

# Análisis Cuantitativo de Datos



## Probabilidad y Muestreo

---

## SEMANA 2

# Contenidos

	<b>Unidad 2: Probabilidad y Muestreo</b>	<b>01</b>
<b>2.1</b>	Introducción a la Teoría de Probabilidades	<b>02</b>
<b>2.2</b>	Variables y Distribuciones de Probabilidad	<b>05</b>
<b>2.3</b>	Muestreo	<b>10</b>
	Conclusiones	<b>14</b>
	<b>Referencias bibliográficas</b>	<b>15</b>

# Unidad 2: Probabilidad y Muestreo



La teoría de probabilidades constituye el fundamento matemático de la inferencia estadística, proporcionando el marco conceptual necesario para cuantificar la incertidumbre y tomar decisiones informadas en presencia de variabilidad. Esta unidad introduce los conceptos fundamentales que permiten la transición de la estadística descriptiva hacia la estadística inferencial.

El muestreo estadístico representa la conexión práctica entre las poblaciones teóricas y las muestras observables, estableciendo los principios que garantizan que las conclusiones extraídas de muestras limitadas sean válidas para poblaciones más amplias. La comprensión de estos conceptos es esencial para el diseño de investigaciones rigurosas y la interpretación correcta de resultados estadísticos.

A lo largo de esta unidad, desarrollaremos las competencias necesarias para evaluar la probabilidad de eventos, trabajar con distribuciones de probabilidad y diseñar estrategias de muestreo que maximicen la validez de las inferencias estadísticas.

## 2.1 Introducción a la Teoría de Probabilidades

La probabilidad cuantifica la posibilidad de ocurrencia de eventos específicos en experimentos aleatorios, proporcionando una medida numérica de la incertidumbre. Esta medida oscila entre 0 (imposibilidad) y 1 (certeza), estableciendo un lenguaje universal para describir fenómenos inciertos.

### Experimento Aleatorio

Proceso que puede repetirse bajo condiciones similares y cuyo resultado no puede predecirse con certeza antes de su realización.

### Espacio Muestral

Conjunto de todos los posibles resultados de un experimento aleatorio. Se denota comúnmente como  $\Omega$  (omega).

### Evento

Subconjunto del espacio muestral. Puede ser simple (un solo resultado) o compuesto (múltiples resultados).

Los axiomas de Kolmogorov establecen los fundamentos matemáticos de la teoría de probabilidades: la probabilidad de cualquier evento es no negativa, la probabilidad del espacio muestral es uno, y la probabilidad de la unión de eventos mutuamente excluyentes es la suma de sus probabilidades individuales.

Concepto	Notación	Rango
Probabilidad de evento A	$P(A)$	$[0, 1]$
Evento seguro	$P(\Omega) = 1$	1
Evento imposible	$P(\emptyset) = 0$	0

# Reglas Fundamentales de Probabilidad

Las reglas de probabilidad establecen relaciones matemáticas que permiten calcular probabilidades de eventos complejos a partir de probabilidades de eventos más simples. Estas reglas son fundamentales para resolver problemas prácticos que involucran múltiples eventos o condiciones.

## Regla de la Adición

Para eventos A y B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Si los eventos son mutuamente excluyentes:

$$P(A \cup B) = P(A) + P(B)$$

## Regla de la Multiplicación

Para eventos independientes:

$$P(A \cap B) = P(A) \times P(B)$$

Para eventos dependientes:

$$P(A \cap B) = P(A) \times P(B|A)$$

### • Eventos Mutuamente Excluyentes

No pueden ocurrir simultáneamente:  $P(A \cap B) = 0$

### • Eventos Independientes

La ocurrencia de uno no afecta al otro:  
 $P(B|A) = P(B)$

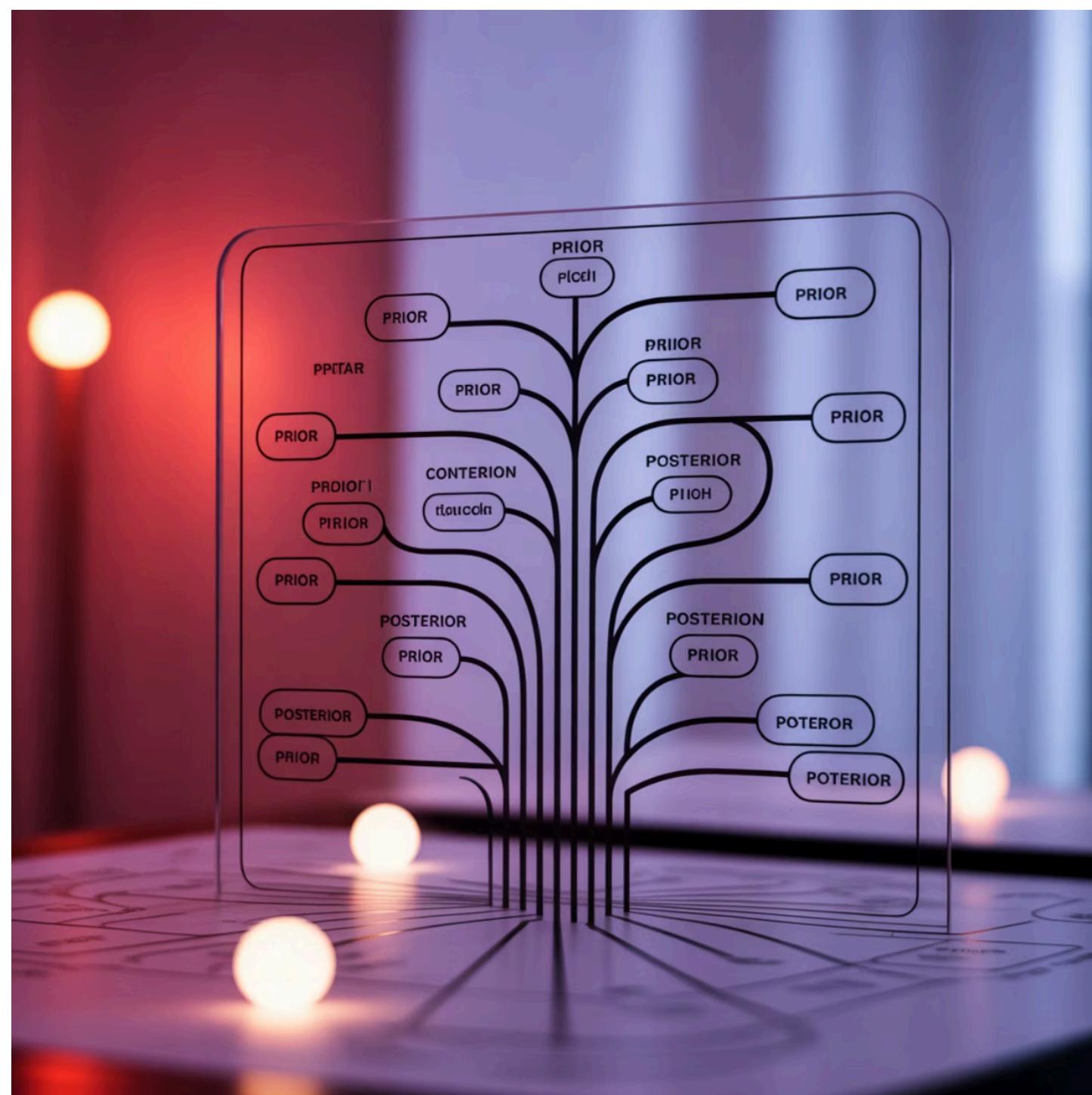
### • Eventos Complementarios

Cubren todo el espacio muestral:  $P(A) + P(A') = 1$

### • Probabilidad Condicional

$P(B|A) = P(A \cap B) / P(A)$ , si  $P(A) > 0$

# Teorema de Bayes



El teorema de Bayes proporciona un método para actualizar probabilidades cuando se obtiene nueva información, constituyendo la base de la inferencia bayesiana. Este teorema es fundamental en aplicaciones que van desde diagnósticos médicos hasta sistemas de recomendación y análisis de riesgo.

La fórmula establece que  $P(A|B) = [P(B|A) \times P(A)] / P(B)$ , donde  $P(A)$  es la probabilidad a priori,  $P(B|A)$  es la verosimilitud, y  $P(A|B)$  es la probabilidad a posteriori.

El teorema permite combinar conocimiento previo (probabilidad a priori) con evidencia observada (verosimilitud) para obtener probabilidades actualizadas (probabilidad a posteriori). Esta capacidad de incorporar nueva información de manera sistemática lo convierte en una herramienta poderosa para la toma de decisiones bajo incertidumbre.

1

2

3

## Probabilidad A Priori

$P(A)$ : Conocimiento inicial sobre la probabilidad del evento antes de observar evidencia.

## Verosimilitud

$P(B|A)$ : Probabilidad de observar la evidencia dado que el evento A es verdadero.

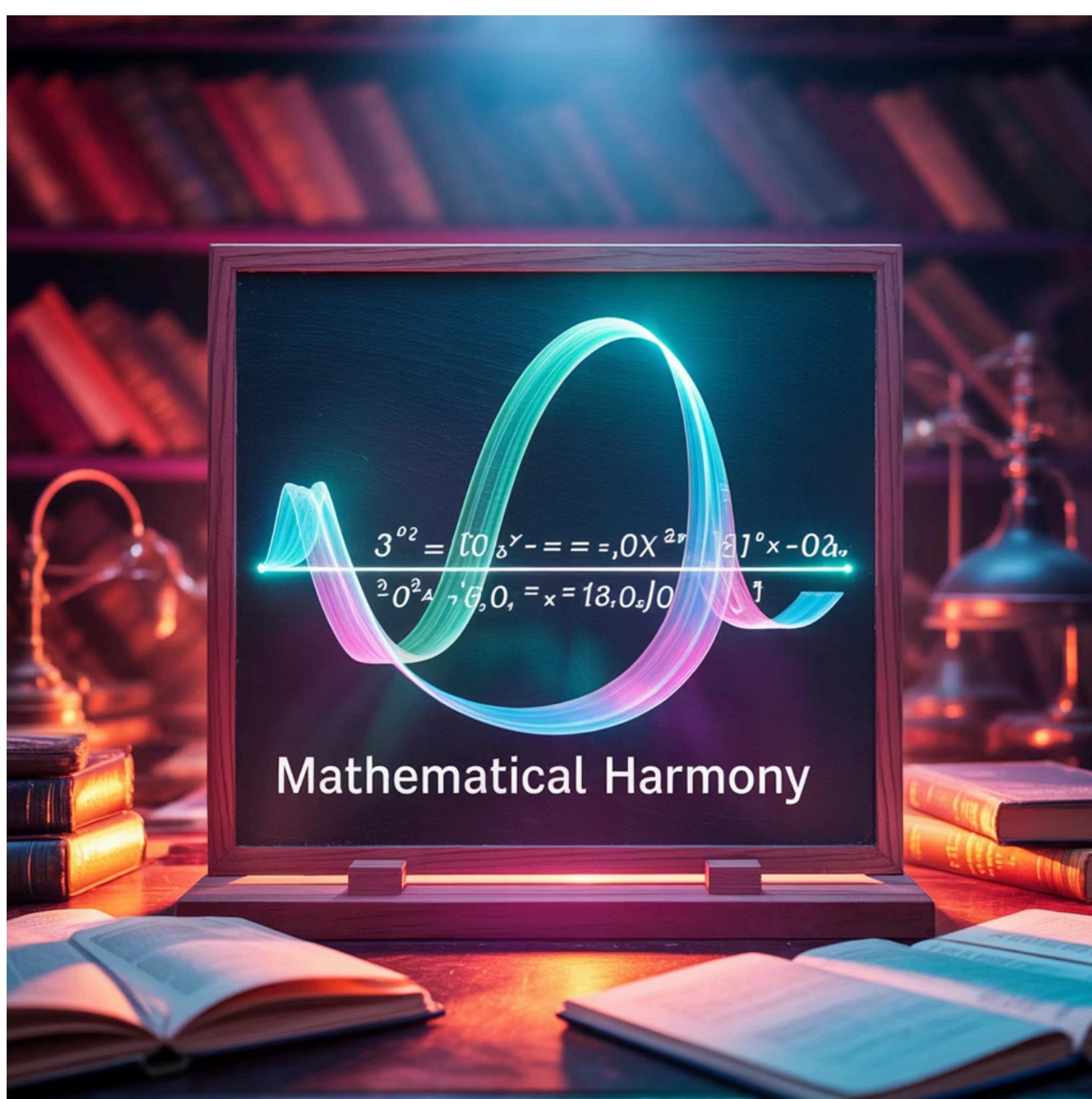
## Probabilidad A Posteriori

$P(A|B)$ : Probabilidad actualizada del evento después de considerar la evidencia observada.

En diagnóstico médico, el teorema de Bayes permite combinar la prevalencia de una enfermedad (probabilidad a priori) con los resultados de pruebas diagnósticas (verosimilitud) para calcular la probabilidad de que un paciente tenga la enfermedad dado un resultado positivo.

## 2.2 Variables y Distribuciones de Probabilidad

Las variables aleatorias proporcionan una función matemática que asigna valores numéricos a los resultados de experimentos aleatorios, facilitando el análisis cuantitativo de fenómenos inciertos. Esta abstracción permite aplicar herramientas matemáticas poderosas para modelar y predecir comportamientos en situaciones de incertidumbre.



Existen dos tipos fundamentales de variables aleatorias: discretas y continuas. Las variables discretas toman valores específicos y separados, típicamente números enteros, mientras que las variables continuas pueden asumir cualquier valor dentro de un intervalo específico del conjunto de números reales.

Las distribuciones de probabilidad describen cómo se distribuyen las probabilidades entre los posibles valores de una variable aleatoria, proporcionando una descripción completa del comportamiento probabilístico del fenómeno estudiado.

### Variables Discretas

Caracterizadas por funciones de masa de probabilidad (PMF) que asignan probabilidades específicas a cada valor posible.

- Número de defectos en producción
- Cantidad de clientes en una cola
- Resultados de encuestas

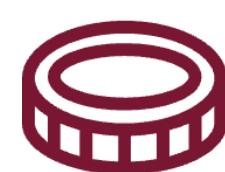
### Variables Continuas

Descritas mediante funciones de densidad de probabilidad (PDF) donde las probabilidades se calculan mediante integrales.

- Tiempo de servicio
- Mediciones físicas
- Variables económicas

# Distribuciones Discretas Fundamentales

Las distribuciones discretas modelan fenómenos donde la variable aleatoria puede tomar únicamente valores específicos y separados. Estas distribuciones son fundamentales para analizar conteos, éxitos en ensayos repetidos y eventos raros en intervalos de tiempo o espacio.



## Distribución Binomial

Modela el número de éxitos ( $k$ ) en  $n$  ensayos independientes con probabilidad constante  $p$ . Parámetros:  $n$  (ensayos) y  $p$  (probabilidad de éxito).

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Aplicaciones: Control de calidad, estudios clínicos, encuestas.

## Distribución de Poisson

Describe eventos raros que ocurren de manera independiente en intervalos fijos de tiempo o espacio. Parámetro:  $\lambda$  (tasa promedio).

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

Aplicaciones: Llegadas a sistemas, defectos por unidad, accidentes.

## Distribución Geométrica

Modela el número de ensayos necesarios hasta obtener el primer éxito. Parámetro:  $p$  (probabilidad de éxito por ensayo).

$$P(X = k) = (1 - p)^{k-1} p$$

Aplicaciones: Tiempo hasta la falla, búsquedas exitosas.

La selección de la distribución apropiada depende de las características específicas del fenómeno: la binomial para ensayos con resultado binario, Poisson para eventos raros, y geométrica para el tiempo hasta el primer éxito.

## Distribución Binomial: Detección de Fraude

**Contexto:** Un modelo de **detección de fraude** tiene un **90% de precisión** (clasifica correctamente transacciones legítimas). **Problema:** Si revisamos **20 transacciones**, ¿cuál es la probabilidad de que **exactamente 18** sean clasificadas correctamente?

## Cálculo

### Parámetros:

- $n$  (total de ensayos) = 20
- $k$  (éxitos deseados) = 18
- $p$  (probabilidad de éxito) = 0.90

**Fórmula:**  $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

**Sustitución:**  $P(X = 18) = \binom{20}{18} (0.90)^{18} (0.10)^2$

## Resultado

La probabilidad de clasificar correctamente **18 de 20 transacciones** es aproximadamente **28.52%**.

## Distribución de Poisson: Tráfico en Servidor Web

**Contexto:** Un servidor web recibe un promedio de **30 solicitudes por minuto** en horas pico.

**Problema:** ¿Cuál es la probabilidad de recibir **exactamente 35 solicitudes** en el próximo minuto?

### Cálculo

**Parámetros:**

- $\lambda$  (tasa promedio) = 30
- $k$  (eventos deseados) = 35

**Fórmula:**  $P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$

**Sustitución:**  $P(X = 35) = \frac{e^{-30} 30^{35}}{35!}$

### Resultado

La probabilidad de recibir **35 solicitudes** en un minuto es aproximadamente **5.16%**.

## Distribución Geométrica: Búsqueda de Lead Calificado

**Contexto:** La probabilidad de que un "lead" (cliente potencial) contactado sea **calificado** es del **5%**.

**Problema:** ¿Cuál es la probabilidad de tener que contactar **exactamente 10 leads** para encontrar el primer lead calificado? (Es decir, los 9 anteriores no fueron calificados).

### Cálculo

**Parámetros:**

- $p$  (probabilidad de éxito) = 0.05
- $k$  (ensayos hasta el primer éxito) = 10

**Fórmula:**  $P(X = k) = (1 - p)^{k-1} p$

**Sustitución:**  $P(X = 10) = (1 - 0.05)^{10-1} (0.05) = P(X = 10) = (0.95)^9 (0.05)$

### Resultado

La probabilidad de encontrar el primer lead calificado al **décimo intento** es aproximadamente **3.15%**.

# Distribuciones Continuas Esenciales

Las distribuciones continuas modelan variables que pueden tomar cualquier valor dentro de un rango específico. Estas distribuciones son fundamentales para describir mediciones, tiempos y muchas variables naturales que exhiben variación continua.

## Distribución Normal

La más importante en estadística, caracterizada por su forma de campana simétrica. Parámetros:  $\mu$  (media) y  $\sigma$  (desviación estándar).

Fundamental por el Teorema del Límite Central y su prevalencia en fenómenos naturales.

## Distribución Uniforme

Todos los valores en un intervalo tienen igual probabilidad. Parámetros:  $a$  y  $b$  (límites del intervalo).

Utilizada en simulaciones y como distribución de referencia.

1

2

3

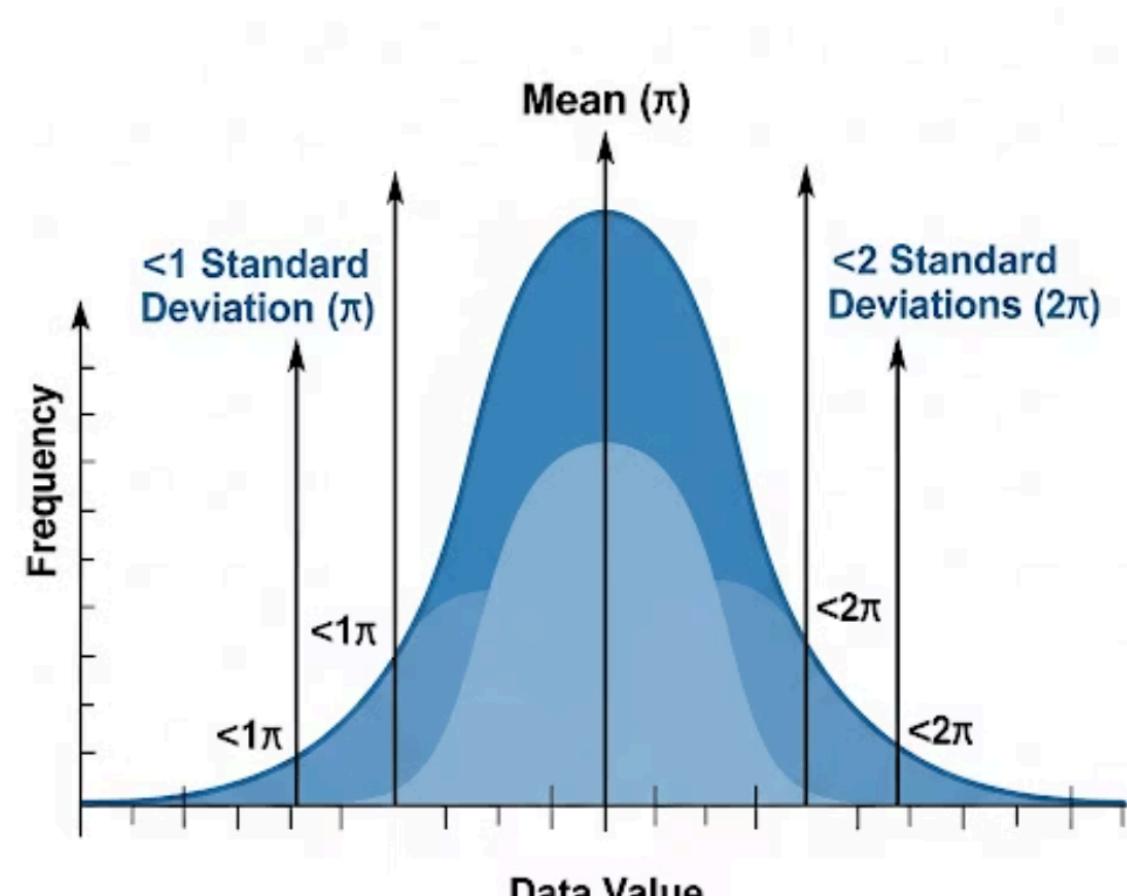
## Distribución Exponencial

Modela tiempos entre eventos en procesos de Poisson. Parámetro:  $\lambda$  (tasa).

Caracterizada por la propiedad de falta de memoria, útil en análisis de supervivencia.

Distribución	Parámetros	Media	Varianza
Normal	$\mu, \sigma$	$\mu$	$\sigma^2$
Exponencial	$\lambda$	$1/\lambda$	$1/\lambda^2$
Uniforme	$a, b$	$(a+b)/2$	$(b-a)^2/12$

# La Distribución Normal: Propiedades y Aplicaciones



La distribución normal ocupa una posición central en la estadística debido a sus propiedades matemáticas excepcionales y su amplia aplicabilidad en fenómenos naturales y sociales. Su importancia se fundamenta en el Teorema del Límite Central, que establece que las medias muestrales tienden hacia la normalidad independientemente de la distribución poblacional original.

La función de densidad normal está completamente determinada por dos parámetros: la media ( $\mu$ ) que determina el centro de la distribución, y la desviación estándar ( $\sigma$ ) que controla su dispersión. La forma característica de campana es simétrica alrededor de la media, con colas que se extienden hacia el infinito sin tocar el eje horizontal.

**68%**

**Primera Desviación**

Datos dentro de  $\mu \pm \sigma$

**95%**

**Segunda Desviación**

Datos dentro de  $\mu \pm 2\sigma$

**99.7%**

**Tercera Desviación**

Datos dentro de  $\mu \pm 3\sigma$

La estandarización mediante la transformación  $Z = (X - \mu)/\sigma$  permite convertir cualquier distribución normal en la distribución normal estándar ( $\mu = 0$ ,  $\sigma = 1$ ), facilitando el cálculo de probabilidades mediante tablas estandarizadas. Esta propiedad es fundamental para la realización de pruebas de hipótesis y el cálculo de intervalos de confianza.

La regla empírica (68-95-99.7) proporciona una herramienta práctica para evaluar rápidamente la normalidad de los datos y identificar valores atípicos en distribuciones aproximadamente normales.

## 2.3 Muestreo

El muestreo estadístico constituye el puente fundamental entre las poblaciones teóricas y la realidad práctica de la investigación. Dado que raramente es posible o económicamente viable estudiar poblaciones completas, el muestreo proporciona métodos sistemáticos para seleccionar subconjuntos representativos que permitan hacer inferencias válidas sobre el conjunto total.

### Población

Conjunto completo de elementos sobre los cuales se desea hacer inferencias. Puede ser finita o infinita, concreta o conceptual

### Muestra

Subconjunto de la población seleccionado para observación y análisis. Debe ser representativa para garantizar inferencias válidas.

### Dato:

es un valor o característica recopilada de un elemento observado.

## Métodos de Muestreo Probabilístico

Los métodos probabilísticos garantizan que cada elemento de la población tenga una probabilidad conocida y no nula de ser seleccionado, permitiendo la aplicación de teoría estadística para cuantificar la precisión de las estimaciones y calcular márgenes de error.

1

**Muestreo aleatorio simple:** Cada elemento tiene igual probabilidad de selección. Método fundamental que sirve como base para otros diseños más complejos.

2

**Muestreo estratificado:** División de la población en estratos homogéneos con muestreo independiente en cada estrato.

3

**Muestreo por conglomerados:** Selección aleatoria de grupos naturales, estudiando todos los elementos de los grupos seleccionados..

4

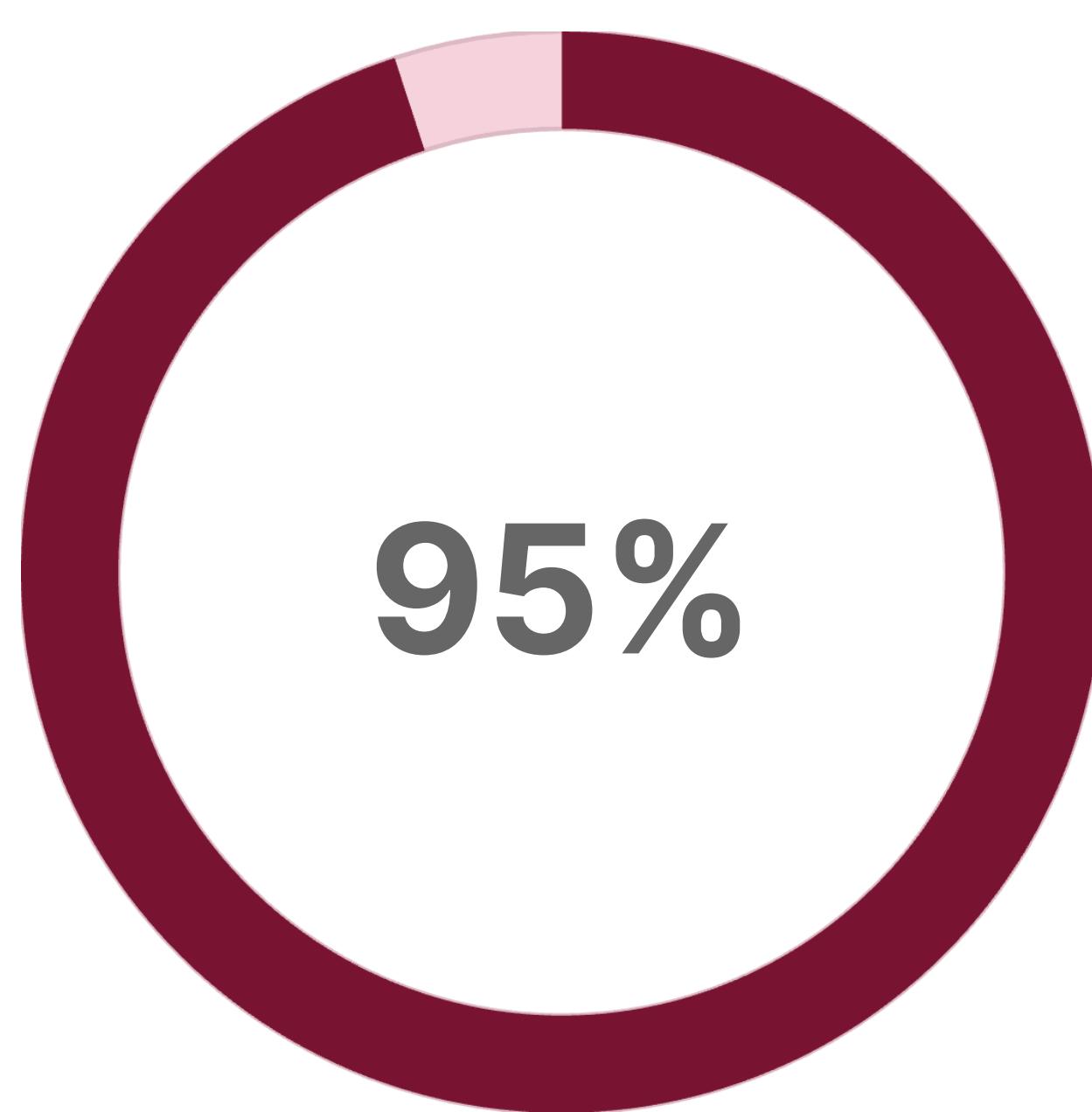
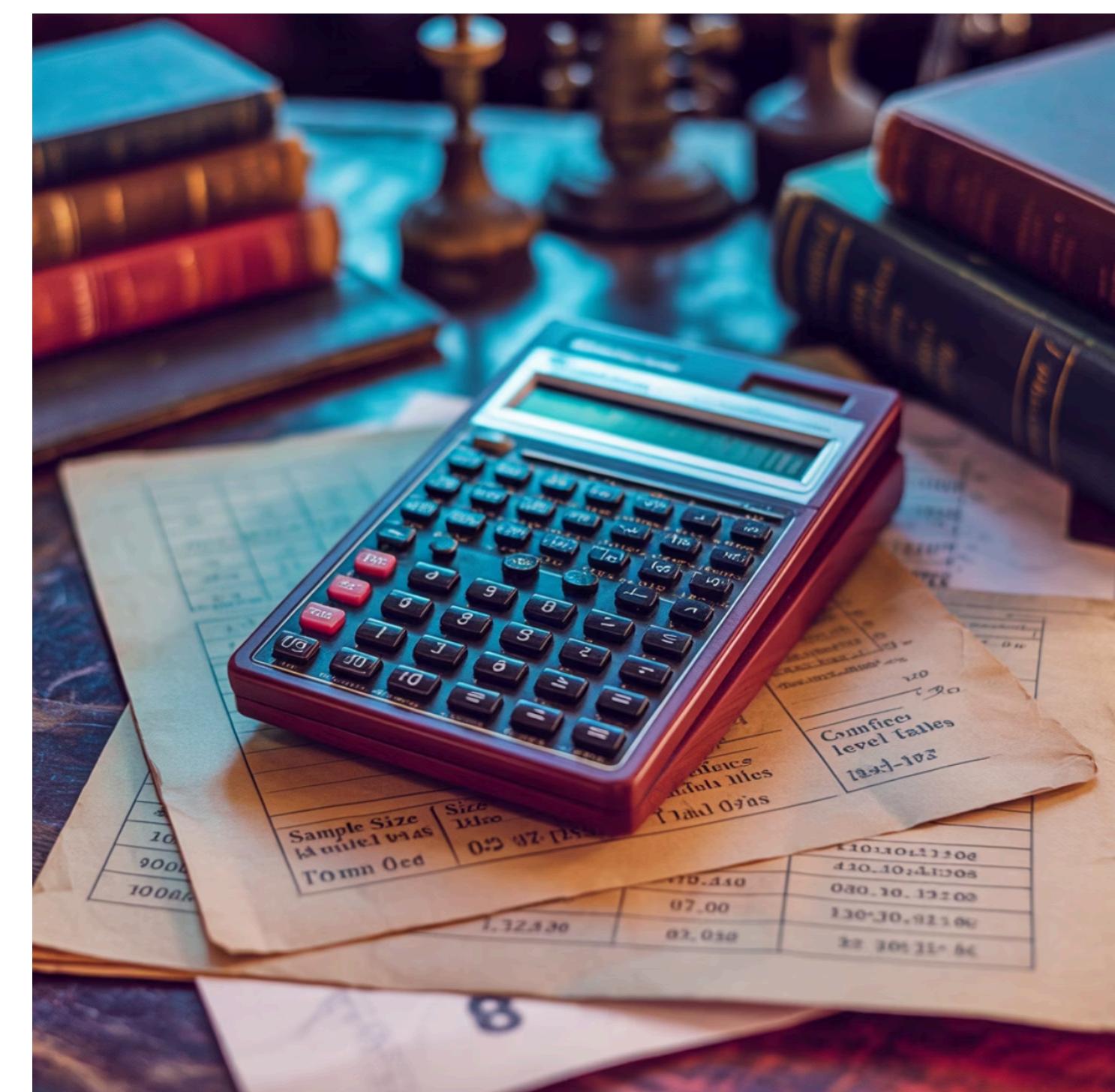
**Muestreo sistemático:** Selección de cada  $k$ -ésimo elemento después de un inicio aleatorio. Eficiente cuando la población está ordenada.

# Tamaño de Muestra y Error Muestral

La determinación del tamaño de muestra apropiado representa una decisión crítica que equilibra la precisión estadística deseada con las limitaciones prácticas de recursos y tiempo. Un tamaño insuficiente puede resultar en estimaciones imprecisas, mientras que un tamaño excesivo desperdicia recursos sin mejoras significativas en la precisión.

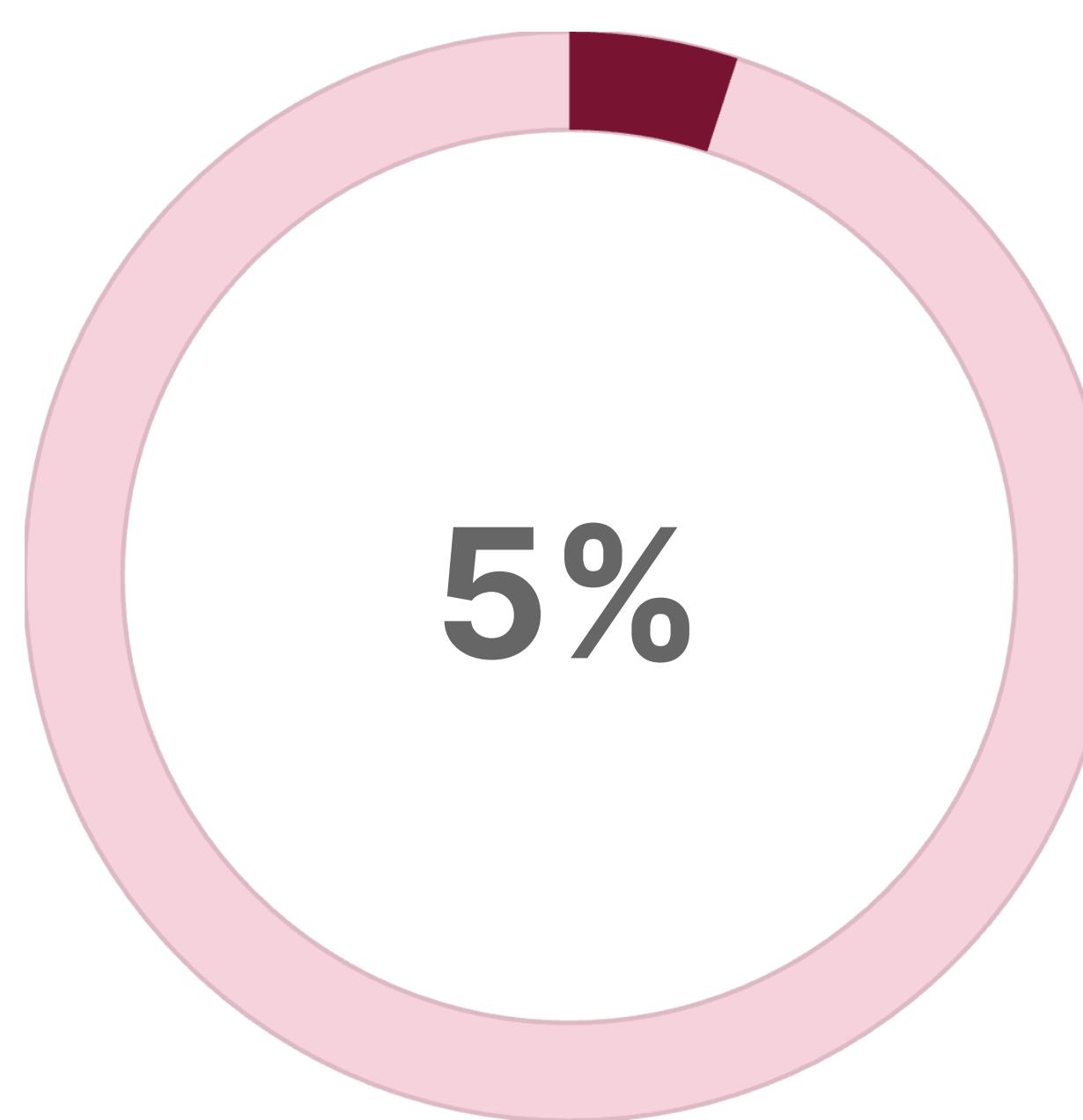
El error muestral refleja la diferencia entre el estadístico muestral y el parámetro poblacional correspondiente. Este error es inherente al proceso de muestreo y puede cuantificarse probabilísticamente, permitiendo establecer márgenes de confianza para las estimaciones.

La fórmula básica para el tamaño de muestra en estimación de proporciones es  $n = (Z^2 \times p \times q) / E^2$ , donde Z es el valor crítico, p la proporción esperada, q = 1-p, y E el margen de error deseado.



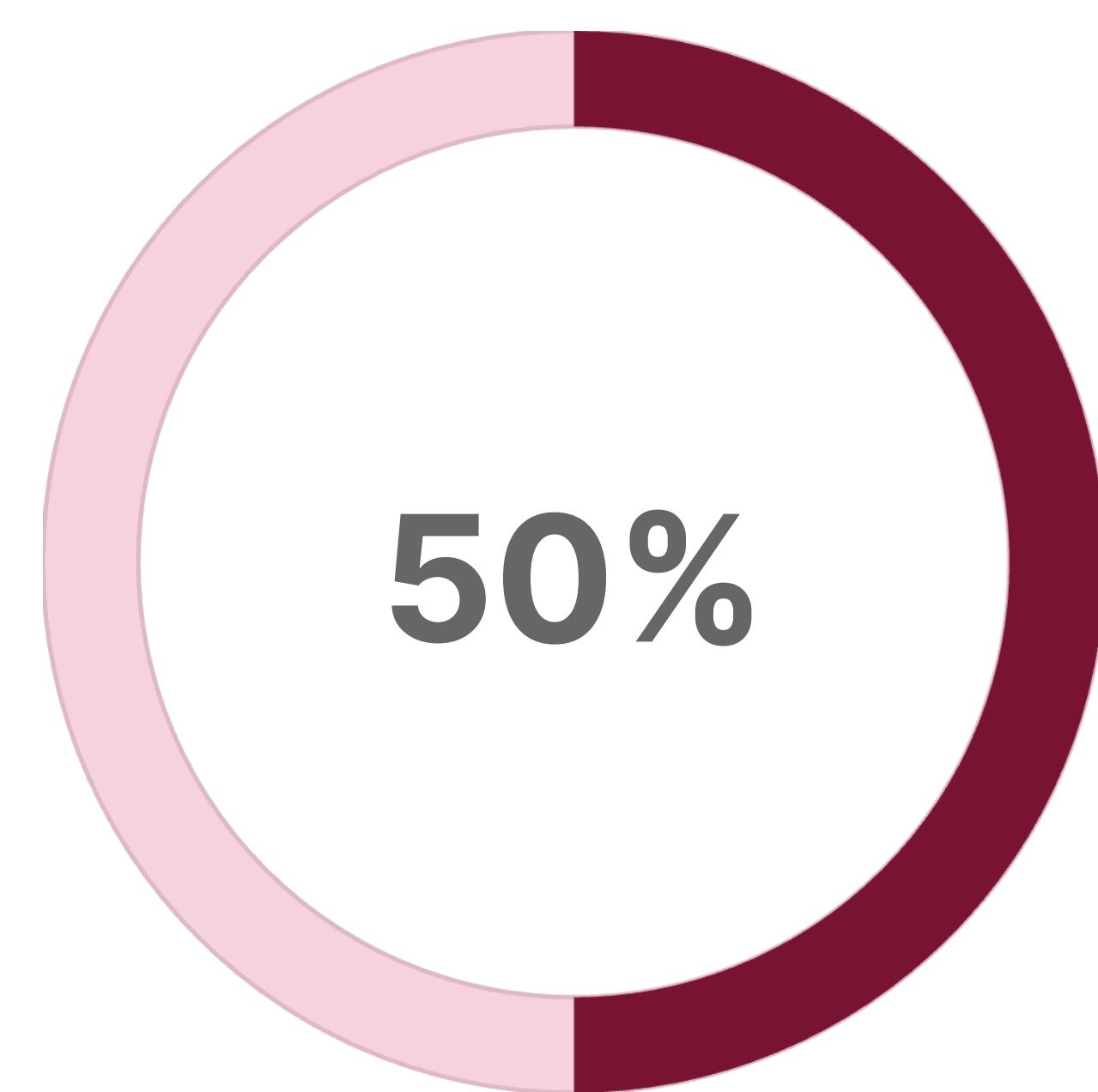
**Nivel de Confianza Típico**

Estándar en investigación académica



**Margen de Error Común**

Balance entre precisión y costo



**Proporción Conservadora**

Maximiza el tamaño muestral requerido

# Distribución Muestral y Teorema del Límite Central

La distribución muestral describe el comportamiento probabilístico de un estadístico (como la media muestral) calculado a partir de todas las posibles muestras de tamaño  $n$  extraídas de una población. Esta distribución es fundamental para la inferencia estadística, ya que permite cuantificar la variabilidad inherente en las estimaciones muestrales.

El Teorema del Límite Central establece que, independientemente de la forma de la distribución poblacional, la distribución de las medias muestrales se aproxima a una distribución normal cuando el tamaño de muestra es suficientemente grande (generalmente  $n \geq 30$ ). Este resultado es revolucionario porque permite aplicar métodos basados en la normalidad sin conocer la distribución poblacional exacta.

## Media de la Distribución Muestral

$E(\bar{X}) = \mu$ : La media de las medias muestrales igual a la media poblacional (estimador insesgado).

## Error Estándar

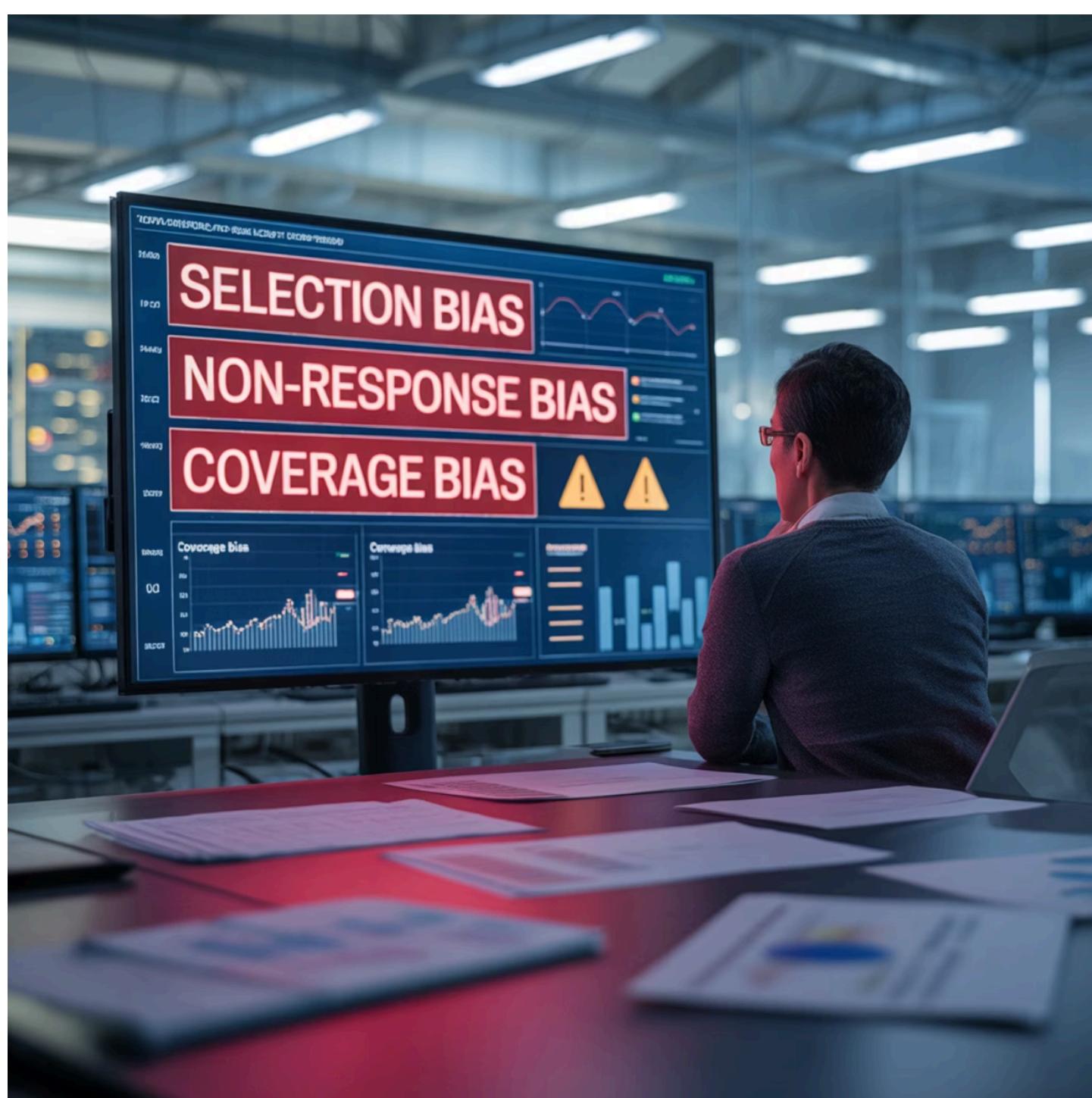
$\sigma_{\bar{X}} = \sigma/\sqrt{n}$ : La desviación estándar de las medias muestrales disminuye con el tamaño de muestra.

## Normalidad Asintótica

Para  $n$  grande,  $\bar{X} \sim N(\mu, \sigma^2/n)$ , permitiendo cálculos probabilísticos basados en la distribución normal.

El error estándar disminuye proporcionalmente a la raíz cuadrada del tamaño de muestra, lo que significa que para reducir el error a la mitad, se requiere multiplicar el tamaño de muestra por cuatro.

# Sesgos en el Muestreo



Los sesgos en el muestreo representan errores sistemáticos que hacen que la muestra no sea representativa de la población objetivo, comprometiendo la validez de las inferencias estadísticas. A diferencia del error muestral aleatorio, los sesgos no se reducen aumentando el tamaño de muestra y pueden invalidar completamente los resultados.

La identificación y prevención de sesgos requiere una planificación cuidadosa del diseño de investigación y una comprensión profunda de las características de la población y el proceso de selección. Los sesgos pueden introducirse en múltiples etapas del proceso de investigación.

## Sesgo de Selección

Ocurre cuando ciertos elementos de la población tienen mayor probabilidad de ser incluidos en la muestra.

**Ejemplo:** Encuestas telefónicas que excluyen hogares sin teléfono fijo.

## Sesgo de No Respuesta

Surge cuando los elementos seleccionados no participan, y su ausencia está relacionada con la variable de interés.

**Ejemplo:** Encuestas sobre ingresos donde los ricos tienden a no responder.

## Sesgo de Supervivencia

Concentración en elementos que "sobrevivieron" hasta el momento del estudio, ignorando los que no lo hicieron.

**Ejemplo:** Estudiar solo empresas exitosas para analizar factores de éxito.

## Sesgo de Autoselección

Los participantes se seleccionan a sí mismos, creando una muestra no representativa.

**Ejemplo:** Encuestas online voluntarias sobre satisfacción laboral.

# Conclusiones de Probabilidad y Muestreo

La unidad de Probabilidad y Muestreo establece los fundamentos teóricos y metodológicos esenciales para la transición hacia la estadística inferencial. Los conceptos desarrollados proporcionan las herramientas matemáticas y conceptuales necesarias para cuantificar la incertidumbre y diseñar investigaciones que produzcan resultados válidos y confiables.

La comprensión profunda de la teoría de probabilidades permite modelar fenómenos inciertos y calcular la probabilidad de eventos complejos. Las distribuciones de probabilidad proporcionan marcos matemáticos para describir la variabilidad en datos y procesos, mientras que el Teorema del Límite Central justifica el uso de métodos normales en una amplia variedad de situaciones prácticas.

## Fundamentos Probabilísticos

Capacidad para aplicar reglas de probabilidad, trabajar con distribuciones y utilizar el teorema de Bayes para actualizar creencias con nueva evidencia.

## Diseño de Muestreo

Competencias para seleccionar métodos de muestreo apropiados, calcular tamaños de muestra y identificar fuentes potenciales de sesgo.

## Inferencia Estadística

Preparación para utilizar distribuciones muestrales en la estimación de parámetros y pruebas de hipótesis que se abordarán en la siguiente unidad.

## Referencias Bibliográficas

1. Cochran, W. G. (1997). *Técnicas de muestreo*. Editorial Limusa.
2. Devore, J. L. (2019). *Probabilidad y estadística para ingeniería y ciencias* (9.<sup>a</sup> ed.). Cengage Learning.
3. Montgomery, D. C., & Runger, G. C. (2018). *Probabilidad y estadística aplicadas a la ingeniería* (7.<sup>a</sup> ed.). John Wiley & Sons.
4. Ross, S. M. (2020). *Introducción a la probabilidad y estadística para ingenieros y científicos* (5.<sup>a</sup> ed.). McGraw-Hill Education.
5. Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. Springer.