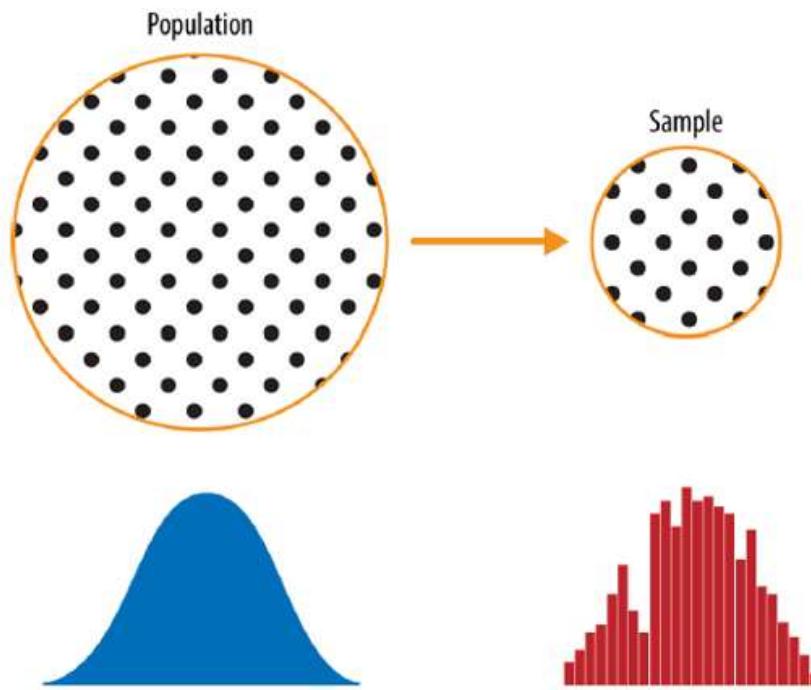


# Capítulo 2 - Distribuciones de datos y muestreo

## 2.1 Población y muestra

La siguiente imagen intenta ilustrar la diferencia entre población y muestra:



Las estadísticas tradicionales se enfocaban mucho en el lado izquierdo, utilizando teoría basada en supuestos fuertes sobre la población. Las estadísticas modernas se han movido hacia el lado derecho, donde tales suposiciones no son necesarias.

## 2.2 Muestreo Aleatorio y Sesgo (Bias) en la Muestra

Una muestra es un subconjunto de datos de un conjunto de datos más grande; los estadísticos llaman a este conjunto de datos más grande la población. Una población en estadística no es lo mismo que en biología: es un conjunto grande, definido (pero a veces teórico o imaginario) de datos.

El muestreo aleatorio es un proceso en el cual cada miembro disponible de la población que se está muestreando tiene la misma posibilidad de ser elegido para la muestra en cada selección. La muestra resultante se llama una **muestra aleatoria simple**.

La calidad de los datos a menudo importa más que la cantidad de datos al hacer una estimación o un modelo basado en una muestra. La calidad de los datos en ciencia de

datos implica completitud, consistencia de formato, limpieza y precisión de los puntos de datos individuales. La estadística además añade la noción de **representatividad**.

## 2.2.1 Términos Clave para el Muestreo Aleatorio

### **Muestra (Sample)**

Un subconjunto de un conjunto de datos más grande.

### **Población**

El conjunto de datos más grande o la idea de un conjunto de datos.

### **N (n)**

El tamaño de la población (muestra).

### **Muestreo aleatorio**

Selección de elementos en una muestra de manera aleatoria.

### **Muestreo estratificado**

División de la población en estratos y muestreo aleatorio dentro de cada estrato. Por ejemplo: estudiantes de un colegio. Se escogen muestras de cada grado/año. 20 Estudiantes de 1er año, 20 estudiantes del 2do año...

### **Estrato (pl., estratos)**

Un subgrupo homogéneo de una población con características comunes. Ejemplo: Estudiantes del primer año.

### **Muestra aleatoria simple**

La muestra que resulta del muestreo aleatorio sin estratificar la población.

### **Sesgo (Bias)**

Error sistemático.

### **Sesgo en la muestra (Sample Bias)**

Una muestra que no representa adecuadamente a la población.

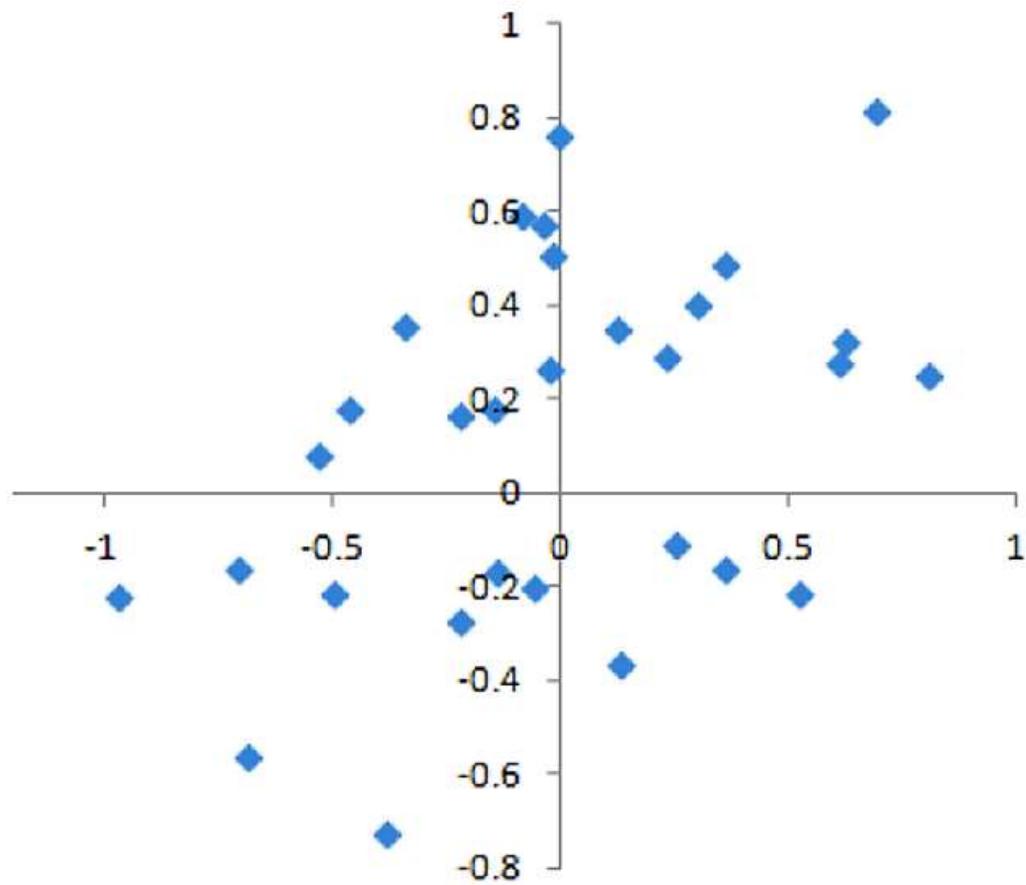
## 2.2.2 Sesgo de Muestreo por Autoselección

Las reseñas de restaurantes, hoteles, cafés, y otros que lees en sitios de redes sociales son propensas al sesgo porque las personas que las envían no son seleccionadas aleatoriamente; más bien, son ellas mismas quienes toman la iniciativa de escribir. Esto conduce a un sesgo de autoselección: las personas motivadas para escribir reseñas pueden haber tenido malas experiencias, pueden tener alguna relación con el establecimiento, o simplemente pueden ser un tipo de persona diferente a las que no escriben reseñas.

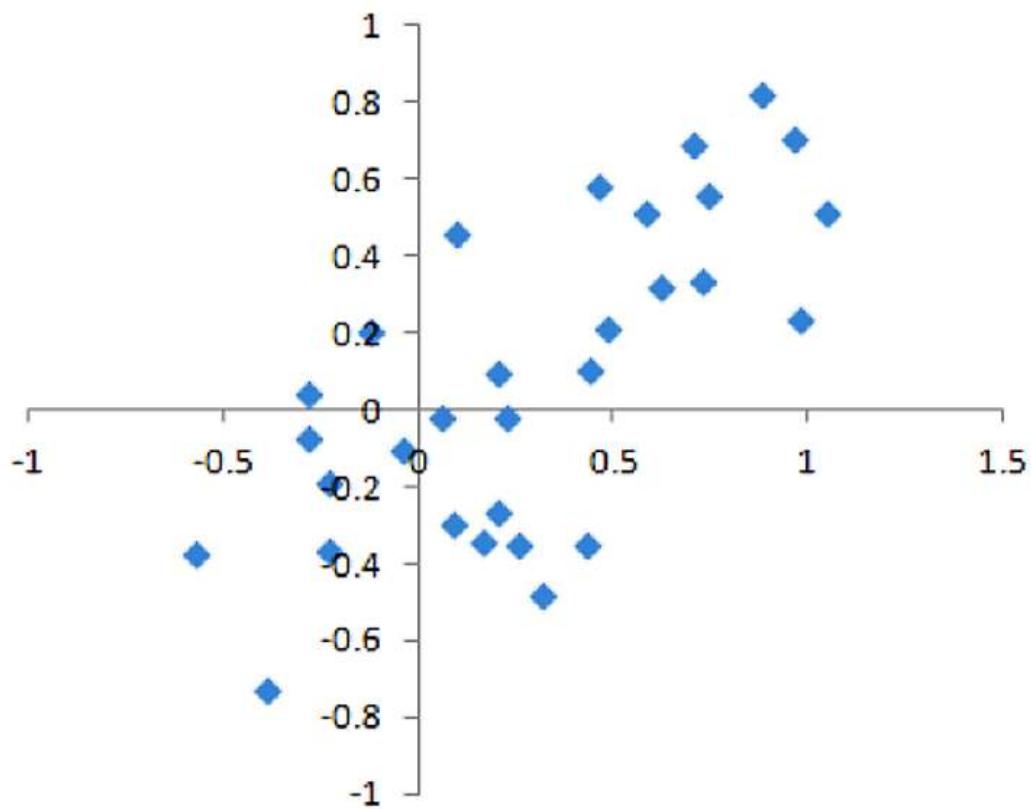
## 2.2.3 Sesgo

El sesgo estadístico se refiere a errores de medición o de muestreo que son sistemáticos y producidos por el proceso de medición o muestreo. Es importante hacer una distinción

entre errores debidos al azar y errores debidos al sesgo. Consideremos el proceso físico de un arma disparando a un blanco. No impactará en el centro absoluto del blanco cada vez, o incluso la mayoría de las veces. Un proceso no sesgado producirá error, pero este será aleatorio y no tenderá fuertemente en ninguna dirección (ver Figura 2.2).



Los resultados mostrados en la Figura 2.3 muestran un proceso sesgado: sigue habiendo error aleatorio en ambas direcciones, x e y, pero también hay un sesgo. Los disparos tienden a caer en el cuadrante superior derecho.



El sesgo se presenta en diferentes formas y puede ser observable o invisible. Cuando un resultado sugiere la presencia de sesgo (por ejemplo, al compararlo con un punto de referencia o valores reales), a menudo es un indicador de que un modelo estadístico o de aprendizaje automático ha sido mal especificado, o que se ha omitido una variable importante.

**Ejercicio 2.1:** *Investigar los tipos de sesgos (Bias) en estadística y aprendizaje automático.* Añade un resumen en este mismo notebook.

**Ejercicio 2.2:** *Investigar la influencia del sesgo en Machine Learning incluye un caso breve en este mismo notebook.*

## 2.2.4 Selección Aleatoria

Para evitar el problema de sesgo en la muestra que llevó al *Literary Digest* a predecir la victoria de Landon sobre Roosevelt, George Gallup optó por métodos seleccionados de manera más científica para lograr una muestra que fuera representativa del electorado votante de los Estados Unidos. Actualmente, existe una variedad de métodos para lograr la representatividad, pero en el corazón de todos ellos se encuentra el muestreo aleatorio.

El muestreo aleatorio no siempre es fácil. La definición adecuada de una población accesible es clave. Supongamos que queremos generar un perfil representativo de

clientes y necesitamos realizar una encuesta piloto. La encuesta necesita ser representativa, pero también es laboriosa.

Primero, necesitamos definir qué es un cliente. Podríamos seleccionar todos los registros de clientes donde la cantidad de compra sea mayor que 0.

Luego, necesitamos especificar un procedimiento de muestreo. Podría ser "seleccionar 100 clientes al azar". Cuando se trata de muestrear un flujo (por ejemplo, transacciones de clientes en tiempo real o visitantes web), las consideraciones de tiempo pueden ser importantes (por ejemplo, un visitante web a las 10 a.m. en un día laborable puede ser diferente de un visitante web a las 10 p.m. en un fin de semana).

En el muestreo estratificado, la población se divide en estratos, y se toman muestras aleatorias de cada estrato. Los encuestadores políticos podrían buscar conocer las preferencias electorales de blancos, negros e hispanos. Una muestra aleatoria simple tomada de la población podría arrojar muy pocos negros e hispanos, por lo que esos estratos podrían tener un peso mayor en el muestreo estratificado para obtener tamaños de muestra equivalentes.

## 2.2.5 Tamaño versus Calidad: ¿Cuándo Importa el Tamaño?

En la era del big data, a veces resulta sorprendente que lo pequeño sea mejor. El tiempo y el esfuerzo dedicados al muestreo aleatorio no solo reducen el sesgo, sino que también permiten prestar más atención a la exploración de datos y a la calidad de los datos. Por ejemplo, los datos faltantes y los valores atípicos pueden contener información útil.

Podría ser prohibitivo rastrear valores faltantes o evaluar valores atípicos en millones de registros, pero hacerlo en una muestra de varios miles de registros puede ser factible. La representación gráfica de datos y la inspección manual se vuelven más difíciles si hay demasiados datos.

### Ejemplo de Importancia de la Calidad versus el Tamaño de los Datos

Imaginemos que una empresa de comercio electrónico quiere analizar la satisfacción de sus clientes después de realizar una compra. La empresa tiene acceso a millones de registros de transacciones, pero solo un pequeño porcentaje de esos clientes ha dejado comentarios o calificaciones sobre su experiencia. La empresa podría estar tentada a utilizar todos los registros disponibles para su análisis, pero aquí es donde surge el problema de la calidad frente al tamaño.

Supongamos que en lugar de utilizar todos los registros, la empresa decide centrarse en una muestra más pequeña, pero representativa, de aquellos clientes que han dejado comentarios detallados y han respondido a encuestas post-compra. Esta muestra podría incluir solo unos pocos miles de clientes, pero los datos son mucho más ricos en detalles: contienen información específica sobre lo que les gustó o no les gustó, problemas que encontraron, y sugerencias para mejorar.

Con esta muestra de alta calidad, la empresa puede identificar patrones específicos de satisfacción o insatisfacción que no serían evidentes en un análisis de todos los registros

de transacciones, que carecen de este nivel de detalle. Por ejemplo, podrían descubrir que los clientes que compran un tipo específico de producto tienen problemas recurrentes con el proceso de envío. Este tipo de información es crucial para mejorar la experiencia del cliente.

En este caso, enfocarse en la calidad de los datos (comentarios detallados) en lugar de la cantidad total de registros permite a la empresa tomar decisiones más informadas y específicas para mejorar la satisfacción del cliente. Aunque la muestra es más pequeña, es mucho más útil que analizar una gran cantidad de datos superficiales que no proporcionan la misma profundidad de conocimiento.

## 2.2.6 Media Muestral versus Media Poblacional

El símbolo  $\bar{x}$  (pronunciado "x-barra") se utiliza para representar la media de una muestra de una población, mientras que  $\mu$  se utiliza para representar la media de una población. ¿Por qué hacer la distinción? La información sobre muestras se observa, y la información sobre grandes poblaciones a menudo se infiere a partir de muestras más pequeñas.

## Ideas Clave

- Incluso en la era del big data, el muestreo aleatorio sigue siendo una herramienta importante en el arsenal del científico de datos.
- El sesgo ocurre cuando las mediciones u observaciones están sistemáticamente en error porque no son representativas de la población completa.
- La calidad de los datos es a menudo más importante que la cantidad de datos, y el muestreo aleatorio puede reducir el sesgo y facilitar la mejora de la calidad que, de otro modo, sería prohibitivamente costosa.

## 2.2.7 Regresión a la Media

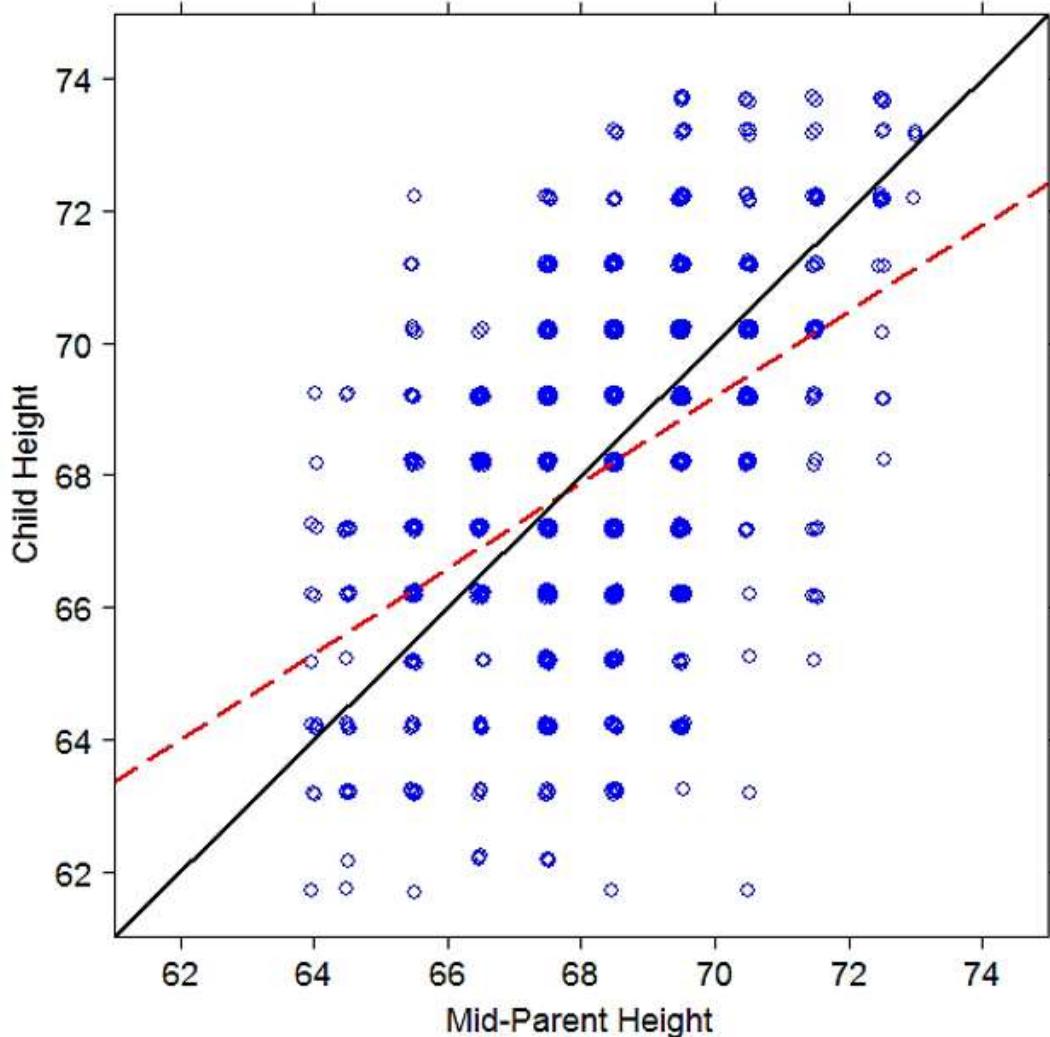
La regresión a la media se refiere a un fenómeno que involucra mediciones sucesivas en una variable dada: las observaciones extremas tienden a ser seguidas por otras más cercanas a la media. Dar un enfoque especial y significado a un valor extremo puede llevar a una forma de sesgo de selección.

Los aficionados al deporte están familiarizados con el fenómeno del "novato del año y la caída del segundo año". Entre los atletas que comienzan su carrera en una temporada dada (la clase de novatos), siempre hay uno que se desempeña mejor que todos los demás. Generalmente, este "novato del año" no tiene un rendimiento tan bueno en su segundo año. ¿Por qué ocurre esto?

En casi todos los deportes principales, al menos aquellos que se juegan con una pelota o disco, hay dos elementos que juegan un papel en el rendimiento general:

- Habilidad
- Suerte

La regresión a la media es una consecuencia de una forma particular de sesgo de selección. Cuando seleccionamos al novato con el mejor desempeño, probablemente estén contribuyendo tanto la habilidad como la buena suerte. En su próxima temporada, la habilidad seguirá ahí, pero muy a menudo la suerte no, por lo que su rendimiento disminuirá—se "regresará" a la media. El fenómeno fue identificado por primera vez por Francis Galton en 1886 [Galton-1886], quien escribió sobre él en relación con tendencias genéticas; por ejemplo, los hijos de hombres extremadamente altos tienden a no ser tan altos como su padre. Ver figura 2.4.



Para investigar la relación entre la estatura de padres e hijos, Galton comparó la estatura de 930 niños que habían alcanzado la edad adulta con la estatura media de sus padres. Para tener en cuenta las diferencias debidas al sexo, multiplicó por 1,08 la estatura de las mujeres.

La figura anterior es una réplica de este gráfico, en el que los círculos azules representan la altura de cada niño comparada con la altura media de sus dos padres (que Galton describió como la "altura media de los padres"). Galton agrupó los resultados en intervalos de 2,5 cm, lo que significa que muchos puntos aparecen uno encima de otro. Siguiendo el planteamiento de Stephen Senn en su artículo Significance sobre Galton, cada punto se ha desplazado una cantidad muy pequeña en ambas direcciones para

separar los puntos superpuestos entre sí, de modo que sea más fácil visualizar dónde hay muchas observaciones trazadas en el mismo punto. Cuando Galton examinó este gráfico descubrió un resultado sorprendente. Si la estatura media de un niño fuera la misma que la de sus padres, cabría esperar que los datos siguieran la línea negra de la figura anterior. Sin embargo, al trazar la línea de mejor ajuste a través de los datos (línea discontinua roja), descubrió que los datos no seguían esta línea negra y que la pendiente del ajuste por mínimos cuadrados era, de hecho, menos pronunciada.

Este fenómeno estadístico se conoce como regresión a la media y se produce cuando se realizan mediciones repetidas. Significa que, en general, las observaciones relativamente altas (o bajas) suelen ir seguidas de otras menos extremas más cercanas a la media real del sujeto. La regresión a la media sigue siendo un fenómeno estadístico importante que a menudo no se tiene en cuenta y que puede dar lugar a conclusiones engañosas. Por ejemplo, las estadísticas oficiales publicadas sobre el impacto de los radares de velocidad sugerían que salvaban una media de 100 vidas al año. Este resultado se basaba en el descenso de accidentes mortales que se había producido desde la instalación de los radares. Sin embargo, los radares de velocidad suelen instalarse después de que se haya producido un número inusualmente alto de accidentes, por lo que, en general, cabría esperar que estos volvieran después a niveles normales. Otro análisis que tuvo en cuenta la regresión a la media descubrió que el 50% del descenso de los accidentes se habría producido tanto si se hubiera instalado un radar de velocidad como si no. Esto pone de relieve que, aunque los radares de velocidad pueden reducir el número de accidentes mortales en carretera, la estimación de la magnitud de su efecto debe hacerse con cuidado.

## Ejemplo en negocios.

Imaginemos una cadena de tiendas de ropa que realiza un seguimiento de las ventas mensuales en todas sus sucursales. Supongamos que una de las tiendas tuvo un mes extraordinariamente bueno, con ventas muy por encima del promedio. La gerencia podría pensar que esta tienda ha encontrado alguna fórmula mágica para el éxito y podrían esperar que continúe con este rendimiento en los meses siguientes.

Sin embargo, es probable que el siguiente mes las ventas de esta tienda vuelvan a un nivel más cercano al promedio de todas las tiendas. Esto puede deberse a varios factores, como una promoción especial, un evento local que aumentó temporalmente el tráfico de clientes, o simplemente buena suerte. Este fenómeno es un ejemplo de regresión a la media: después de un rendimiento extremo (en este caso, ventas muy altas), es más probable que los resultados posteriores sean más cercanos a la media general.

## Ejemplo en aprendizaje automático

Supongamos que se entrena un modelo de clasificación para predecir si un cliente comprará o no un producto basado en ciertos datos de comportamiento. Durante el proceso de evaluación, se realiza una prueba en un conjunto de datos de validación, y el

modelo obtiene una precisión extremadamente alta, mucho mayor que la obtenida en otras pruebas anteriores.

Es tentador pensar que el modelo es excepcionalmente bueno, pero esta alta precisión podría ser un resultado fortuito debido a la particularidad de ese conjunto de validación (quizás el conjunto de datos era más fácil de predecir por casualidad). Al probar el modelo en otros conjuntos de datos adicionales o en datos nuevos, es probable que la precisión vuelva a un nivel más cercano al promedio obtenido anteriormente. Este es un caso de regresión a la media en aprendizaje automático: después de un rendimiento extremo, los resultados posteriores tienden a ser más cercanos a la media de los resultados anteriores.

**Ejercicio 2.3** Supongamos que tienes los resultados de dos exámenes de matemáticas de un grupo de 100 estudiantes. El primer examen se realizó al inicio del semestre y el segundo examen al final del semestre. Queremos analizar si existe una regresión a la media entre los resultados de estos dos exámenes.

Observa los resultados en el gráfico (ejecuta el código). Compara los estudiantes que obtuvieron puntuaciones extremadamente altas o bajas en el primer examen con sus puntuaciones en el segundo examen. ¿Notas alguna tendencia hacia el promedio en sus puntuaciones?

```
import matplotlib.pyplot as plt
import numpy as np

# Simulación de resultados de exámenes
np.random.seed(42)
exam1_scores = np.random.normal(70, 15, 100) # Puntuaciones del primer examen
exam2_scores = exam1_scores * 0.5 + np.random.normal(35, 10, 100) # Puntuaciones del segundo examen con regresión a la media

# Crear el gráfico
plt.figure(figsize=(8, 6))
plt.scatter(exam1_scores, exam2_scores, color='blue', alpha=0.6)
plt.plot([30, 110], [30, 110], color='red', linestyle='--') # Línea de referencia (sin regresión)
plt.title('Regresión a la Media en Resultados de Exámenes')
plt.xlabel('Puntuación en el Primer Examen')
plt.ylabel('Puntuación en el Segundo Examen')
plt.grid(True)
plt.show()
```

a. ¿Qué observas en el gráfico en términos de regresión a la media?

b. ¿Por qué crees que los estudiantes que obtuvieron puntuaciones extremadamente altas o bajas en el primer examen tienden a acercarse más al promedio en el segundo examen?

c. ¿Cómo podría este concepto aplicarse en otras áreas, como en el análisis de rendimiento en deportes o en la predicción de ventas?

## Ideas Clave

- Especificar una hipótesis y luego recopilar datos siguiendo los principios de aleatorización y muestreo aleatorio asegura contra el sesgo.
- La regresión a la media, que significa "volver atrás", es distinta del método de modelado estadístico de regresión lineal, en el cual se estima una relación lineal entre variables predictoras y una variable de resultado.

## 2.3 Distribución Muestral de una Estadística

El término "distribución muestral de una estadística" se refiere a la distribución de una estadística muestral sobre muchas muestras tomadas de la misma población. Gran parte de la estadística clásica se ocupa de hacer inferencias de (pequeñas) muestras a (muy grandes) poblaciones.

### Términos Clave para la Distribución Muestral

#### **Estadística muestral**

Una métrica calculada para una muestra de datos extraída de una población más grande.

#### **Distribución de datos**

La distribución de frecuencias de valores individuales en un conjunto de datos.

#### **Distribución muestral**

La distribución de frecuencias de una estadística muestral sobre muchas muestras o resamples.

#### **Teorema del límite central**

La tendencia de la distribución muestral a adoptar una forma normal a medida que aumenta el tamaño de la muestra.

#### **Error estándar**

La variabilidad (desviación estándar) de una estadística muestral en muchas muestras (no debe confundirse con la desviación estándar, que por sí sola se refiere a la variabilidad de valores individuales de datos).

## Explicaciones Adicionales

Típicamente, se toma una muestra con el objetivo de medir algo (con una estadística muestral) o modelar algo (con un modelo estadístico o de aprendizaje automático).

Dado que nuestra estimación o modelo se basa en una muestra, podría tener errores; podría ser diferente si tomáramos una muestra diferente. Por lo tanto, nos interesa saber qué tan diferente podría ser: una preocupación clave es la variabilidad muestral. Si

tuviéramos muchos datos, podríamos tomar muestras adicionales y observar directamente la distribución de una estadística muestral.

Es importante distinguir entre la distribución de los puntos de datos individuales, conocida como la distribución de datos, y la distribución de una estadística muestral, conocida como la distribución muestral.

Esto se ilustra en un ejemplo usando el ingreso anual de los solicitantes de préstamos de LendingClub.

La tabla 2.1 muestra algunos registros de datos de préstamos personales de LendingClub. LendingClub es un líder en préstamos entre personas (peer-to-peer lending), donde grupos de inversionistas otorgan préstamos personales a individuos.

**Table 2.1. Algunos registros y columnas de datos de préstamos de LendingClub**

Outcome	Loan amount	Income	Purpose	Years employed	Home ownership	State
Paid off	10000	79100	debt_consolidation	11	MORTGAGE	NV
Paid off	9600	48000	moving	5	MORTGAGE	TN
Paid off	18800	120036	debt_consolidation	11	MORTGAGE	MD
Default	15250	232000	small_business	9	MORTGAGE	CA
Paid off	17050	35000	debt_consolidation	4	RENT	MD
Paid off	5500	43000	debt_consolidation	4	RENT	KS

A continuación se presenta una explicación de lo que representa cada columna en la tabla de datos de LendingClub:

- **Outcome:** Indica el resultado del préstamo. Puede tener dos valores principales:
  - **Paid off:** Significa que el préstamo fue pagado en su totalidad por el prestatario.
  - **Default:** Significa que el prestatario incumplió y no pudo pagar el préstamo según lo acordado.
- **Loan amount:** Representa la cantidad de dinero (en dólares) que fue prestada al individuo.
- **Income:** Indica el ingreso anual del prestatario en dólares. Este dato se utiliza para evaluar la capacidad del prestatario para devolver el préstamo.
- **Purpose:** Describe el propósito o razón por la cual el prestatario solicitó el préstamo. Ejemplos comunes incluyen:
  - **debt\_consolidation:** El prestatario solicitó el préstamo para consolidar deudas existentes.
  - **moving:** El préstamo fue solicitado para cubrir gastos de mudanza.
  - **small\_business:** El prestatario solicitó el préstamo para financiar un pequeño negocio.

- **Years employed:** Indica la cantidad de años que el prestatario ha estado empleado. Este dato es relevante para evaluar la estabilidad laboral del prestatario.
- **Home ownership:** Describe el estado de propiedad de la vivienda del prestatario. Los valores comunes incluyen:
  - **MORTGAGE:** El prestatario tiene una hipoteca sobre su vivienda.
  - **RENT:** El prestatario vive en una vivienda alquilada.
- **State:** Indica el estado de los Estados Unidos en el que reside el prestatario, usando abreviaturas de dos letras (por ejemplo, **NV** para Nevada, **CA** para California).

Estas columnas proporcionan información clave que se puede utilizar para evaluar la solvencia del prestatario y el riesgo asociado al préstamo.

Una vez claro el significado de la tabla, vamos a tomar tres muestras de estos datos:

- una muestra de 1,000 valores: simplemente se seleccionarán 1,000 registros individuales de la tabla de datos de préstamos.
- una muestra de 1,000 medias de 5 valores: en lugar de tomar 1,000 registros individuales, se tomarán grupos de 5 registros y se calculará la media (promedio) de un valor específico en cada grupo (por ejemplo, la media del ingreso). Este proceso se repetirá hasta tener 1,000 medias, cada una calculada a partir de un grupo de 5 valores. Este enfoque nos permite observar cómo se comporta la media de pequeños grupos de datos en lugar de observar solo los valores individuales.
- una muestra de 1,000 medias de 20 valores: similar al anterior, pero en lugar de tomar grupos de 5 registros, se tomarán grupos de 20 registros. Nuevamente, se calculará la media para cada grupo y se repetirán los cálculos hasta obtener 1,000 medias, cada una basada en 20 valores. Este tipo de muestra nos da una idea de cómo la media de un tamaño de muestra más grande se comporta en comparación con la media de muestras más pequeñas (en este caso, de 5 valores).

Luego, se hace el histograma de cada muestra para producir la gráfica correspondiente.

```
In [19]: import pandas as pd
import numpy as np
from scipy import stats
from sklearn.utils import resample

import seaborn as sns
import matplotlib.pyplot as plt

loans_income = pd.read_csv('loans_income.csv').squeeze('columns')

sample_data = pd.DataFrame({
    'income': loans_income.sample(1000),
    'type': 'Data',
})

sample_mean_05 = pd.DataFrame({
    'income': [loans_income.sample(5).mean() for _ in range(1000)],
    'type': 'Mean of 5',
})
```

```

    })

sample_mean_20 = pd.DataFrame({
    'income': [loans_income.sample(20).mean() for _ in range(1000)],
    'type': 'Mean of 20',
})

results = pd.concat([sample_data, sample_mean_05, sample_mean_20])
print(results.head())

```

	income	type
42954	38236.0	Data
3781	53000.0	Data
17576	90000.0	Data
42662	52000.0	Data
8768	65000.0	Data

In [20]: `print(results.tail())`

	income	type
995	45286.65	Mean of 20
996	63048.00	Mean of 20
997	73369.50	Mean of 20
998	69334.95	Mean of 20
999	64893.70	Mean of 20

**Ejercicio 2.4:** Explique el código de arriba. Con otro dataset de su preferencia repita el ejercicio de arriba.

In [21]: `# Guardar 'results' en un archivo CSV llamado 'results.csv'  
# results.to_csv('results.csv', index=False)`

Ahora vamos a obtener los histogramas:

```

In [31]: import seaborn as sns
import matplotlib.pyplot as plt

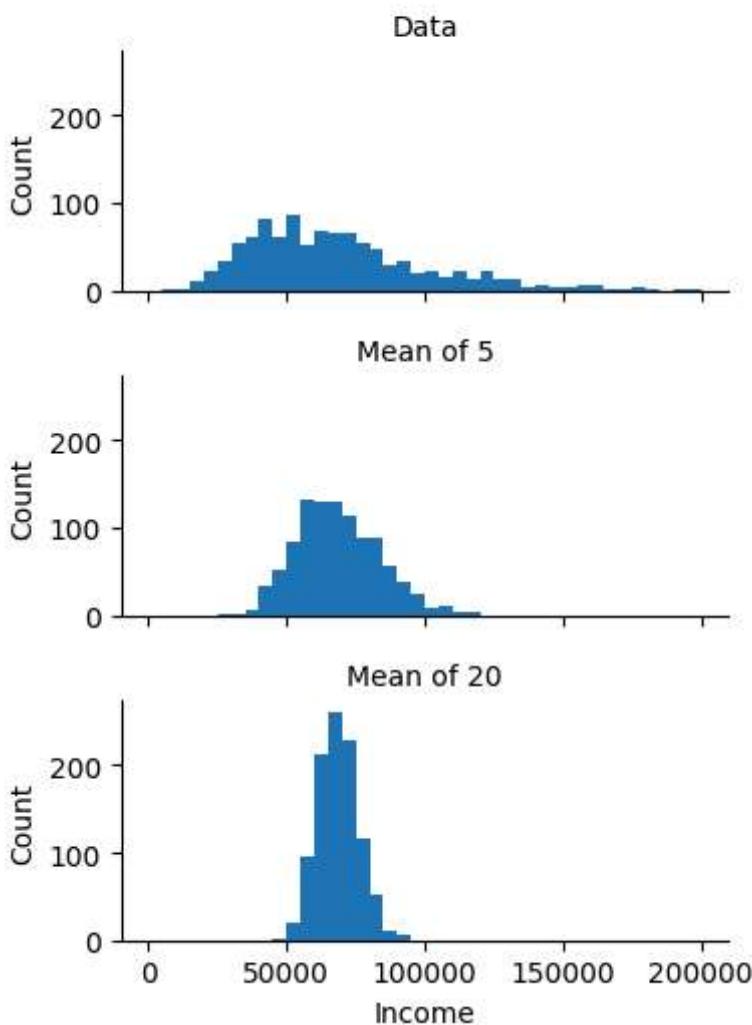
g = sns.FacetGrid(results, col='type', col_wrap=1,
                   height=2, aspect=2)
g.map(plt.hist, 'income', range=[0, 200000], bins=40)
g.set_axis_labels('Income', 'Count')
g.set_titles('{col_name}')

g.fig.suptitle('Figura 2.5', fontsize=14)

plt.tight_layout(rect=[0, 0, 1, 0.95])

plt.show()

```

**Figura 2.5**

**Ejercicio 2.5:** Explique el código dado arriba. Con el mismo dataset usado en el ejercicio 2.4 haga los plots

El histograma de los valores de datos individuales está ampliamente distribuido y sesgado hacia valores más altos, como es de esperar con los datos de ingresos. Los histogramas de las medias de 5 y 20 valores son cada vez más compactos y tienen una forma más similar a una campana.

## 2.4 Teorema del límite central

El fenómeno que acabamos de describir se denomina el teorema del límite central. Este establece que las medias obtenidas de múltiples muestras se asemejarán a la conocida curva normal en forma de campana, incluso si la población de origen no está distribuida normalmente, siempre que el tamaño de la muestra sea lo suficientemente grande y la desviación de los datos respecto a la normalidad no sea demasiado grande. El teorema del límite central permite que se utilicen fórmulas de aproximación normal, como la distribución t por ejemplo.

## Error Estándar

El error estándar es una métrica única que resume la variabilidad en la distribución muestral de una estadística. El error estándar puede estimarse usando una estadística basada en la desviación estándar ( $s$ ) de los valores de la muestra y el tamaño de la muestra ( $n$ ):

$$\text{Error estándar} = SE = \frac{s}{\sqrt{n}}$$

A medida que aumenta el tamaño de la muestra, el error estándar disminuye, lo que corresponde a lo observado en la Figura 2.5.

Considera el siguiente enfoque para medir el error estándar:

1. Recoge una serie de nuevas muestras de la población.
2. Para cada nueva muestra, calcula la estadística (por ejemplo, la media).
3. Calcula la desviación estándar de las estadísticas calculadas en el paso 2; utiliza esto como tu estimación del error estándar.

En la práctica, este enfoque de recolectar nuevas muestras para estimar el error estándar no suele ser muy eficiente. En su lugar, puedes utilizar remuestreo **bootstrap**. En la estadística moderna, el bootstrap se ha convertido en la forma más usada de estimar el error estándar. Puede utilizarse para prácticamente cualquier estadística y no depende del teorema del límite central u otros supuestos de distribución.

## Desviación Estándar versus Error Estándar

No confundir la desviación estándar (que mide la variabilidad de puntos de datos individuales) con el error estándar (que mide la variabilidad de una métrica de la muestra).

## Ideas Clave

- La distribución de frecuencias de una estadística muestral nos dice cómo esa métrica podría variar de una muestra a otra.
- Esta distribución muestral puede estimarse a través del bootstrap o mediante fórmulas que dependen del teorema del límite central.
- Una métrica clave que resume la variabilidad de una estadística muestral es su error estándar.

## Bootstrap

Es una técnica de remuestreo que se utiliza para estimar la distribución muestral de una estadística, como la media o los parámetros de un modelo, a partir de una única muestra de datos. El bootstrap es una técnica estadística que consiste en tomar muestras repetidas (con reemplazo) de un conjunto de datos observado para poder estimar la distribución de una estadística muestral. Esto se hace sin necesidad de hacer

suposiciones fuertes sobre la distribución original de los datos (por ejemplo, no es necesario asumir que los datos se distribuyen normalmente).

## Términos Clave para el Bootstrap

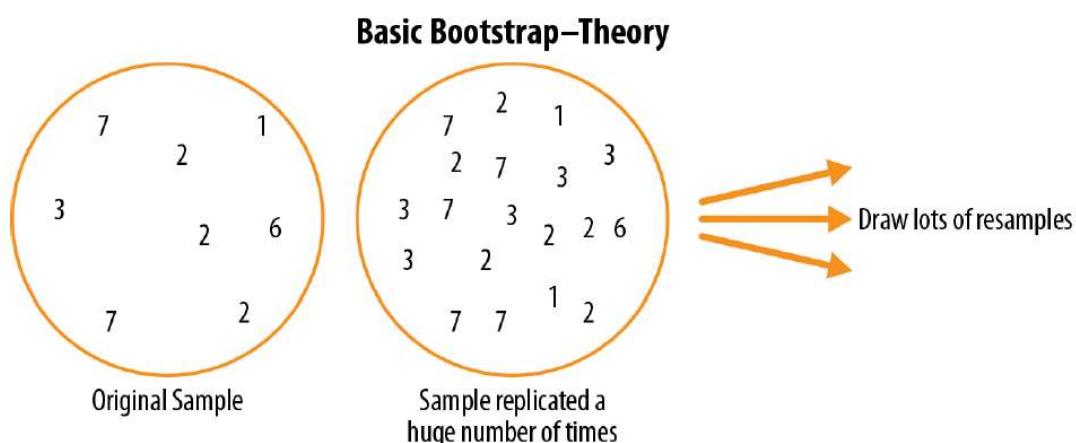
### Bootstrap Sample

Una muestra tomada con reemplazo de un conjunto de datos observado.

### Remuestreo (Resampling)

El proceso de tomar muestras repetidas de datos observados; incluye tanto procedimientos de bootstrap como de permutación (shuffling).

Conceptualmente, se puede imaginar el bootstrap como replicar la muestra original miles o millones de veces para que tengas una población hipotética que encarna todo el conocimiento de tu muestra original (simplemente es más grande). Luego puedes tomar muestras de esta población hipotética con el propósito de estimar una distribución muestral; ver la Figura 2.6.



## ¿Cómo funciona el Bootstrap?

El proceso de bootstrap se puede resumir en los siguientes pasos:

### 1. Toma de muestras con reemplazo:

Se toma una muestra del conjunto de datos original. Lo importante aquí es que esta muestra se toma con reemplazo, lo que significa que después de seleccionar un dato, este se vuelve a colocar en la población para que pueda ser seleccionado de nuevo. Esto permite crear variaciones en las muestras que se parecen a la población original.

### 2. Cálculo de la estadística:

Para cada una de estas muestras generadas, se calcula la estadística de interés, como la media, la desviación estándar, o un parámetro de un modelo.

### 3. Repetición del proceso:

Este proceso de muestreo y cálculo se repite muchas veces (denotado como ( R ) iteraciones). Cada iteración produce una nueva estimación de la estadística.

### 4. Análisis de los resultados:

Con el conjunto de estadísticas obtenidas de todas las iteraciones, se pueden calcular métricas adicionales como la desviación estándar de las estadísticas obtenidas (lo que da una estimación del error estándar), se pueden construir intervalos de confianza, o visualizar la distribución muestral con un histograma o un diagrama de caja.

Los principales paquetes de Python no proporcionan implementaciones del enfoque bootstrap. Sin embargo, se puede implementar utilizando el método `resample` de scikit-learn:

In [33]:

```
results = []
for nrepeat in range(1000):
    sample = resample(loans_income)
    results.append(sample.median())
results = pd.Series(results)
print('Bootstrap Statistics:')
print(f'original: {loans_income.median()}')
print(f'bias: {results.mean() - loans_income.median()}')
print(f'std. error: {results.std()}')
```

```
Bootstrap Statistics:
original: 62000.0
bias: -76.83099999999831
std. error: 216.62910956311967
```

#### 1. Original: 62000.0

- Este valor es la mediana original de los ingresos ( `loans_income` ) antes de aplicar el bootstrap. La mediana es el valor que separa la mitad superior de los ingresos de la mitad inferior. En este caso, la mediana de los ingresos en la muestra original es de **62,000**.

#### 2. Bias: -76.83099999999831

- El sesgo (**bias**) se calcula como la diferencia entre la media de las medianas obtenidas en las iteraciones del bootstrap y la mediana original de los datos. En este caso, el sesgo es de aproximadamente **-76.83**, lo que indica que, en promedio, las medianas calculadas a partir de las muestras bootstrap son ligeramente menores que la mediana original. Un sesgo negativo sugiere que las muestras tienden a subestimar la mediana original.

En el contexto del bootstrap, el **sesgo** (bias) se refiere a la diferencia sistemática entre la estimación obtenida a partir de los datos muestrales (en este caso, las medianas obtenidas a través del bootstrap) y el valor observado en los datos originales (la mediana original de los ingresos).

En la salida proporcionada, el sesgo obtenido significa que, en promedio, las medianas calculadas a partir de las 1,000 muestras bootstrap son **76.830 unidades menores** que la mediana original de la muestra de ingresos (`loans_income`).

#### Dirección del Sesgo:

- El signo negativo (-) indica la dirección del sesgo. En este caso, sugiere que las muestras bootstrap tienden a subestimar la mediana verdadera de la población. Si el sesgo fuera positivo, implicaría una sobreestimación sistemática.

#### Magnitud del Sesgo:

- La magnitud del sesgo (76.830) nos da una idea de cuánto difieren, en promedio, las medianas bootstrap de la mediana original. Aunque un sesgo de -76.830 podría parecer pequeño en relación con la mediana original de 62,000, es importante en contextos donde la precisión es crítica.

El sesgo puede surgir por varias razones, incluyendo:

- **Naturaleza de la Muestra Original:** Si la muestra original no es completamente representativa de la población subyacente, el bootstrap podría reflejar y amplificar ese sesgo.
- **Distribución Asimétrica de los Datos:** Si los datos están sesgados (por ejemplo, con una cola larga a la derecha), la mediana en las muestras bootstrap podría estar sistemáticamente desviada hacia un lado en relación con la mediana original.

#### 3. Std. error: 216.62910956311967

- El error estándar (**std. error**) mide la variabilidad de las medianas obtenidas a partir del bootstrap. Es una estimación de la desviación estándar de la distribución muestral de la mediana. En este caso, el error estándar es de aproximadamente **216.63**, lo que proporciona una idea de cuánto podrían variar las medianas si se tomaran diferentes muestras de la misma población.

**Ejercicio 2.6:** Explicar el código de bootstrap dado arriba. Con otra dataset de tu preferencia repite el código de arriba e interpreta resultados.

## Próxima clase: Intervalos de confianza y principales distribuciones estadísticas.

In [ ]: