para Calcular IC

Supongamos que estamos controlando la calidad de un producto que es fabricado en una empresa. Para esto, se realiza un control de calidad seleccionado aleatoriamente 50 producto, los cuales son inspeccionados uno a uno en busca de fallos.

Los resultados son registrados con un 0 cuando no hay falla, y con un 1 cuando sí la hay. Es así que se obtiene el siguiente registro:

$$x=[00001001001000011000000000100011000000010000001000]$$

En este caso, podemos resumir el conjunto de valores diciendo que la tasa media de fallos es de 0.2, es decir, un 20%.

Bootstrapping para Calcular IC

Sin embargo, ¿cómo es de fiable este valor? Puede haber sido fruto del azar por haber elegido esos 50 productos específicos. Si se escogieran otros 50 productos, seguramente salga un registro de fallas distinto y con eso la media podría variar. Hay una incertidumbre intrínseca a la media que hemos calculado. Esto se debe a que esta media, no es la tasa media de fallos real (la media de la población) sino la media calculada a partir de una muestra (la media muestral).

Nuestro objetivo será calcular la incertidumbre de dicha estimación usando un intervalo de confianza. Para ello, usaremos dos métodos: el analítico y el computacional (bootstrapping), de manera de comparar sus ventajas e inconvenientes.

Bootstrapping para Calcular IC

Opción 1: Método Analítico

Se cuenta con la siguiente información.

X : variable aleatoria proporción de productos con falla

$X \rightarrow N(X, \mu, \sigma\text{^}2)$

Muestra:

n=50

p=20% fallos

Población:

σ=??

μ=??

Confianza:

$(1-\alpha)$=0.95

Bootstrapping para calcular IC

Opción 1: Método Analítico

Para determinar el intervalo de confianza, utilizaremos la fórmula del intervalo de confianza para una proporción.

$$CI_{(1-\alpha)} = p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Valor normal de Z

$$Z_{0.025} = -1.96$$

$$CI_{95\%} = 0.2 \pm 1.96\sqrt{\frac{0.2(1-0.2)}{50}}$$

$$CI_{95\%} = 0.2 \pm 0.11$$

$$CI_{95\%} = [\,0.09, 0.31\,]$$

Nótese que hemos calculado el intervalo de confianza en base a una fórmula que no es tan sencilla de comprender, a menos que contemos con profundos conocimientos estadísticos.

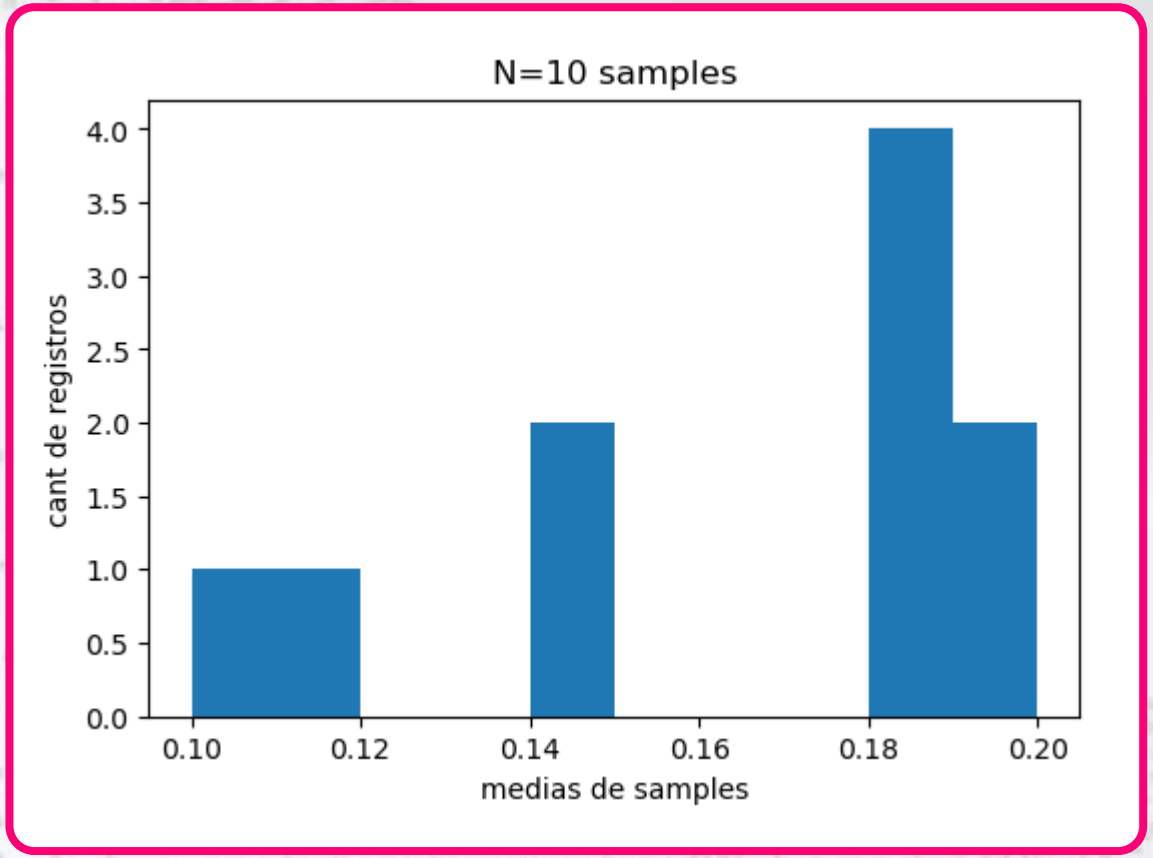
Bootstrapping para Calcular IC

Opción 2: Método Computacional (bootstrapping)
Para determinar el intervalo de confianza, aplicaremos el procedimiento explicado anteriormente. Partiremos de la muestra original x, y a partir de ella, obtendremos nuevas muestras con reemplazo del mismo tamaño, x', en donde calcularemos la media. Este procedimiento lo repetiremos N veces.

```
# registro original de valores
registro = '0000100100100001100000000010001100000001000000001000'


# transformamos a nd.array de zeros y unos (valores int)
x = np.array([int(i) for i in list(registro)])


# numero de sampleos
N = 10

x_samples_10 = np.random.choice(x, size=(N, x.size), replace=True)
x_samples_means_10 = np.mean(x_samples_10, axis=1)
```

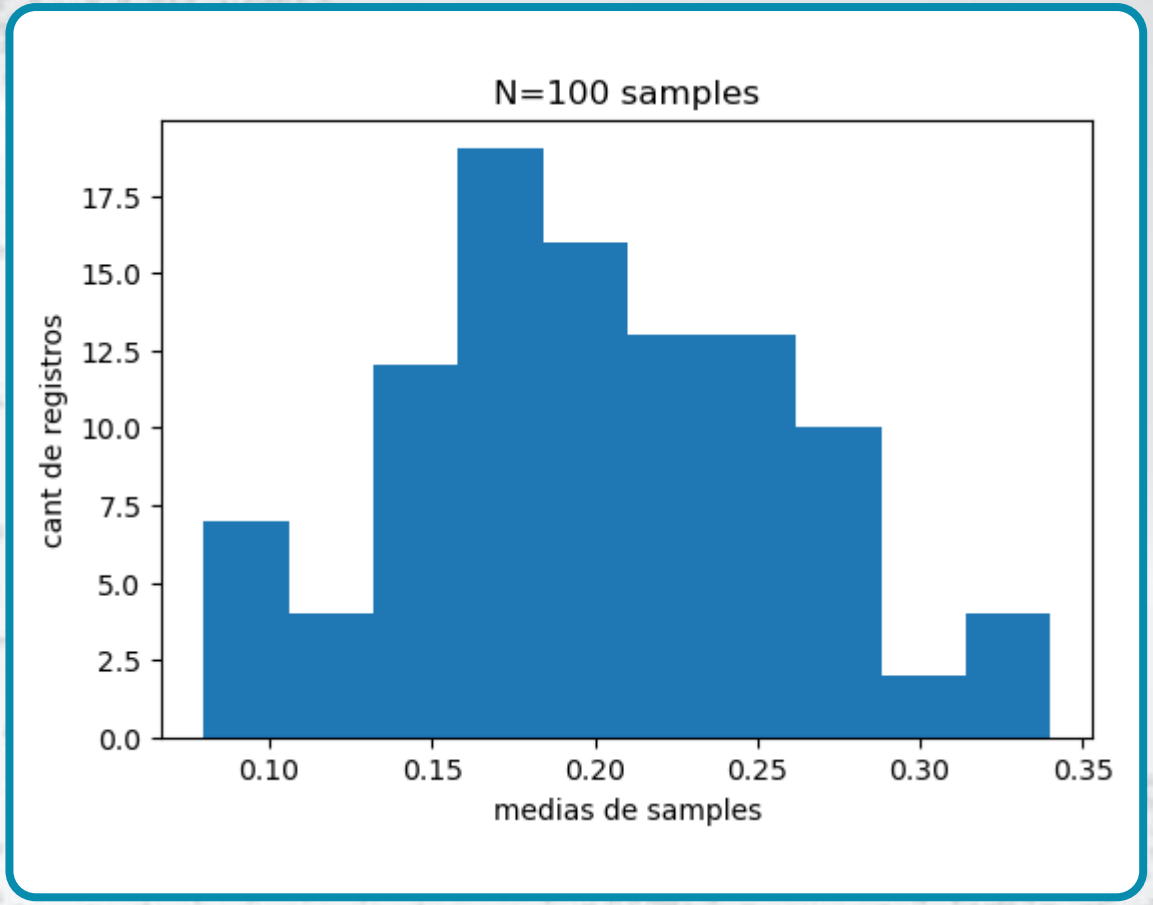
Bootstrapping para Calcular IC

Opción 2: Método Computacional (bootstrapping)
Este es el histograma obtenido a partir de las medias de 10 muestras.



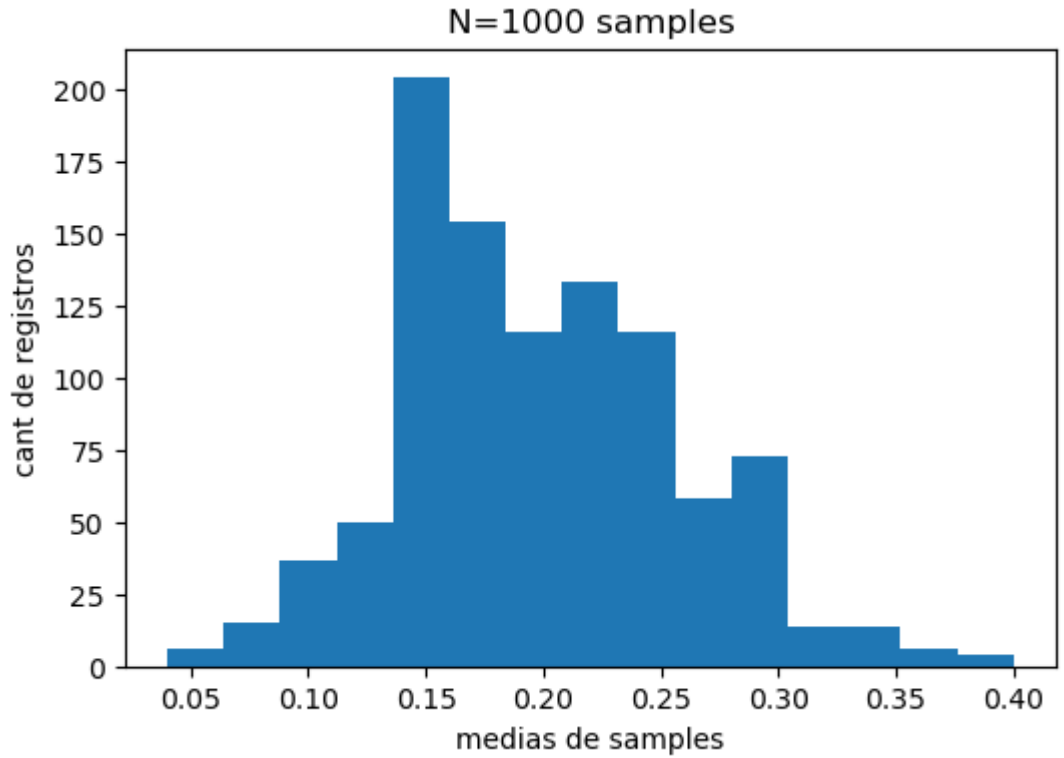
Bootstrapping para calcular IC

Opción 2: Método Computacional (bootstrapping)
Este es el histograma obtenido a partir de las medias de 100 muestras.



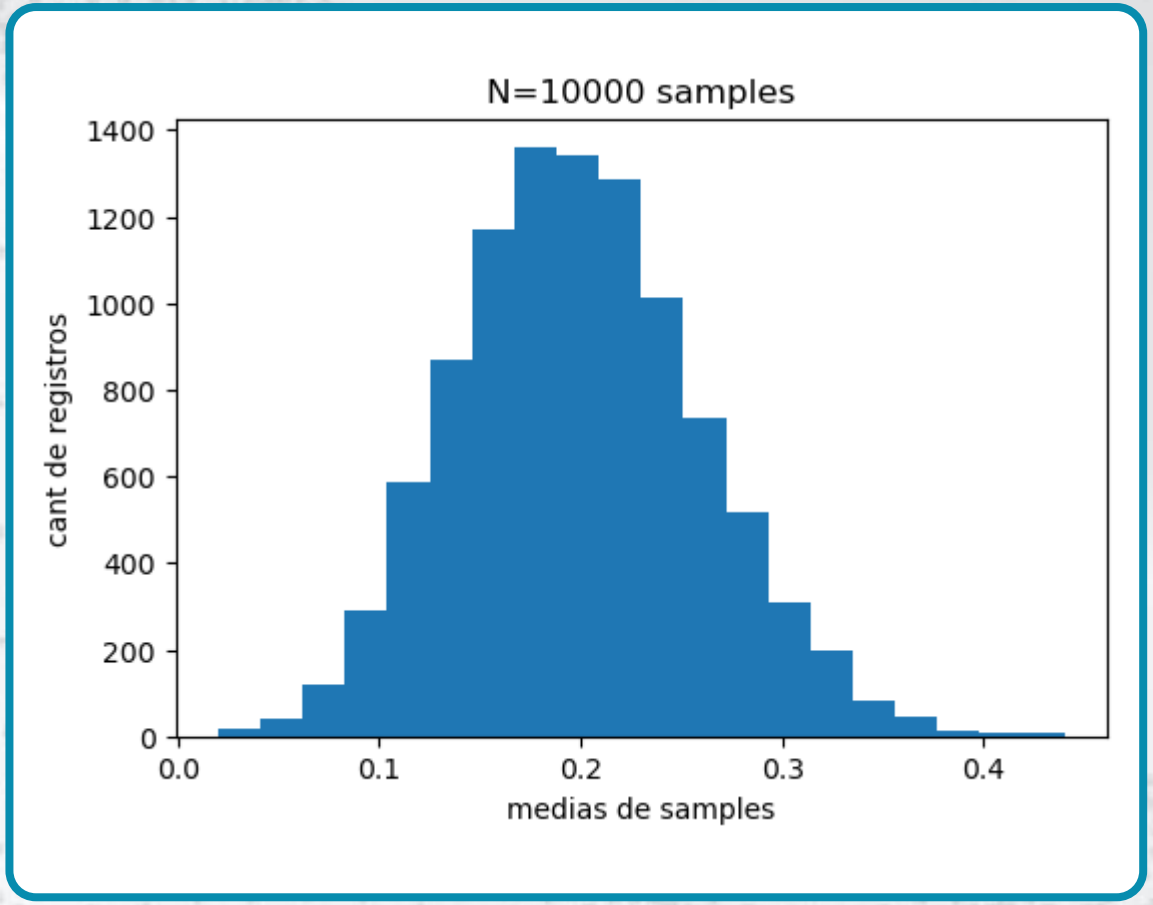
Bootstrapping para Calcular IC

Opción 2: Método Computacional (bootstrapping)
Este es el histograma obtenido a partir de las medias de 1.000 muestras.



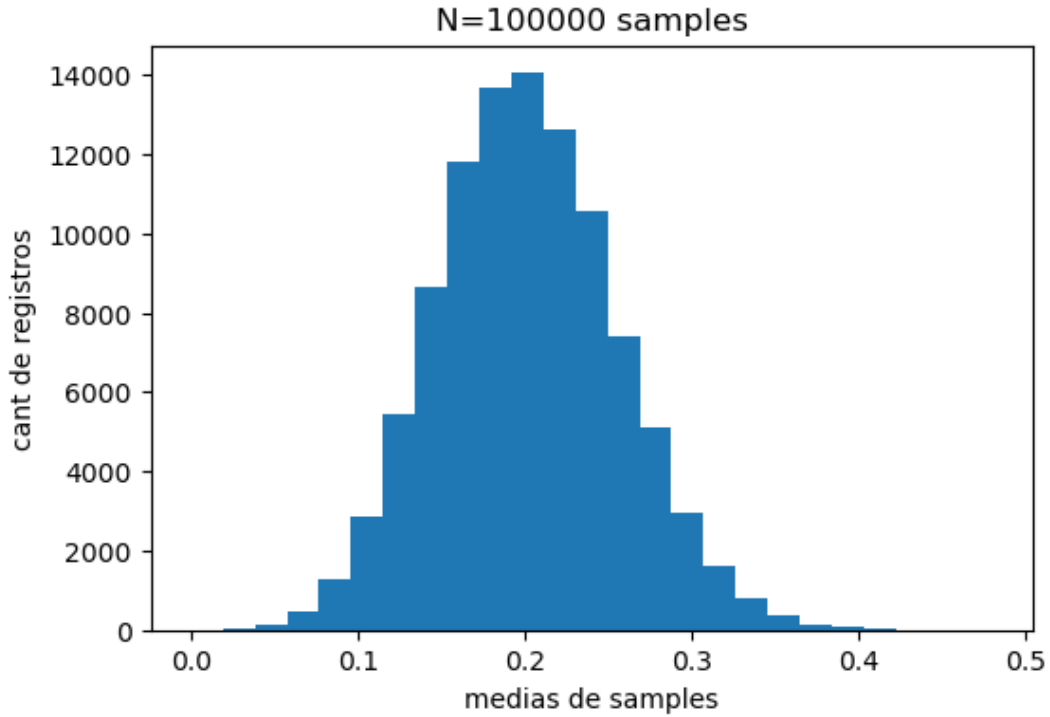
Bootstrapping para Calcular IC

Opción 2: Método Computacional (bootstrapping)
Este es el histograma obtenido a partir de las medias de 10.000 muestras.



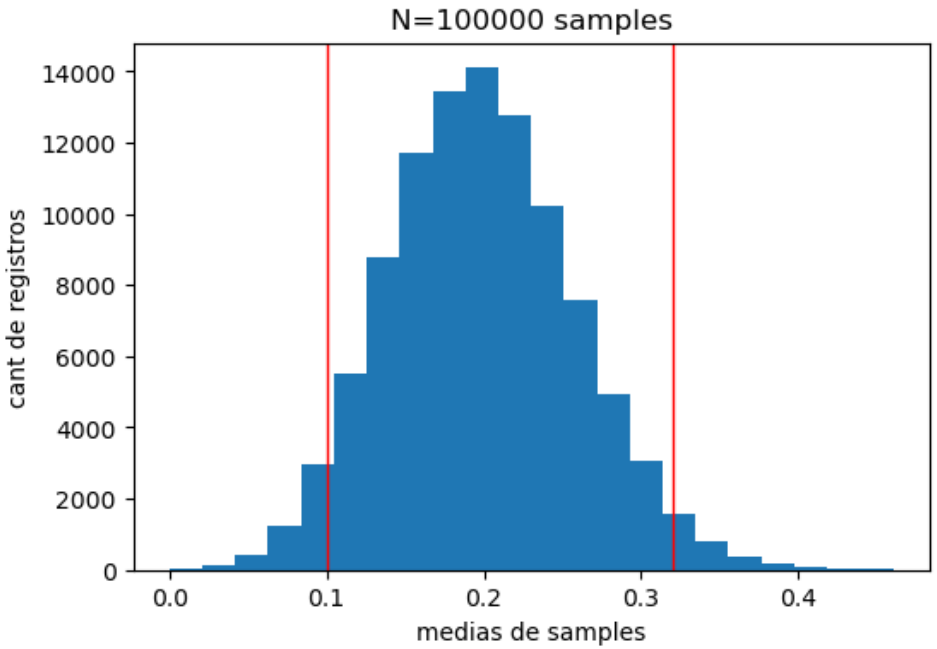
Bootstrapping para calcular IC

Opción 2: Método Computacional (bootstrapping)
Este es el histograma obtenido a partir de las medias de 100.000 muestras



Bootstrapping para calcular IC

Opción 2: Método Computacional (bootstrapping)
Con los 100.000 resampleos, calculemos el intervalo de confianza.




```
ic = np.percentile(x_samples_means_10, [2.5, 97.5])
ic
```

```
array([0.1 , 0.32])
```

En este caso, el intervalo de confianza calculado es:

$$CI_{95\%} = [\ 0.1, 0.32\ ]$$

Comparación de Resultados

A continuación, se resumen los resultados obtenidos con cada método:

Método Analítico $CI_{95\%} = [\, 0.09, 0.31 \,]$

Método Computacional $CI_{95\%} = [\, 0.1, 0.32 \,]$

Como se puede observar, los intervalos son muy parecidos. Sin embargo, a diferencia del método analítico, no hemos tenido que saber las fórmulas ni hacer ninguna suposición sobre la distribución muestral de los datos. Sólo tuvimos que codificar algunas líneas de código para extraer una muestra aleatoria con reemplazo y utilizar un loop iterativo.

Esta técnica es extraordinariamente potente, ya que podemos extraer cualquier intervalo de confianza que queramos simplemente cambiando los percentiles a calcular.

Boostrapping para Calcular Intervalos de Confianza

Existen varios métodos para estimar intervalos de confianza mediante el uso de bootstrapping:

- Intervalos basados en distribución normal (normal bootstrap intervals).
- Intervalos basados en percentiles (percentile bootstrap intervals).
- Intervalos basados en distribución Student's (bootstrap Student's t intervals).
- Bias-Corrected and Accelerated Bootstrap Method (BCA).
- Intervalos empíricos (empirical bootstrap intervals).

Dudas y consultas
¡Gracias!

KIBERNUM