ABP MÓDULO 5
MARTIN KOCK

Lección 1: Método científico y estadística

1. Planteamiento del problema

Una mina subterránea presenta eventos de daño dinámico y estallidos de roca. Como consultoría de Data Science en geotecnia, proponemos un estudio inferencial con datos simulados para cuantificar relaciones entre sismicidad, esfuerzos y condiciones operacionales, estimar incertidumbre con intervalos de confianza y validar hipótesis mediante pruebas de significancia, con el fin de apoyar decisiones de control de riesgo

Problema de investigación:

Evaluar si los turnos con ocurrencia de rockburst presentan un esfuerzo estimado promedio mayor que los turnos sin rockburst, cuantificando el tamaño del efecto e incertidumbre mediante inferencia estadística; variables operacionales y geotécnicas adicionales se registran para describir el contexto y controlar variabilidad.

2) Pregunta de investigación

- ¿La media de stress_mpa es mayor en turnos con rockburst que sin rockburst?

3) Hipótesis (nula y alternativa)

Hipótesis A (medias):

- H0: La media del esfuerzo estimado (stress_mpa) es igual en turnos con rockburst y sin rockburst.
- H1: La media del esfuerzo estimado (stress_mpa) es mayor en turnos con rockburst que en turnos sin rockburst.

4) Variables relevantes (cuantitativas y cualitativas)

Variable objetivo (respuesta):

- **rockburst** (cualitativa nominal/binaria): 0 = no ocurrió; 1 = ocurrió.

Variables explicativas cuantitativas:

- **stress_mpa** (continua): esfuerzo estimado en la labor (MPa). Variable principal para contraste de hipótesis.
- **ppv_mm_s** (continua): velocidad pico de partícula (Peak Particle Velocity, mm/s) registrada en el turno.
- **depth_m** (continua): profundidad de la labor (metros).

- **gsi** (discreta): índice de resistencia geológica (Geological Strength Index), valor entero de 0 a 100.
- **seismic_events_count** (discreta): cantidad de eventos sísmicos registrados durante el turno.

Variables cualitativas:

- **sector** (nominal): zona geomecánica de operación (categorías: A, B, C).
- **shift** (nominal/binaria): turno de trabajo (categorías: Día, Noche).
- **id_turno** (identificador): correlativo único del registro (variable auxiliar).

5) Enfoque del método científico

1. Observación del fenómeno (rockburst) y formulación del problema.
2. Revisión conceptual (geomecánica: esfuerzo, sismicidad inducida, calidad de roca, influencia operacional).
3. Formulación de hipótesis testables sobre medias.
4. Diseño de estudio y definición de variables/mediciones.
5. Recolección o **simulación** de datos (en este proyecto se generará un dataset sintético en Python).
6. Análisis inferencial: probabilidades, selección de distribuciones, TLC, intervalos de confianza y tests de hipótesis.
7. Conclusión: aceptación/rechazo de H0 según valor-p y estimación con IC; recomendaciones operacionales y de monitoreo.

6) Diseño preliminar del estudio

Tipo de estudio: observacional analítico.

Población objetivo: todos los turnos operacionales de la mina (o de los sectores definidos) durante un periodo de referencia.

Muestra/dataset: 200 registros, cada fila representa un turno en un sector, con variables geotécnicas/operacionales y ocurrencia de `rockburst`.

Unidad de análisis: turno–sector (1 fila = 1 turno en 1 sector)".

Tipo de muestreo: muestreo estratificado por sector (A/B/C) con selección aleatoria de turnos dentro de cada estrato

Lección 2: Probabilidad y estadística

1) Espacio muestral

Espacio muestral (Omega): todos los registros turno-sector del periodo de referencia. Cada resultado es un turno en un sector, con sus mediciones (stress_mpa, seismic_events_count, shift, etc.) y si ocurrió o no rockburst.

2) Definición de eventos aleatorios

Se definen los siguientes eventos:

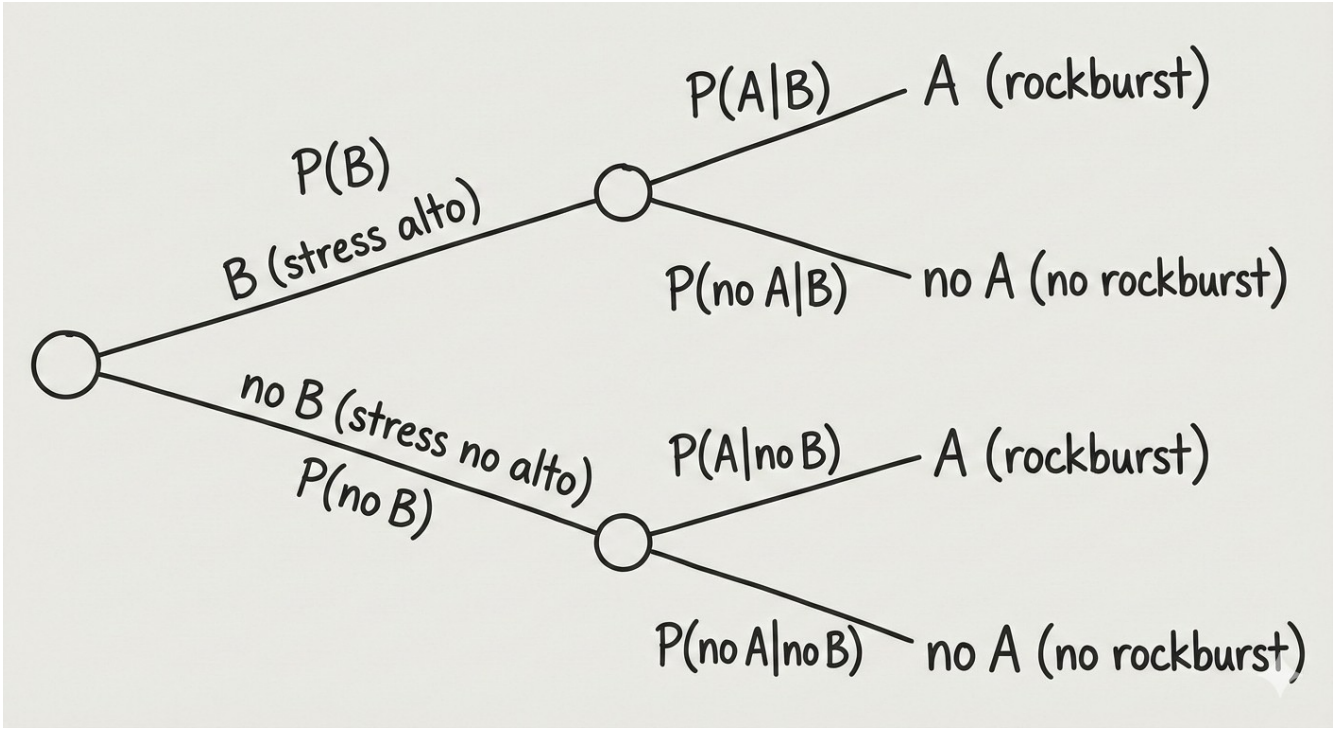
A: ocurre rockburst (rockburst = 1).
B: stress alto (stress_mpa >= s0).
C: alta sismicidad (seismic_events_count >= k).
D: turno noche (shift = "noche").

Notas:

- s0 y k son umbrales que se fijan antes del análisis.

3) Árbol de probabilidad

- Rama 1: Stress alto con probabilidad P(B)
    - Subrama 1: Ocurre rockburst dado stress alto con probabilidad P(A dado B)
    - Subrama 2: no ocurre rockburst dado stress alto con probabilidad P(no A dado B)
- Rama 2: no B (stress no alto) con probabilidad P(no B)
    - Subrama 1: ocurre rockburst dado stress no alto con probabilidad P(A dado no B)
    - Subrama 2: no ocurre rockburst dado stress no alto con probabilidad P(no A dado no B)



4) Tipo de muestreo

Se aplicará muestreo estratificado por sector (A, B, C). Dentro de cada sector se seleccionan turnos de forma aleatoria para asegurar representación de todas las zonas geomecánicas y reducir sesgo por heterogeneidad espacial.

5) Simulación del diseño muestral

Se simulará un dataset con 200 registros. Cada fila representa un turno en un sector (unidad de análisis: turno-sector).

Variables mínimas a incluir para esta lección:
sector, shift, stress_mpa, seismic_events_count, rockburst.

Ejemplo de asignación por estrato:
aproximadamente 67 registros por sector (A, B y C).

6) Cálculo de probabilidades básicas

--- Umbrales definidos ---

Umbral Stress Alto (B): >= 43.28 MPa

Umbral Alta Sismicidad (C): >= 4 eventos

--- Probabilidades Marginales ---

P(A) [Rockburst]: 0.0600 (12/200)

P(no A) [No Rockburst]: 0.9400

--- Intersección P(A ∩ B) ---

Probabilidad de Rockburst Y Stress Alto: 0.0500

--- Unión P(A U C) ---

Probabilidad de Rockburst O Alta Sismicidad: 0.4450

--- Probabilidades Condicionales (Para el Árbol) ---

P(B) [Stress Alto]: 0.2500

P(A | B) [Rockburst dado Stress Alto]: 0.2000

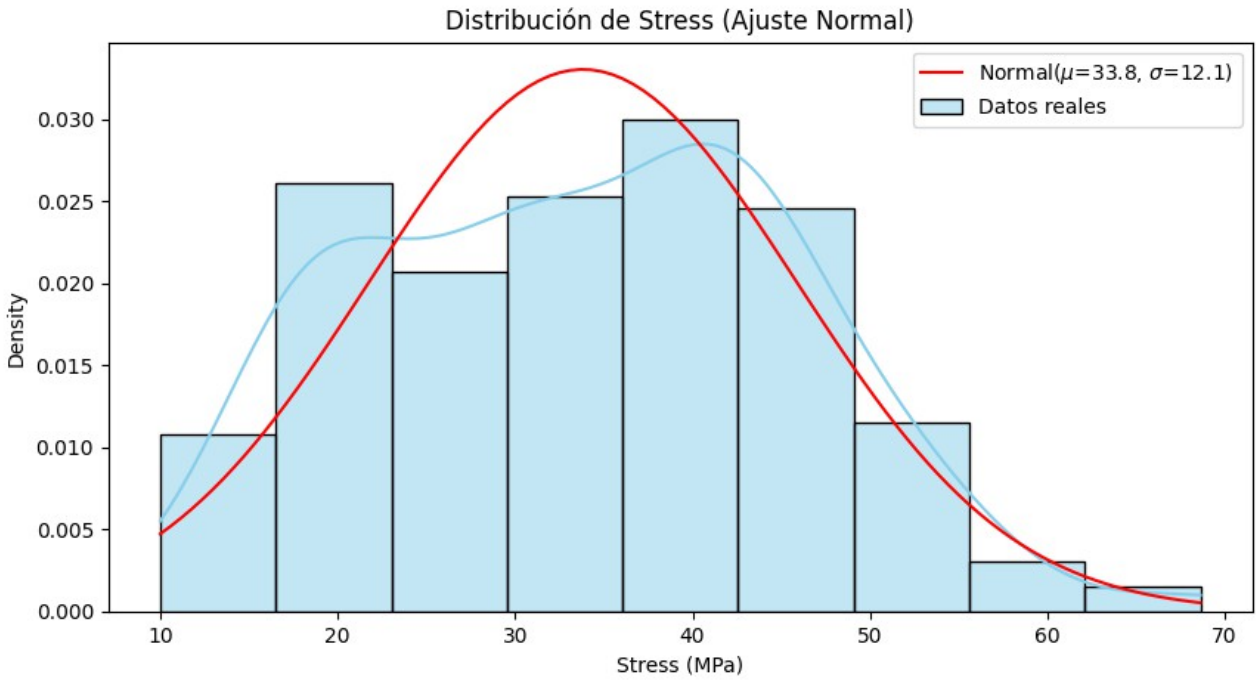
P(A | no B) [Rockburst dado Stress no alto]: 0.0133

Lección 3: Distribución de Probabilidad

3.1. Identificación y justificación de distribuciones

Variable Continua: stress_mpa (Esfuerzo estimado)

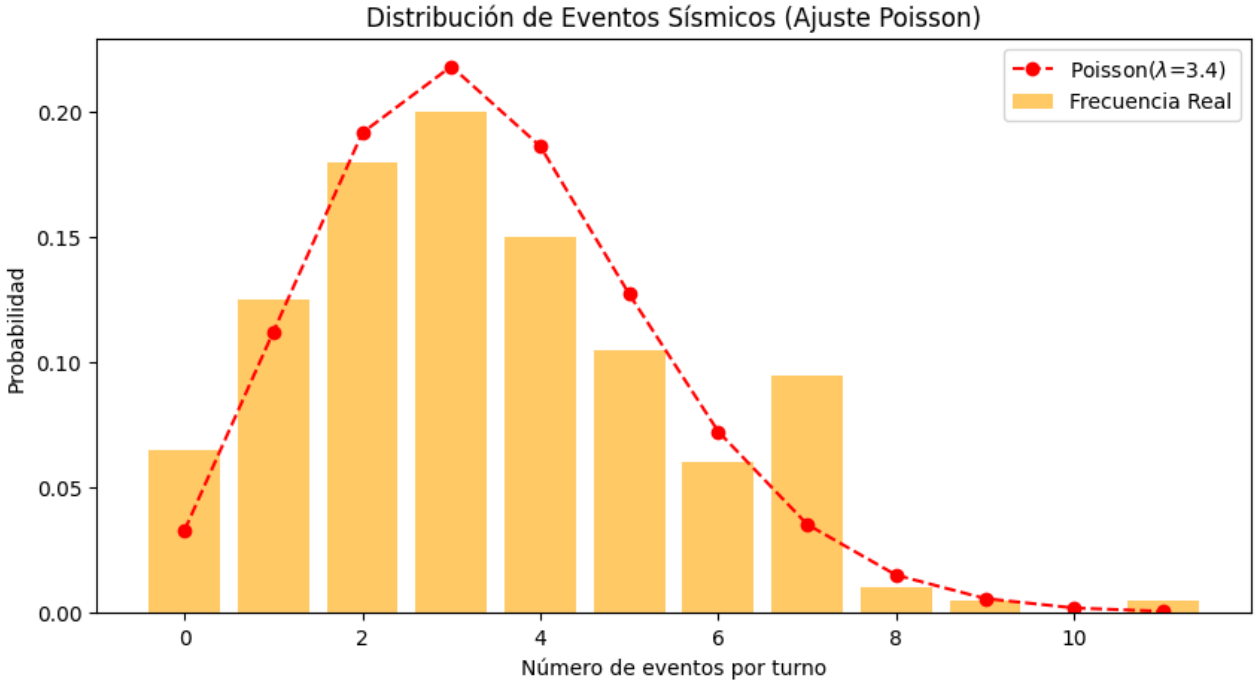
- **Distribución:** Normal (Gaussiana).
- **Justificación:** El esfuerzo en macizos rocosos es el resultado de la suma de múltiples factores independientes (carga litostática, tectónica, heterogeneidad local). Según el Teorema Central del Límite, esta convergencia de efectos aleatorios tiende a generar una distribución simétrica alrededor de una media, lo cual se confirmó visualmente con el histograma de los datos simulados.



-

Variable Discreta 1: seismic_events_count (Conteo de eventos)

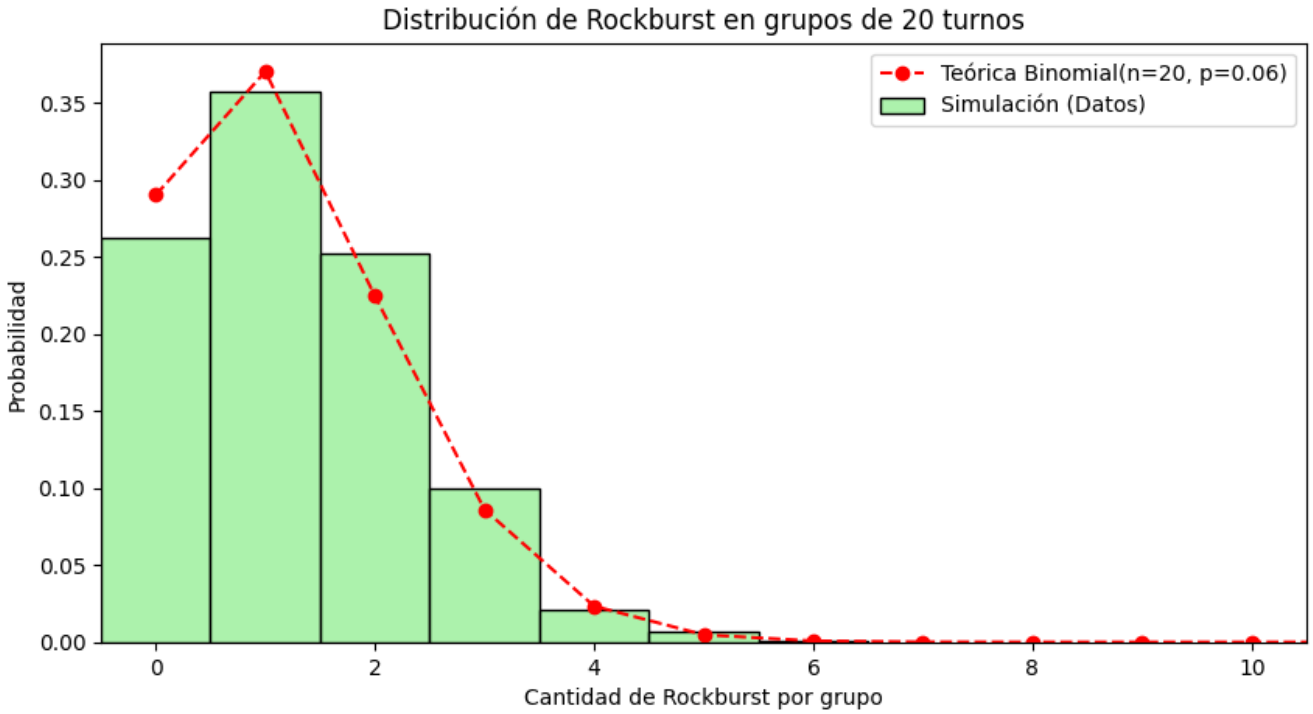
- **Distribución:** Poisson.
- **Justificación:** Representa el conteo de eventos independientes (sismos) que ocurren en un intervalo fijo de tiempo (un turno). Al graficar la frecuencia observada, se verificó el ajuste a la curva teórica de Poisson, caracterizada por su asimetría positiva y cola hacia la derecha.



-

Variable Discreta 2: rockburst (Ocurrencia agrupada)

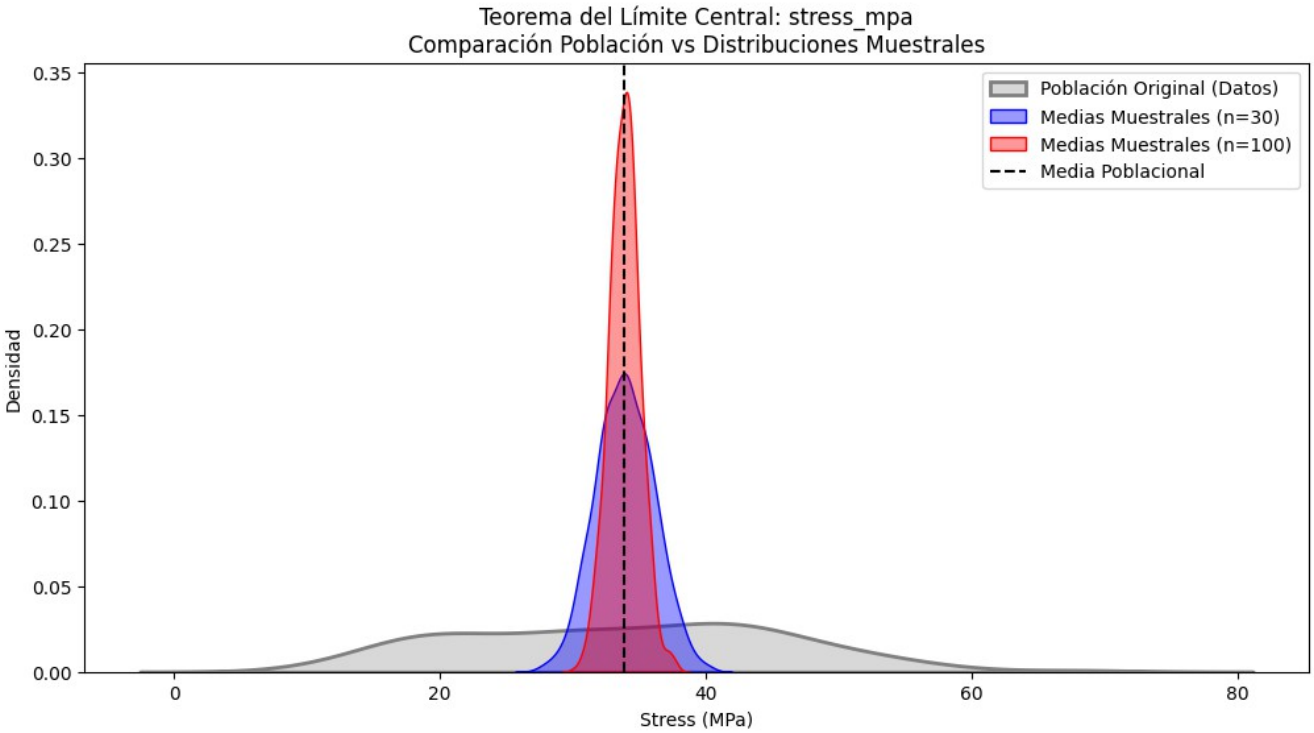
- **Distribución:** Binomial.
- **Justificación:** La variable original es binaria (Bernoulli). Al analizar la ocurrencia de rockburst en grupos de n=20 turnos mediante simulación, la distribución de frecuencias del número de éxitos (eventos de rockburst) se ajustó correctamente a una distribución Binomial B(n, p).



Lección 4: Distribución muestral y Teorema del Límite Central (TLC)

1. Generación de distribuciones muestrales y verificación del TLC

Para verificar empíricamente el Teorema del Límite Central, se realizó una simulación de muestreo repetido sobre la variable poblacional stress_mpa (con media mu = 33.82 MPa y desviación sigma = 12.08 MPa). Se generaron 1000 muestras (medias) aleatorias mediante la técnica de remuestreo con reemplazo (bootstrapping) a partir del dataset original para dos tamaños distintos: n=30 y n=100.



2. Comparación de distribuciones y análisis de dispersión
Los resultados obtenidos (ver gráfico y tabla de simulación) confirman los postulados del TLC:

- **Centralidad:** Las medias de las distribuciones muestrales (33.92 para n=30 y 33.88 para n=100) convergen casi exactamente a la media poblacional (33.82), demostrando que la media muestral es un estimador insesgado.
- **Normalidad:** A diferencia de la distribución poblacional original (curva gris), que presenta mayor dispersión e irregularidad, las distribuciones de las medias muestrales (curvas azul y roja) adoptan una forma de campana simétrica perfecta (distribución Normal).
- **Dispersión (Error Estándar):** Se observa una reducción significativa de la variabilidad al aumentar el tamaño de muestra. El error estándar simulado bajó de 2.18 (con n=30) a 1.18 (con n=100). Esto coincide con la teoría, que indica que el error estándar disminuye proporcionalmente a la raíz cuadrada del tamaño de muestra, validando que muestras más grandes entregan estimaciones más precisas de la media poblacional.

Lección 5: Inferencia e intervalos de confianza para la media

5.1. Intervalos de confianza para dos variables clave

Variable: stress_mpa (Media Muestral: 33.82 MPa)

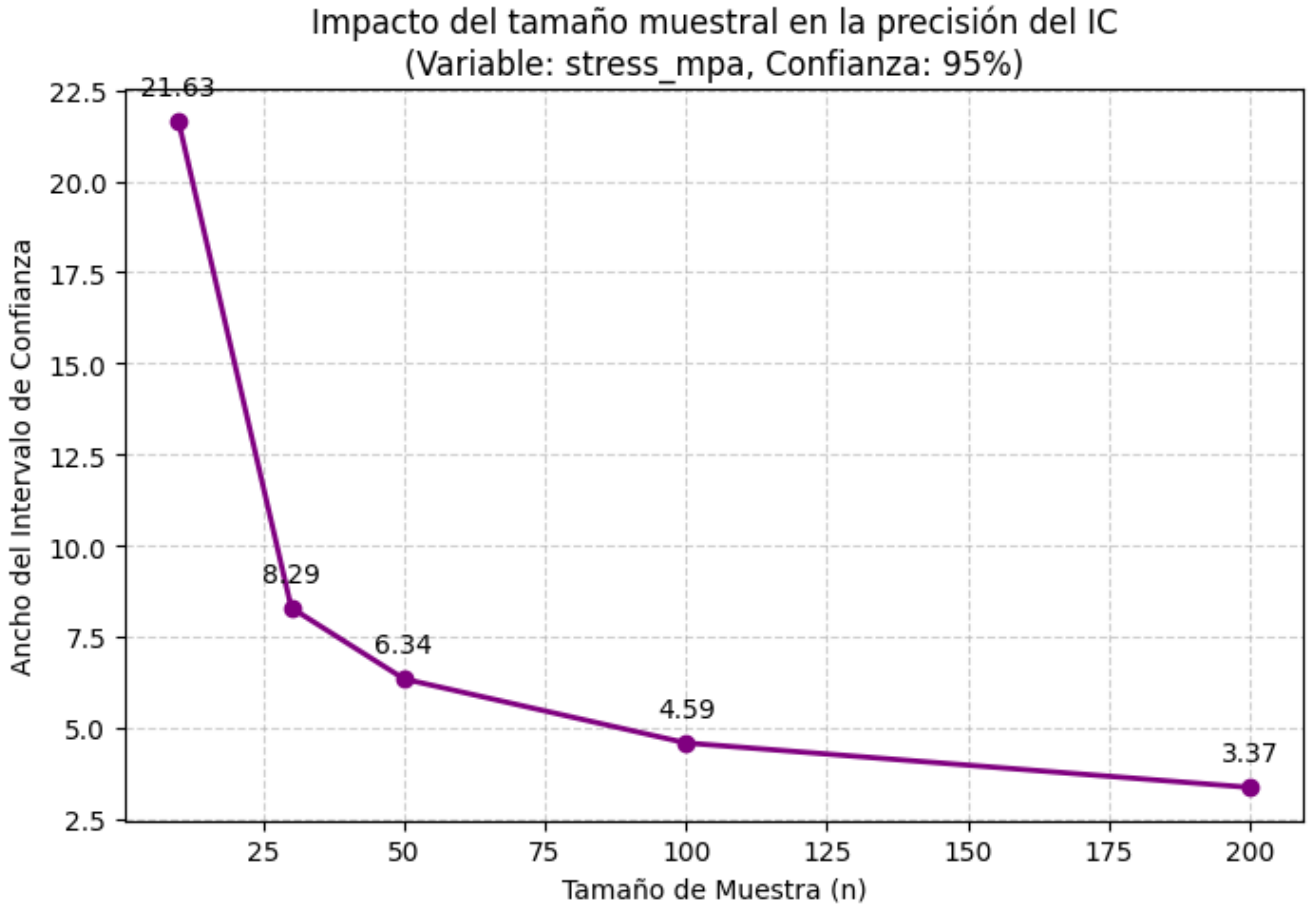
- **IC 90%:** [32.41, 35.23]. Existe un 90% de confianza de que la media real del esfuerzo poblacional se encuentra en este rango.
- **IC 95%:** [32.14, 35.51]. Al aumentar la exigencia de confianza, el intervalo se ensancha ligeramente (ancho de 3.37 MPa).
- **IC 99%:** [31.60, 36.04]. Para garantizar un 99% de certeza de contener el parámetro

poblacional, el margen de error aumenta, resultando en el intervalo más amplio (4.44 MPa).

Variable: ppv_mm_s (Media Muestral: 9.95 mm/s)

- **IC 90%:** [7.04, 12.87].
- **IC 95%:** [6.48, 13.43].
- **IC 99%:** [5.37, 14.54]. Se observa una mayor dispersión relativa en esta variable, lo que se traduce en intervalos proporcionalmente más anchos respecto a la media.

5.2. Impacto del tamaño muestral en la precisión



El análisis de sensibilidad demuestra cómo la precisión de la estimación depende críticamente del tamaño de la muestra (n):

- **Con n=10:** El intervalo es extremadamente ancho (21.63 MPa), lo que implica una incertidumbre muy alta e inutilidad práctica para la toma de decisiones.
- **Con n=50:** El ancho se reduce drásticamente a 6.34 MPa, mejorando significativamente la calidad de la estimación.
- **Con n=200:** Se alcanza la mayor precisión observada, con un ancho de solo 3.37 MPa.

Conclusión: Existe una relación inversa no lineal entre n y el ancho del intervalo. Duplicar el tamaño de muestra reduce el error de estimación, pero con rendimientos decrecientes, validando la importancia de diseñar un plan de muestreo con un n suficiente (como n=200) para obtener intervalos de confianza

aceptables.

Lección 6: Test de significancia y conclusiones

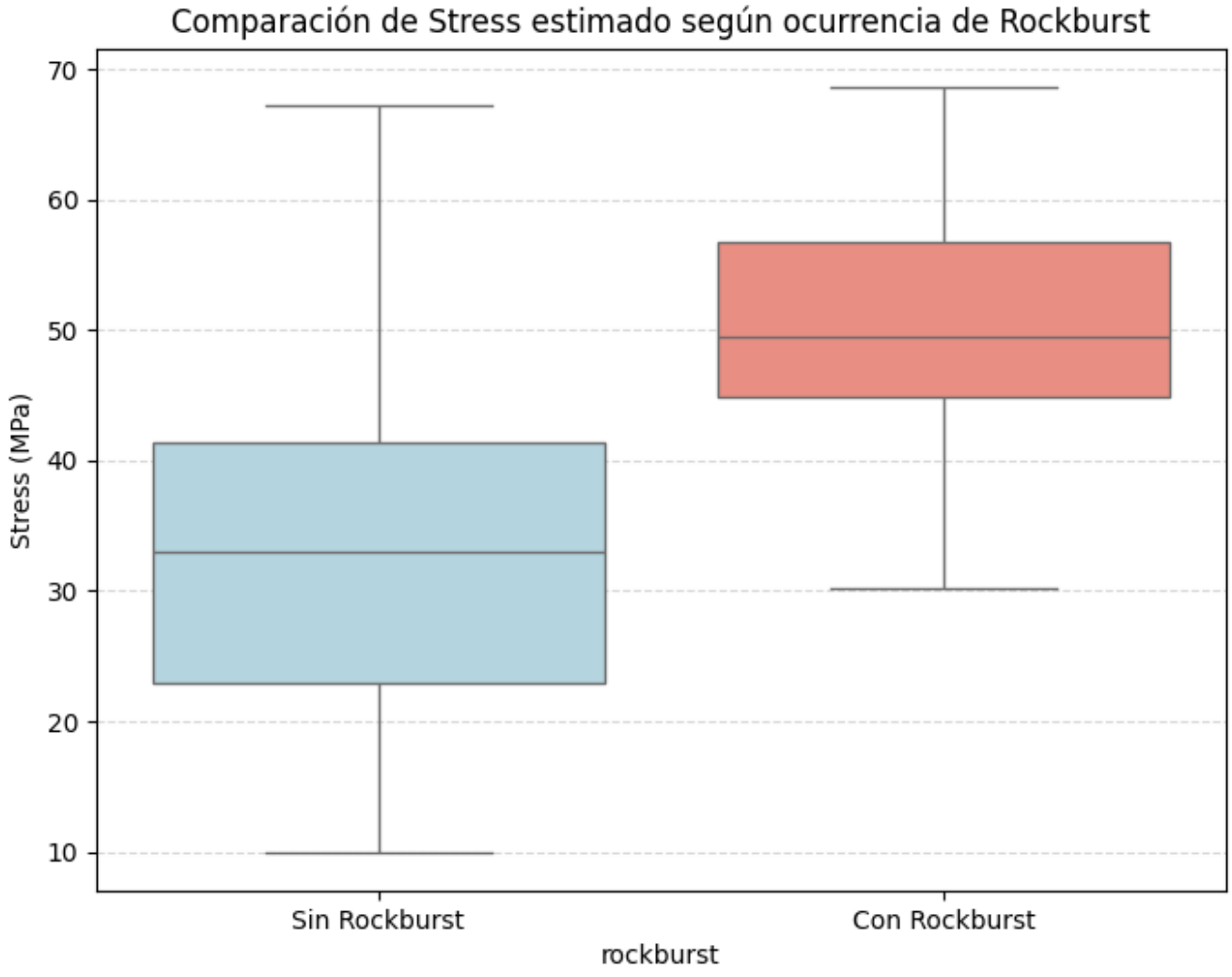
6.1. Formulación de la prueba de hipótesis

Para validar si el esfuerzo del macizo rocoso es un factor determinante en la ocurrencia de eventos sísmicos mayores, se planteó la siguiente prueba de hipótesis unilateral:

- **H0 (Nula):** La media del esfuerzo estimado (stress_mpa) en turnos con rockburst es menor o igual a la de turnos sin rockburst.
- **H1 (Alternativa):** La media del esfuerzo estimado en turnos con rockburst es mayor que en turnos sin rockburst.
- **Nivel de significancia (alpha):** 0.05.

6.2. Resultados del test t-Student

Se compararon dos muestras independientes: un grupo de control (sin rockburst, n=188) y el grupo de interés (con rockburst, n=12). Los estadísticos descriptivos muestran una clara diferencia visual (ver diagrama de cajas), con una media de 49.35 MPa para el grupo con eventos versus 32.83 MPa para el grupo sin eventos.



El test estadístico arrojó un **valor-t de 5.0756** y un **valor-p de 0.0001168**.

6.3. Decisión y análisis de errores

Dado que el valor-p (0.00011) es significativamente menor que alpha (0.05), se rechaza la Hipótesis Nula. Esto implica que la diferencia observada no es producto del azar.

- **Análisis de Error Tipo I:** Existe una probabilidad muy baja (0.01%) de haber rechazado H0 incorrectamente (afirmar que hay relación cuando no la hay). En el contexto operativo, este riesgo es aceptable dado el beneficio de la prevención.
- **Análisis de Error Tipo II:** Al ser una prueba potente, el riesgo de no detectar una relación real se minimiza, lo cual es crítico en seguridad minera para no pasar por alto condiciones de peligro.

6.4. Conclusión final

Existe evidencia estadística suficiente para afirmar que los turnos con ocurrencia de rockburst presentan niveles de esfuerzo significativamente mayores. Por tanto, la variable 'stress_mpa' es un predictor válido y se recomienda implementar protocolos de alerta temprana cuando las mediciones superen el umbral crítico identificado.