Exámen Final SOL114

Instrucciones: Responde a las preguntas proporcionadas, asegurándote de justificar tus respuestas con cálculos y razonamientos claros.

1. Un exclusivo club social admite solo a personas ricas y de "buena presencia". Para ser admitido en el club, los solicitantes deben cumplir con dos condiciones: estar en el 10% más alto de la distribución de ingresos y tener un nivel de belleza por encima de la media general.

- Las variables Y y X representan ingresos y belleza, respectivamente.
- La variable W mide si un individuo es admitido en el club social o no, donde W=1 indica que es admitido, y W=0 indica que no lo es.
- Y sigue una distribución normal con una media (μ) de 500,000 pesos y una desviación estándar (σ) de 600,000 pesos.
- X sigue una distribución normal estándar, $X \sim \text{Normal}(0,1)$.
- Las variables ingresos (Y) y belleza (X) son independientes entre sí.

Preguntas:

a) ¿Cuál es el nivel mínimo de ingresos necesario para ser admitido en el club?

Para ser admitido en el club, se debe estar en el 10% más alto de la distribución de ingresos. Esto significa que el nivel mínimo de ingresos (Y_{minimo}) necesario para ser admitido se calcula como:

Figure 1: Club Social

$$Y_{\texttt{minimo}} = \mu + z_{0.90} \cdot \sigma$$

Donde:

- μ es la media de ingresos (500,000 pesos).
- $z_{0.90}$ es el valor z correspondiente al percentil 90 de una distribución normal estándar (aproximadamente 1.28).
- σ es la desviación estándar de ingresos (600,000 pesos).

Calculando $Y_{\texttt{minimo}}$:

$$Y_{\texttt{minimo}} = 500,000 + 1.28 \cdot 600,000 = 1,268,000$$

Por lo tanto, el nivel mínimo de ingresos necesario para ser admitido en el club es de 1,268,000 pesos.

b) ¿Cuál es el nivel mínimo de belleza necesario para ser admitido en el club?

Para ser admitido en el club, no se especifica un nivel mínimo de belleza en el enunciado. Sin embargo, se requiere que el nivel de belleza (X) esté por encima de la media general, lo que se traduce en ($X > 0$). Por lo tanto, se espera que la belleza (X) sea mayor que cero para ser considerado para la admisión en el club.

c) ¿Cuál es la probabilidad de ser admitido en el club?

La probabilidad de ser admitido en el club se calcula como el producto de dos probabilidades independientes:

- Probabilidad de estar en el 10% más alto de ingresos: Esta probabilidad corresponde al percentil 90 de la distribución de ingresos, que ya calculamos en el punto (a).
- Probabilidad de tener un nivel de belleza por encima de la media general: Dado que X sigue una distribución normal estándar, la probabilidad de estar por encima de la media general es $P(X > 0) = 0.5$.

La probabilidad de ser admitido es el producto de estas dos probabilidades:

$$P(\texttt{Admitido}) = P(\texttt{Ingresos en el 10pct más alto}) \cdot P(\texttt{Belleza por encima de la media general})$$

$$P(\texttt{Admitido}) = 0.1 \cdot 0.5 = 0.05$$

Por lo tanto, la probabilidad de ser admitido en el club es del 5%.

d) ¿Cuál es la distribución de probabilidad que caracteriza a la variable W?

La variable W mide si un individuo es admitido en el club o no, y toma valores 1 (admitido) o 0 (no admitido). Dado que la probabilidad de ser admitido es 0.05 y la probabilidad de no ser admitido es complementaria (1 - 0.05 = 0.95), la distribución de probabilidad de W se puede resumir de la siguiente manera:

$$P(W = 1) = 0.05$$

$$P(W = 0) = 0.95$$

e) ¿Cuál es la media y la varianza de la variable W?

La media de la variable W se calcula como la suma ponderada de sus posibles valores:

$$E(W) = 1 \cdot P(W = 1) + 0 \cdot P(W = 0) = 0.05$$

La varianza de la variable W se calcula como la suma ponderada de las desviaciones cuadráticas de los valores posibles respecto a la media:

$$Var(W) = (1-E(W))^2 \cdot P(W = 1) + (0-E(W))^2 \cdot P(W = 0) = 0.05 \cdot 0.95 + 0.95 \cdot 0.05 = 0.0475$$

Por lo tanto, la media de la variable W es 0.05 y la varianza es 0.0475, teniendo en cuenta que X sigue una distribución normal estándar con media 0.

2. Si X es una variable aleatoria continua, con función de densidad $f(x)$, identifica y explica brevemente las siguientes expresiones:

a). .... $= \int_{-\infty}^{x} f(x)dx$, en palabras ....

$F(x) = \int_{-\infty}^{x} f(x)dx$: Esta expresión representa la Función de Distribución Acumulativa (FDA) de la variable aleatoria X . En palabras, se describe como la probabilidad de que la variable aleatoria X sea menor o igual a un valor dado x . Es una medida de cuánta probabilidad se ha acumulado hasta el punto x .

b). .... $= \int_{-\infty}^{\infty+} x{\cdot}f(x)dx$, en palabras ....

$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$: Esta expresión denota el Valor Esperado de la variable aleatoria X . En términos simples, es el valor promedio esperado de X. Esta integral pondera cada valor posible que X puede tomar por su probabilidad y suma todos estos productos, proporcionando una medida central de la distribución de X .

c). .... $= \int_{-\infty}^{\infty+} (x{-}E[X])^2{\cdot}f(x)dx$, en palabras ....

$Var[X] = \int_{-\infty}^{\infty} (x - E[X])^2 \cdot f(x)dx$: Esta fórmula representa la Varianza de la variable aleatoria X. Se define como el valor esperado del cuadrado de la desviación de X de su media $E[X]$. En palabras, es una medida de cuánto se espera que los valores de X se desvíen de la media. La varianza es un indicador importante de la dispersión o el grado de variación de los valores de X alrededor de su media.

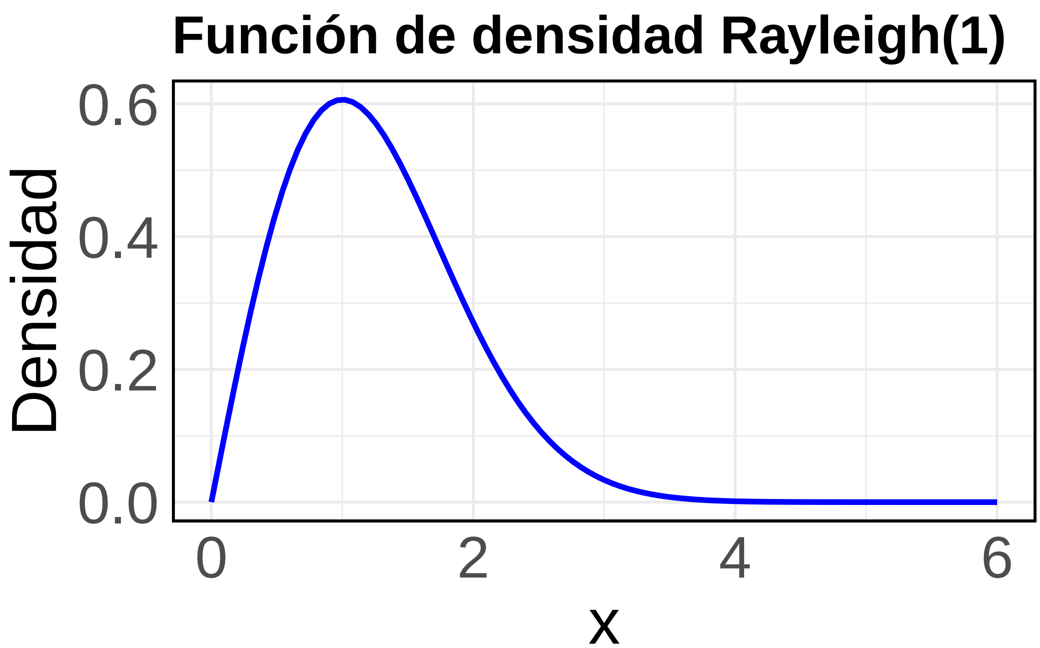
3. Supongamos que tenemos 1000 variables aleatorias independientes $X_1, X_2, ..., X_{1000}$, cada una siguiendo una distribución Rayleigh(1), es decir: $X_i \sim \text{Rayleigh}(1)$. La función de densidad de probabilidad (PDF) de la distribución Rayleigh(1) es:

$$f(x) = x \cdot e^{-x^2/2}, \quad x \geq 0$$

con parámentros

- Media: $\mu = \sqrt{\frac{\pi}{2}}$
- Varianza: $\sigma^2 = 2 - \frac{\pi}{2}$

Visualmente:



Preguntas:

Si definimos una nueva variable aleatoria Y como el promedio de estas 1000 variables aleatorias:

$$Y = \frac{1}{10000} \sum_{i=1}^{10000} X_i$$

a) Según el Teorema del Límite Central (CTL), ¿cuál es la distribución aproximada de Y? Especifica la media y la varianza de esta distribución aproximada.

El Teorema del Límite Central establece que, en condiciones muy generales, la suma (o el promedio) de una gran cantidad de variables aleatorias independientes e idénticamente distribuidas se aproxima a una distribución normal, independientemente de la forma de la distribución original de las variables. Esto se cumple siempre y cuando el número de variables sea lo suficientemente grande, lo cual es el caso con 1000 variables.

Dado que tenemos 1000 variables aleatorias $X_1, X_2, \ldots, X_{10000}$ que siguen una distribución Rayleigh(1) y son independientes entre sí, podemos aplicar el TLC para encontrar la distribución aproximada de su promedio Y.

La distribución Rayleigh(1) tiene las siguientes características: - Media (μ): $\sqrt{\frac{\pi}{2}}$ - Varianza (σ^2): $2-\frac{\pi}{2}$

Cuando calculamos el promedio de estas 1000 variables, la media de la distribución resultante será igual a la media de la distribución original (Rayleigh), ya que el promedio no cambia la ubicación del centro de la distribución. Sin embargo, la varianza de la distribución resultante será igual a la varianza de la distribución original dividida por el número de variables (1000 en este caso), debido a la propiedad de la varianza de sumas de variables aleatorias independientes.

Por lo tanto, la media (μ_Y) de Y será:

$$\mu_Y = \sqrt{\frac{\pi}{2}}$$

y la varianza (σ_Y^2) de Y será:

$$\sigma_Y^2 = \frac{2-\frac{\pi}{2}}{10000}$$

La distribución de Y se aproxima a una distribución normal con estos parámetros. Ahora, calculemos los valores numéricos exactos de la media y la varianza de Y.

La distribución aproximada de Y, según el Teorema del Límite Central, es una distribución normal con los siguientes parámetros:

- Media (μ_Y): aproximadamente 1.2533
- Varianza (σ_Y^2): aproximadamente 0.0000429

Esto significa que el promedio de estas 10000 variables aleatorias independientes, cada una siguiendo una distribución Rayleigh(1), se aproxima a una distribución normal con una media de alrededor de 1.2533 y una varianza de alrededor de 0.0000429.

4. Un sociólogo toma una muestra aleatoria de la variable X y reporta los siguientes estadísticos:

Estadístico	Valor
Tamaño de la muestra (n)	576
Media de la muestra ($\bar{x}$)	10
Varianza muestral naive (σ^2_{naive})	4

Preguntas:

a) ¿Por qué $\bar{x}$ es una variable aleatoria y no simplemente un número?

$\bar{x}$ es considerado una variable aleatoria porque su valor depende de la muestra específica que se selecciona de la población. Cada vez que se toma una muestra diferente, es probable que se obtenga un valor diferente de $\bar{x}$. Por lo tanto, $\bar{x}$ tiene una distribución de probabilidad asociada que refleja la variabilidad en estas muestras.

b) Dado que $\bar{x}$ es una variable aleatoria, ¿cuál es su distribución?. Explíca la forma de la distribución y sus parámetros (pista: media y desviación estándar).

Aplicando el Teorema del Límite Central, si el tamaño de la muestra es grande ($n > 30$), la distribución de $\bar{x}$ se aproxima a una distribución normal con:

- Media: Igual a la media poblacional μ.
- Desviación Estándar (Error Estándar): Se calcula ajustando la varianza muestral "naive" y luego dividiendo por la raíz cuadrada del tamaño de la muestra.

La fórmula sería $\frac{\sqrt{\sigma^2_{naive} \cdot \frac{n}{n-1}}}{\sqrt{n}}$.

Para los datos dados ($\sigma^2_{naive} = 4$, $n = 576$), el cálculo sería:

$$\text{Error Estándar} = \frac{\sqrt{4 \cdot \frac{576}{575}}}{\sqrt{576}} = 0.0834$$

c) ¿Que otro nombre recibe la desviación estándar de un estimador como $\bar{x}$ y qué mide?

La desviación estándar de $\bar{x}$, en este contexto, se conoce como Error Estándar. Mide la variabilidad o dispersión esperada de $\bar{x}$ como estimador de la media poblacional μ. Es una medida clave de la precisión de $\bar{x}$ y es esencial para calcular intervalos de confianza y realizar pruebas de hipótesis.

El error estándar de $\bar{x}$, después de ajustar la varianza muestral para obtener una estimación insesgada, es aproximadamente 0.0834. Este valor mide la variabilidad o dispersión esperada del estimador $\bar{x}$, proporcionando una indicación de la precisión de $\bar{x}$ como estimador de la media poblacional.

5. En un estudio se investiga el efecto de un programa de realidad virtual (RV) en la reducción de la ansiedad social en jóvenes adultos. El programa utiliza entornos virtuales para simular situaciones sociales desafiantes y enseñar habilidades de afrontamiento.

- Diseño del Experimento: Se seleccionan dos grupos de jóvenes adultos que han reportado niveles moderados de ansiedad social.
- Grupo de Tratamiento (GT): Participantes que usan el programa de RV durante 8 semanas, con sesiones de 30 minutos tres veces por semana.
- Grupo de Control (GC): Participantes que no usan el programa de RV y mantienen sus rutinas habituales.
- Medición: La ansiedad social se mide utilizando la Escala de Ansiedad Social (EAS), una herramienta validada con puntuaciones que van de 0 (sin ansiedad) a 100 (ansiedad extremadamente alta). Las mediciones se realizan antes y después del período de intervención.

Los datos recopilados al final del estudio son los siguientes:

Grupo	Tamaño de la Muestra	promedio de EAS ($\bar{X}$)	Varianza de EAS (s^2)
Tratamiento (1)	200	35	12
Control (2)	200	50	15

- Pregunta de Investigación:: ¿El uso de un programa de realidad virtual para la práctica de habilidades sociales reduce significativamente los niveles de ansiedad social en comparación con un grupo de control que no utiliza esta tecnología?

Preguntas:

a) Calcula y compara los intervalos de confianza al 95% para los promedios de ansiedad social de cada grupo.

La fórmula para el intervalo de confianza de un promedio es:

$$\bar{x} \pm z \times \sqrt{\frac{s^2}{n}}$$

donde: - $\bar{x}$ es la media de la muestra. - z es el valor crítico para el 95% de confianza, que es aproximadamente 1.96 para una distribución normal. - s^2 es la varianza de la muestra. - n es el tamaño de la muestra.

Grupo de Tratamiento (GT)

- Media ($\bar{x}_1$) = 35
- Varianza (s_1^2) = 12
- Tamaño de la muestra (n_1) = 200

El error estándar para GT es:

$$\sqrt{\frac{s_1^2}{n_1}} = \sqrt{\frac{12}{200}} \approx 0.245$$

El intervalo de confianza es:

$$35 \pm 1.96 \times 0.245 \approx (35 \pm 0.48) \approx (34.52, 35.48)$$

Grupo de Control (GC)

- Media ($\bar{x}_2$) = 50
- Varianza (s_2^2) = 15
- Tamaño de la muestra (n_2) = 200

El error estándar para GC es:

$$\sqrt{\frac{s_2^2}{n_2}} = \sqrt{\frac{15}{200}} \approx 0.273$$

El intervalo de confianza es:

$$50 \pm 1.96 \times 0.273 \approx (50 \pm 0.54) \approx (49.46, 50.54)$$

Los intervalos de confianza para los promedios de cada grupo muestran claramente que no se superponen, lo que indica una diferencia significativa entre los dos grupos.

b) Establece un intervalo de confianza al 95% para la diferencia en los promedios de ansiedad de cada grupo.

La fórmula para el intervalo de confianza de la diferencia entre dos medias es:

$$(\bar{x}_1 - \bar{x}_2) \pm z \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

La diferencia de medias es $35 - 50 = -15$. El error estándar de la diferencia es:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{12}{200} + \frac{15}{200}} \approx 0.366$$

El intervalo de confianza para la diferencia es:

$$-15 \pm 1.96 \times 0.366 \approx (-15 \pm 0.72) \approx (-15.72, -14.28)$$

El intervalo de confianza para la diferencia entre los grupos muestra que la reducción de la ansiedad social en el grupo de tratamiento es significativamente mayor en comparación con el grupo de control.

c) Realiza un test de hipótesis (a dos colas) para la diferencia en los promedios de ansiedad de cada grupo, con un nivel de significancia del 5%.

Para el test de hipótesis, establecemos:

- Hipótesis nula (H_0): No hay diferencia en los promedios de ansiedad ($\mu_1 = \mu_2$).
- Hipótesis alternativa (H_A): Hay una diferencia en los promedios de ansiedad ($\mu_1 \neq \mu_2$).

Usamos un test Z para la diferencia entre dos medias, ya que el tamaño de la muestra es suficientemente grande. El valor Z se calcula como:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Calculando el valor Z:

$$Z = \frac{-15}{0.366} \approx -40.98$$

El valor Z calculado es aproximadamente -40.98. Dado que este valor es mucho más bajo que -1.96, el valor p asociado es efectivamente 0, indicando que la diferencia es estadísticamente significativa al nivel del 5%.

- En base a tus hallazgos, redacta conclusiones sustantivas sobre las similitudes o diferencias en los niveles de ansiedad de ambos grupos.

Los resultados sugieren que el uso del programa de realidad virtual para la práctica de habilidades sociales reduce significativamente los niveles de ansiedad social en comparación con el grupo de control que no utiliza esta tecnología. Estos hallazgos apoyan la eficacia del programa de RV en la mejora de las habilidades sociales y la reducción de la ansiedad social en jóvenes adultos.

6. La consultora CALLEMP lleva a cabo una encuesta en zonas rurales y urbanas del país para medir la intención de voto en un plebiscito con dos opciones: "A favor" y "En contra". La siguiente tabla de contingencia resume los resultados obtenidos:

	Rural	Urbano
"A favor"	50	60
"En contra"	50	140

Preguntas:

a) Realiza un test de χ^2 para testear la asociación entre el tipo zona y la intención de voto. Explicita la hipótesis que se testea y todos los pasos involucrados en el cálculo. Provee una interpretación sustantiva del resultado de este test.

- Hipótesis Nula (H_0): No hay asociación entre el tipo de zona (rural o urbano) y la intención de voto (a favor o en contra).
- Hipótesis Alternativa (H_A): Existe una asociación entre el tipo de zona y la intención de voto.

Para calcular el estadístico de chi-cuadrado ( ) paso a paso y explicitar las distribuciones marginales, primero necesitas calcular las distribuciones marginales para las filas y columnas de tu tabla de contingencia. Luego, puedes usar estas distribuciones marginales para calcular el . Aquí está el cálculo paso a paso:

Primero, calculemos las sumas marginales para las filas y columnas de la tabla:

- Suma marginal para "A favor" = 50 + 60 = 110
- Suma marginal para "En contra" = 50 + 140 = 190
- Suma marginal para "Rural" = 50 + 50 = 100
- Suma marginal para "Urbano" = 60 + 140 = 200
- Suma total de encuestados = 100 + 200 = 300 (la suma total de encuestados en tu tabla)

Ahora, calculemos las frecuencias esperadas para cada celda en la tabla de contingencia. Las frecuencias esperadas se calculan utilizando la fórmula:

$$\text{Frecuencia esperada} = \frac{\text{Suma marginal de fila} \times \text{Suma marginal de columna}}{\text{Suma total de encuestados}}$$

Para "A favor" en "Rural":

$$\text{Frecuencia esperada} = \frac{110 \times 100}{300} = 36.67$$

Para "En contra" en "Rural":

$$\text{Frecuencia esperada} = \frac{190 \times 100}{300} = 63.33$$

Para "A favor" en "Urbano":

$$\text{Frecuencia esperada} = \frac{110 \times 200}{300} = 73.33$$

Para "En contra" en "Urbano":

$$\text{Frecuencia esperada} = \frac{190 \times 200}{300} = 126.67$$

Ahora, calculamos el estadístico de chi-cuadrado (   ) utilizando la fórmula:

$$\chi^2 = \sum \frac{(\text{Observado} - \text{Esperado})^2}{\text{Esperado}}$$

Donde la suma se realiza sobre todas las celdas de la tabla de contingencia. En este caso, tendrás cuatro términos en la suma correspondientes a cada celda.

Para "A favor" en "Rural":

$$\chi_1^2 = \frac{(50 - 36.67)^2}{36.67} \approx 4.85$$

Para "En contra" en "Rural":

$$\chi_2^2 = \frac{(50 - 63.33)^2}{63.33} \approx 2.81$$

Para "A favor" en "Urbano":

$$\chi_3^2 = \frac{(60 - 73.33)^2}{73.33} \approx 2.42$$

Para "En contra" en "Urbano":

$$\chi_4^2 = \frac{(140 - 126.67)^2}{126.67} \approx 1.4$$

Paso 4: Calcular el estadístico de chi-cuadrado total (total)

Finalmente, sumamos todos los términos calculados en el Paso 3 para obtener el estadístico de chi-cuadrado total:

$$\chi^2 = 4.85 + 2.81 + 2.42 + 1.4 \approx 11.48$$

Este es el valor del estadístico de chi-cuadrado (   ). Puedes compararlo con una tabla de valores críticos de chi-cuadrado o utilizar una función estadística para calcular el p-valor y determinar si hay una asociación significativa entre las variables "A favor" y "En contra" en "Rural" y "Urbano".

El valor p obtenido de la tabla de distribución de Chi-Cuadrado para 1 grado de libertad = **0.0007023**.

Un valor de χ^2 significativamente alto (valor p muy bajo) sugiere que rechazamos la hipótesis nula, indicando una asociación entre el tipo de zona y la intención de voto.

b) En base a los datos calcula las proporciones muestrales que estiman las siguientes cantidades: $P(\text{En contra} \mid \text{Rural})$ y $P(\text{En contra} \mid \text{Urbano})$. Llama a estas proporciones $\hat{p}_{\text{R}}$ y $\hat{p}_{\text{U}}$, respectivamente.

- $\hat{p}_{\text{R}} = P(\text{En contra} \mid \text{Rural})$ - Número de personas "En contra" en Rural: **50** - Total de personas en Rural: **100** - $\hat{p}_{\text{R}} = \frac{50}{100} = 0.5$
- $\hat{p}_{\text{U}} = P(\text{En contra} \mid \text{Urbano})$ - Número de personas "En contra" en Urbano: **140** - Total de personas en Urbano: **200** - $\hat{p}_{\text{U}} = \frac{140}{200} = 0.7$

c) Realiza un test de diferencia en proporciones para evaluar la siguiente hipótesis: $P(\text{En contra} \mid \text{Rural}) > P(\text{En contra} \mid \text{Urbano})$. Provee una interpretación sustantiva del resultado de este test.

Para evaluar si existe una diferencia significativa en las proporciones de personas que están "En contra" del plebiscito entre las zonas rurales y urbanas, realizamos el siguiente test estadístico:

- Hipótesis Nula (H_0): No hay diferencia en las proporciones de personas que están "En contra" del plebiscito entre las zonas rurales y urbanas, es decir, $P(\text{En contra} \mid \text{Rural}) = P(\text{En contra} \mid \text{Urbano})$.
- Hipótesis Alternativa (H_A): $P(\text{En contra} \mid \text{Rural}) > P(\text{En contra} \mid \text{Urbano})$
- Error Estándar:

La fórmula para el error estándar de la diferencia en proporciones es:

$$SE = \sqrt{\frac{p_R(1-p_R)}{n_R} + \frac{p_U(1-p_U)}{n_U}}$$

Al sustituir los valores, obtenemos:

$$SE = \sqrt{\frac{0.5 \times 0.5}{100} + \frac{0.7 \times 0.3}{200}} = 0.0707$$

- Z-score:
- Calculamos el Z-score para evaluar la diferencia entre las proporciones:
$$Z = \frac{p_R - p_U}{SE}$$
- Sustituyendo los valores, tenemos:
$$Z = \frac{0.5 - 0.7}{0.0707} = -2.827$$
- Valor p:
  - Considerando que la hipótesis alternativa es que $P(\text{En contra} \mid \text{Rural}) > P(\text{En contra} \mid \text{Urbano})$, el valor-p para un test de una cola se calcula como: $p = P(Z > -2.827)$
  - Encontramos que: $p = 0.9976507$

Con un valor p de 0.9976507, que es significativamente mayor que un nivel de significancia típico (como 0.05), no tenemos suficiente evidencia para rechazar la hipótesis nula. En otras palabras, basándonos en nuestros datos y el análisis estadístico realizado, podemos afirmar que si la hipótesis nula fuera verdaderas - es decir, si no existiera diferencia en las proporciones de personas "En contra" del plebiscito entre las zonas rurales y urbanas - obtener un valor en el test igual o mayor a -2.827 es un resultado altamente probable.

7. Un reconocido sociólogo ha desarrollado una teoría general para analizar la estabilidad social, representada por la ecuación:

$$y = \sqrt{\pi} + 10 * \mathtt{sin}(\mathtt{ln}\, x)$$

En esta fórmula:

- y representa el índice de estabilidad social, midiendo aspectos como cohesión y satisfacción en la sociedad.
- x simboliza el nivel de desigualdad socioeconómica, abarcando factores como ingresos y acceso a la educación.

La constante $\sqrt{\pi}$ refleja los elementos universales en todas las sociedades, mientras que $\sin(\ln x)$ muestra cómo pequeñas variaciones en la desigualdad pueden influir significativamente en la estabilidad social.

Esta teoría sugiere una relación no lineal entre desigualdad socioeconómica y estabilidad social, proporcionando un marco para comprender cómo la igualdad puede impactar positivamente en la dinámica social.

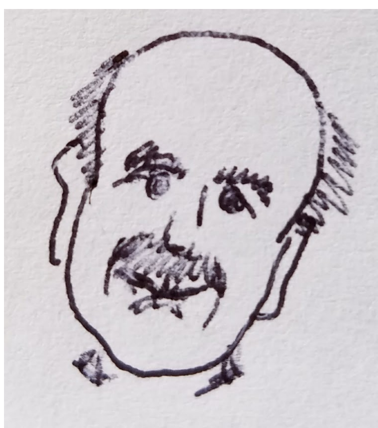
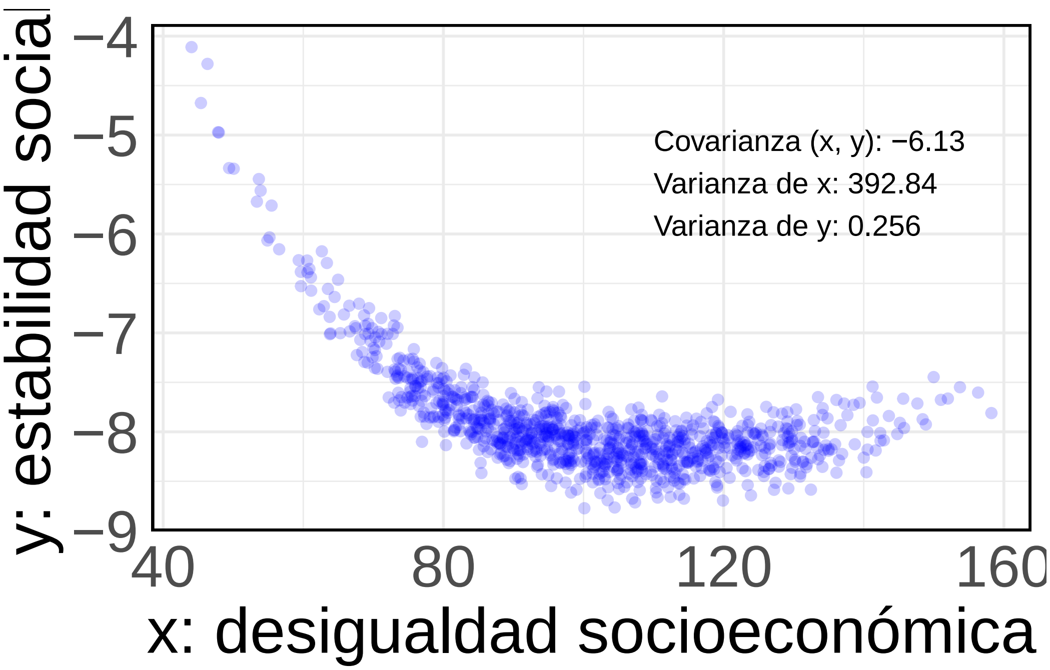


Figure 2: Theoretician

Para testear su teoría el sociólogo recolectó datos de 1000 sociedad distintas, obteniendo los resultados descritos en el siguiente gráfico:



Preguntas:

a) ¿Que relación observas entre desigualdad y estabilidad social?

El gráfico generado a partir de los datos recolectados muestra una relación negativa y no lineal entre la desigualdad socioeconómica (x) y la estabilidad social (y). Esto es consistente con la teoría, en la cual el término $10 * \sin(\ln x)$ indica que cambios en la desigualdad socioeconómica pueden tener efectos fluctuantes y no lineales en la estabilidad social.

b) Calcula el coeficiente de correlación de Pearson para medir la dirección y fortaleza de esta relación. Discute si el coeficiente de correlación de Pearson es o no una medida apropiada para capturar la relación entre ambas variables.

Para calcular el coeficiente de correlación de Pearson, utilizamos los valores de covarianza entre x e y y las varianzas de x e y proporcionados:

- Covarianza entre x e y ($\text{cov}(x,y)$): -6.13
- Varianza de x (σ_x^2): 392.84
- Varianza de y (σ_y^2): 0.256

El coeficiente de correlación de Pearson (r) se calcula como:

$$r = \frac{\mathbf{cov}(x, y)}{\sigma_x \sigma_y}$$

donde σ_x y σ_y son las desviaciones estándar de x e y, respectivamente. Calculamos estas desviaciones estándar tomando la raíz cuadrada de las varianzas:

- $\sigma_x = \sqrt{392.84}$
- $\sigma_y = \sqrt{0.256}$

Entonces, el coeficiente de correlación de Pearson es:

$$r = \frac{-6.13}{\sqrt{392.84} \times \sqrt{0.256}} \approx -0.61$$

El coeficiente de correlación de Pearson, basado en los datos proporcionados, es aproximadamente **-0.61**. Este valor indica una correlación negativa moderadamente fuerte entre la desigualdad socioeconómica y el índice de estabilidad social.

Dado que la relación teórica entre desigualdad socioeconómica y estabilidad social es no lineal, el coeficiente de correlación de Pearson puede no ser la medida más adecuada para capturar la totalidad de esta relación. Mientras que Pearson es útil para identificar y medir relaciones lineales, puede no capturar completamente las relaciones oscilantes o más complejas, como las sugeridas por la función seno en la teoría.

c) ¿Por qué es util calcular el coeficiente de correlación si ya sabemos que la covarianza en negativa? Comenta sobre el signo y magnitud de la correlación obtenida.

Mientras que la covarianza proporciona una medida del grado de relación lineal entre dos variables, su valor depende de las escalas de las variables y, por lo tanto, puede ser difícil de interpretar. El coeficiente de correlación de Pearson, por otro lado, estandariza esta relación en un rango entre -1 y 1, lo que facilita la interpretación y comparación con otras relaciones.

El coeficiente de correlación no solo confirma la dirección de la relación (positiva o negativa, como en la covarianza) sino que también indica la fuerza de la relación. Un valor cercano a -1 o 1 indica una relación fuerte, mientras que un valor cercano a 0 indica una relación débil. Por esta razón el coeficiente de correlación permite comparar la fuerza de las relaciones lineales entre diferentes pares de variables en estudios similares o diferentes, independientemente de sus escalas. Respecto al signo y magnitud de la correlación obtenida:

El signo negativo (-0.61) en nuestra correlación indica una relación inversa entre desigualdad socioeconómica y estabilidad social. Esto significa que, en general, un aumento en la desigualdad socioeconómica está asociado con una disminución en la estabilidad social. La magnitud de -0.61 sugiere una relación moderadamente fuerte. Esto implica que, aunque hay una asociación significativa entre las variables, hay otros factores que también pueden estar influyendo en la estabilidad social.