CLASE 4
INTRODUCCIÓN A LOS MODELOS MULTIVARIANTES

Gabriel Sotomayor

Abril 2024

OBJETIVO DE LA CLASE

Introducir el uso de modelos en ciencias sociales.

Recordar algunos conceptos estadísticos fundamentales.

¿POR QUÉ USAMOS MODELOS EN CIENCIAS SOCIALES?

¿POR QUÉ USAMOS MODELOS EN CIENCIAS SOCIALES?

Los modelos son formalismos lógicos o matemáticos que buscan describir la realidad.

Nos permiten simultáneamente capturar la complejidad de la realidad social (y los datos con que contamos) y reducirla, de manera de hacerla inteligible: **producir conocimiento (vincular nuestros datos a un contexto teórico mayor)**.

Permiten:

Formalizar, dando precisión y permitiendo poner teorías a prueba.

Develar relaciones entre variables y mecanismo causales

Predecir

Simular

¿POR QUÉ USAMOS MODELOS EN CIENCIAS SOCIALES?

Pensemos algunos ejemplos de relaciones sociales complejas que podemos comprender a partir de modelos.

MODELOS EXPLORATORIOS Y MODELOS CONFIRMATORIOS

- **Modelos Exploratorios**: son utilizados para identificar patrones, relaciones o agrupaciones entre variables sin hipótesis previas. Un ejemplo clásico es el análisis de componentes principales (PCA), que reduce la dimensionalidad de los datos para descubrir estructuras subyacentes. En este curso veremos análisis factorial confirmatorio (AFE).
- **Modelos Confirmatorios:** testean hipótesis específicas basadas en la teoría o investigaciones previas, como el análisis de sendero o modelos de ecuaciones estructurales. Estos modelos buscan confirmar la existencia de relaciones teóricamente postuladas entre variables.

COVARIANZA Y CORRELACIÓN

MEDIDAS DE DISPERSIÓN

Describen la variabilidad de los datos de una distribución.

Varianza: es el promedio de las distancias de los casos al promedio, tomando en cuenta los signos (eleva al cuadrado todas las distancias al promediarlas)

$$S^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Desviación estándar: Es la raíz cuadrada de la varianza. Es la que **mejor da cuenta de la dispersión** (es decir de las distancias de los casos al promedio)

-

$$s_x = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

COVARIANZA

La covarianza da cuenta del a variación conjunta de dos variables respecto de sus medias.

Puede tomar valores positivos, dando cuenta de una relación directa (por ejemplo a mayor educación, mayores ingresos) o valores negativos, dando cuenta de una relación inversa (por ejemplo a menores horas de trabajo, mayor satisfacción con la vida).

$$Cov(X,Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

CORRELACIÓN

La correlación (correlación de Pearson) corresponde a un valor estandarizado de la covarianza que puede tomar valores ente -1 y 1. La correlación se considera más descriptiva que la covarianza debido a su naturaleza estandarizada, lo que permite comparar la fuerza de la relación lineal entre diferentes pares de variables independientemente de sus unidades de medida.

$$r = \cos(\alpha) = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

CORRELACIONES POLICÓRICAS Y TETRACÓRICAS

Las correlaciones policóricas y tetracóricas son estadísticosque se utilizan específicamente para analizar la relación entre variables categóricas ordenadas y binarias, respectivamente. Estas medidas de correlación son especialmente relevantes en las ciencias sociales, donde a menudo las variables de interés no son continuas sino que se expresan en categorías con un orden inherente o son dicotómicas.

En ambos casos se asume que detrás de estas variables hay variables latentes continuas que están relacionadas linealmente.

Este enfoque es particularmente útil en cuestionarios o encuestas donde las respuestas se dan en escalas Likert (por ejemplo, de "muy en desacuerdo" a "muy de acuerdo").

INFERENCIA

¿PORQUÉ USAMOS MUESTRAS?

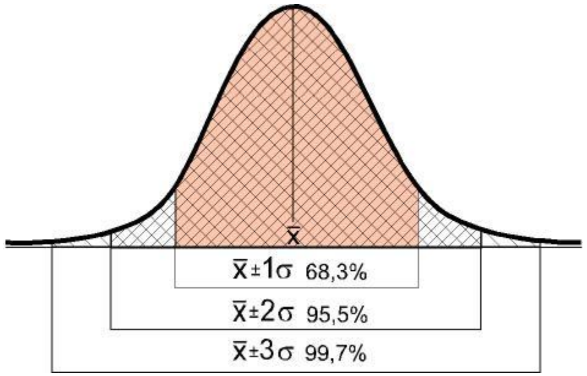
Usamos muestras porque nos permiten obtener datos **representativos** de una población a partir de una selección aleatoria de casos, una muestra. Este proceso se llama **estimación:** la estimación de un parámetro poblacional a partir de un estadístico muestral.

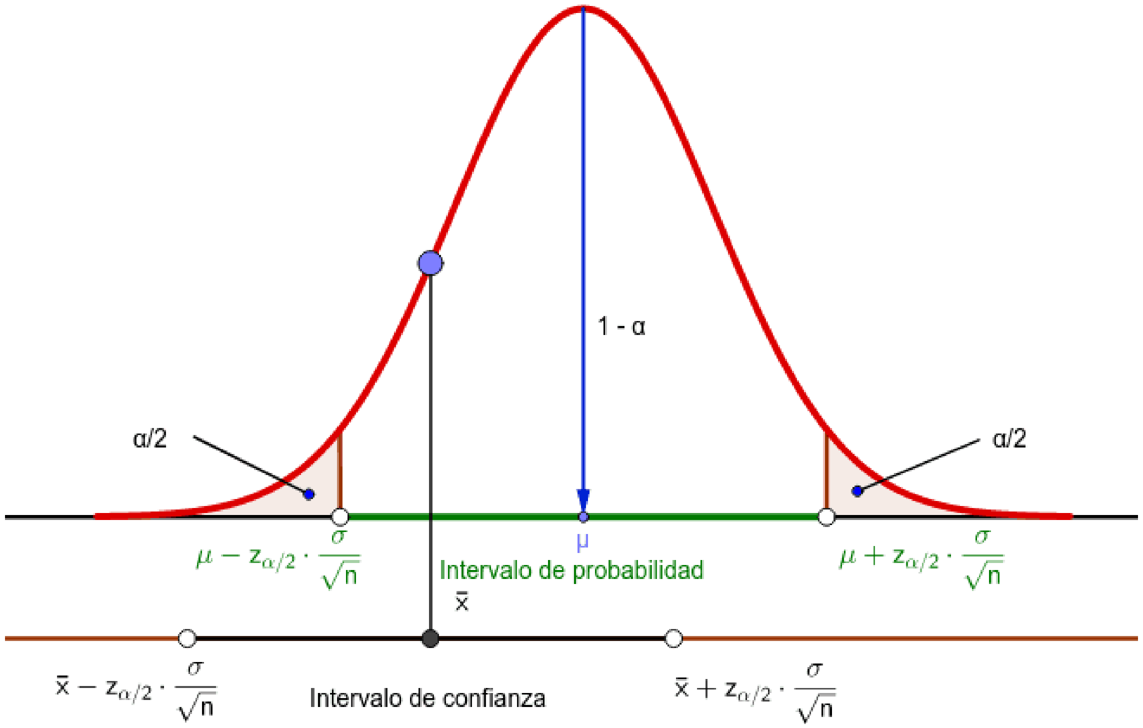
¿Porqué es esto posible?

Sabemos que la representatividad es posible, gracias a dos leyes estadísticas esenciales:

- **El teorema del límite central** -> La distribución de medias muestrales extraídas de forma aleatoria de una población, se aproxima a la distribución normal a medida que aumenta el tamaño de la muestra.
- **La ley de los grandes números** -> la diferencia entre el estadístico muestral y el parámetro poblacional tiende a 0 cuando el tamaño de la muestra tiende al infinito.

DISTRIBUCIONES MUESTRALES





A partir del Teorema del Límite central, vimos que la media de las distintas muestras que podemos extraer de una población se distribuye de forma normal, con muestras suficientemente grandes (<50).

$$N(\mu; \frac{\sigma}{\sqrt{n}})$$

INTERVALOS DE CONFIANZA

Estimación de punto: Puede ser útil cuando solo tenemos la media (o proporción) muestral y no contamos con más información, sin embargo no sabemos la precisión (el error) de nuestra medición.

Estimación de Intervalo: Podemos calcularlo a partir de las propiedades de la distribución muestral de las medias, usando los estadísticos muestrales como estimadores de los parámetros poblacionales para calcular el error estándar.

Medias

$$\boxed{IC = \bar{x} \pm Z_{1-\alpha/2} * \frac{s}{\sqrt{n}}}$$

Proporciones

$$\boxed{IC = \hat{p} \pm Z_{1-\alpha/2} * \sqrt{\frac{p(1-p)}{n}}}$$

Error estándar

PRUEBAS DE HIPÓTESIS

Una **hipótesis estadística** es una conjetura acerca de un parámetro de la población.

Contrastar o probamos la hipótesis formulada respecto de la población con la información que obtenemos a partir de una muestra.

–Una **prueba de hipótesis,** es un procedimiento estándar para probar una aseveración acerca de una propiedad de la población.

Se debe definir previamente la hipótesis nula H0 y la hipótesis alternativa Ha.

Procedimiento similar en todas las pruebas de hipótesis

– Se calcula el estadístico en la muestra.

– Se compara el estadístico con la distribución que se daría si la hipótesis nula fuese cierta.

– Si el estadístico tiene un valor que, dada la distribución de contraste, resulta muy improbable, rechazamos la hipótesis nula.

– Si el estadístico tiene un valor que, dada la distribución de contraste, es probable, no rechazamos la hipótesis nula.