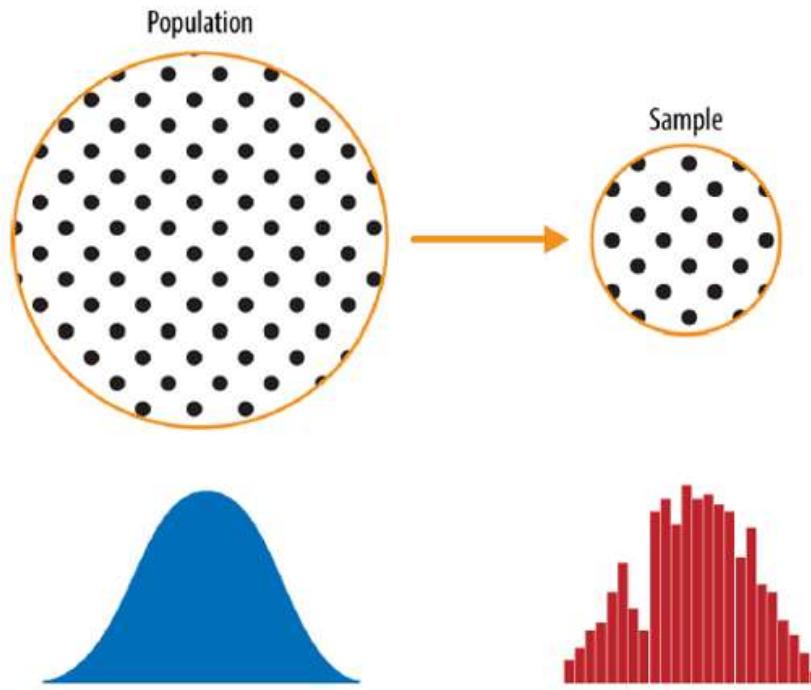


Capítulo 2 - Distribuciones de datos y muestreo

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.

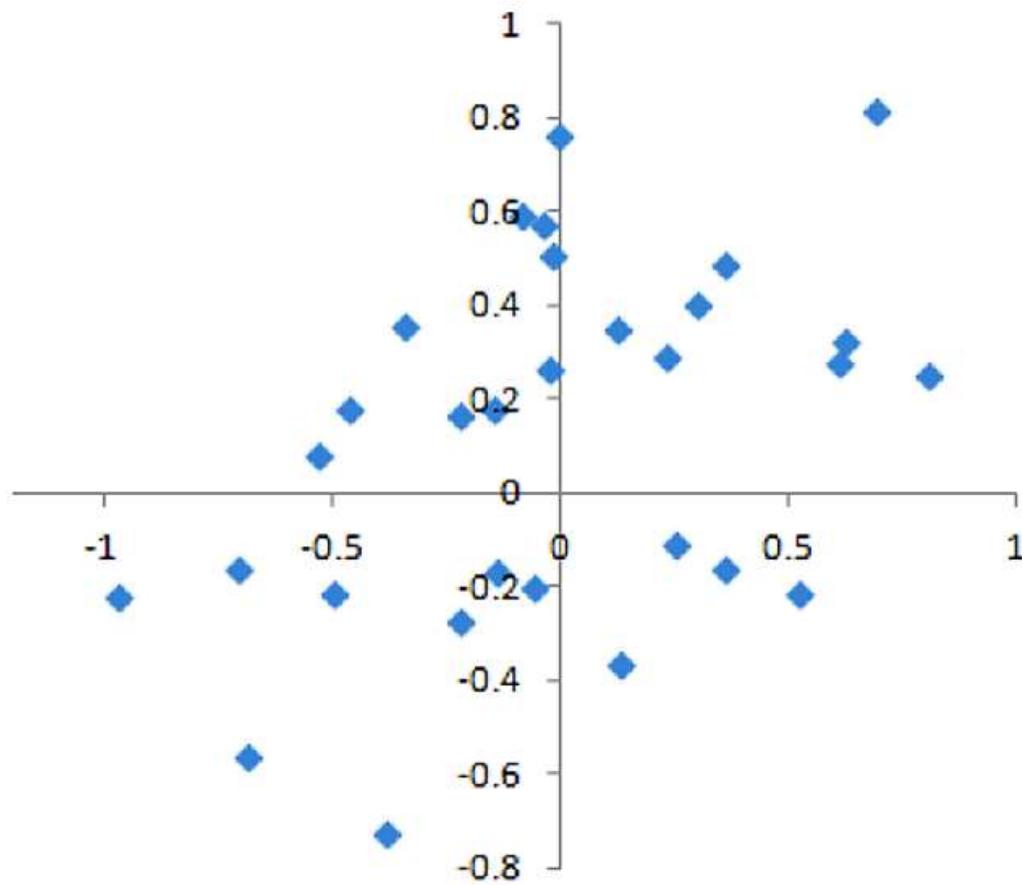
2.2.2 Sesgo de Muestreo por Autoselección

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.

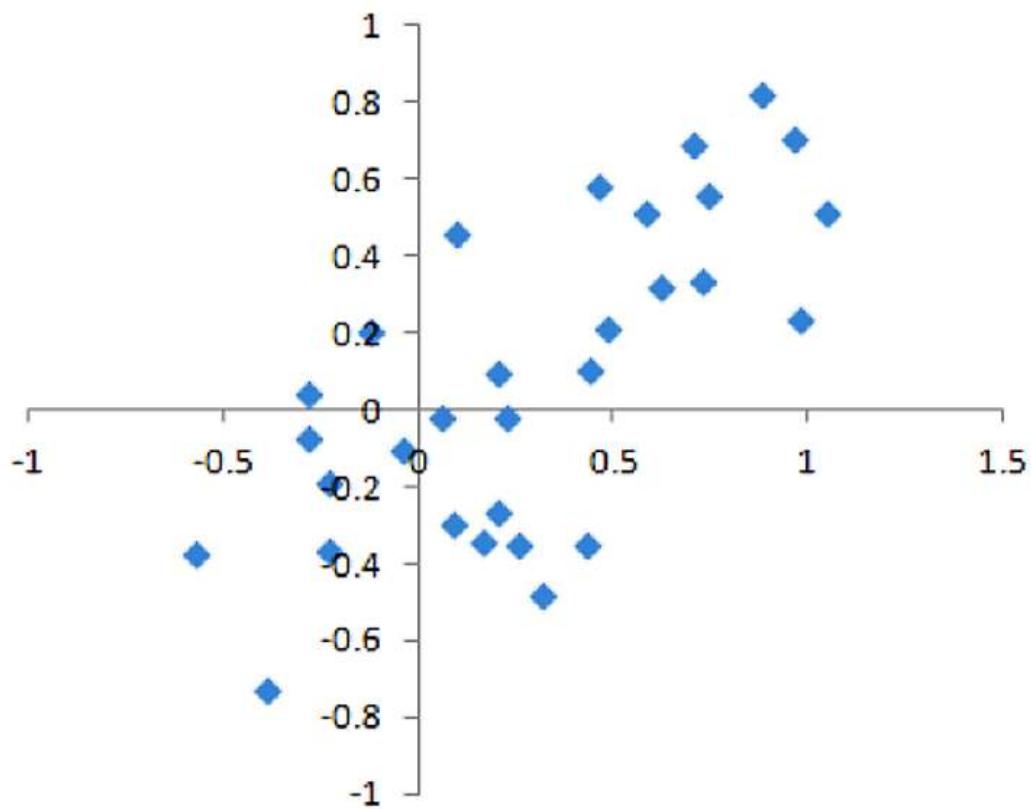
2.2.3 Sesgo

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 μ 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

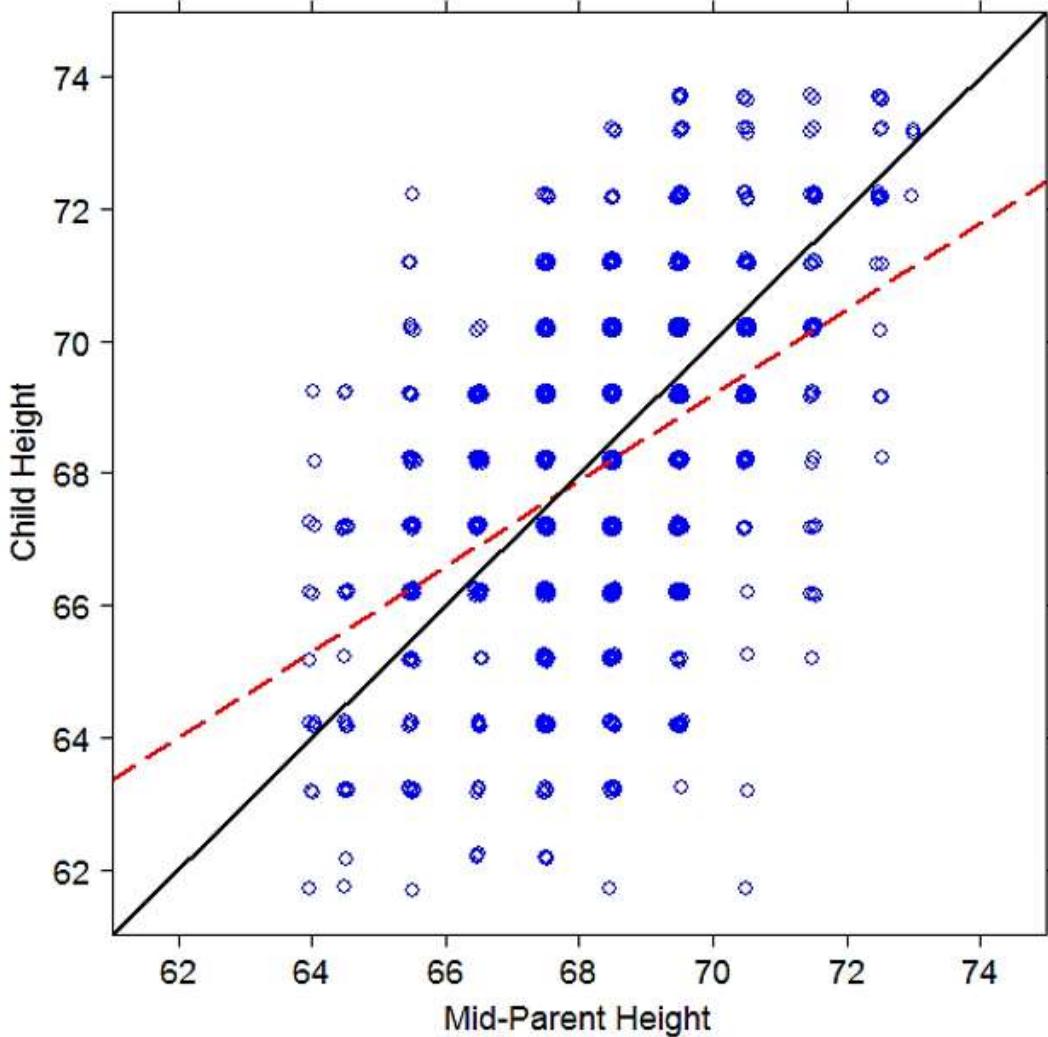
La regresión a la media se refiere a un fenómeno que involucra mediciones sucesivas en una variable dada: las observaciones extremas tienden a ser seguidas por otras más cercanas a la media. Dar un enfoque especial y significado a un valor extremo puede llevar a una forma de sesgo de selección.

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 [6]: 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
45293	72000.0	Data
42993	93000.0	Data
29339	87500.0	Data
37329	69370.0	Data
7852	35000.0	Data

In [7]: `print(results.tail())`

	income	type
995	65165.9	Mean of 20
996	84587.0	Mean of 20
997	77984.5	Mean of 20
998	72194.7	Mean of 20
999	70188.1	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 [8]: import seaborn as sns
import matplotlib.pyplot as plt

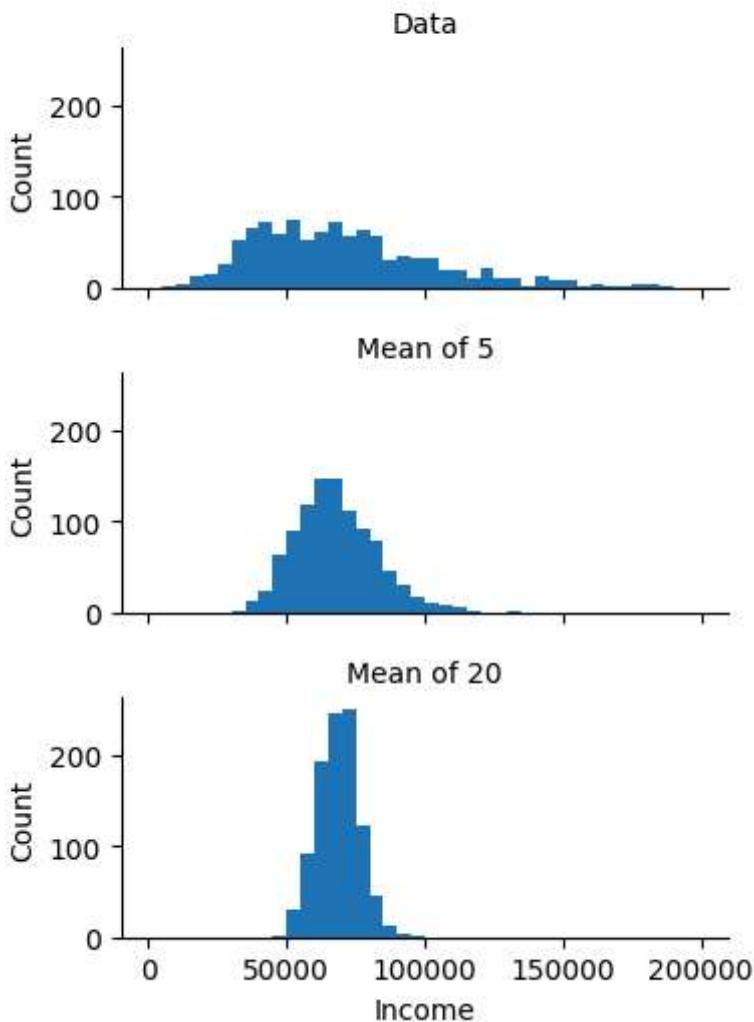
g = sns.FacetGrid(results, col='type', col_wrap=1,
                   height=2, aspect=2)
g.map(plt.hist, 'income', range=[0, 200000], bins=40)
g.set_axis_labels('Income', 'Count')
g.set_titles('{col_name}')

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.

2.6 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

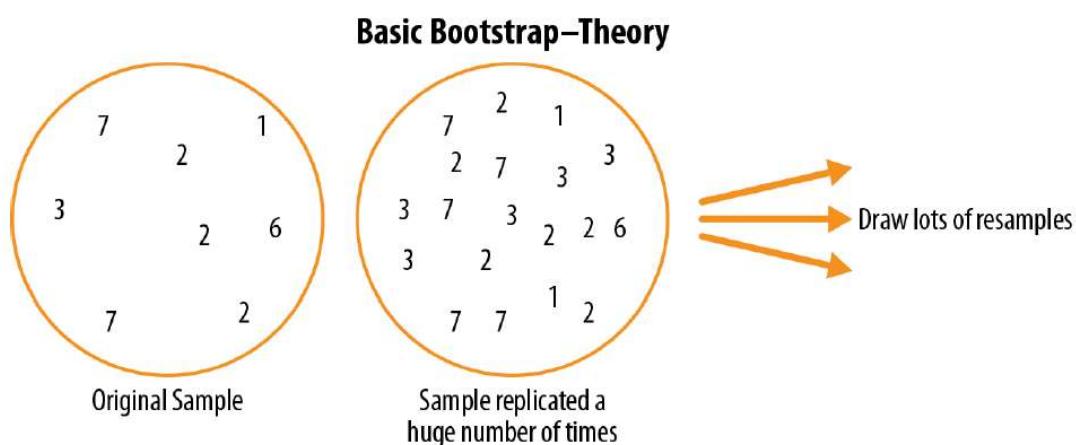
Bootstrap Sample

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 [9]:

```
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: -79.29099999999744
std. error: 227.78612191337808
```

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: -79.29099999999744

- 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 **-79.29099999999744**, 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 (79.30) 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: 227.78612191337808

- 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 **227.78612191337808**, 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.

2.7 Intervalos de confianza.

Las tablas de frecuencias, histogramas, diagramas de caja (boxplots) y errores estándar son formas de entender el posible error en una estimación muestral. Los intervalos de confianza también ayudan a encontrar dichos errores.

Existe una aversión natural a la incertidumbre; las personas (especialmente los expertos) dicen "No sé" con demasiada poca frecuencia. Los analistas y gerentes, aunque reconocen la incertidumbre, tienden a depositar una fe excesiva en una estimación cuando se presenta como un solo número (una estimación puntual). Presentar una

estimación no como un solo número, sino como un rango. Los intervalos de confianza hacen esto de una manera fundamentada en los principios del muestreo estadístico.

Un intervalo de confianza es un rango de valores, derivado de datos muestrales, que se utiliza para estimar un parámetro desconocido de una población. Este rango viene acompañado de un nivel de confianza que indica la probabilidad de que el intervalo contenga el verdadero valor del parámetro poblacional.

En términos prácticos, si se construye un intervalo de confianza del 95% a partir de una muestra, esto significa que si se tomaran muchas muestras similares y se construyeran intervalos de confianza para cada una, aproximadamente el 95% de esos intervalos contendrían el verdadero valor del parámetro poblacional.

Por ejemplo, si calculas un intervalo de confianza del 95% para la media de una población, el intervalo te proporcionará un rango de valores dentro del cual es razonable esperar que caiga la verdadera media poblacional, con un 95% de certeza.

In [2]:

```
import numpy as np
import scipy.stats as stats

np.random.seed(42)
data = np.random.normal(loc=50, scale=10, size=100) # Media = 50, Desviación est

mean = np.mean(data)

sem = stats.sem(data) # sem = standard error of the mean

confidence_interval = stats.t.interval(0.95, len(data)-1, loc=mean, scale=sem)

ci_lower = float(confidence_interval[0])
ci_upper = float(confidence_interval[1])

print(f"Media de la muestra: {mean:.2f}")
print(f"Intervalo de confianza del 95%: ({ci_lower:.2f}, {ci_upper:.2f})")
```

Media de la muestra: 48.96

Intervalo de confianza del 95%: (47.16, 50.76)

Media de la muestra: La media de los datos generados es 48.96. Esto es la estimación puntual de la media de la población.

Intervalo de Confianza del 95%: El intervalo de confianza del 95% para la media está entre 47.16 y 50.76. Esto significa que, si tomamos muchas muestras de la misma población y calculamos un intervalo de confianza del 95% para cada una, aproximadamente el 95% de esos intervalos contendrán la verdadera media poblacional. En este caso, podemos estar un 95% seguros de que la verdadera media de la población de la cual proviene la muestra se encuentra entre 47.16 y 50.76.

Ejercicio 2.7: Explicar cada linea del código. Repite el ejercicio con otra muestra de datos y usando un intervalo de confianza del 98%

Ejercicio 2.8. Repetir el ejemplo anterior pero esta vez utilice statsmodels en vez de scipy. Compare resultados.

Dado un tamaño de muestra (n), y una estadística muestral de interés, el algoritmo para un intervalo de confianza bootstrap es el siguiente:

1. Tomar una muestra aleatoria de tamaño (n) con reemplazo de los datos (un remuestreo).
2. Registrar la estadística de interés para el remuestreo.

Estadística de interés: Puede ser cualquier medida que estés estudiando, como la media, la mediana, la desviación estándar, etc. Después de tomar la muestra con reemplazo en el paso 1, calculas esta estadística para la muestra seleccionada.

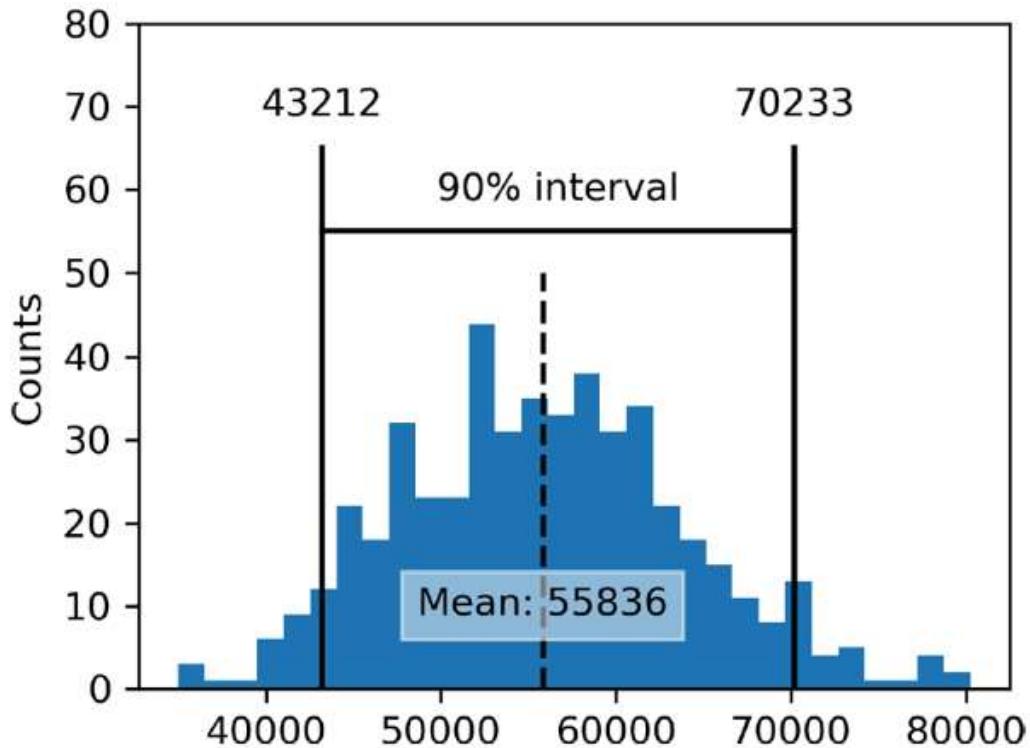
3. Repetir los pasos 1-2 muchas veces ((R) veces). Cuantas más iteraciones hagas, más precisa será la estimación del intervalo de confianza.
4. Para un intervalo de confianza de $x\%$, recortar $\left[\frac{100-x}{2}\right]\%$ de los resultados de los (R) remuestreos desde cada extremo de la distribución.

Recorte: Después de tener todas las estadísticas calculadas en los (R) remuestreos, ordenas estas estadísticas de menor a mayor. Luego, para un intervalo de confianza de $x\%$, recortas los valores extremos, es decir, eliminas el $\left[\frac{100-x}{2}\right]\%$ de los valores más bajos y el mismo porcentaje de los valores más altos. Por ejemplo, para un intervalo de confianza del 95%, recortarías el 2.5% inferior y el 2.5% superior de las estadísticas obtenidas de los R remuestreos.

5. Los puntos de recorte son los extremos de un intervalo de confianza bootstrap de $x\%$.

Puntos de recorte: Después de recortar los extremos, los valores más bajo y más alto que quedan forman los extremos del intervalo de confianza. Es decir, el intervalo de confianza estará entre estos dos puntos, y con un $x\%$ de confianza, podemos decir que este intervalo contiene la verdadera estadística poblacional.

La Figura 2.9 muestra un intervalo de confianza del 90% para el ingreso anual promedio de los solicitantes de préstamos, basado en una muestra de 20 en la que la media fue de \$55.836.



In []:

```
In [10]: print(loans_income.mean())
np.random.seed(seed=3)
# create a sample of 20 Loan income data
sample20 = resample(loans_income, n_samples=20, replace=False)
print(sample20.mean())
results = []
for nrepeat in range(500):
    sample = resample(sample20)
    results.append(sample.mean())
results = pd.Series(results)

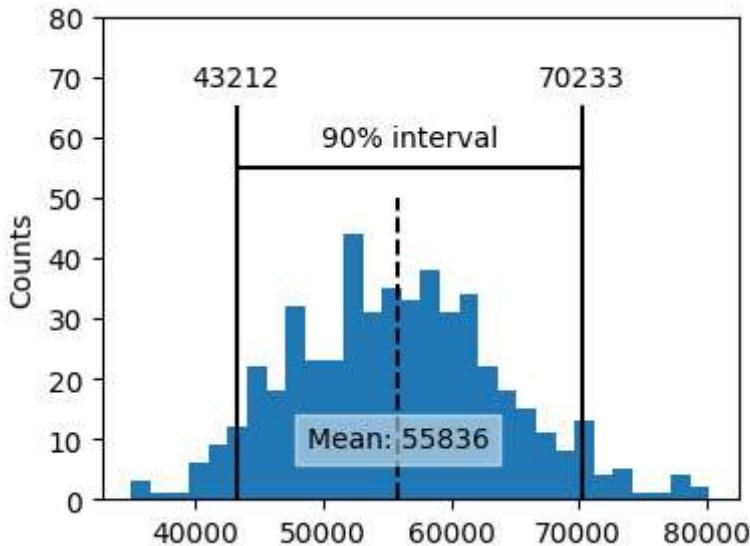
confidence_interval = list(results.quantile([0.05, 0.95]))
ax = results.plot.hist(bins=30, figsize=(4, 3))
ax.plot(confidence_interval, [55, 55], color='black')
for x in confidence_interval:
    ax.plot([x, x], [0, 65], color='black')
    ax.text(x, 70, f'{x:.0f}', horizontalalignment='center', verticalalignment='center')
ax.text(sum(confidence_interval) / 2, 60, '90% interval',
       horizontalalignment='center', verticalalignment='center')

meanIncome = results.mean()
ax.plot([meanIncome, meanIncome], [0, 50], color='black', linestyle='--')
ax.text(meanIncome, 10, f'Mean: {meanIncome:.0f}', bbox=dict(facecolor='white', edgecolor='white', alpha=0.5),
        horizontalalignment='center', verticalalignment='center')
ax.set_xlim(0, 80)
ax.set_ylabel('Counts')

plt.tight_layout()
plt.show()
```

68760.51844

55734.1



El bootstrap es una herramienta general que se puede utilizar para generar intervalos de confianza para la mayoría de las estadísticas o parámetros de modelos. Los libros de texto de estadística y el software, con raíces en más de medio siglo de análisis estadístico sin computadoras, también hacen referencia a intervalos de confianza generados por fórmulas, especialmente la distribución t.

Por supuesto, lo que realmente nos interesa cuando tenemos un resultado muestral es: "¿Cuál es la probabilidad de que el valor verdadero esté dentro de un cierto intervalo?" Esta no es realmente la pregunta que responde un intervalo de confianza, pero termina siendo la forma en que la mayoría de las personas interpretan la respuesta.

El porcentaje asociado con el intervalo de confianza se denomina nivel de confianza. Cuanto mayor sea el nivel de confianza, más amplio será el intervalo. Esto sucede porque, al querer estar más seguro de que el intervalo contiene el valor verdadero, se necesita un rango más amplio para cubrir todas las posibles variaciones en los datos muestrales. Por ejemplo, un intervalo de confianza del 99% será más amplio que un intervalo de confianza del 95%, porque queremos estar más seguros (99% en lugar de 95%) de que el intervalo contiene el valor verdadero, y para lograr esa mayor seguridad, ampliamos el intervalo.

Además, cuanto más pequeña sea la muestra, más amplio será el intervalo (es decir, mayor será la incertidumbre). Esto sucede porque con menos datos, hay más incertidumbre acerca de la estimación del parámetro verdadero. Para compensar esa incertidumbre, se necesita un intervalo más amplio.

Ambos aspectos tienen sentido: cuanto más seguro quieras estar, y cuanto menos datos tienes, más amplio debe ser el intervalo de confianza para estar suficientemente seguro de capturar el valor verdadero. Un tamaño de muestra pequeño significa que tienes menos información sobre la población, lo que introduce más variabilidad e incertidumbre en tus estimaciones. Para tener la misma seguridad (por ejemplo, un 95% de confianza), el intervalo debe ser más amplio para asegurarse de que incluye el valor verdadero, dado que hay más margen de error. Por el contrario, con un tamaño de

muestra más grande, tienes más información, lo que reduce la incertidumbre y permite calcular un intervalo de confianza más estrecho.

Ejemplo para Ilustrar el Concepto

Imagina que estás midiendo la altura promedio de estudiantes en una escuela. Si quieres estar 95% seguro de que el intervalo de confianza contiene la altura promedio verdadera, el intervalo podría ser algo así como de 160 cm a 170 cm. Si decides que necesitas estar 99% seguro, podrías necesitar ampliar el intervalo a algo como de 155 cm a 175 cm para capturar todas las posibles variaciones. Si tomas la altura de solo 5 estudiantes (una muestra pequeña), podrías obtener un intervalo de confianza de 150 cm a 180 cm debido a la mayor incertidumbre. Pero si tomas la altura de 50 estudiantes (una muestra más grande), podrías obtener un intervalo más preciso, digamos de 160 cm a 170 cm.

Ideas Clave

- Los intervalos de confianza son la forma típica de presentar estimaciones como un rango de intervalos.
- Cuantos más datos tengas, menos variable será una estimación muestral.
- Cuanto menor sea el nivel de confianza que puedas tolerar, más estrecho será el intervalo de confianza.
- El bootstrap es una forma efectiva de construir intervalos de confianza.

Ejercicio 2.9. Investigar las aplicaciones y casos de uso de Bootstrapping en Machine Learning. Mostrar los resultados en un resumen.

Ejercicio 2.10 Replicar en un jupyter notebook los resultados del siguiente artículo: <https://cienciadedatos.net/documentos/pystats04-bootstrapping-python> Nota: Con replicar los códigos sería suficiente para entregar; replicar la teoría es opcional aunque por lo menos debes leerla si realmente quieras entender la importancia del Bootstrapping

2.8 Distribución Normal

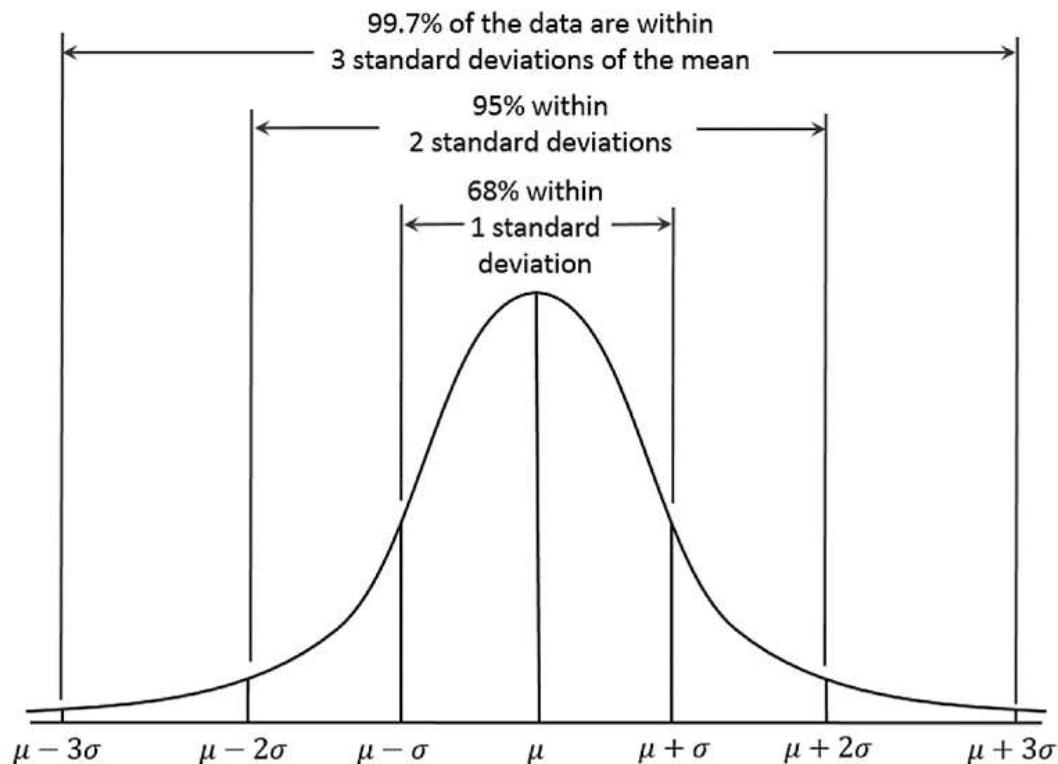
La distribución normal en forma de campana es icónica en la estadística tradicional. El hecho de que las distribuciones de las estadísticas muestrales a menudo tengan forma normal la ha convertido en una herramienta poderosa en el desarrollo de fórmulas matemáticas que aproximan esas distribuciones.

Términos Clave para la Distribución Normal

- **Error:** La diferencia entre un punto de datos y un valor predicho o promedio.
- **Estandarizar:** Restar la media y dividir por la desviación estándar.
- **Puntuación z (z-score):** El resultado de estandarizar un punto de datos individual.

- **Normal estándar:** Una distribución normal con media = 0 y desviación estándar = 1.
- **Gráfico QQ (QQ-Plot):** Un gráfico para visualizar qué tan cercana está una distribución muestral a una distribución especificada, por ejemplo, la distribución normal.

En una distribución normal (Figura 2.8), el 68% de los datos se encuentra dentro de una desviación estándar de la media, y el 95% se encuentra dentro de dos desviaciones estándar.



En el centro de la distribución se encuentra en la media μ . Es el valor promedio, donde la curva alcanza su punto más alto. A medida que te alejas de la media, la probabilidad disminuye simétricamente a ambos lados.

La desviación estándar σ mide la dispersión de los datos con respecto a la media. En la imagen, se muestran las desviaciones estándar desde la media hacia ambos lados:

- $(\mu - \sigma)$ y $(\mu + \sigma)$: una desviación estándar a la izquierda y derecha de la media.
- $(\mu - 2\sigma)$ y $(\mu + 2\sigma)$: dos desviaciones estándar.
- $(\mu - 3\sigma)$ y $(\mu + 3\sigma)$: tres desviaciones estándar.

Supongamos que tenemos los siguientes datos sobre las calificaciones de 10 estudiantes en un examen:

[60, 70, 80, 90, 85, 75, 95, 85, 80, 100]

La **media** μ , es 82 y la desviación estándar σ es 11.23 . La desviación estándar nos dice cómo se dispersan los datos con respecto a la media.

- **1 desviación estándar:** Aproximadamente el 68% de los datos estarán dentro del rango $(\mu - \sigma)$ y $(\mu + \sigma)$, es decir, entre $82 - 11.23 = 70.77$ y $82 + 11.23 = 93.23$.
- **2 desviaciones estándar:** Aproximadamente el 95% de los datos estarán dentro de 2 desviaciones estándar, es decir, entre $(82 - 2(11.23)) = 59.54$ y $(82 + 2(11.23)) = 104.46$.
- **3 desviaciones estándar:** Aproximadamente el 99.7% de los datos estarán dentro de 3 desviaciones estándar, es decir, entre $(82 - 3(11.23)) = 48.31$ y $(82 + 3(11.23)) = 115.69$.

Probabilidades dentro de las Desviaciones Estándar

La figura dada arriba muestra que en una distribución normal, la mayoría de los datos se concentran cerca de la media, y a medida que te alejas, los datos se vuelven menos frecuentes.

- **68% dentro de una desviación estándar:**
 - Aproximadamente el **68%** de los datos se encuentra dentro de una desviación estándar de la media $(\mu - \sigma)$ a $(\mu + \sigma)$. Esto significa que si la variable sigue una distribución normal, 68 de cada 100 valores estarán dentro de este rango alrededor de la media.
- **95% dentro de dos desviaciones estándar:**
 - Aproximadamente el **95%** de los datos se encuentra dentro de dos desviaciones estándar de la media $(\mu - 2\sigma)$ a $(\mu + 2\sigma)$. En otras palabras, 95 de cada 100 valores caerán dentro de este rango, lo que hace de este intervalo un buen estimador para capturar casi todos los valores.
- **99.7% dentro de tres desviaciones estándar:**
 - Casi todo el conjunto de datos (el **99.7%**) cae dentro de tres desviaciones estándar de la media $(\mu - 3\sigma)$ a $(\mu + 3\sigma)$. Esto significa que es muy raro encontrar datos fuera de este rango. Cualquier valor que se encuentre fuera de este rango puede ser considerado un **outlier** o valor atípico.

La mayor parte de los datos brutos en general — no están distribuidas normalmente. La utilidad de la distribución normal proviene del hecho de que muchas estadísticas están distribuidas normalmente en su **distribución muestral**.

Aun así, las suposiciones de normalidad son generalmente un último recurso, utilizadas cuando las distribuciones de probabilidad empíricas o las distribuciones bootstrap no están disponibles.

La distribución normal también se conoce como **distribución gaussiana**, en honor a Carl Friedrich Gauss, un prolífico matemático alemán de finales del siglo XVIII y principios del siglo XIX. Otro nombre previamente utilizado para la distribución normal fue la **distribución de errores**.

El desarrollo de Gauss de la distribución normal surgió de su estudio de los errores en las mediciones astronómicas, los cuales se encontraron distribuidos normalmente.

Ejemplo

Supongamos que trabajas como analista de datos en una empresa de servicios financieros. Quieres analizar los ingresos mensuales de los clientes actuales para entender mejor su distribución. Estos datos son importantes porque afectan el perfil de riesgo de los clientes y su capacidad para tomar préstamos o pagar productos financieros. Se requiere los ingresos mensuales de los clientes para determinar si siguen una distribución normal y ver cómo se distribuyen alrededor de la media. También queremos identificar si la mayor parte de los clientes tienen ingresos dentro de un rango específico (por ejemplo, alrededor de la media) y si hay outliers (clientes con ingresos muy bajos o muy altos).

```
In [6]: # Importar las librerías necesarias
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm

# Generar algunos datos de ejemplo (ingresos mensuales en USD de clientes)
# Suponemos que los ingresos siguen aproximadamente una distribución normal
np.random.seed(42)
income_data = np.random.normal(loc=5000, scale=1500, size=1000) # Media = 5000

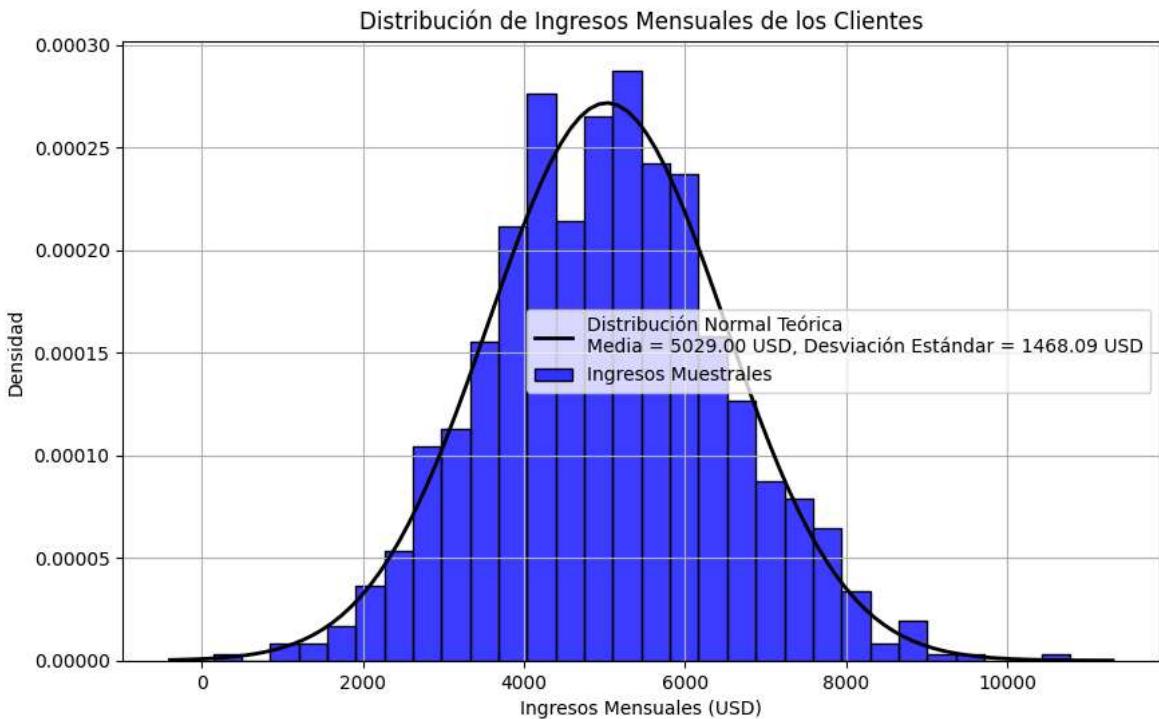
# Calcular la media y la desviación estándar de los ingresos generados
mean_income, std_income = np.mean(income_data), np.std(income_data)

# Crear un histograma de los datos de ingresos con un ajuste de la curva de dist
plt.figure(figsize=(10, 6))
sns.histplot(income_data, bins=30, kde=False, color='blue', stat='density', label='Datos Reales')
xmin, xmax = plt.xlim()

# Generar los valores de la curva de distribución normal
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mean_income, std_income)

# Graficar la curva de la distribución normal teórica
plt.plot(x, p, 'k', linewidth=2, label=f'Distribución Normal Teórica\nMedia = {mean_income:.2f}, Desviación Estándar = {std_income:.2f}')

# Añadir etiquetas y título
plt.title('Distribución de Ingresos Mensuales de los Clientes')
plt.xlabel('Ingresos Mensuales (USD)')
plt.ylabel('Densidad')
plt.legend()
plt.grid(True)
plt.show()
```



Plot con las primeras tres desviaciones típicas

```
In [10]: # Importar las librerías necesarias
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm

# Generar algunos datos de ejemplo (ingresos mensuales en USD de clientes)
np.random.seed(42)
income_data = np.random.normal(loc=5000, scale=1500, size=1000) # Media = 5000

# Calcular la media y desviación estándar de los ingresos generados
mean_income, std_income = np.mean(income_data), np.std(income_data)

# Calcular las primeras tres desviaciones estándar
one_std_range = (mean_income - std_income, mean_income + std_income)
two_std_range = (mean_income - 2*std_income, mean_income + 2*std_income)
three_std_range = (mean_income - 3*std_income, mean_income + 3*std_income)

# Crear el nuevo plot con las desviaciones estándar resaltadas
plt.figure(figsize=(10, 6))

# Histograma de los datos de ingresos
sns.histplot(income_data, bins=30, kde=False, color='skyblue', stat='density', l

# Generar los valores de la curva de distribución normal
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mean_income, std_income)

# Graficar la curva de la distribución normal teórica
plt.plot(x, p, 'k', linewidth=2, label=f'Distribución Normal Teórica\nMedia = {m

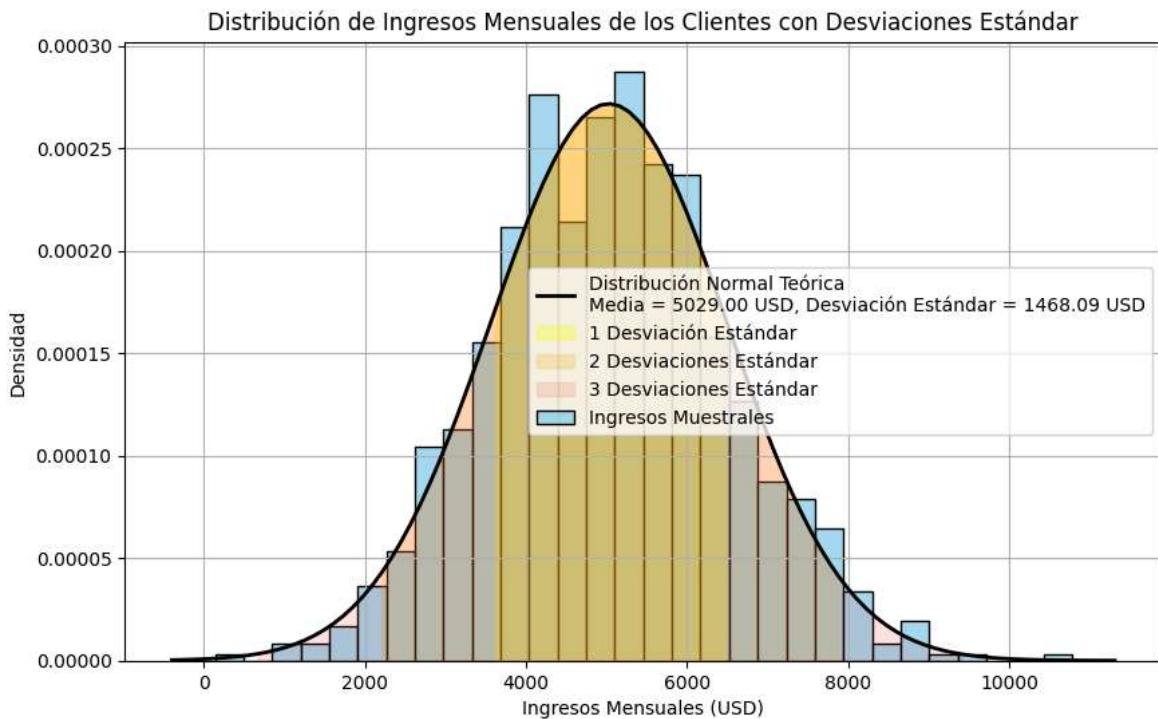
# Colorear las áreas de una, dos y tres desviaciones estándar
plt.fill_between(x, 0, p, where=(x >= one_std_range[0]) & (x <= one_std_range[1])
plt.fill_between(x, 0, p, where=(x >= two_std_range[0]) & (x <= two_std_range[1])
```

```

plt.fill_between(x, 0, p, where=(x >= three_std_range[0]) & (x <= three_std_range[2]), color='blue', alpha=0.2)

# Añadir etiquetas y título
plt.title('Distribución de Ingresos Mensuales de los Clientes con Desviaciones Estándar')
plt.xlabel('Ingresos Mensuales (USD)')
plt.ylabel('Densidad')
plt.legend()
plt.grid(True)
plt.show()

```



Donde, **Densidad** en el eje vertical se calcula a partir de la **función de densidad de probabilidad (PDF)** de una distribución normal $N(\mu, \sigma^2)$ está dada por la fórmula:

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

Donde:

- ($f(x)$) es la densidad de probabilidad en un valor (x),
- (μ) es la media de la distribución,
- (σ) es la desviación estándar,
- (e) es la constante de Euler (aproximadamente 2.718),
- (π) es la constante pi (aproximadamente 3.1416).

Ejercicio 2.11. Con los datos del ejemplo anterior, calcule con python la primera, segunda y la tercera desviación estandar. ¿Hay outliers en los datos?

2.9 Distribución Normal Estándar

La **distribución normal estándar** es un tipo específico de distribución normal en la que:

- La **media** μ es igual a 0.
- La **desviación estándar** σ es igual a 1.

Esto significa que todos los valores en la distribución normal estándar se expresan en términos de cuántas desviaciones estándar se encuentran por encima o por debajo de la media. La distribución normal estándar se utiliza para simplificar comparaciones entre diferentes conjuntos de datos que han sido estandarizados (o normalizados), convirtiendo los valores originales en **z-scores**. La distribución normal a veces se llama **distribución-z**

Fórmula para el z-score:

Para convertir cualquier valor x de una distribución normal en un **z-score** en la distribución normal estándar, se usa la siguiente fórmula:

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

Donde:

- z es el z-score o puntaje estandarizado,
- x es el valor original,
- μ es la media de la distribución,
- σ es la desviación estándar.

Los z-scores indican cuántas desviaciones estándar un valor (x) está alejado de la media. La **distribución normal estándar** es fundamental en estadística, ya que permite comparar diferentes conjuntos de datos y realizar inferencias estadísticas.

QQ-Plots (Quantile-Quantile Plots)

Un **QQ-Plot** (o diagrama quantil-quantil) se utiliza para determinar visualmente cuán cerca está una muestra de una distribución especificada, en este caso, la distribución normal.

La **Figura 2.9** muestra un **QQ-Plot** para una muestra de 100 valores generados aleatoriamente a partir de una distribución normal; como era de esperar, los puntos siguen de cerca la línea. Si los puntos caen aproximadamente en la línea diagonal, entonces se puede considerar que la distribución de la muestra es cercana a la normal.

In [13]:

```

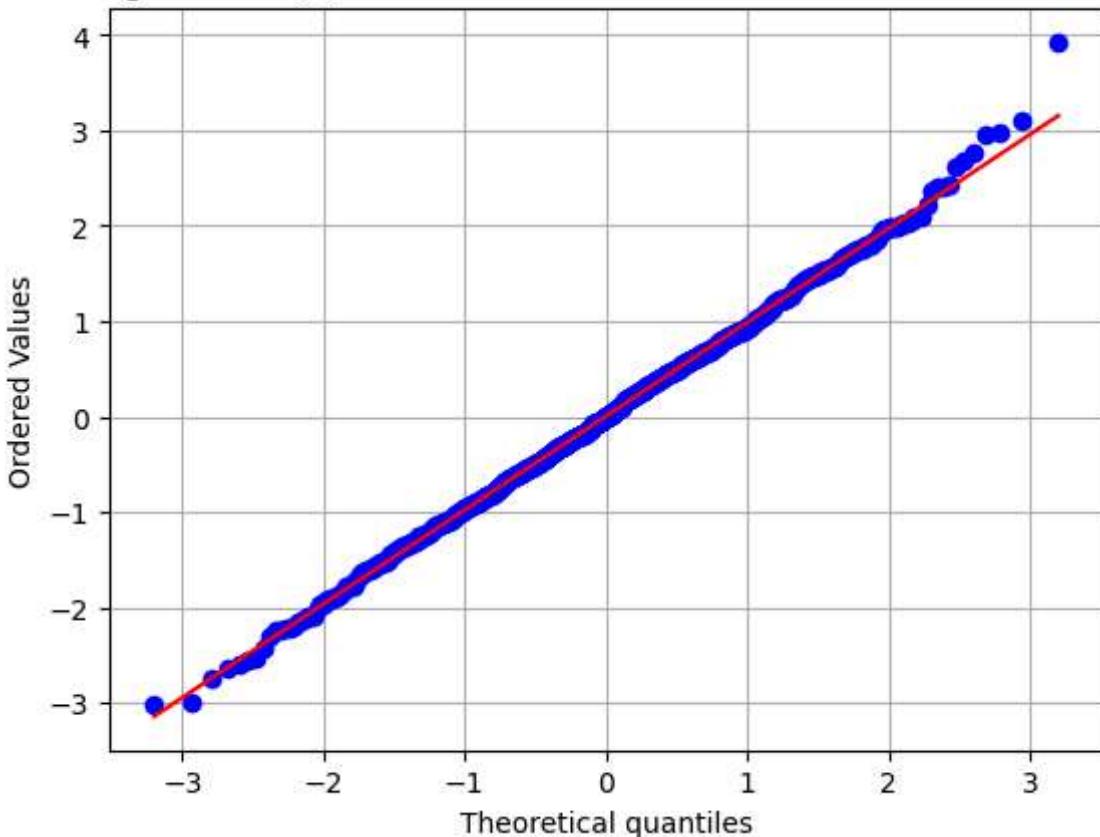
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt

# Generar datos de ejemplo a partir de una distribución normal
data = np.random.normal(0, 1, 1000)

# Crear un QQ-Plot para comparar con la distribución normal
stats.probplot(data, dist="norm", plot=plt)
plt.title("Figura 2.9. QQ-Plot de los datos contra la distribución normal")
plt.grid(True)
plt.show()

```

Figura 2.9. QQ-Plot de los datos contra la distribución normal



Los **QQ-Plots** se utilizan tambien para detectar desviaciones de la distribución teórica, lo cual puede indicar la presencia de **colas largas, asimetría o outliers**.

Para construir un **QQ-Plot** (diagrama cuantil-cuantil), se utilizan **cuantiles de la distribución teórica y cuantiles de los datos observados**. El tipo de cuantil específico depende de la distribución teórica que se esté utilizando para la comparación. Los cuantiles empleados reflejan la posición relativa de los datos dentro de ambas distribuciones, y el proceso es el siguiente:

1. Ordenar los datos observados:

- Primero, se **ordenan** los datos observados de menor a mayor.
- A cada dato se le asigna un **cuantil** que corresponde a su posición en la muestra. En general, si tienes n datos, el i -ésimo dato se asigna al cuantil correspondiente a $\frac{i}{n+1}$ para evitar asignar cuantiles extremos (como 0 o 1).

2. Cuantiles de la distribución teórica:

- A continuación, se generan los **cuantiles teóricos** de la distribución con la cual se están comparando los datos (normal, t-student, exponencial, etc.). Por ejemplo, si estás comparando con una distribución normal, se generan los cuantiles teóricos de esa distribución.

3. Graficar:

- En el eje x , se colocan los **cuantiles teóricos** de la distribución de referencia.
- En el eje y , se colocan los **cuantiles observados** de los datos ordenados.

Si los datos observados siguen la misma distribución que la distribución teórica, los puntos se alinearán aproximadamente en una línea recta.

¿Qué cuantil se usa?

El cuantil que se utiliza para cada punto en el gráfico depende de su **posición relativa** dentro del conjunto de datos. Si tienes n datos observados, entonces:

- Para el valor más pequeño (primer dato ordenado), se toma un cuantil bajo, como $\frac{1}{n+1}$.
- Para el valor más grande (último dato), se toma un cuantil cercano a 1, como $\frac{n}{n+1}$.

Fórmula para los Cuantiles Observados:

Para cada dato en la muestra de tamaño n , el cuantil asociado a la posición i -ésima en los datos observados se aproxima mediante:

$$q_{\text{observado}} = \frac{i}{n + 1}$$

Este valor se utiliza para seleccionar el cuantil correspondiente de la distribución teórica.

Ejemplo:

Supongamos que tienes una muestra de 100 datos:

- El primer dato ordenado tendrá un cuantil de $\frac{1}{101}$ (aproximadamente 0.0099).
- El segundo dato ordenado tendrá un cuantil de $\frac{2}{101}$ (aproximadamente 0.0198).
- El dato en la posición 50 tendrá un cuantil de $\frac{50}{101}$ (aproximadamente 0.495).

Estos cuantiles se comparan con los cuantiles correspondientes de la distribución teórica (por ejemplo, la distribución normal).

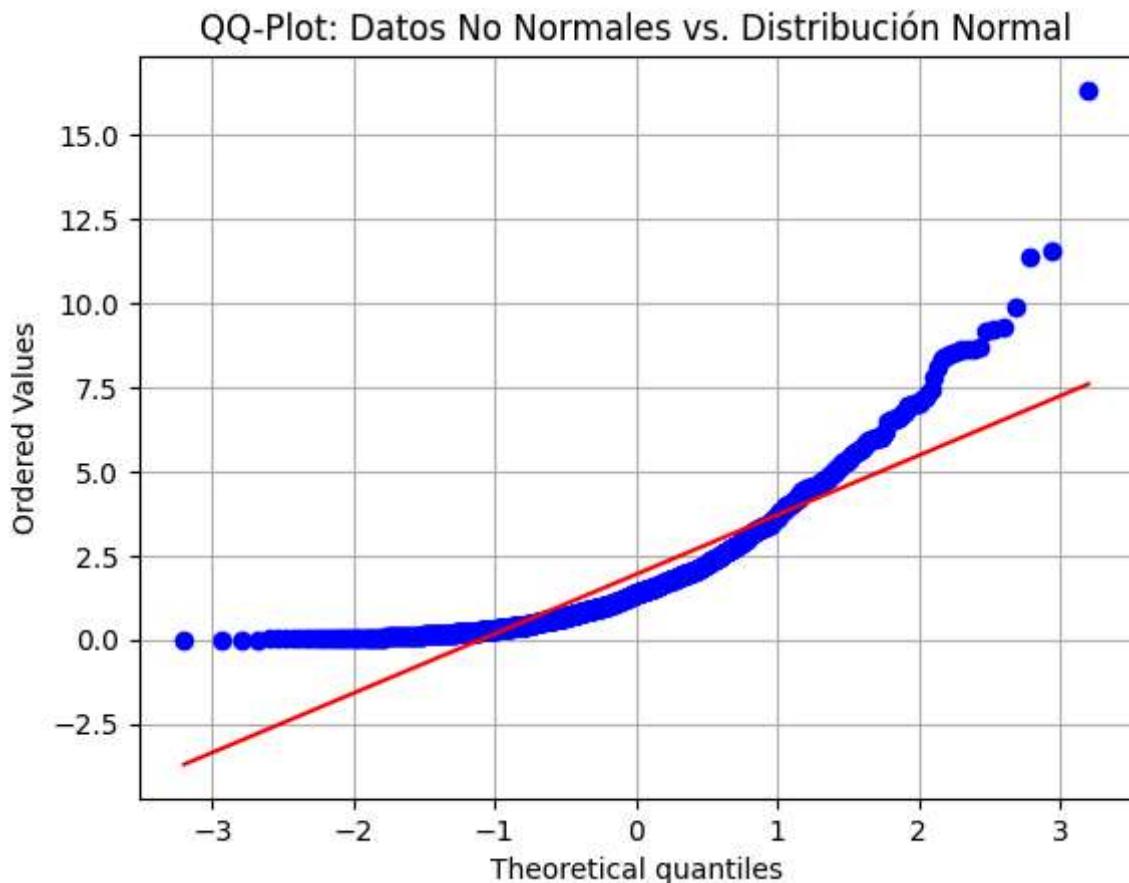
Ejemplo de una distribución de datos que no sigue a la normal

```
In [14]: # Importar las librerías necesarias
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
```

```
# Generar datos de ejemplo que NO sigan una distribución normal (distribución ex
np.random.seed(42)
data_non_normal = np.random.exponential(scale=2, size=1000)

# Crear el QQ-Plot para comparar Los datos no normales con una distribución norm
stats.probplot(data_non_normal, dist="norm", plot=plt)

# Añadir un título
plt.title("QQ-Plot: Datos No Normales vs. Distribución Normal")
plt.grid(True)
plt.show()
```



Convertir los datos a z-score (es decir, estandarizar o normalizar los datos) no hace que los datos sigan una distribución normal. Simplemente coloca los datos en la misma escala que la distribución normal estándar, a menudo para fines de comparación.

Ideas Clave:

- La distribución normal fue esencial para el desarrollo histórico de la estadística, ya que permitió la aproximación matemática de la incertidumbre y la variabilidad.
- Aunque los datos en bruto generalmente no siguen una distribución normal, los errores a menudo sí lo hacen, al igual que los promedios y totales en muestras grandes.
- Para convertir los datos a z-scores, restas la media de los datos y divides por la desviación estándar; luego puedes comparar los datos con una distribución normal.

2.10 Distribuciones de Colas Largas

A pesar de la importancia histórica de la distribución normal en la estadística, y en contraste con lo que sugiere el nombre, los datos generalmente no siguen una distribución normal.

Términos Clave para las Distribuciones de Colas Largas:

- **Cola (tail):** La porción larga y estrecha de una distribución de frecuencias, donde ocurren valores relativamente extremos con baja frecuencia.
- **Sesgo (skew):** Cuando una cola de la distribución es más larga que la otra.

Aunque la distribución normal suele ser apropiada y útil con respecto a la distribución de errores y las estadísticas muestrales, típicamente no caracteriza la distribución de los datos en bruto. A veces, la distribución está muy sesgada (asimétrica), como en el caso de los datos de ingresos; o la distribución puede ser discreta, como en los datos binomiales. Tanto las distribuciones simétricas como las asimétricas pueden tener colas largas. Las colas de una distribución corresponden a los valores extremos (pequeños y grandes). Las colas largas, y la precaución contra ellas, son ampliamente reconocidas en el trabajo práctico. Nassim Taleb ha propuesto la **teoría del cisne negro**, que predice que los eventos anómalos, como un colapso del mercado de valores, son mucho más probables de lo que predice la distribución normal.

La teoría del cisne negro sugiere que los eventos extremadamente raros y anómalos, conocidos como "cisnes negros", ocurren con mucha mayor frecuencia de lo que predice la distribución normal. Estos eventos tienen un impacto significativo y a menudo son impredecibles o inesperados.

Características de los Cisnes Negros: Son sorpresivos: Los cisnes negros son eventos inesperados para quienes los observan. Antes de que ocurran, se consideran improbables o imposibles bajo el paradigma de la distribución normal, que asigna baja probabilidad a los eventos extremos.

Tienen un impacto masivo: Cuando ocurren, los cisnes negros tienen consecuencias enormes y desestabilizan sistemas o sectores enteros. Un ejemplo claro es el colapso del mercado de valores de 2008 o los ataques del 11 de septiembre de 2001.

Se racionalizan a posteriori: Despues de que ocurre un cisne negro, las personas tienden a racionalizarlo y a intentar explicarlo como si fuera predecible, lo que Taleb llama el sesgo retrospectivo.

En lugar de confiar exclusivamente en la distribución normal, Taleb sugiere que debemos utilizar modelos que reconozcan la posibilidad de colas largas, como la **distribución de Pareto** o la **distribución de colas pesadas**.

Un buen ejemplo para ilustrar la naturaleza de colas largas en los datos son los rendimientos de las acciones. Vamos a observar la distribución de los valores de las acciones de Netflix.

In [23]:

```

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

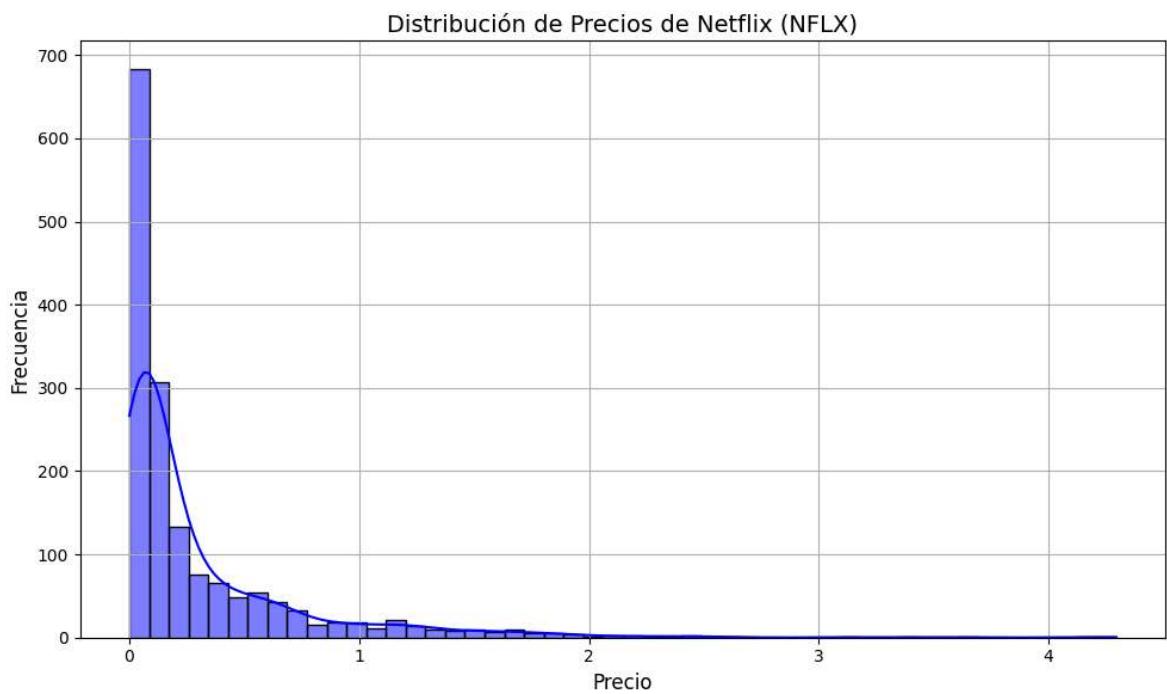
# Cargar los datos del archivo CSV (ajusta la ruta según corresponda)
sp500_px = pd.read_csv('sp500_data.csv.gz')

# Extraer los precios de NFLX (Netflix)
nflx = sp500_px['NFLX']

# Filtrar los datos positivos para evitar problemas con logaritmos
nflx = nflx[nflx > 0]

# Visualizar la distribución de los precios de Netflix
plt.figure(figsize=(10, 6))
sns.histplot(nflx, bins=50, kde=True, color='blue')
plt.title('Distribución de Precios de Netflix (NFLX)', fontsize=14)
plt.xlabel('Precio', fontsize=12)
plt.ylabel('Frecuencia', fontsize=12)
plt.grid(True)
plt.tight_layout()
plt.show()

```



In [20]:

```

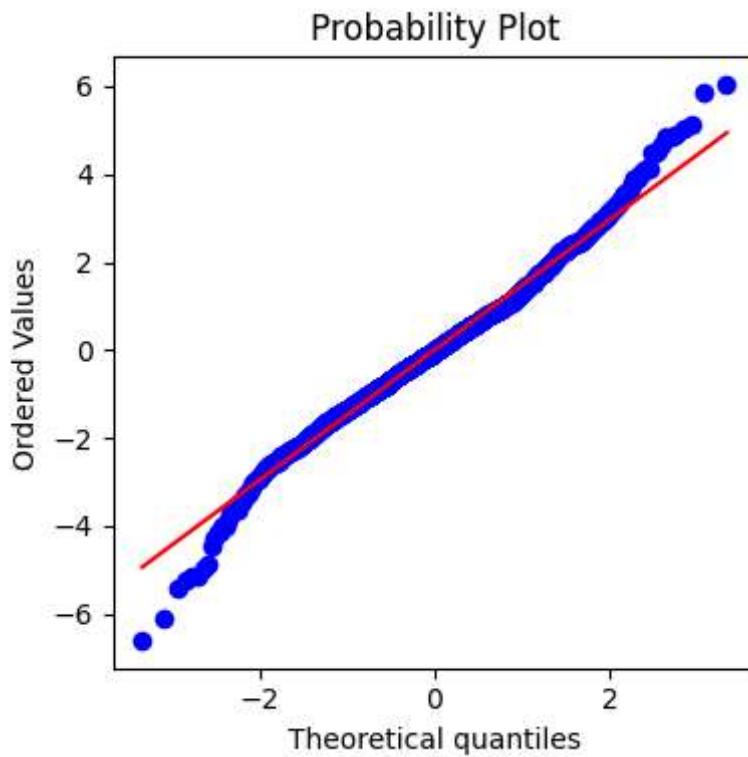
import pandas as pd
sp500_px = pd.read_csv('sp500_data.csv.gz')

nflx = sp500_px.NFLX
nflx = np.diff(np.log(nflx[nflx>0]))

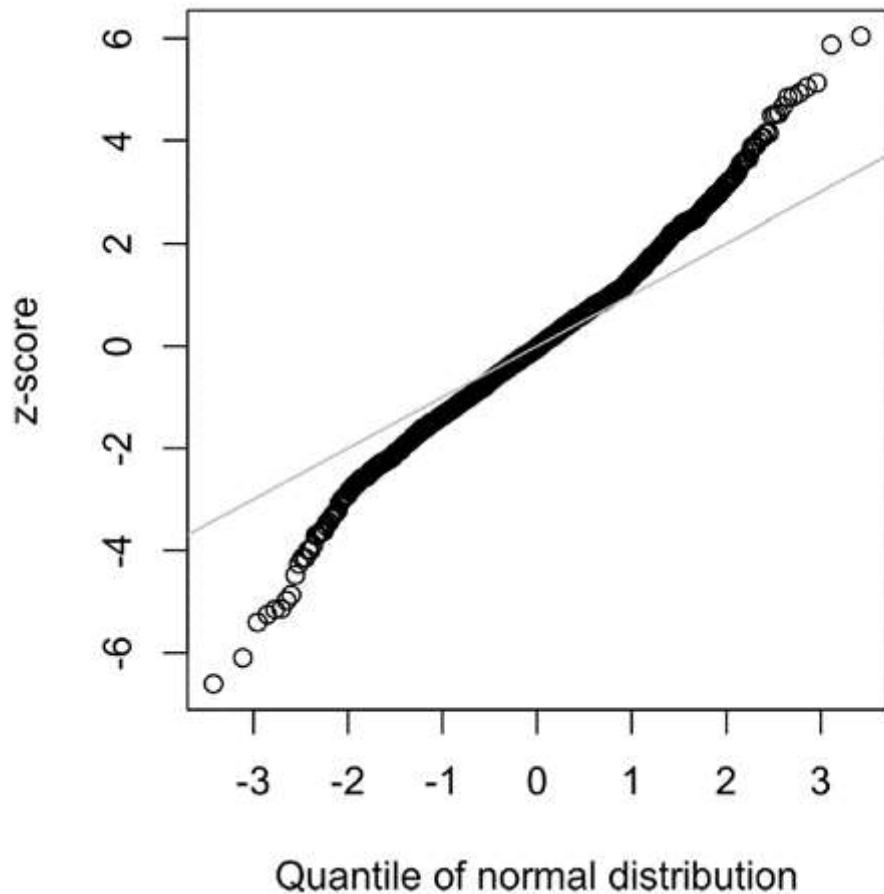
fig, ax = plt.subplots(figsize=(4, 4))
stats.probplot(nflx, plot=ax)

plt.tight_layout()
plt.show()

```



Mientras que la grafica hecha con R se ve asi:



En contraste con la Figura 2.9, los puntos están muy por debajo de la línea para los valores bajos y muy por encima de la línea para los valores altos, lo que indica que los datos no siguen una distribución normal. Esto significa que es mucho más probable observar valores extremos de lo que se esperaría si los datos siguieran una distribución

normal. La Figura de arriba muestra otro fenómeno común: los puntos están cerca de la línea para los datos que se encuentran dentro de una desviación estándar de la media. Tukey se refiere a este fenómeno como datos que son "normales en el centro" pero que tienen colas mucho más largas.

Existe una gran cantidad de literatura estadística sobre la tarea de ajustar distribuciones estadísticas a datos observados. Debemos tener cuidado con un enfoque excesivamente centrado en los datos para esta tarea, que es tanto un arte como una ciencia. Los datos son variables y, a menudo, son consistentes, en apariencia, con más de una forma y tipo de distribución. Generalmente, es necesario aplicar conocimientos tanto del dominio como de la estadística para determinar qué tipo de distribución es la adecuada para modelar una situación dada.

Por ejemplo, podríamos tener datos sobre el nivel de tráfico de internet en un servidor durante muchos períodos consecutivos de cinco segundos. Es útil saber que la mejor distribución para modelar "eventos por periodo de tiempo" es la distribución de Poisson.

2.11 Distribución t de Student

La distribución t es una distribución con forma normal, excepto que tiene colas un poco más gruesas y largas. Se utiliza extensamente para representar distribuciones de estadísticas muestrales. Las distribuciones de medias muestrales suelen tener una forma similar a la distribución t, y existe una familia de distribuciones t que difieren según el tamaño de la muestra. Cuanto mayor sea la muestra, más se asemeja la distribución t a una distribución normal.

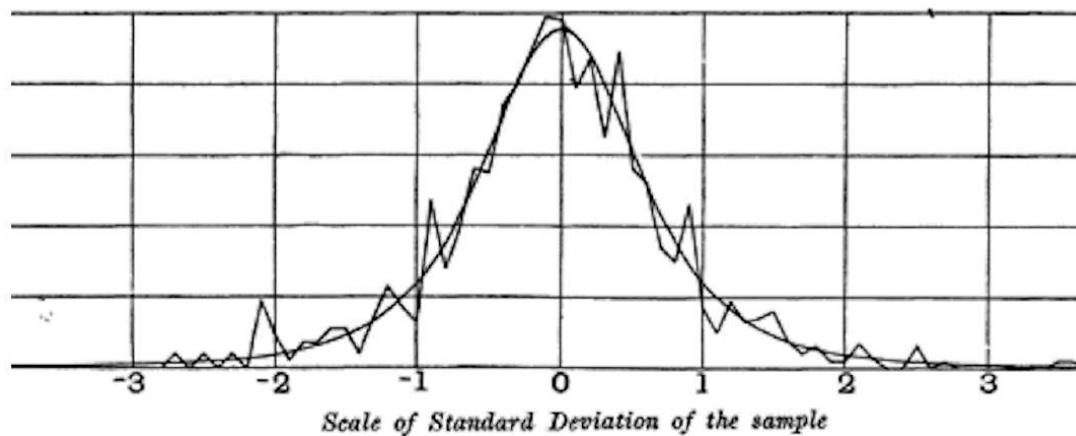
Términos Clave para la Distribución t de Student

- **n:** Tamaño de la muestra.
- **Grados de libertad:** Un parámetro que permite a la distribución t ajustarse a diferentes tamaños de muestra, estadísticas y números de grupos.

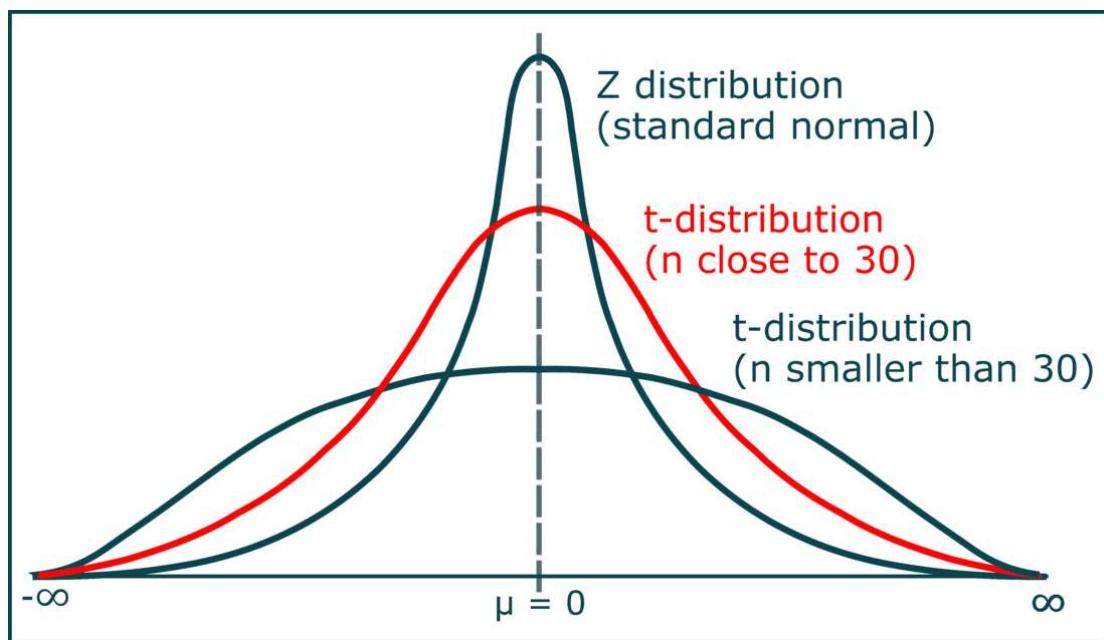
La distribución t se llama a menudo **t de Student** porque fue publicada en 1908 en la revista Biometrika por **W. S. Gosset** bajo el seudónimo de "Student". El empleador de Gosset, la cervecería Guinness, no quería que sus competidores supieran que estaba utilizando métodos estadísticos, por lo que insistió en que Gosset no usara su nombre en el artículo.

Gosset quería responder a la pregunta: "¿Cuál es la distribución muestral de la media de una muestra, extraída de una población más grande?". Comenzó con un experimento de remuestreo: tomando muestras aleatorias de 4 datos de un conjunto de 3,000 medidas de la altura de criminales y la longitud del dedo medio izquierdo. (Siendo esta la era de la eugenios, había mucho interés en los datos sobre criminales y en descubrir correlaciones entre las tendencias criminales y los atributos físicos o psicológicos). Gosset representó los resultados estandarizados (los puntajes-z) en el eje x y la frecuencia en el eje y. Por separado, derivó una función, ahora conocida como t de

Student, y ajustó esta función sobre los resultados de la muestra, trazando la Figura 2-13).



Comparación entre la distribución normal standard y la distribución t para distintos tamaños de muestra



Se pueden comparar varias estadísticas diferentes, después de la estandarización, con la distribución t para estimar intervalos de confianza teniendo en cuenta la variación muestral. Considera una muestra de tamaño (n) para la cual se ha calculado la media muestral (\bar{x}). Si (s) es la desviación estándar muestral, un intervalo de confianza del 90% alrededor de la media muestral se da por:

$$\bar{x} \pm t_{n-1,0.05} \cdot \frac{s}{\sqrt{n}}$$

donde $t_{n-1,0.05}$ es el valor de la estadística t, con $n - 1$ grados de libertad, que "recorta" el 5% de la distribución t en cada extremo. La distribución t se ha utilizado como referencia para la distribución de una media muestral, la diferencia entre dos medias muestrales, parámetros de regresión y otras estadísticas.

Esta ecuación se utiliza para calcular un **intervalo de confianza** para la **media de una población** basada en una muestra. El intervalo de confianza nos da un rango dentro del cual creemos que caerá la verdadera media poblacional con un nivel de confianza determinado (en este caso, el 90%).

- (\bar{x}) es el punto central, es decir, el valor de la media muestral.
- $(t_{n-1,0.05} \cdot \frac{s}{\sqrt{n}})$ es el **margen de error**. Este margen de error depende tanto de la **desviación estándar de la muestra** como del tamaño de la muestra y la distribución t, que ajusta el intervalo de confianza en función de los **grados de libertad**.
- Un intervalo de confianza del 90% significa que, si tomamos muchas muestras del mismo tamaño (n), aproximadamente el 90% de los intervalos construidos de esta manera contendrían la verdadera **media poblacional**.

Si hubiera habido potencia computacional disponible en 1908, la estadística sin duda habría dependido mucho más de métodos de remuestreo intensivos en computación desde el principio. Al carecer de computadoras, los estadísticos recurrieron a las matemáticas y funciones como la distribución t para aproximar distribuciones muestrales. La capacidad de computación permitió experimentos de remuestreo prácticos en la década de 1980, pero para entonces, el uso de la distribución t y distribuciones similares ya estaba profundamente arraigado en los libros de texto y el software.

La precisión de la distribución t al representar el comportamiento de una estadística muestral requiere que la distribución de esa estadística para esa muestra tenga una forma similar a una distribución normal. Resulta que las estadísticas muestrales a menudo están distribuidas normalmente, incluso cuando los datos de la población subyacente no lo están (un hecho que llevó a la aplicación generalizada de la distribución t). Esto nos devuelve al fenómeno conocido como el teorema del límite central.

¿Qué necesitan saber los científicos de datos sobre la distribución t y el teorema del límite central? No mucho. La distribución t se utiliza en la inferencia estadística clásica, pero no es tan central para los propósitos de la ciencia de datos. Comprender y cuantificar la incertidumbre y la variación son importantes para los científicos de datos, pero el remuestreo empírico con bootstrap puede responder la mayoría de las preguntas sobre el error muestral.

Sin embargo, los científicos de datos encontrarán rutinariamente estadísticas t en los resultados del software estadístico y en procedimientos estadísticos en R o Python, por ejemplo, en pruebas A/B y regresiones, por lo que es útil estar familiarizado con su propósito.

2.12 Distribución Binomial

Los resultados binarios de sí/no (binomiales) son fundamentales en el análisis de datos, ya que a menudo representan el resultado de una decisión u otro proceso: comprar/no

comprar, hacer clic/no hacer clic, sobrevivir/morir, etc. En el centro de la comprensión de la distribución binomial está la idea de un conjunto de **pruebas**, donde cada prueba tiene dos posibles resultados con probabilidades definidas.

Por ejemplo, lanzar una moneda 10 veces es un experimento binomial con 10 pruebas, cada una con dos posibles resultados (cara o cruz). Estos resultados de sí/no o 0/1 se denominan **resultados binarios**, y no necesitan tener probabilidades 50/50. Cualquier conjunto de probabilidades que sume 1.0 es válido. En estadística, es convencional llamar al resultado "1" el **resultado de éxito**; también es común asignar el "1" al resultado menos frecuente. El término éxito no implica necesariamente que el resultado sea deseable o beneficioso, sino que tiende a indicar el resultado de interés. Por ejemplo, los impagos de préstamos o las transacciones fraudulentas son eventos relativamente poco comunes que podríamos querer predecir, por lo que se denominan "1s" o "éxitos".

Términos clave para la Distribución Binomial

- **Prueba:** Un evento con un resultado discreto (por ejemplo, lanzar una moneda).
- **Éxito:** El resultado de interés en una prueba.
 - **Sinónimo:** "1" (en contraposición al "0").
- **Binomial:** Tener dos resultados.
 - **Sinónimos:** sí/no, 0/1, binario.
- **Prueba binomial:** Una prueba con dos resultados.
 - **Sinónimo:** Prueba de Bernoulli.
- **Distribución binomial:** Distribución del número de éxitos en (x) pruebas.
 - **Sinónimo:** Distribución de Bernoulli.

La distribución binomial es la distribución de frecuencias del número de éxitos ((x)) en un número determinado de pruebas ((n)) con una probabilidad específica ((p)) de éxito en cada prueba. Existe una familia de distribuciones binomiales, dependiendo de los valores de (n) y (p). La distribución binomial respondería a una pregunta como:

Si la probabilidad de que un clic se convierta en una venta es 0.02, ¿cuál es la probabilidad de observar 0 ventas en 200 clics?

El módulo `scipy.stats` implementa una amplia variedad de distribuciones estadísticas. Para la distribución binomial, se pueden usar las funciones `stats.binom.pmf` y `stats.binom.cdf`.

Ejemplo de Distribución Binomial en Data Science: Compró o No Compró

Supongamos que estamos realizando un análisis de comportamiento de compra en un sitio web. Queremos modelar la probabilidad de que un cliente realice una compra después de hacer clic en un anuncio. Basándonos en datos históricos, sabemos que la tasa de conversión (probabilidad de compra) es del 3% (es decir, (p = 0.03)). Vamos a modelar este escenario usando una **distribución binomial**, donde cada clic es una "prueba" con dos posibles resultados: **compró (1)** o **no compró (0)**.

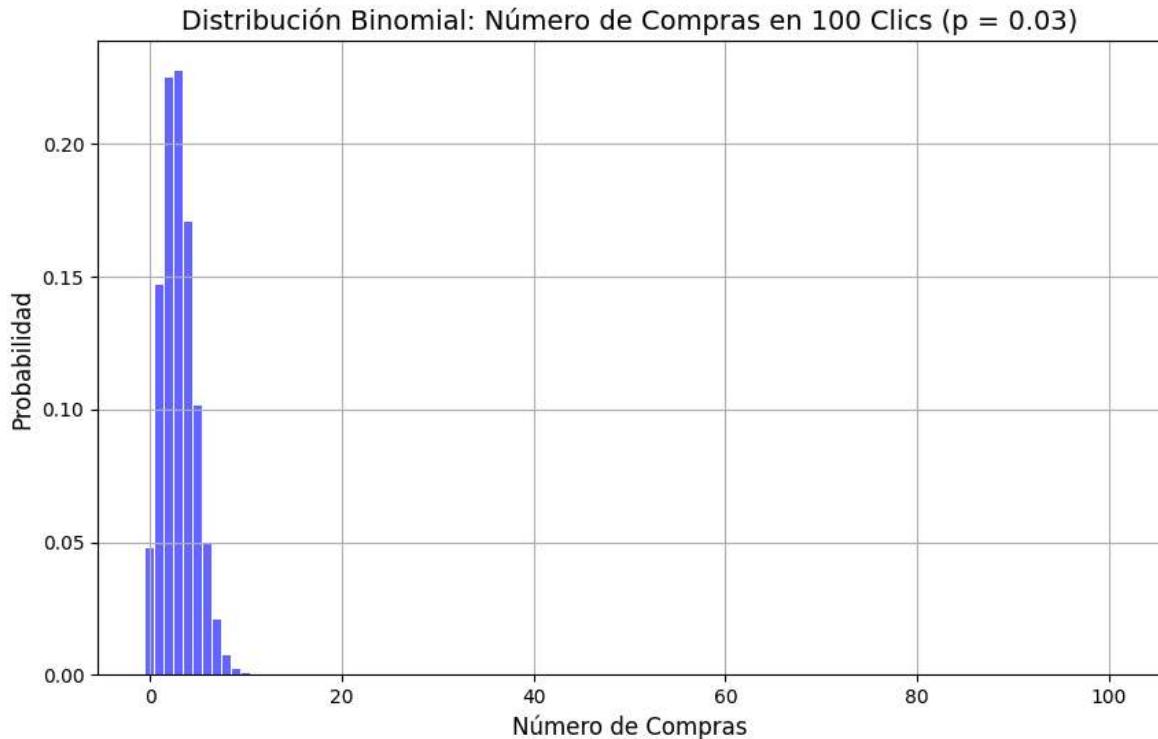
En este caso, utilizaremos la distribución binomial para simular el número de compras que ocurren en un conjunto de 100 clics, y visualizaremos la distribución de las posibles compras.

```
In [33]: import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import binom

# Parámetros del problema
n = 100 # Número de clics
p = 0.03 # Probabilidad de que un clic resulte en compra (tasa de conversión)

# Generar la distribución binomial
x = np.arange(0, n+1)
binomial_dist = binom.pmf(x, n, p)

# Graficar la distribución binomial
plt.figure(figsize=(10, 6))
plt.bar(x, binomial_dist, color='blue', alpha=0.6)
plt.title('Distribución Binomial: Número de Compras en 100 Clics (p = 0.03)', fontsize=14)
plt.xlabel('Número de Compras', fontsize=12)
plt.ylabel('Probabilidad', fontsize=12)
plt.grid(True)
plt.show()
```



Eje X (Número de Compras):

El eje horizontal representa el **número de compras** que podríamos observar en 100 clics. Esto va desde 0 compras (ningún cliente compra) hasta un número máximo que podría estar cerca del valor esperado.

Eje Y (Probabilidad):

El eje vertical muestra la **probabilidad** de observar exactamente ese número de compras para cada valor en el eje X. Cada barra representa la probabilidad de obtener un número específico de compras en 100 clics.

Valor Esperado (pico cerca de 3 compras):

Dado que la tasa de conversión (probabilidad de compra) es del 3%, el valor esperado de compras es:

$$E[X] = n \times p = 100 \times 0.03 = 3$$

Esto significa que, en promedio, esperamos alrededor de **3 compras en 100 clics**. En el gráfico, observarás que el pico de la distribución está alrededor de **3 compras**, lo que indica que este es el resultado más probable.

Forma de la Distribución:

La distribución binomial es **asimétrica** (ligeramente sesgada a la derecha), lo que significa que:

- Las probabilidades disminuyen rápidamente a medida que el número de compras se aleja del valor esperado.
- Aunque el valor más probable es alrededor de **3 compras**, todavía existe una pequeña probabilidad de observar valores más altos (4, 5, o incluso más compras) o valores más bajos (1 o 0 compras).

Colas de la Distribución:

Las colas de la distribución muestran que la probabilidad de observar un número de compras muy alto o muy bajo es baja, pero no imposible.

- Por ejemplo, la probabilidad de observar **0 compras** es mayor que la probabilidad de observar **10 compras**, ya que el valor esperado 3 compras**.

Conclusión:

- El **pico** alrededor de **3 compras** indica que este es el número más probable de compras que observaremos en 100 clics.
- Los resultados se distribuyen principalmente alrededor de este valor, pero hay una pequeña probabilidad de observar más o menos compras (entre **0 y 6 compras**, por ejemplo).
- El gráfico es útil para entender la **variabilidad** en los resultados posibles, proporcionando una visión probabilística de cuántas compras podemos esperar** con ($n = 100$) y ($p = 0.03$).

¿Cómo usar el modelo binomial para predecir futuras compras?

1. Determinar el número de pruebas (clics futuros):

Primero, necesitas saber cuántos clics futuros esperas tener, ya que este será el número de pruebas en la predicción.

2. Definir la probabilidad de éxito (p):

En este caso, la probabilidad de éxito es la tasa de conversión, que ya has calculado como ($p = 0.03$) (3% de los clics resultan en compras).

3. Realizar las predicciones usando la distribución binomial:

Usando la distribución binomial, puedes calcular la probabilidad de observar un número determinado de compras en los clics futuros. Puedes hacerlo para cualquier número de compras, ya sea el valor más probable (esperado) o cualquier otro valor.

Imagina que quieras predecir cuántas compras podrías observar si tienes **200 clics futuros** (es decir, $n = 200$).

In [34]:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import binom

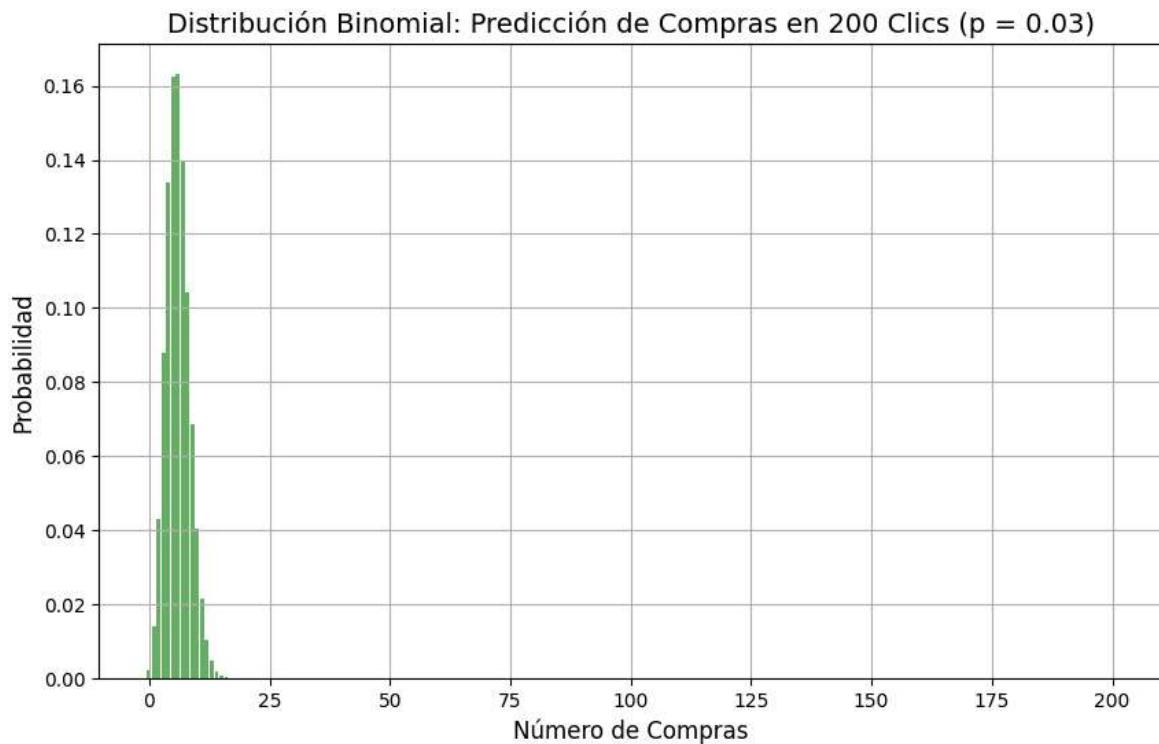
# Parámetros del modelo binomial
n_futuro = 200 # Número de clics futuros
p = 0.03 # Probabilidad de compra (tasa de conversión)

# Generar la distribución binomial para futuras compras
x_futuro = np.arange(0, n_futuro + 1)
binom_dist_futuro = binom.pmf(x_futuro, n_futuro, p)

# Valor esperado (n * p)
valor_esperado = n_futuro * p
print(f"El número esperado de compras en {n_futuro} clics es: {valor_esperado:.2f}")

# Graficar la distribución binomial de las futuras compras
plt.figure(figsize=(10, 6))
plt.bar(x_futuro, binom_dist_futuro, color='green', alpha=0.6)
plt.title('Distribución Binomial: Predicción de Compras en 200 Clics (p = 0.03)')
plt.xlabel('Número de Compras', fontsize=12)
plt.ylabel('Probabilidad', fontsize=12)
plt.grid(True)
plt.show()
```

El número esperado de compras en 200 clics es: 6.00



Ejercicio 2.12 Explicar el código anterior.

Ejercicio 2.13. Investigar las aplicaciones de las distribuciones: Chi-cuadrado, F, Poisson y Weibull en Data Science. Para cada distribución desarrolle un ejemplo con scipy o stastmodels.

2.13 Distribución Chi-cuadrado

La **distribución chi-cuadrado** es una de las distribuciones más importantes en la inferencia estadística y es utilizada en pruebas de hipótesis y análisis de datos categóricos. Se utiliza para comparar las frecuencias observadas en una muestra con las frecuencias esperadas, con el fin de determinar si existen diferencias significativas entre ellas.

Pruebas de hipótesis: Una prueba de hipótesis es un procedimiento estadístico que se utiliza para tomar decisiones o hacer inferencias sobre una población basada en una muestra de datos. Se usa comúnmente para evaluar una afirmación o suposición sobre un parámetro de la población (como la media, la proporción, la varianza, etc.) y determinar si la evidencia de la muestra respalda o refuta dicha afirmación.

Propiedades de la Distribución Chi-Cuadrado

- No es simétrica:** A diferencia de otras distribuciones como la normal, la distribución chi-cuadrado es asimétrica, con una cola hacia la derecha.

2. **Dependencia de los grados de libertad:** La forma de la distribución depende de los **grados de libertad (df)**. A medida que los grados de libertad aumentan, la distribución se vuelve más simétrica.

3. **Usos principales:**

- Pruebas de **independencia**. Se utiliza para verificar si dos variables categóricas son independientes. Por ejemplo, para verificar si la promoción en un trabajo es independiente del género de los empleados.
- Pruebas de **bondad de ajuste** (goodness-of-fit). Se utiliza para determinar si las frecuencias observadas de una variable categórica coinciden con las frecuencias esperadas bajo una distribución teórica.
- Pruebas para comparar **varianzas**.

La estadística chi-cuadrado se calcula con la fórmula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

donde:

- O_i son los valores observados.
- E_i son los valores esperados.
- La suma se realiza sobre todas las categorías.

1. **Calcular las frecuencias esperadas E_i :** Se calculan en función de los totales de las filas y columnas de la tabla de contingencia:

$$E_i = \frac{(\text{Total de la fila}) \times (\text{Total de la columna})}{\text{Total general}}$$

2. **Restar las frecuencias esperadas de las observadas** ($O_i - E_i$): Para cada categoría, se calcula la diferencia entre las frecuencias observadas y esperadas.

3. **Elevar al cuadrado la diferencia** ($(O_i - E_i)^2$): Esta diferencia se eleva al cuadrado.

4. **Dividir por las frecuencias esperadas** $\frac{(O_i - E_i)^2}{E_i}$: El valor obtenido en el paso anterior se divide por las frecuencias esperadas.

5. **Sumar los valores obtenidos para todas las categorías:** Finalmente, se suman los resultados obtenidos para todas las celdas de la tabla.

Ejemplo:

Supongamos que tienes una tabla con las siguientes **frecuencias observadas** y **frecuencias esperadas**:

	Compró	No Compró	Total
Hombre	50	100	150

	Compró	No Compró	Total
Mujer	30	70	100
Total	80	170	250

Las **frecuencias esperadas** son:

	Compró	No Compró	Total
Hombre	48	102	150
Mujer	32	68	100
Total	80	170	250

La **tabla de frecuencias esperadas** se calcula bajo la **hipótesis nula** de que las variables son **independientes**. Estas frecuencias esperadas nos indican los valores que esperaríamos observar en cada celda de la tabla **si no hubiera relación** entre las dos variables que estás analizando. Se calculan usando los totales de las filas y columnas de la tabla de contingencia.

Usando los datos del ejemplo, la tabla de frecuencias observadas muestra los datos observados: cuántos **hombres** y **mujeres** compraron o no compraron.

Cálculo de las Frecuencias Esperadas:

1. Para hombres que compraron:

$$E_{\text{hombre, compró}} = \frac{\text{Total fila hombre} \times \text{Total columna compró}}{\text{Total general}} = \frac{150 \times 80}{250} = 48$$

2. Para hombres que no compraron:

$$E_{\text{hombre, no compró}} = \frac{\text{Total fila hombre} \times \text{Total columna no compró}}{\text{Total general}} = \frac{150 \times 170}{250}$$

3. Para mujeres que compraron:

$$E_{\text{mujer, compró}} = \frac{\text{Total fila mujer} \times \text{Total columna compró}}{\text{Total general}} = \frac{100 \times 80}{250} = 32$$

4. Para mujeres que no compraron:

$$E_{\text{mujer, no compró}} = \frac{\text{Total fila mujer} \times \text{Total columna no compró}}{\text{Total general}} = \frac{100 \times 170}{250} = 68$$

Tabla de Frecuencias Esperadas:

	Compró	No Compró	Total
Hombre	48	102	150
Mujer	32	68	100

	Compró	No Compró	Total
Total	80	170	250

Las **frecuencias esperadas** representan los valores que esperaríamos observar en las celdas de la tabla **si no hubiera relación** entre las variables, es decir, si el género no afectara la decisión de compra.

Ahora calculamos la estadística chi-cuadrado para cada celda:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Para la primera celda (hombres que compraron):

$$\frac{(50 - 48)^2}{48} = \frac{(2)^2}{48} = \frac{4}{48} = 0.0833$$

Para la segunda celda (hombres que no compraron):

$$\frac{(100 - 102)^2}{102} = \frac{(-2)^2}{102} = \frac{4}{102} = 0.0392$$

Para la tercera celda (mujeres que compraron):

$$\frac{(30 - 32)^2}{32} = \frac{(-2)^2}{32} = \frac{4}{32} = 0.125$$

Para la cuarta celda (mujeres que no compraron):

$$\frac{(70 - 68)^2}{68} = \frac{(2)^2}{68} = \frac{4}{68} = 0.0588$$

Suma total:

$$\chi^2 = 0.0833 + 0.0392 + 0.125 + 0.0588 = 0.3063$$

Este es el valor de la **estadística chi-cuadrado** para este ejemplo.

Ejemplo: Prueba de Independencia Chi-Cuadrado

Supongamos que queremos verificar si existe una relación entre el **género** y si una persona **compra o no un producto**. Tenemos la siguiente tabla de datos (frecuencias observadas):

Género	Compró	No Compró	Total
Hombre	50	100	150
Mujer	30	70	100
Total	80	170	250

Hipótesis Nula H_0 : No existe una relación significativa entre el género y si una persona compra o no un producto. Esto significa que asumimos que el género y la decisión de compra **son** independientes.

Hipótesis Alternativa H_1 : Existe una relación significativa entre el género y si una persona compra o no un producto. Esto implica que el género y la decisión de compra **no** son independientes, es decir, que el género afecta la probabilidad de que una persona compre o no un producto.

Usaremos la prueba chi-cuadrado para determinar si existe una relación significativa entre el **género** y la **decisión de compra**.

In [1]:

```
import numpy as np
import scipy.stats as stats

# Crear La tabla de contingencia (frecuencias observadas)
observed = np.array([[50, 100], # Hombre
                     [30, 70]]) # Mujer

# Realizar La prueba chi-cuadrado
chi2, p, dof, expected = stats.chi2_contingency(observed)

# Mostrar resultados
print(f'Estadística chi-cuadrado: {chi2:.4f}')
print(f'Valor p: {p:.4f}')
print(f'Grados de libertad: {dof}')
print('Frecuencias esperadas:')
print(expected)
```

Estadística chi-cuadrado: 0.1723

Valor p: 0.6780

Grados de libertad: 1

Frecuencias esperadas:

[[48. 102.]
[32. 68.]]

Interpretación del Output:

- **Estadística chi-cuadrado: 0.1723:** La estadística chi-cuadrado (diferencia entre las frecuencias observadas en tus datos y las frecuencias esperadas). Cuanto mayor sea el valor, mayor será la diferencia.
- **Valor p: 0.6780:** Si el valor p es pequeño (por ejemplo menor a 0.05), rechazamos la hipótesis nula y concluimos que hay una relación significativa entre género y decisión de compra. Como regla general, si $p > 0.05$, no rechazamos la hipótesis nula (se acepta la hipótesis nula). En este caso, $p = 0.6780$ es mucho mayor que 0.05, por lo que **no rechazamos la hipótesis nula**, concluyendo que no hay una relación significativa entre las variables.
- **Grados de libertad: 1:** Los **grados de libertad** en una prueba chi-cuadrado de independencia se calculan como:

$$\text{Grados de libertad} = (\text{número de filas} - 1) \times (\text{número de columnas} - 1)$$

En este caso, la tabla tiene **2 filas** (por ejemplo, género: hombre/mujer) y **2 columnas** (compró/no compró). Entonces, los grados de libertad serían:

$$(2 - 1) \times (2 - 1) = 1$$

Esto indica que el número de parámetros libres que pueden variar en tu tabla es **1**, lo cual es típico en tablas 2×2 .

- **Frecuencias esperadas:**

$$\begin{bmatrix} 48 & 102 \\ 32 & 68 \end{bmatrix}$$

Estas son las frecuencias que esperaríamos observar si las variables fueran independientes. Esto significa que, bajo la hipótesis nula:

- En la primera fila, esperabas que **48 personas compraran y 102 no compraran**.
- En la segunda fila, esperabas que **32 personas compraran y 68 no compraran**.

Las **frecuencias observadas** se comparan con estas frecuencias esperadas para calcular la **estadística chi-cuadrado**.

```
In [8]: import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

# Crear la tabla de contingencia (frecuencias observadas)
observed = np.array([[50, 100], # Hombre
                     [30, 70]]) # Mujer

# Realizar la prueba chi-cuadrado
chi2, p, dof, expected = stats.chi2_contingency(observed)

# Mostrar resultados
print(f'Estadística chi-cuadrado: {chi2:.4f}')
print(f'Valor p: {p:.4f}')
print(f'Grados de libertad: {dof}')
print('Frecuencias esperadas:')
print(expected)

# Gráfica de la distribución chi-cuadrado con 1 grado de libertad
x = np.linspace(0, 10, 500) # Valores en el eje x
chi2_dist = stats.chi2.pdf(x, df=dof) # Función de densidad de probabilidad (PD)

# Crear el gráfico
plt.figure(figsize=(8, 6))
plt.plot(x, chi2_dist, label=f'Distribución chi-cuadrado (df={dof})', color='blue')

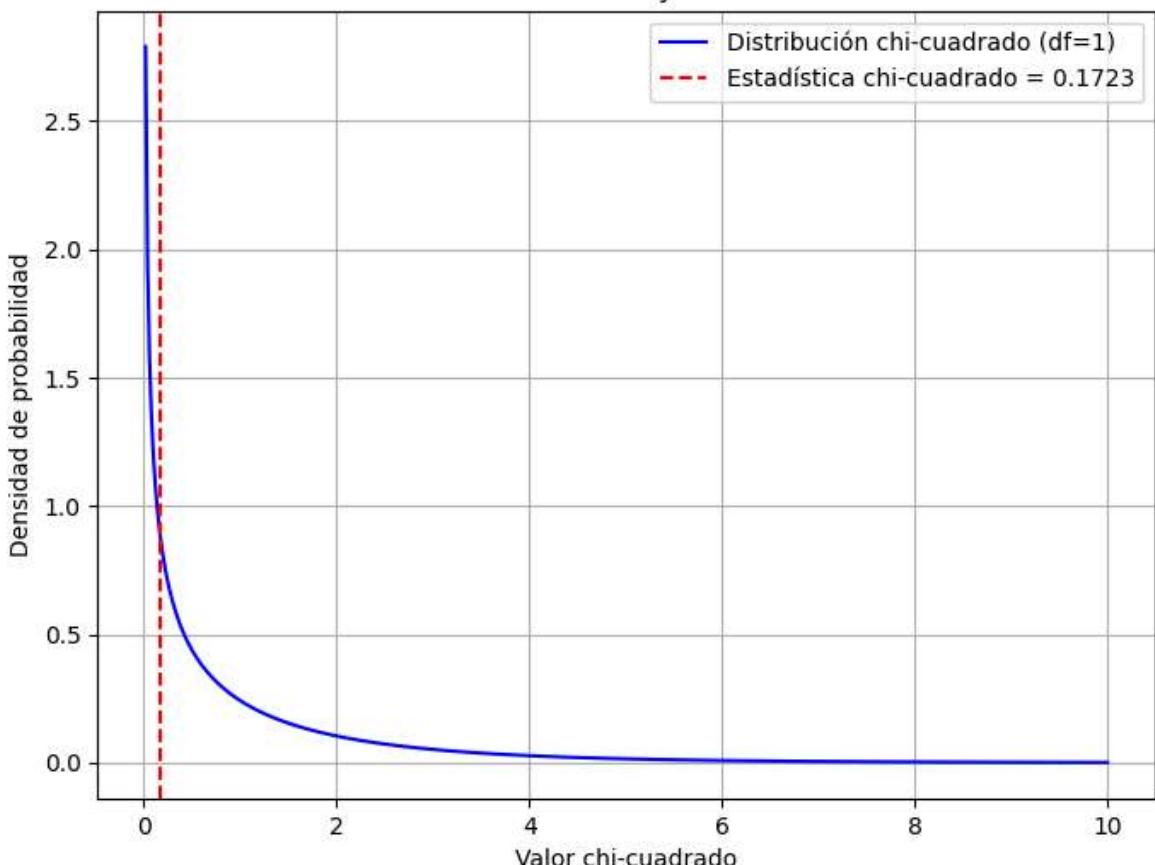
# Línea vertical en la estadística chi-cuadrado obtenida
plt.axvline(chi2, color='red', linestyle='--', label=f'Estadística chi-cuadrado')

# Añadir etiquetas y Leyenda
plt.title('Distribución Chi-Cuadrado y Estadística Obtenida')
plt.xlabel('Valor chi-cuadrado')
plt.ylabel('Densidad de probabilidad')
plt.legend()
```

```
# Mostrar el gráfico
plt.grid(True)
plt.show()
```

Estadística chi-cuadrado: 0.1723
 Valor p: 0.6780
 Grados de libertad: 1
 Frecuencias esperadas:
 [[48. 102.]]
 [32. 68.]]

Distribución Chi-Cuadrado y Estadística Obtenida



La gráfica te ayuda a visualizar qué tan "extrema" es la estadística chi-cuadrado obtenida en relación con la distribución de referencia. Si el valor cae en una zona de la cola derecha, indica que es poco probable bajo la hipótesis nula, y podrías rechazarla. En este caso, el valor de chi-cuadrado es muy pequeño, lo que sugiere que las diferencias observadas y esperadas no son lo suficientemente grandes como para rechazar la hipótesis nula.

Visualización de una Distribución Chi-Cuadrado

En este ejemplo, vamos a generar una distribución chi-cuadrado usando Python y visualizar su forma. Esto te permitirá observar cómo varía la distribución a medida que cambian los **grados de libertad (df)**.

La distribución chi-cuadrado se utiliza con diferentes grados de libertad según la cantidad de variables en el análisis, y la forma de la distribución cambia significativamente dependiendo de los **grados de libertad**.

```
In [7]: import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chi2

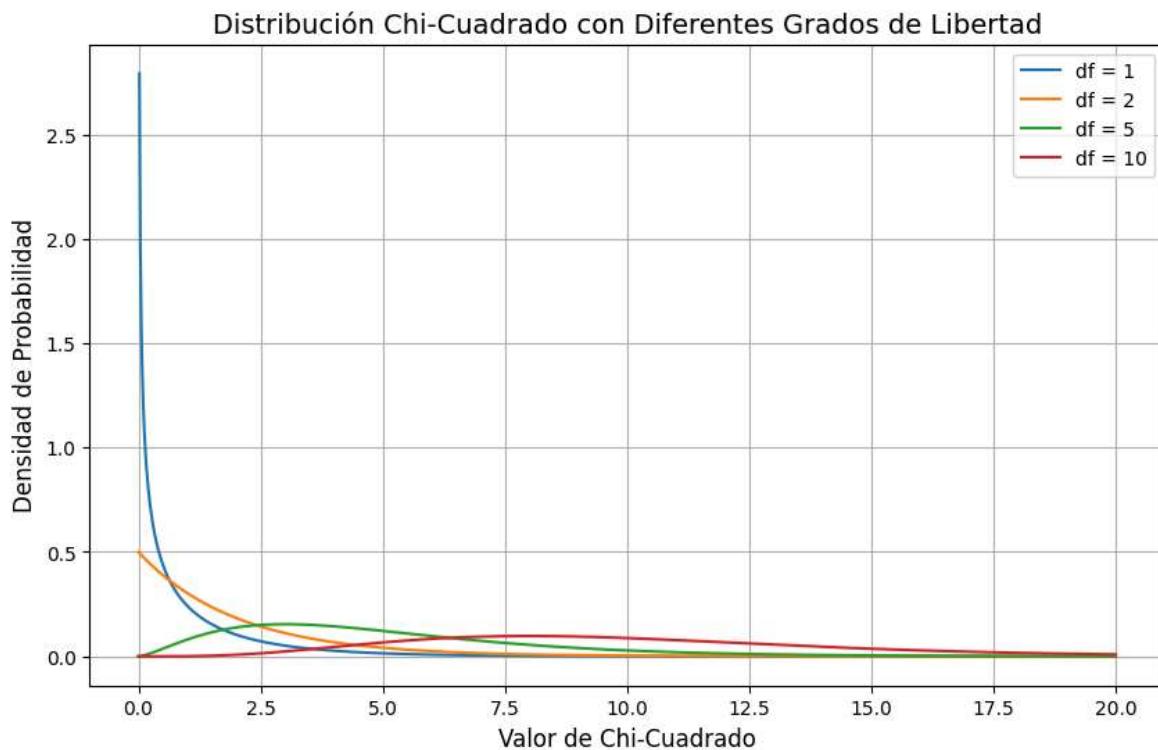
# Grados de Libertad para Las distribuciones chi-cuadrado
df_values = [1, 2, 5, 10]

# Valores del eje X (rango para chi-cuadrado)
x = np.linspace(0, 20, 1000)

# Crear la gráfica
plt.figure(figsize=(10, 6))

# Generar y graficar La distribución chi-cuadrado para diferentes grados de Libre
for df in df_values:
    y = chi2.pdf(x, df) # Generar La función de densidad de probabilidad (pdf)
    plt.plot(x, y, label=f'df = {df}')

# Añadir título y etiquetas
plt.title('Distribución Chi-Cuadrado con Diferentes Grados de Libertad', fontsize=14)
plt.xlabel('Valor de Chi-Cuadrado', fontsize=12)
plt.ylabel('Densidad de Probabilidad', fontsize=12)
plt.grid(True)
plt.legend()
plt.show()
```



Ejercicio 2.14

En una tienda de tecnología, se registran los siguientes datos sobre la **preferencia de sistema operativo** entre hombres y mujeres:

Género	Windows	MacOS	Linux	Total
Hombre	60	30	10	100

Género	Windows	MacOS	Linux	Total
Mujer	40	35	5	80
Total	100	65	15	180

Realiza una prueba chi-cuadrado para verificar si hay una relación significativa entre el **género** y la **preferencia de sistema operativo**. ¿Es significativa la relación entre estas dos variables? Usa **Python** para resolver el ejercicio y grafica los resultados obtenidos.

2.14 Distribución F

Las estadísticas F se generan automáticamente en R y Python como parte de las rutinas de regresión y ANOVA.

La **distribución F**, también conocida como distribución de Fisher-Snedecor, es una distribución estadística utilizada principalmente en:

- **Análisis de varianza (ANOVA)** para comparar las medias de más de dos grupos. Permite comparar la variabilidad entre grupos en un experimento, determinando si las diferencias observadas son atribuibles a factores aleatorios o a características intrínsecas de los grupos.
- **Regresión lineal** para evaluar la calidad de ajuste del modelo de regresión.
- **Diseño de experimentos**: Ayuda a seleccionar el tamaño de muestra adecuado para un experimento, garantizando que se tenga suficiente potencia estadística para detectar diferencias significativas.

El valor del estadístico F

El valor F, resultado de la prueba F de Snedecor, se representa como $F = \text{Varianza 1} / \text{Varianza 2}$, donde Varianza 1 y Varianza 2 corresponden a las varianzas de los dos conjuntos de datos que se comparan. **Valores F altos**: Indican que la variabilidad del primer conjunto de datos es mayor que la del segundo, lo que podría sugerir la existencia de diferencias significativas entre ambos. **Valores F bajos**: Sugieren que la variabilidad de ambos conjuntos de datos es similar, lo que no evidencia diferencias significativas.

El valor F en Machine Learning

La distribución F también juega un papel importante en la selección de modelos. Algoritmos como la regresión lineal o la selección de características utilizan el valor F para evaluar la relevancia de las variables predictoras, descartando aquellas que no aportan información significativa al modelo.

Imaginemos un estudio que compara el rendimiento académico de dos grupos de estudiantes: uno que recibió un método de enseñanza innovador y otro que siguió el método tradicional.

La distribución F se puede utilizar para determinar si el método innovador tuvo un impacto significativamente positivo en el rendimiento de los estudiantes.

Otro ejemplo podría ser un modelo de aprendizaje automático que predice el precio de las viviendas. En este caso se podría utilizar la distribución F para seleccionar las características más relevantes, como el tamaño, la ubicación o la cantidad de habitaciones, para mejorar la precisión del modelo.

La **distribución F** tiene dos tipos de grados de libertad:

1. **Grados de libertad entre los grupos** df_1 : Relacionados con el número de grupos.
2. **Grados de libertad dentro de los grupos** df_2 : Relacionados con el número total de observaciones menos el número de grupos.

Propiedades de la Distribución F:

- La distribución F es asimétrica y siempre positiva.
- La cola derecha es larga, lo que significa que la estadística F puede tomar valores altos.
- Se utiliza para comparar varianzas y evaluar si un modelo ajusta bien a los datos.

Fórmula de la Estadística F.

La estadística F se calcula como:

$$F = \frac{\text{Variabilidad entre grupos}}{\text{Variabilidad dentro de los grupos (residual)}}$$

Donde:

- La **variabilidad entre grupos** se calcula como la suma de las diferencias entre las medias de los grupos y la media total.
- La **variabilidad dentro de los grupos** se calcula como la suma de las diferencias dentro de cada grupo.

Ejemplo: Evaluación de un Nuevo Método Educativo.

Se quiere evaluar si un nuevo método educativo, más innovador, mejora el rendimiento de los estudiantes en comparación con los métodos tradicionales. Para ello, se recogen datos de tres grupos de estudiantes que fueron sometidos a diferentes métodos de enseñanza.

Descripción de los Grupos:

- **Grupo 1**: Estudiantes que recibieron la enseñanza tradicional.
- **Grupo 2**: Estudiantes que recibieron un método educativo intermedio.
- **Grupo 3**: Estudiantes que recibieron el nuevo método educativo más innovador.

Planteamiento de Hipótesis

- **Hipótesis Nula H_0** : No hay diferencias significativas en el rendimiento académico entre los tres grupos. Es decir, los tres métodos de enseñanza producen resultados similares.

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

- **Hipótesis Alternativa H_A** : Al menos uno de los métodos de enseñanza produce un rendimiento académico significativamente diferente.

$$H_A : \mu_1 \neq \mu_2 \neq \mu_3$$

Datos de Rendimiento Académico:

- **Grupo 1 (Método Tradicional)**: [65, 70, 68, 72, 66]
- **Grupo 2 (Método Intermedio)**: [75, 78, 74, 80, 76]
- **Grupo 3 (Nuevo Método Innovador)**: [85, 88, 90, 87, 89]

In [11]:

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

# Datos de rendimiento académico
grupo1 = [65, 70, 68, 72, 66] # Método tradicional
grupo2 = [75, 78, 74, 80, 76] # Método intermedio
grupo3 = [85, 88, 90, 87, 89] # Nuevo método innovador

# Realizar ANOVA
f_stat, p_value = stats.f_oneway(grupo1, grupo2, grupo3)

# Mostrar los resultados
print(f'Estadística F: {f_stat:.4f}')
print(f'Valor p: {p_value:.4f}')

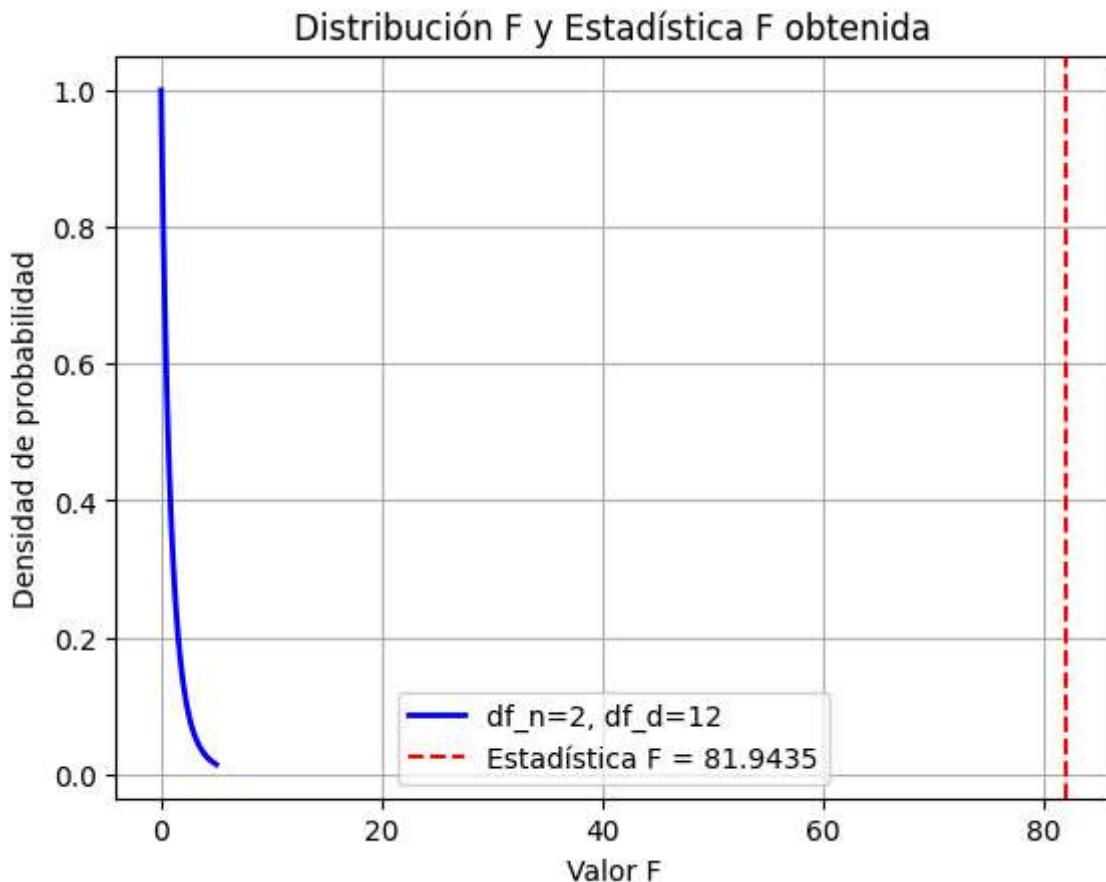
# Graficar la distribución F con Los grados de Libertad
dfn = 2 # Grados de Libertad del numerador (entre grupos)
dfd = len(grupo1) + len(grupo2) + len(grupo3) - 3 # Grados de Libertad del denominador

x = np.linspace(0, 5, 500)
f_dist = stats.f.pdf(x, dfn, dfd)

plt.plot(x, f_dist, 'b-', lw=2, label=f'df_n={dfn}, df_d={dfd}')
plt.axvline(f_stat, color='red', linestyle='--', label=f'Estadística F = {f_stat}')
plt.title('Distribución F y Estadística F obtenida')
plt.xlabel('Valor F')
plt.ylabel('Densidad de probabilidad')
plt.legend()
plt.grid(True)
plt.show()
```

Estadística F: 81.9435

Valor p: 0.0000



Un valor F de 81.9435 es extremadamente alto, lo que indica que las diferencias entre las medias de los tres grupos son mucho mayores que la variabilidad dentro de los grupos. Esto sugiere que la diferencia en los rendimientos académicos entre los grupos no se debe simplemente al azar.

Dado que el valor p es menor que 0.05 (nivel típico de significancia), rechazamos la hipótesis nula. Esto significa que hay una diferencia significativa en el rendimiento académico entre los tres grupos.

Los resultados sugieren que al menos uno de los métodos educativos tiene un impacto significativo en el rendimiento académico de los estudiantes, ya que la hipótesis nula (que afirma que todas las medias son iguales) ha sido rechazada.

Ejercicio 2.15

Imagina que realizas un estudio similar, pero esta vez comparando cuatro grupos de estudiantes con diferentes metodologías:

- Grupo 1 (Método Tradicional): [55, 60, 58, 62, 59]
- Grupo 2 (Método Visual): [65, 68, 64, 69, 66]
- Grupo 3 (Método Innovador): [75, 78, 76, 79, 77]
- Grupo 4 (Método Basado en Proyectos): [85, 88, 86, 89, 87]

Pregunta: ¿Hay diferencias significativas en el rendimiento académico entre los cuatro grupos? Interpretar el significado del valor de F y p. Interpreta el gráfico también. Elabore conclusiones y recomendaciones.

In []: