

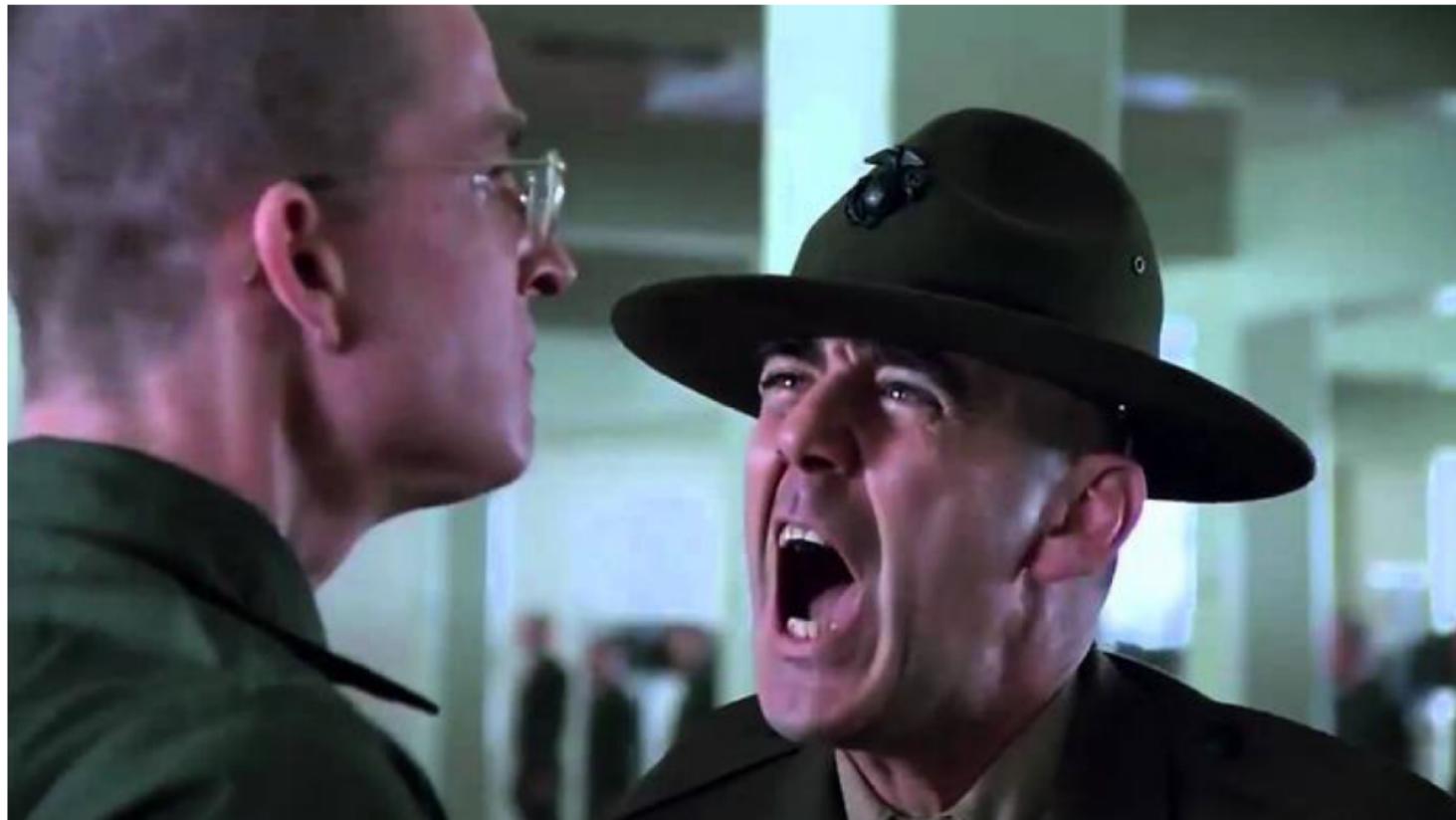
Crash Course: Bases de Estadística y Econometría

Paúl J. Corcuerá

Informática para Economistas
Universidad de Piura
Ciclo 2024-II

**Nota basada en material del curso de Jonathan Roth y Peter Hull en Brown University.*

¿Así que quieres saber de machine learning? Primero las bases



¿Qué es Econometría?

→ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas económicas usando data

Algunas preguntas que nos interesan:

¿Qué es Econometría?

→ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas económicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
- ¿Cuál va a ser la tasa de desempleo el otro año?

¿Qué es Econometría?

→ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas económicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
 - **Preg. Descriptiva:** pregunta cómo las cosas son (o fueron) en la realidad
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
- ¿Cuál va a ser la tasa de desempleo el otro año?

¿Qué es Econometría?

→ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas económicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
 - **Preg. Descriptiva:** pregunta cómo las cosas son (o fueron) en la realidad
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
 - **Preg. Causal:** ¿Qué hubiera pasado en un mundo contrafactual?
- ¿Cuál va a ser la tasa de desempleo el otro año?

¿Qué es Econometría?

→ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas económicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
 - **Preg. Descriptiva:** pregunta cómo las cosas son (o fueron) en la realidad
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
 - **Preg. Causal:** ¿Qué hubiera pasado en un mundo contrafactual?
- ¿Cuál va a ser la tasa de desempleo el otro año?
 - **Preg. Predicción:** ¿Qué pasará el otro año?

¿Qué es Econometría?

→ Es la caja de herramientas estadísticas que usamos los economistas para responder preguntas económicas usando data

Algunas preguntas que nos interesan:

- ¿Ha aumentado la desigualdad económica desde los años 60?
 - **Preg. Descriptiva:** pregunta cómo las cosas son (o fueron) en la realidad
- ¿Cómo afecta el aumento de salario mínimo en el empleo?
 - **Preg. Causal:** ¿Qué hubiera pasado en un mundo contrafactual?
- ¿Cuál va a ser la tasa de desempleo el otro año?
 - **Preg. Predicción:** ¿Qué pasará el otro año?

Por lo general, los economistas nos concentraremos en las primeras dos, con un énfasis en preguntas causales

¿Por qué es difícil responder estas preguntas?

- Para preg. descriptivas: solo observamos una **muestra** de individuos, no la **población** completa
 - Ejemplo: queremos saber la proporción de empleo informal pero solo observamos empleo e informalidad de la Encuesta Nacional de Hogares

¿Por qué es difícil responder estas preguntas?

- Para preg. descriptivas: solo observamos una **muestra** de individuos, no la **población** completa
 - Ejemplo: queremos saber la proporción de empleo informal pero solo observamos empleo e informalidad de la Encuesta Nacional de Hogares
- El mejor escenario:
Nuestra muestra es **aleatoriamente** seleccionada de la población
 - Por ej., los nombres de los trabajadores que vemos en la encuesta fueron sacados de un sombrero al azar donde estaban todos los posibles trabajadores
 - Si este es el caso, debemos tomar en cuenta que por simples chances podemos tener una muestra con diferentes características que la población

¿Por qué es difícil responder estas preguntas?

- Para preg. descriptivas: solo observamos una **muestra** de individuos, no la **población** completa
 - Ejemplo: queremos saber la proporción de empleo informal pero solo observamos empleo e informalidad de la Encuesta Nacional de Hogares
- El mejor escenario:
Nuestra muestra es **aleatoriamente** seleccionada de la población
 - Por ej., los nombres de los trabajadores que vemos en la encuesta fueron sacados de un sombrero al azar donde estaban todos los posibles trabajadores
 - Si este es el caso, debemos tomar en cuenta que por simples chances podemos tener una muestra con diferentes características que la población
- El peor escenario: nuestra muestra es *no representativa* de la población que nos interesa
 - Por ej., los trabajadores que son formales tenían muchas mayor probabilidad de responder la encuesta

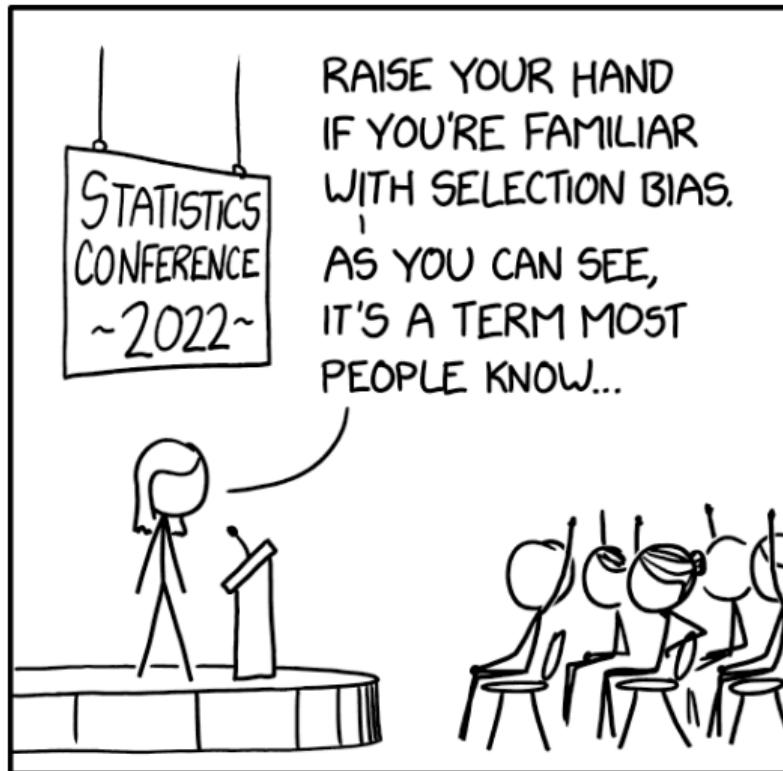


- En 1948, el Chicago Tribune escribió que Thomas Dewey ganaba a Harry Truman en la elección presidencial, basada en una encuesta a votantes



- En 1948, el Chicago Tribune escribió que Thomas Dewey ganaba a Harry Truman en la elección presidencial, basada en una encuesta a votantes
- Pero su encuesta fue por teléfono. En ese año solo la gente rica tenía teléfonos: muestra ≠ población → ¡resultados engañosos!

Sesgo de selección se refiere a contextos como el que mencionamos anteriormente, donde la muestra no es sacada aleatoriamente desde la población de interés



¿Por qué responder estas preguntas es difícil? (Parte II)

- Responder preguntas causales es incluso *más difícil* que las descriptivas. ¿Por qué?

¿Por qué responder estas preguntas es difícil? (Parte II)

- Responder preguntas causales es incluso *más difícil* que las descriptivas. ¿Por qué?
- Las preg. causales envuelven tanto un componente descriptivo (*¿cómo es el outcome en la realidad?*) y un componente **contrafactual** (*¿cómo hubieran sido las cosas bajo un tratamiento diferente?*)

¿Por qué responder estas preguntas es difícil? (Parte II)

- Responder preguntas causales es incluso *más difícil* que las descriptivas. ¿Por qué?
- Las preg. causales envuelven tanto un componente descriptivo (*¿cómo es el outcome en la realidad?*) y un componente **contrafactual** (*¿cómo hubieran sido las cosas bajo un tratamiento diferente?*)
- Ejemplo: *¿Cuál es el efecto causal en sus salarios de ir a UdeP en vez de la U. Pacífico?*
 - Preg. Descriptiva: *¿Cuánto ganan los alumnos de UdeP después de graduarse?*
 - Preg. Contrafactual: *¿Cuánto *hubieran* ganado los alumnos de UdeP después de graduarse si *hubieran estudiado en la UP*?*

¿Por qué responder estas preguntas es difícil? (Parte II)

- Responder preguntas causales es incluso *más difícil* que las descriptivas. ¿Por qué?
- Las preg. causales envuelven tanto un componente descriptivo (*¿cómo es el outcome en la realidad?*) y un componente **contrafactual** (*¿cómo hubieran sido las cosas bajo un tratamiento diferente?*)
- Ejemplo: *¿Cuál es el efecto causal en sus salarios de ir a UdeP en vez de la U. Pacífico?*
 - Preg. Descriptiva: *¿Cuánto ganan los alumnos de UdeP después de graduarse?*
 - Preg. Contrafactual: *¿Cuánto *hubieran* ganado los alumnos de UdeP después de graduarse si *hubieran estudiado en la UP*?*
- Preg. Contrafactuales no se pueden contestar con data solamente. ¡Necesitamos supuestos para aprender de ellos!

Partir el problema

- Al pensar en preg. causales, suele ser mejor partir el problema en dos
- **Identificación:** ¿qué podemos aprender de los parámetros que nos importan (efectos causales) si tuvieramos *data observable* de toda la población
 - Necesitamos hacer supuestos sobre cómo los outcomes observados se relacionan con los outcomes que hubieran sucedido bajo distinto tratamiento
- **Estadística:** ¿qué podemos aprender sobre la población desde la muestra finita que tenemos en la práctica?
 - Necesitamos entender el proceso de cómo se generó la data para la población completa

Mapa conceptual para entender estos pasos

- **Muestra:** la data observable
 - Encuesta a alumnos graduados de UdeP y UP sobre sus salarios

Mapa conceptual para entender estos pasos

- **Muestra:** la data observable
 - Encuesta a alumnos graduados de UdeP y UP sobre sus salarios
- **Estimador:** una función de la data en la muestra
 - Diferencia en salario promedio entre UdeP y UP en la encuesta

Mapa conceptual para entender estos pasos

- **Muestra:** la data observable
 - Encuesta a alumnos graduados de UdeP y UP sobre sus salarios
- **Estimador:** una función de la data en la muestra
 - Diferencia en salario promedio entre UdeP y UP en la encuesta
- **Estimando:** función de la data de la *población*
 - Diferencia en la *media* salarial entre UdeP y UP

Mapa conceptual para entender estos pasos

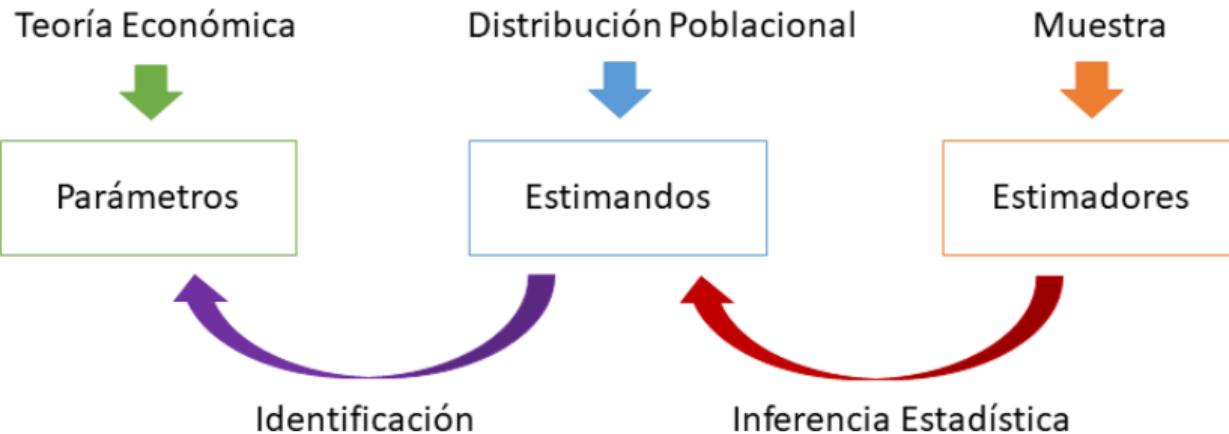
- **Muestra:** la data observable
 - Encuesta a alumnos graduados de UdeP y UP sobre sus salarios
- **Estimador:** una función de la data en la muestra
 - Diferencia en salario promedio entre UdeP y UP en la encuesta
- **Estimando:** función de la data de la *población*
 - Diferencia en la *media* salarial entre UdeP y UP
- **Parámetro (estructural) objetivo:** lo que realmente nos importa conocer
 - Efecto causal de ir a UdeP relativo a UP en salarios

Mapa conceptual para entender estos pasos

- **Muestra:** la data observable
 - Encuesta a alumnos graduados de UdeP y UP sobre sus salarios
- **Estimador:** una función de la data en la muestra
 - Diferencia en salario promedio entre UdeP y UP en la encuesta
- **Estimando:** función de la data de la *población*
 - Diferencia en la *media* salarial entre UdeP y UP
- **Parámetro (estructural) objetivo:** lo que realmente nos importa conocer
 - Efecto causal de ir a UdeP relativo a UP en salarios
- El proceso para ir al *estimando* desde el estimador construido con tu *muestra* se llama **estimación/inferencia estadística**.

Mapa conceptual para entender estos pasos

- **Muestra:** la data observable
 - Encuesta a alumnos graduados de UdeP y UP sobre sus salarios
- **Estimador:** una función de la data en la muestra
 - Diferencia en salario promedio entre UdeP y UP en la encuesta
- **Estimando:** función de la data de la *población*
 - Diferencia en la *media* salarial entre UdeP y UP
- **Parámetro (estructural) objetivo:** lo que realmente nos importa conocer
 - Efecto causal de ir a UdeP relativo a UP en salarios
- El proceso para ir al *estimando* desde el estimador construido con tu *muestra* se llama **estimación/inferencia estadística**.
- El proceso para aprender del *parámetro* desde el *estimando* es llamado **identificación**.



$$\theta = E[X_i]/E[Y_i]$$

$$E[X_i]$$

$$E[Y_i]$$

$$\sum X_i/n$$

$$\sum Y_i/n$$

Agreguemosle un poquito de mate...

- Introduciremos la notación de **potential outcomes**
 - ¡Conceptualmente es muy útil para pensar acerca de causalidad!

Agreguemosle un poquito de mate...

- Introduciremos la notación de **potential outcomes**
 - ¡Conceptualmente es muy útil para pensar acerca de causalidad!
- D_i = indicator si recibo el tratamiento (1 si UdeP, 0 si UP)

Agreguemosle un poquito de mate...

- Introduciremos la notación de **potential outcomes**
 - ¡Conceptualmente es muy útil para pensar acerca de causalidad!
- D_i = indicator si recibo el tratamiento (1 si UdeP, 0 si UP)
- $Y_i(1)$ = outcome si tratado = salarios de UdeP
- $Y_i(0)$ = outcome si control = salarios de UP

Agreguemosle un poquito de mate...

- Introduciremos la notación de **potential outcomes**
 - ¡Conceptualmente es muy útil para pensar acerca de causalidad!
- D_i = indicator si recibo el tratamiento (1 si UdeP, 0 si UP)
- $Y_i(1)$ = outcome si tratado = salarios de UdeP
- $Y_i(0)$ = outcome si control = salarios de UP
- El outcome observado Y_i es $Y_i(1)$ si $D_i = 1$ y $Y_i(0)$ si $D_i = 0$. (Y_i es tu salario *actual*)

Agreguemosle un poquito de mate...

- Introduciremos la notación de **potential outcomes**
 - ¡Conceptualmente es muy útil para pensar acerca de causalidad!
- D_i = indicator si recibo el tratamiento (1 si UdeP, 0 si UP)
- $Y_i(1)$ = outcome si tratado = salarios de UdeP
- $Y_i(0)$ = outcome si control = salarios de UP
- El outcome observado Y_i es $Y_i(1)$ si $D_i = 1$ y $Y_i(0)$ si $D_i = 0$. (Y_i es tu salario *actual*)
- Podemos reescribir el outcome observado como $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

- Ejemplo de muestra: (Y_i, D_i) para $i = 1, \dots, N$. Data de salarios y donde fuiste a la universidad

- Ejemplo de muestra: (Y_i, D_i) para $i = 1, \dots, N$. Data de salarios y donde fuiste a la universidad
- Ejemplo de estimador:
 - Diferencia en la media muestral de salarios de gente que fue a UdeP, y de gente que fue a UP

$$\underbrace{\frac{1}{N_1} \sum_{i:D_i=1} Y_i}_{\text{Salario prom de UdeP}} - \underbrace{\frac{1}{N_0} \sum_{i:D_i=0} Y_i}_{\text{Salario prom de UP}}$$

- Ejemplo de muestra: (Y_i, D_i) para $i = 1, \dots, N$. Data de salarios y donde fuiste a la universidad
- Ejemplo de estimador:
 - Diferencia en la media muestral de salarios de gente que fue a UdeP, y de gente que fue a UP

$$\underbrace{\frac{1}{N_1} \sum_{i:D_i=1} Y_i}_{\text{Salario prom de UdeP}} - \underbrace{\frac{1}{N_0} \sum_{i:D_i=0} Y_i}_{\text{Salario prom de UP}}$$

- Ejemplo estimando:
 - Diferencia en media poblacional salarial entre gente que fue UdeP y UP

$$\underbrace{E[Y_i | D_i = 1]}_{\text{Media salarial de UdeP}} - \underbrace{E[Y_i | D_i = 0]}_{\text{Media salarial de UP}}$$

- Ejemplo de muestra: (Y_i, D_i) para $i = 1, \dots, N$. Data de salarios y donde fuiste a la universidad
- Ejemplo de estimador:
 - Diferencia en la media muestral de salarios de gente que fue a UdeP, y de gente que fue a UP

$$\underbrace{\frac{1}{N_1} \sum_{i:D_i=1} Y_i}_{\text{Salario prom de UdeP}} - \underbrace{\frac{1}{N_0} \sum_{i:D_i=0} Y_i}_{\text{Salario prom de UP}}$$

- Ejemplo estimando:
 - Diferencia en media poblacional salarial entre gente que fue UdeP y UP

$$\underbrace{E[Y_i|D_i = 1]}_{\text{Media salarial de UdeP}} - \underbrace{E[Y_i|D_i = 0]}_{\text{Media salarial de UP}}$$

- Ejemplo parámetro objetivo:
 - Efecto causal de ir a UdeP en salarios:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Salarios de UdeP para gente de UdeP en la pob.}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para gente de UdeP en la pob.}}.$$

¿Por qué es difícil la identificación causal?

- Experimento mental: imagina que tenemos data de salarios para *todos* los graduados de UdeP y UP
- Podemos aprender esto de la data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Salarios de UdeP para alumnos UdeP}}$$

and

$$\underbrace{E[Y_i(0)|D_i = 0]}_{\text{Salarios de UP para alumnos UP}}$$

¿Por qué es difícil la identificación causal?

- Experimento mental: imagina que tenemos data de salarios para *todos* los graduados de UdeP y UP
- Podemos aprender esto de la data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Salarios de UdeP para alumnos UdeP}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Salarios de UP para alumnos UP}}$$

- El efecto causal de ir a UdeP dentro de los graduados UdeP es

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Salario de UdeP para alumnos UdeP}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para alumnos UdeP}}$$

¿Por qué es difícil la identificación causal?

- Experimento mental: imagina que tenemos data de salarios para *todos* los graduados de UdeP y UP
- Podemos aprender esto de la data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Salarios de UdeP para alumnos UdeP}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Salarios de UP para alumnos UP}}$$

- El efecto causal de ir a UdeP dentro de los graduados UdeP es

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Salario de UdeP para alumnos UdeP}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para alumnos UdeP}}$$

- La data no me dice $\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para alumnos UdeP}}$. ¿Por qué no?

¿Por qué es difícil la identificación causal?

- Experimento mental: imagina que tenemos data de salarios para *todos* los graduados de UdeP y UP
- Podemos aprender esto de la data:

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Salarios de UdeP para alumnos UdeP}} \quad \text{and} \quad \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Salarios de UP para alumnos UP}}$$

- El efecto causal de ir a UdeP dentro de los graduados UdeP es

$$\underbrace{E[Y_i(1)|D_i = 1]}_{\text{Salario de UdeP para alumnos UdeP}} - \underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para alumnos UdeP}}$$

- La data no me dice $\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para alumnos UdeP}}$. ¿Por qué no?

- ¡Porque jamás podremos ver la vida de los alumnos de UdeP si hubieran hipotéticamente ido a UP!

- Una idea para arreglar este problema es asumir que:

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para alumnos UdeP}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Salarios de UP para alumnos UP}}$$

- ¿Por qué esto puede darnos la respuesta incorrecta?

- Una idea para arreglar este problema es asumir que:

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para alumnos UdeP}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Salarios de UP para alumnos UP}}$$

- ¿Por qué esto puede darnos la respuesta incorrecta?
- Porque los alumnos UdeP pueden ser diferentes de los alumnos UP en muchas maneras que afectarian sus salarios (irrespectivamente de a cual de las dos universidades fueron)
 - Habilidad académica, entorno e ingresos familiares, metas laborales, etc.
- Estas diferencias se les llaman *variables omitidas* o, en inglés, *confounding factors*

Experimentos

- El estándar de oro para aprender sobre causalidad son los llamados *randomized controlled trial (RCT)*, aka experimentos
- Imaginen que podemos aleatorizar quienes van a que universidad (asumamos que solo existen UdeP y UP por simplicidad)
- Como la universidad es aleatoriamente asignada, lo único distinto entre la gente de UdeP y UP es la universidad a la que atendieron
- Por tanto,

$$\underbrace{E[Y_i(0)|D_i = 1]}_{\text{Salarios de UP para alumnos UdeP}} = \underbrace{E[Y_i(0)|D_i = 0]}_{\text{Salarios de UP para alumnos UP}}$$

dado que eliminamos los *confounding factors*

Pero correr experimentos suele ser difícil/imposible

- Desgraciadamente, por temas de ética no podemos aleatorizar quién va a qué universidad
- De la misma manera, no puedes convencer al gobierno de que aleatorice donde subir el salario mínimo, u otras políticas
- En pocas palabras, aleatorizar no solo es difícil sino que puede ser inmoral dado que hablamos de sujetos humanos
 - “¿Cuál es el efecto causal de que fallezca algún familiar?”

Pero correr experimentos suele ser difícil/imposible

- Desgraciadamente, por temas de ética no podemos aleatorizar quién va a qué universidad
- De la misma manera, no puedes convencer al gobierno de que aleatorice donde subir el salario mínimo, u otras políticas
- En pocas palabras, aleatorizar no solo es difícil sino que puede ser inmoral dado que hablamos de sujetos humanos
 - “¿Cuál es el efecto causal de que fallezca algún familiar?”
- Un gran componente de su rama de econometría trata sobre herramientas que podemos usar los economistas cuando los experimentos no son posibles

Hacia esto vamos...

- **Revisión de probabilidad/estadística.** Para tener un mismo lenguaje matemático para poder discutir
 - ① *Estimación/Inferencia Estadística*: como la muestra se relaciona con la población de interés
 - ② *Identification*: como las características observables de la población se relacionan a parámetros (causales) que nos importan

Hacia esto vamos...

- **Revisión de probabilidad/estadística.** Para tener un mismo lenguaje matemático para poder discutir
 - ① *Estimación/Inferencia Estadística*: como la muestra se relaciona con la población de interés
 - ② *Identification*: como las características observables de la población se relacionan a parámetros (causales) que nos importan
- **Regresión lineal y predicción:** Discutiremos el método de mínimos cuadrados (en inglés, Ordinary Least Squares, OLS), la forma de estimación más utilizada por economistas.
¿Cuándo funciona y cuándo falla?

Hacia esto vamos...

- **Revisión de probabilidad/estadística.** Para tener un mismo lenguaje matemático para poder discutir
 - ① *Estimación/Inferencia Estadística*: como la muestra se relaciona con la población de interés
 - ② *Identification*: como las características observables de la población se relacionan a parámetros (causales) que nos importan
- **Regresión lineal y predicción:** Discutiremos el método de mínimos cuadrados (en inglés, Ordinary Least Squares, OLS), la forma de estimación más utilizada por economistas.
¿Cuándo funciona y cuándo falla?
- **Regularización:** Discutiremos brevemente el método LASSO para lidiar con el caso en que hacemos una predicción utilizando muchísimas características (*alta dimensionalidad*)

Outline

1. Media Condicional
2. Muestreo Aleatorio
3. Test de Hipótesis e Inferencia
4. Teoría Asintótica

Variable aleatoria y CDF

- Una **variable aleatoria** es un resumen numérico de un proceso aleatorio (formalmente, es una función de los elementos del espacio muestral)
 - Por ej. X es el número de “caras” que vemos al tirar una moneda 10 veces; $X = 1$ si llueve, 0 de lo contrario.

Variable aleatoria y CDF

- Una **variable aleatoria** es un resumen numérico de un proceso aleatorio (formalmente, es una función de los elementos del espacio muestral)
 - Por ej. X es el número de “caras” que vemos al tirar una moneda 10 veces; $X = 1$ si llueve, 0 de lo contrario.
- Una variable aleatoria real se caracteriza por su **distribución de función acumulada** (CDF),

$$F(x) = \Pr(X \leq x)$$

que dice la probabilidad de que X tome el valor x o algun valor menor

Variable aleatoria y CDF

- Una **variable aleatoria** es un resumen numérico de un proceso aleatorio (formalmente, es una función de los elementos del espacio muestral)
 - Por ej. X es el número de “caras” que vemos al tirar una moneda 10 veces; $X = 1$ si llueve, 0 de lo contrario.
- Una variable aleatoria real se caracteriza por su **distribución de función acumulada** (CDF),

$$F(x) = \Pr(X \leq x)$$

que dice la probabilidad de que X tome el valor x o algun valor menor

- Nota: minusculas “ x ” suele denotar realizaciones (es decir, números no aleatorios) de variables aleatorias como “ X ” ...

Esperanza Condicional

En economía nos interesa conocer la **esperanza condicional**

- ¿Cuál es el promedio Y cuando $X = x$ (por ej. cual es mi salario si mi universidad fue UdeP)?

Esperanza Condicional

En economía nos interesa conocer la **esperanza condicional**

- ¿Cuál es el promedio Y cuando $X = x$ (por ej. cual es mi salario si mi universidad fue UdeP)?
- La función de esperanza condicional (CEF):

$$E[Y | X = x] = \sum_{y \in \mathbb{Y}} y p(y | x) \text{ si es discreta}$$

$$E[Y | X = x] = \int_{y \in \mathbb{Y}} y f(y | x) dy \text{ si es continua}$$

Esperanza Condicional

En economía nos interesa conocer la **esperanza condicional**

- ¿Cuál es el promedio Y cuando $X = x$ (por ej. cual es mi salario si mi universidad fue UdeP)?

- La función de esperanza condicional (CEF):

$$E[Y | X = x] = \sum_{y \in \mathbb{Y}} y p(y | x) \text{ si es discreta}$$

$$E[Y | X = x] = \int_{y \in \mathbb{Y}} y f(y | x) dy \text{ si es continua}$$

- A veces se escribe $E[Y | X]$ para el CEF aleatorio evaluado en X (var. aleatoria)

Esperanza Condicional

En economía nos interesa conocer la **esperanza condicional**

- ¿Cuál es el promedio Y cuando $X = x$ (por ej. cual es mi salario si mi universidad fue UdeP)?
- La función de esperanza condicional (CEF):
$$E[Y | X = x] = \sum_{y \in \mathbb{Y}} y p(y | x) \text{ si es discreta}$$
$$E[Y | X = x] = \int_{y \in \mathbb{Y}} y f(y | x) dy \text{ si es continua}$$
- A veces se escribe $E[Y | X]$ para el CEF aleatorio evaluado en X (var. aleatoria)
- Decimos que Y es *independiente en media* de X cuando $E[Y | X = x] = E[Y]$ para todo x

Esperanza Condicional

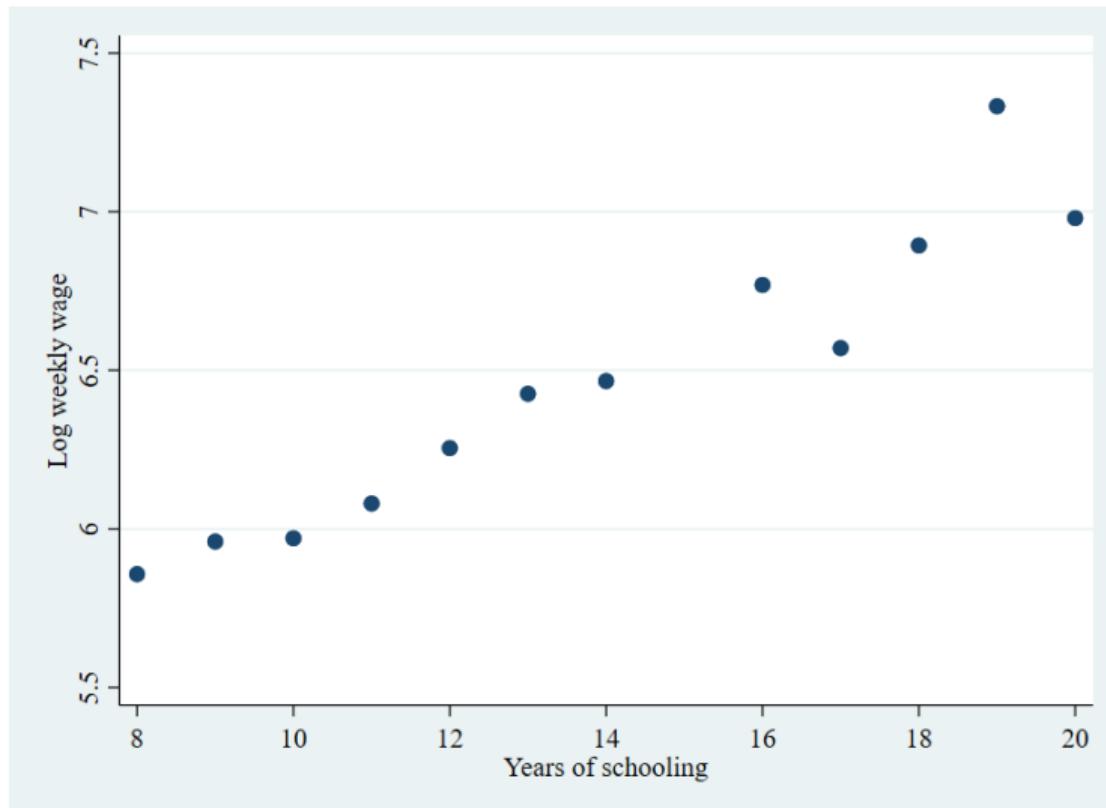
En economía nos interesa conocer la **esperanza condicional**

- ¿Cuál es el promedio Y cuando $X = x$ (por ej. cual es mi salario si mi universidad fue UdeP)?
- La función de esperanza condicional (CEF):
 $E[Y | X = x] = \sum_{y \in \mathbb{Y}} y p(y | x)$ si es discreta
 $E[Y | X = x] = \int_{y \in \mathbb{Y}} y f(y | x) dy$ si es continua
- A veces se escribe $E[Y | X]$ para el CEF aleatorio evaluado en X (var. aleatoria)
- Decimos que Y es *independiente en media* de X cuando $E[Y | X = x] = E[Y]$ para todo x
- Condicionar en X hace que sea una función constante: por ej.

$$E[f(X) + g(X)|X = x] = f(x) + g(x)E[Y|X = x] \text{ para cualquier } f(\cdot), g(\cdot)$$

Esperanza Condicional

CEF del (logaritmo) del ingreso anual dado años de educación



Ley de Esperanzas Iteradas (LIE)

Uno de los resultados más importantes que usamos es esta ley llamada en inglés **Law of Iterated Expectations** (LIE)

Veamos un ejemplo. Supongamos que queremos calcular la altura promedio de personas en Perú. Entonces la LIE dice que podemos:

- 1) Calcular el promedio de altura en hombres.
- 2) Calcular el promedio de altura en mujeres.
- 3) Calcular un promedio ponderado entre la altura hombres y mujeres (proporcional a la fracción de personas que son mujeres)

Ley de Esperanzas Iteradas (LIE)

Uno de los resultados más importantes que usamos es esta ley llamada en inglés **Law of Iterated Expectations** (LIE)

Veamos un ejemplo. Supongamos que queremos calcular la altura promedio de personas en Perú. Entonces la LIE dice que podemos:

- 1) Calcular el promedio de altura en hombres.
- 2) Calcular el promedio de altura en mujeres.
- 3) Calcular un promedio ponderado entre la altura hombres y mujeres (proporcional a la fracción de personas que son mujeres)

Matemáticamente estamos diciendo lo siguiente

$$E[\text{altura}] = P(\text{mujer})E[\text{altura}|\text{mujer}] + P(\text{hombre})E[\text{altura}|\text{hombre}]$$

Ley de Esperanzas Iteradas (LIE)

Uno de los resultados más importantes que usamos es esta ley llamada en inglés **Law of Iterated Expectations** (LIE)

Veamos un ejemplo. Supongamos que queremos calcular la altura promedio de personas en Perú. Entonces la LIE dice que podemos:

- 1) Calcular el promedio de altura en hombres.
- 2) Calcular el promedio de altura en mujeres.
- 3) Calcular un promedio ponderado entre la altura hombres y mujeres (proporcional a la fracción de personas que son mujeres)

Matemáticamente estamos diciendo lo siguiente

$$\begin{aligned} E[\text{altura}] &= P(\text{mujer})E[\text{altura}|\text{mujer}] + P(\text{hombre})E[\text{altura}|\text{hombre}] \\ &= E[E[\text{altura}|\text{sexo}]] \end{aligned}$$

Ley de Esperanzas Iteradas (LIE)

La versión formal de la **Ley de Esperanzas Iteradas** (LIE) es:

$$E[Y] = E[E[Y | X]]$$

Ley de Esperanzas Iteradas (LIE)

La versión formal de la **Ley de Esperanzas Iteradas** (LIE) es:

$$E[Y] = E[E[Y | X]]$$

Note que la esperanza de la izquierda usa la distribución de Y ($p(y)$); la esperanza de la parte fuera de la derecha usa $p(x)$, mientras que la esperanza interna usa la distribución condicional $p(y | x)$

Independencia en Media vs No Correlacionado

LIE muestra que independencia en media ($E[Y | X] = E[Y]$) implica que las variables no están correlacionadas

$$\text{Corr}(X, Y) \propto E[(X - E[X])(Y - E[Y])]$$

Independencia en Media vs No Correlacionado

LIE muestra que independencia en media ($E[Y | X] = E[Y]$) implica que las variables no están correlacionadas

$$\begin{aligned} \text{Corr}(X, Y) &\propto E[(X - E[X])(Y - E[Y])] \\ &= E[E[(X - E[X])(Y - E[Y]) | X]] \end{aligned}$$

Independencia en Media vs No Correlacionado

LIE muestra que independencia en media ($E[Y | X] = E[Y]$) implica que las variables no están correlacionadas

$$\begin{aligned} \text{Corr}(X, Y) &\propto E[(X - E[X])(Y - E[Y])] \\ &= E[E[(X - E[X])(Y - E[Y]) | X]] \\ &= E[(X - E[X])E[Y - E[Y] | X]] \end{aligned}$$

Independencia en Media vs No Correlacionado

LIE muestra que independencia en media ($E[Y | X] = E[Y]$) implica que las variables no están correlacionadas

$$\begin{aligned} \text{Corr}(X, Y) &\propto E[(X - E[X])(Y - E[Y])] \\ &= E[E[(X - E[X])(Y - E[Y]) | X]] \\ &= E[(X - E[X])E[Y - E[Y] | X]] \\ &= E[(X - E[X])(E[Y | X] - E[Y])] \end{aligned}$$

Independencia en Media vs No Correlacionado

LIE muestra que independencia en media ($E[Y | X] = E[Y]$) implica que las variables no están correlacionadas

$$\begin{aligned} \text{Corr}(X, Y) &\propto E[(X - E[X])(Y - E[Y])] \\ &= E[E[(X - E[X])(Y - E[Y]) | X]] \\ &= E[(X - E[X])E[Y - E[Y] | X]] \\ &= E[(X - E[X])(E[Y | X] - E[Y])] \\ &= 0, \text{ when } E[Y | X] = E[Y] \end{aligned}$$

¡Asegúrense de entender cada paso!

El converso de esta proposición no es cierto: estar no correlacionado no implica que sean independientes en media

- Por supuesto, independencia \implies independencia en media (pero no a la otra dirección \Leftarrow)

Outline

1. Media Condicional ✓
2. Muestreo Aleatorio
3. Test de Hipótesis e Inferencia
4. Teoría Asintótica

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población
- Baseline: observamos una muestra *representativa* de tamaño N que es *independiente e idénticamente distribuida (iid)* por ej. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población
- Baseline: observamos una muestra *representativa* de tamaño N que es *independiente e idénticamente distribuida (iid)* por ej. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$
 - *Independiente*: Y_i es independiente de Y_j para todo $i \neq j$

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población
- Baseline: observamos una muestra *representativa* de tamaño N que es *independiente e idénticamente distribuida (iid)* por ej. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$
 - *Independiente*: Y_i es independiente de Y_j para todo $i \neq j$
 - *Idénticamente Distribuido*: Y_i y Y_j tienen la misma distribución para todo i, j

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población
- Baseline: observamos una muestra *representativa* de tamaño N que es *independiente e idénticamente distribuida (iid)* por ej. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$
 - *Independiente*: Y_i es independiente de Y_j para todo $i \neq j$
 - *Idénticamente Distribuido*: Y_i y Y_j tienen la misma distribución para todo i, j
 - *Representativo*: La distribución de Y_i es la misma que la distribución de la población de interés

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población
- Baseline: observamos una muestra *representativa* de tamaño N que es *independiente e idénticamente distribuida (iid)* por ej. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$
 - *Independiente*: Y_i es independiente de Y_j para todo $i \neq j$
 - *Idénticamente Distribuido*: Y_i y Y_j tienen la misma distribución para todo i, j
 - *Representativo*: La distribución de Y_i es la misma que la distribución de la población de interés
- Data *iid* y representativa es un caso base que es relativamente fácil de analizar, pero es importante saber que no necesariamente se cumple en la práctica

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población
- Baseline: observamos una muestra *representativa* de tamaño N que es *independiente e idénticamente distribuida (iid)* por ej. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$
 - *Independiente*: Y_i es independiente de Y_j para todo $i \neq j$
 - *Idénticamente Distribuido*: Y_i y Y_j tienen la misma distribución para todo i, j
 - *Representativo*: La distribución de Y_i es la misma que la distribución de la población de interés
- Data *iid* y representativa es un caso base que es relativamente fácil de analizar, pero es importante saber que no necesariamente se cumple en la práctica
 - si muestreamos gente dentro de un mismo hogar, esas observaciones no son independientes

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población
- Baseline: observamos una muestra *representativa* de tamaño N que es *independiente e idénticamente distribuida (iid)* por ej. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$
 - *Independiente*: Y_i es independiente de Y_j para todo $i \neq j$
 - *Idénticamente Distribuido*: Y_i y Y_j tienen la misma distribución para todo i, j
 - *Representativo*: La distribución de Y_i es la misma que la distribución de la población de interés
- Data *iid* y representativa es un caso base que es relativamente fácil de analizar, pero es importante saber que no necesariamente se cumple en la práctica
 - si muestreamos gente dentro de un mismo hogar, esas observaciones no son independientes
 - si muestreamos de manera estratificada dentro de cada departamento, no es identicamente distribuido

Defiendo una muestra

- Para formalizar el trabajo de la inferencia estadística debemos especificar cómo la data observada se saca de la población
- Baseline: observamos una muestra *representativa* de tamaño N que es *independiente e idénticamente distribuida (iid)* por ej. $\mathbf{Y} = [Y_1, Y_2, \dots, Y_N]'$
 - *Independiente*: Y_i es independiente de Y_j para todo $i \neq j$
 - *Idénticamente Distribuido*: Y_i y Y_j tienen la misma distribución para todo i, j
 - *Representativo*: La distribución de Y_i es la misma que la distribución de la población de interés
- Data *iid* y representativa es un caso base que es relativamente fácil de analizar, pero es importante saber que no necesariamente se cumple en la práctica
 - si muestreamos gente dentro de un mismo hogar, esas observaciones no son independientes
 - si muestreamos de manera estratificada dentro de cada departamento, no es identicamente distribuido
 - En el ejemplo de Dewey v. Truman no era representativo!

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Calculemos la media y varianza de $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_i Y_i\right] =$$

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Calculemos la media y varianza de $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_i Y_i\right] = \frac{1}{N} \sum_i E[Y_i] =$$

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Calculemos la media y varianza de $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_i Y_i\right] = \frac{1}{N} \sum_i E[Y_i] = \mu$$

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Calculemos la media y varianza de $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_i Y_i\right] = \frac{1}{N} \sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N} \sum_i Y_i\right) =$$

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Calculemos la media y varianza de $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_i Y_i\right] = \frac{1}{N} \sum_i E[Y_i] = \mu$$

$$\text{Var}(\hat{\mu}) = \text{Var}\left(\frac{1}{N} \sum_i Y_i\right) = \frac{1}{N^2} \sum_i \text{Var}(Y_i) =$$

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Calculemos la media y varianza de $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_i Y_i\right] = \frac{1}{N} \sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N} \sum_i Y_i\right) = \frac{1}{N^2} \sum_i Var(Y_i) = \sigma^2/N,$$

donde $\sigma^2 = Var(Y_i)$

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Calculemos la media y varianza de $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_i Y_i\right] = \frac{1}{N} \sum_i E[Y_i] = \mu$$

$$Var(\hat{\mu}) = Var\left(\frac{1}{N} \sum_i Y_i\right) = \frac{1}{N^2} \sum_i Var(Y_i) = \sigma^2/N,$$

donde $\sigma^2 = Var(Y_i)$

- Ecuación (1) dice que $\hat{\mu}$ es *insesgado*: su media es μ

Media y Varianza de Promedio Muestral

Suponga que queremos saber sobre la media poblacional $\mu = E[Y_i]$ de una muestra representativa *iid* \mathbf{Y} de tamaño N

- Un estimador natural es el *promedio muestral*: $\hat{\mu} = \frac{1}{N} \sum_i Y_i$
- $\hat{\mu}$ es una función de la data aleatoria \mathbf{Y} . Por lo tanto es una variable aleatoria, tiene una distribución (también llamada “distribución muestral”)

Calculemos la media y varianza de $\hat{\mu}$

$$E[\hat{\mu}] = E\left[\frac{1}{N} \sum_i Y_i\right] = \frac{1}{N} \sum_i E[Y_i] = \mu$$

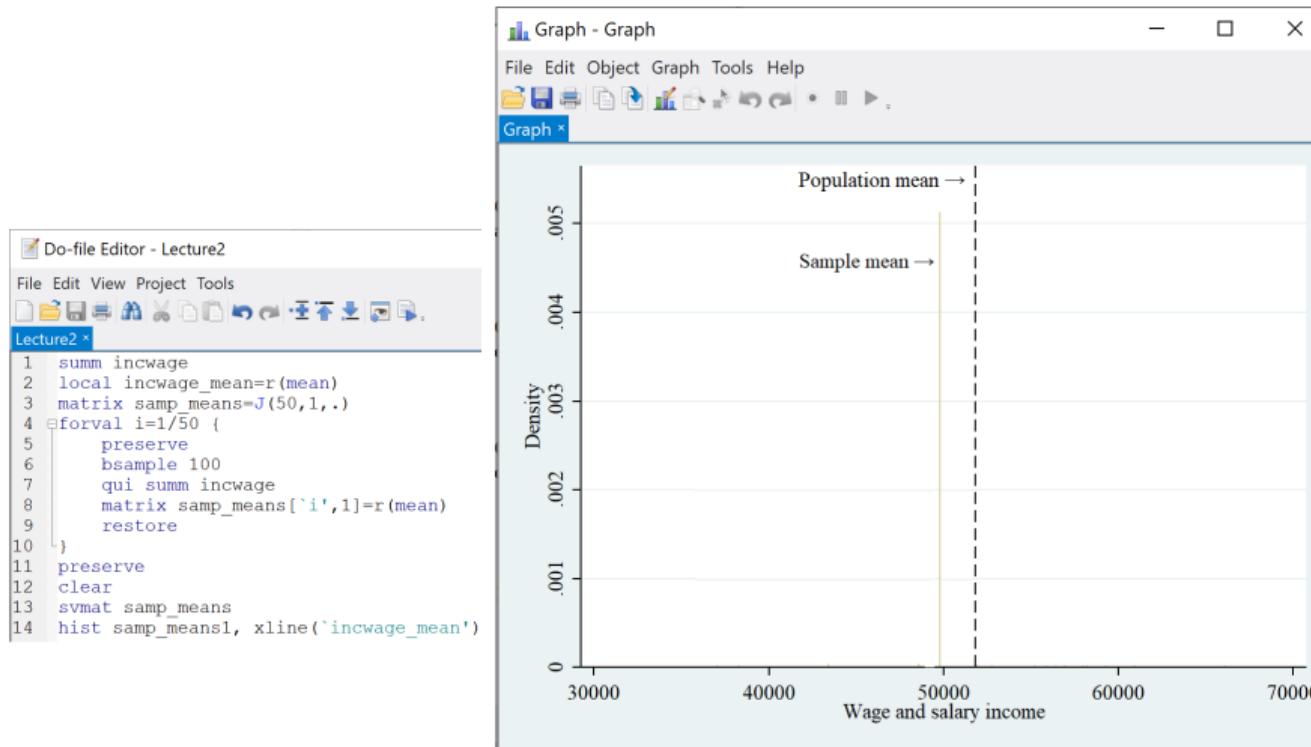
$$Var(\hat{\mu}) = Var\left(\frac{1}{N} \sum_i Y_i\right) = \frac{1}{N^2} \sum_i Var(Y_i) = \sigma^2/N,$$

donde $\sigma^2 = Var(Y_i)$

- Ecuación (1) dice que $\hat{\mu}$ es *insesgado*: su media es μ
- Ecuación (2) dice que la desviación estándar de $\hat{\mu}$ desde su media μ se achica a medida que N crece (\approx *consistencia*)

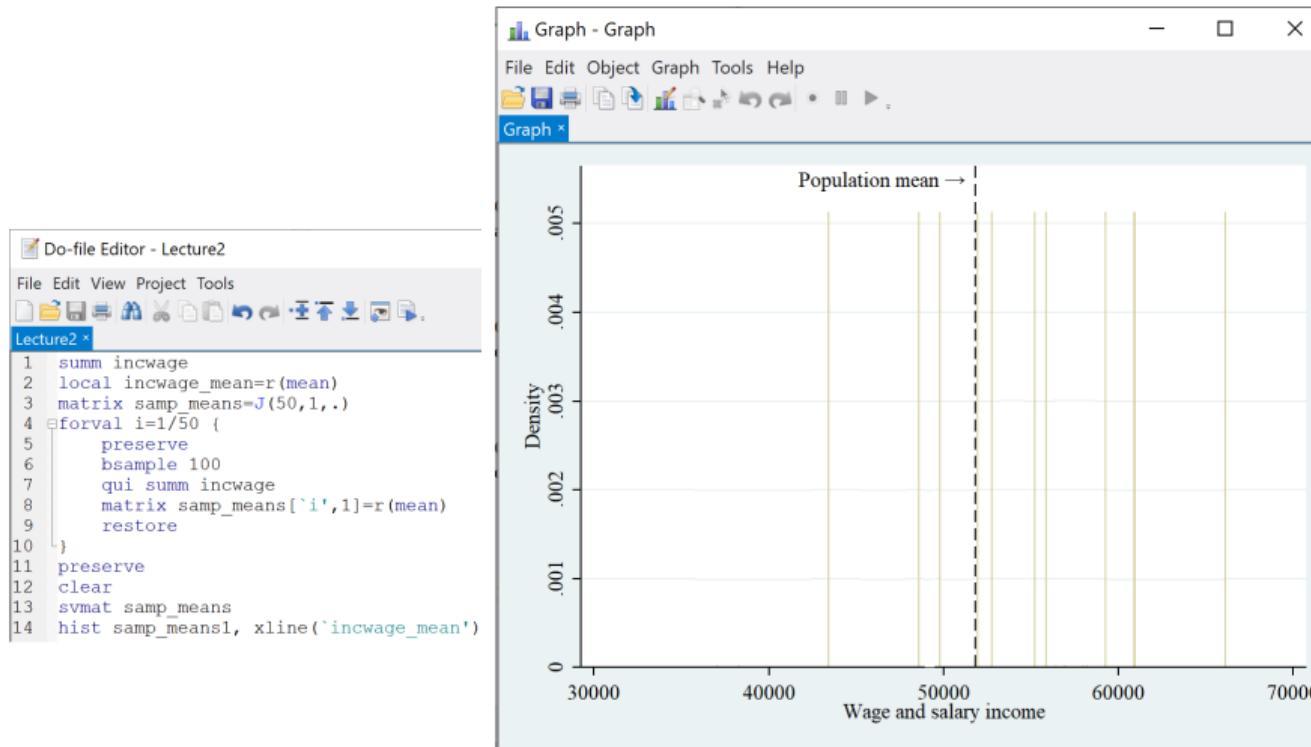
Muestreo Aleatorio y Promedio Muestral

Simulen insesgadez y consistencia para estimar la media de ingresos:



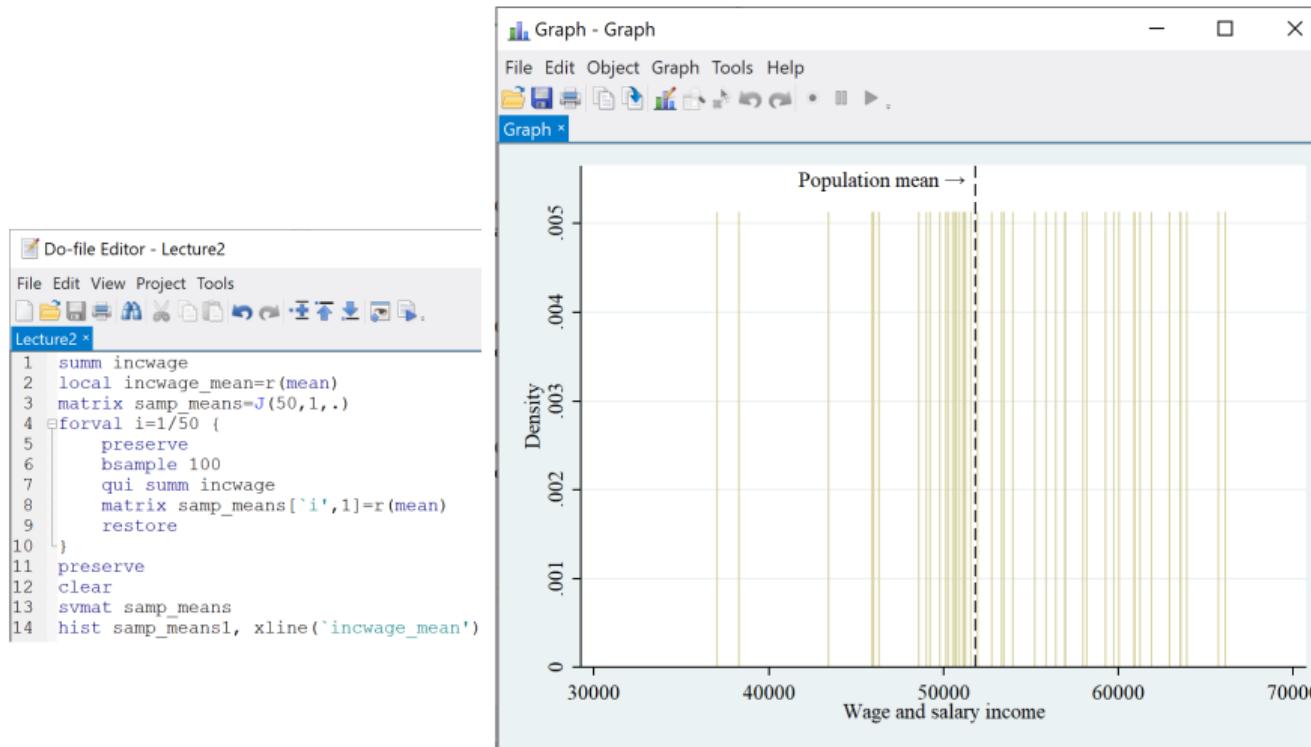
Simulations of Random Sampling

Simulen insesgadez y consistencia para estimar la media de ingresos:



Random Sampling and Sample Means

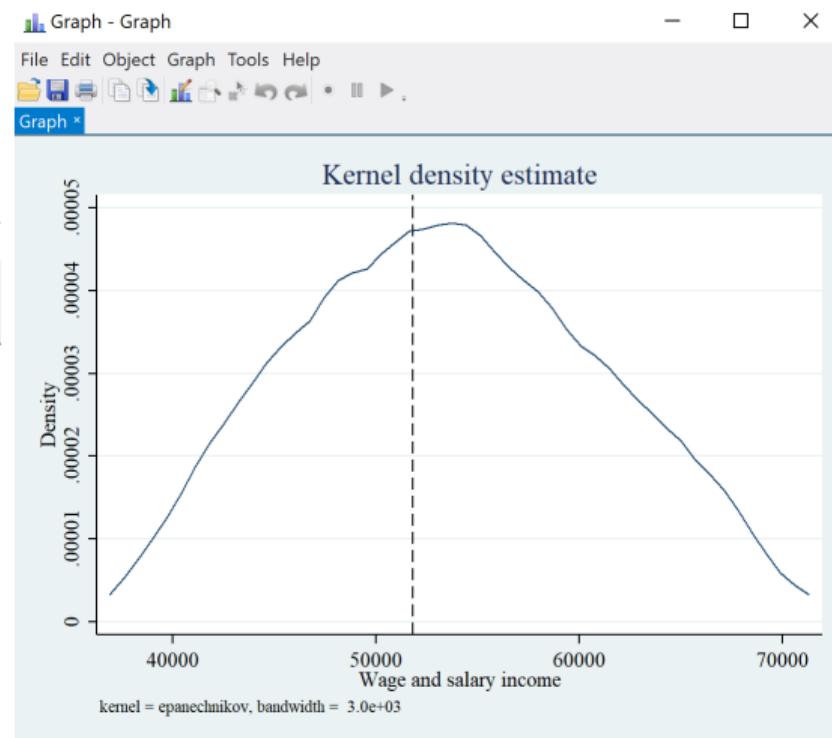
Simulen insesgadez y consistencia para estimar la media de ingresos:



Muestreo Aleatorio y Promedio Muestral

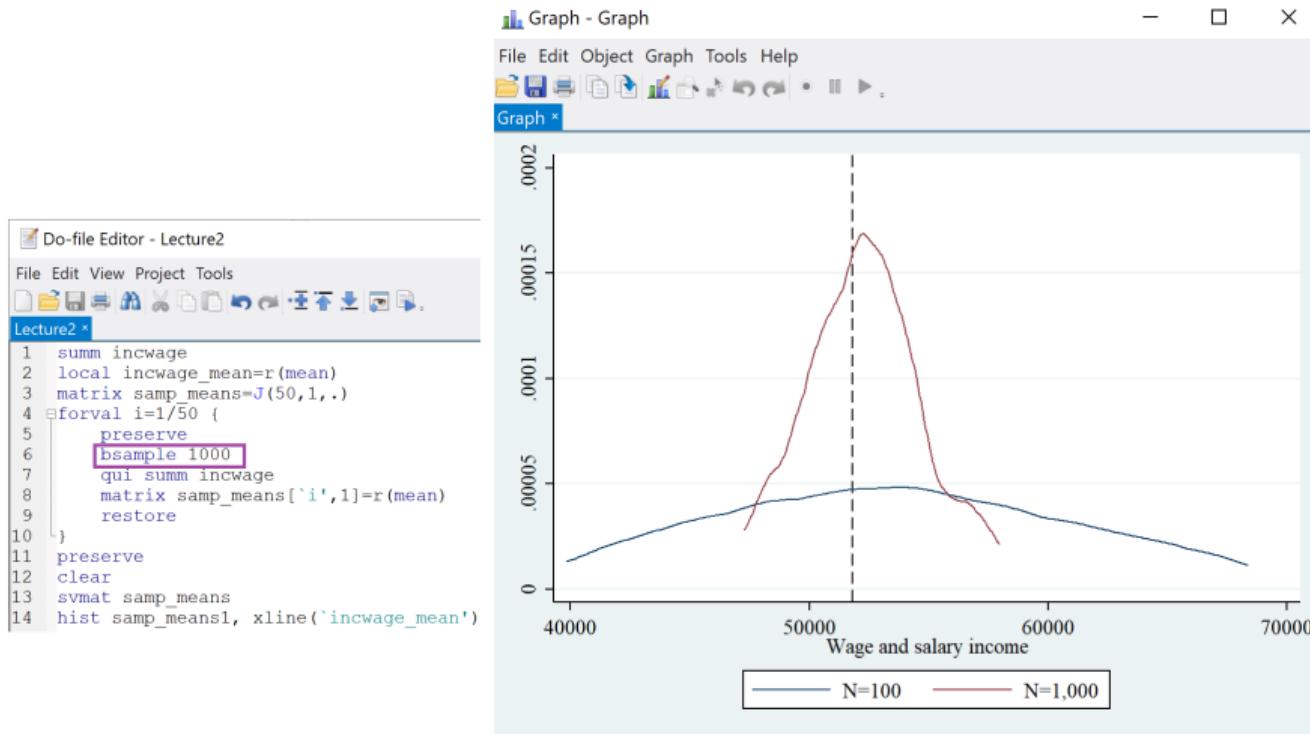
Simulen insesgadez y consistencia para estimar la media de ingresos:

```
Do-file Editor - Lecture2
File Edit View Project Tools
Lecture2 *
1  summ incwage
2  local incwage_mean=r(mean)
3  matrix samp_means=J(50,1,.)
4  forval i=1/50 {
5    preserve
6    bsample 100
7    qui summ incwage
8    matrix samp_means[`i',1]=r(mean)
9    restore
10 }
11 preserve
12 clear
13 svmat samp_means
14 hist samp_means, xline(`incwage_mean')
```



Muestreo Aleatorio y Promedio Muestral

Simulen insesgadez y consistencia para estimar la media de ingresos:



Muestreo Aleatorio y Promedio Muestral

Entonces, dos razones por las que $\hat{\mu} = \frac{1}{N} \sum_i Y_i$ es un buen estimador de $\mu = E[Y_i]$:

- Es insesgado: $E[\hat{\mu}] = \mu$
- Su varianza se encoge a 0 a medida que la muestra crece: $\lim_{N \rightarrow \infty} \text{Var}(\hat{\mu}) = 0$

Outline

1. Media Condicional ✓
2. Muestreo Aleatorio ✓
3. Test de Hipótesis e Inferencia
4. Teoría Asintótica

Test de Hipótesis - Intro

- Vimos que cuando N es largo, el promedio $\hat{\mu}$ se acerca mucho a la media μ
- ¿Pero qué significa estar “cerca”?
- Si el promedio es \$50,000, ¿es razonable pensar que la media verdadera puede ser \$55,000? ¿Y qué tal \$70,000?

Test de Hipótesis - Intro

- Vimos que cuando N es largo, el promedio $\hat{\mu}$ se acerca mucho a la media μ
- ¿Pero qué significa estar “cerca”?
- Si el promedio es \$50,000, ¿es razonable pensar que la media verdadera puede ser \$55,000? ¿Y qué tal \$70,000?
- **Test de hipótesis** nos ayuda a formalizar el concepto de “cercanía”.

Test de Hipótesis - Intro

- Vimos que cuando N es largo, el promedio $\hat{\mu}$ se acerca mucho a la media μ
- ¿Pero qué significa estar “cerca”?
- Si el promedio es \$50,000, ¿es razonable pensar que la media verdadera puede ser \$55,000? ¿Y qué tal \$70,000?
- **Test de hipótesis** nos ayuda a formalizar el concepto de “cercanía”.
- Nos dice que tan probable es ver un promedio muestral de \$50,000 si la media verdadera es \$55,000, \$70,000, etc.

Overview de Test de Hipótesis

- ① Especificamos la **hipótesis nula** que la media poblacional toma cierto valor, $H_0 : \mu = \mu_0$.
 - Por ej. La media poblacional es \$55,000 significaría que $H_0 : \mu = 55,000$

Overview de Test de Hipótesis

- ① Especificamos la **hipótesis nula** que la media poblacional toma cierto valor, $H_0 : \mu = \mu_0$.
 - Por ej. La media poblacional es \$55,000 significaría que $H_0 : \mu = 55,000$
- ② Calculamos que tan probable es observar $\hat{\mu}$ por lo menos tan lejos como observamos de μ_0 si la nula es verdadera. Esto se llama un **p-value**

Overview de Test de Hipótesis

- ① Especificamos la **hipótesis nula** que la media poblacional toma cierto valor, $H_0 : \mu = \mu_0$.
 - Por ej. La media poblacional es \$55,000 significaría que $H_0 : \mu = 55,000$
- ② Calculamos que tan probable es observar $\hat{\mu}$ por lo menos tan lejos como observamos de μ_0 si la nula es verdadera. Esto se llama un **p-value**
- ③ **Rechazamos** si el p-value es pequeño, i.e.: es poco probable que observemos un $\hat{\mu}$ tan lejano de μ_0 si la nula fuera cierta

Overview de Test de Hipótesis

- ① Especificamos la **hipótesis nula** que la media poblacional toma cierto valor, $H_0 : \mu = \mu_0$.
 - Por ej. La media poblacional es \$55,000 significaría que $H_0 : \mu = 55,000$
- ② Calculamos que tan probable es observar $\hat{\mu}$ por lo menos tan lejos como observamos de μ_0 si la nula es verdadera. Esto se llama un **p-value**
- ③ **Rechazamos** si el p-value es pequeño, i.e.: es poco probable que observemos un $\hat{\mu}$ tan lejano de μ_0 si la nula fuera cierta
 - El umbral más común es $\alpha = 0.05$

Overview de Test de Hipótesis

- ① Especificamos la **hipótesis nula** que la media poblacional toma cierto valor, $H_0 : \mu = \mu_0$.
 - Por ej. La media poblacional es \$55,000 significaría que $H_0 : \mu = 55,000$
- ② Calculamos que tan probable es observar $\hat{\mu}$ por lo menos tan lejos como observamos de μ_0 si la nula es verdadera. Esto se llama un **p-value**
- ③ **Rechazamos** si el p-value es pequeño, i.e.: es poco probable que observemos un $\hat{\mu}$ tan lejano de μ_0 si la nula fuera cierta
 - El umbral más común es $\alpha = 0.05$
- ④ Podemos construir un **intervalo de confianza** (CI) que colecciona todos los posibles valores de μ_0 que *no podemos rechazar* de esta manera

Overview de Test de Hipótesis

- ① Especificamos la **hipótesis nula** que la media poblacional toma cierto valor, $H_0 : \mu = \mu_0$.
 - Por ej. La media poblacional es \$55,000 significaría que $H_0 : \mu = 55,000$
- ② Calculamos que tan probable es observar $\hat{\mu}$ por lo menos tan lejos como observamos de μ_0 si la nula es verdadera. Esto se llama un **p-value**
- ③ **Rechazamos** si el p-value es pequeño, i.e.: es poco probable que observemos un $\hat{\mu}$ tan lejano de μ_0 si la nula fuera cierta
 - El umbral más común es $\alpha = 0.05$
- ④ Podemos construir un **intervalo de confianza** (CI) que colecciona todos los posibles valores de μ_0 que *no podemos rechazar* de esta manera
 - El CI, por construcción, contiene el verdadero valor μ en 95% de las realizaciones de la data cuando $\alpha = 0.05$

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido

- Trabajemos sobre este caso especial en todos sus pasos: $\hat{\mu} \sim N(\mu, \sigma^2/N)$ con σ^2 known.

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido

- Trabajemos sobre este caso especial en todos sus pasos: $\hat{\mu} \sim N(\mu, \sigma^2/N)$ con σ^2 known.
- ¿Por qué considerar este caso?
 - $N(\mu, \sigma^2/N)$ es la exacta distribución de $\hat{\mu}$ si $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido

- Trabajemos sobre este caso especial en todos sus pasos: $\hat{\mu} \sim N(\mu, \sigma^2/N)$ con σ^2 known.
- ¿Por qué considerar este caso?
 - $N(\mu, \sigma^2/N)$ es la exacta distribución de $\hat{\mu}$ si $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$
 - En un rato mostraremos que incluso si Y_i no es normal, $\hat{\mu}$ es aprox. normal para N grande

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido

- Trabajemos sobre este caso especial en todos sus pasos: $\hat{\mu} \sim N(\mu, \sigma^2/N)$ con σ^2 known.
- ¿Por qué considerar este caso?
 - $N(\mu, \sigma^2/N)$ es la exacta distribución de $\hat{\mu}$ si $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$
 - En un rato mostraremos que incluso si Y_i no es normal, $\hat{\mu}$ es aprox. normal para N grande
 - También mostraremos que con N grande podemos estimar σ^2 arbitrariamente bien

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido

- Trabajemos sobre este caso especial en todos sus pasos: $\hat{\mu} \sim N(\mu, \sigma^2/N)$ con σ^2 known.
- ¿Por qué considerar este caso?
 - $N(\mu, \sigma^2/N)$ es la exacta distribución de $\hat{\mu}$ si $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$
 - En un rato mostraremos que incluso si Y_i no es normal, $\hat{\mu}$ es aprox. normal para N grande
 - También mostraremos que con N grande podemos estimar σ^2 arbitrariamente bien
- Suponga que queremos testear $H_0 : \mu = \mu_0$ (por ej. media de ingresos es \$55,000)
- Definimos $\hat{t} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{N}}$, y note que bajo la nula $\hat{t} \sim N(0, 1)$

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido

- Trabajemos sobre este caso especial en todos sus pasos: $\hat{\mu} \sim N(\mu, \sigma^2/N)$ con σ^2 known.
- ¿Por qué considerar este caso?
 - $N(\mu, \sigma^2/N)$ es la exacta distribución de $\hat{\mu}$ si $Y_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$
 - En un rato mostraremos que incluso si Y_i no es normal, $\hat{\mu}$ es aprox. normal para N grande
 - También mostraremos que con N grande podemos estimar σ^2 arbitrariamente bien
- Suponga que queremos testear $H_0 : \mu = \mu_0$ (por ej. media de ingresos es \$55,000)
- Definimos $\hat{t} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{N}}$, y note que bajo la nula $\hat{t} \sim N(0, 1)$
 - La distribución de \hat{t} es sobre muestras repetidas provenientes de la muestra (Y_1, \dots, Y_N) .

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido (cont.)

- Bajo la nula, $H_0 : \mu = \mu_0$, $\hat{t} \sim N(0, 1)$.

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido (cont.)

- Bajo la nula, $H_0 : \mu = \mu_0$, $\hat{t} \sim N(0, 1)$.
- ¿Qué es $Pr(|\hat{t}| > t)$ para algún $t \geq 0$?

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido (cont.)

- Bajo la nula, $H_0 : \mu = \mu_0$, $\hat{t} \sim N(0, 1)$.
- ¿Qué es $Pr(|\hat{t}| > t)$ para algún $t \geq 0$?

$$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t)$$

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido (cont.)

- Bajo la nula, $H_0 : \mu = \mu_0$, $\hat{t} \sim N(0, 1)$.
- ¿Qué es $Pr(|\hat{t}| > t)$ para algún $t \geq 0$?

$$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) =$$

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido (cont.)

- Bajo la nula, $H_0 : \mu = \mu_0$, $\hat{t} \sim N(0, 1)$.

- ¿Qué es $Pr(|\hat{t}| > t)$ para algún $t \geq 0$?

$$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) = 1 - (\Phi(t) - \Phi(-t))$$

- Definimos el *p-value* para la nula como $H_0 : \mu = \mu_0$ as

$$p(\hat{t}) = 1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|))$$

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido (cont.)

- Bajo la nula, $H_0 : \mu = \mu_0$, $\hat{t} \sim N(0, 1)$.

- ¿Qué es $Pr(|\hat{t}| > t)$ para algún $t \geq 0$?

$$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) = 1 - (\Phi(t) - \Phi(-t))$$

- Definimos el *p-value* para la nula como $H_0 : \mu = \mu_0$ as

$$p(\hat{t}) = 1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)) = 1 - \left(\Phi\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}}\right) - \Phi\left(\frac{-|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}}\right) \right)$$

Test de Hipótesis con $\hat{\mu}$ Normalmente Distribuido (cont.)

- Bajo la nula, $H_0 : \mu = \mu_0$, $\hat{t} \sim N(0, 1)$.

- ¿Qué es $Pr(|\hat{t}| > t)$ para algún $t \geq 0$?

$$Pr(|\hat{t}| > t) = 1 - Pr(|\hat{t}| \leq t) = 1 - Pr(-t \leq \hat{t} \leq t) = 1 - (\Phi(t) - \Phi(-t))$$

- Definimos el *p-value* para la nula como $H_0 : \mu = \mu_0$ as

$$\begin{aligned} p(\hat{t}) &= 1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|)) = 1 - \left(\Phi\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}}\right) - \Phi\left(\frac{-|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}}\right) \right) \\ &= 2 \left(1 - \Phi\left(\frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}}\right) \right) \end{aligned}$$

- Intuitivamente, *p* es la probabilidad de observar un $|\hat{t}|$ por lo menos así de grande si la nula es verdadera

Ilustración de Construcción de P-Value

Standard Normal PDF (mean zero, unit std. dev.)

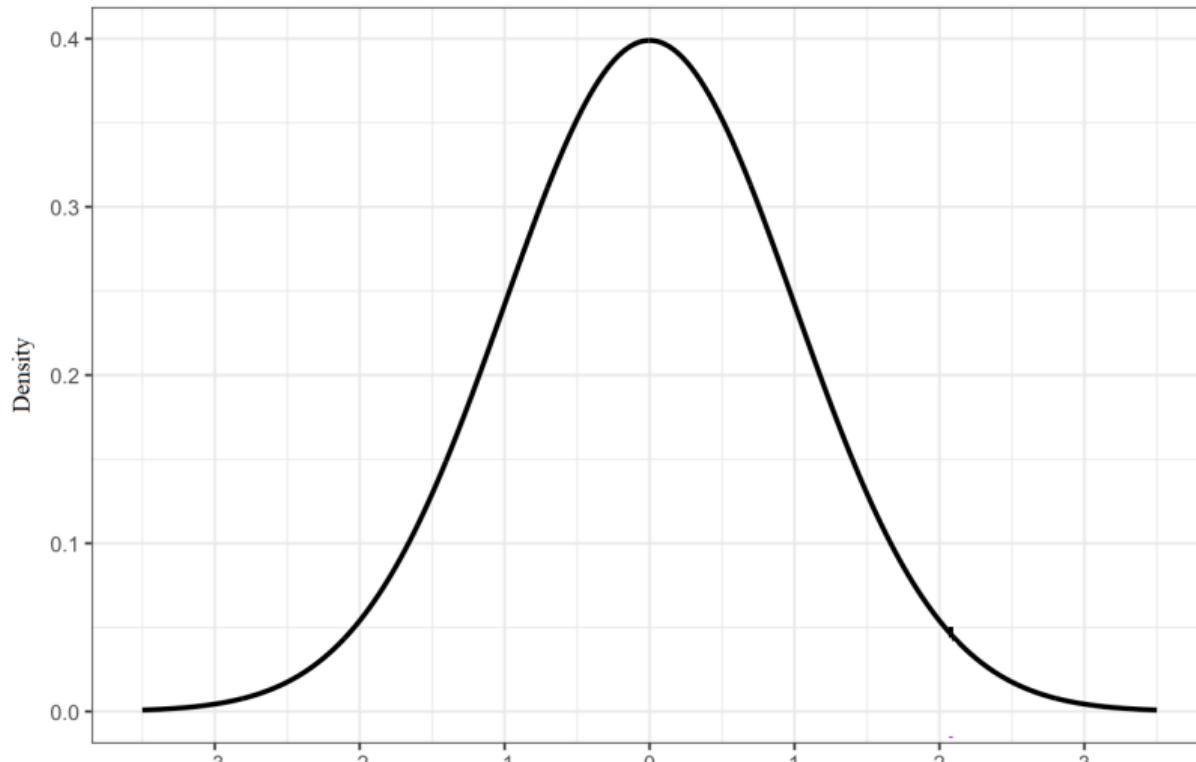


Ilustración de Construcción de P-Value

Normalized realization of the random estimator

$$\left| \frac{\hat{\mu}(x)^{obs} - \mu_0}{\sqrt{\sigma^2(x)/N}} \right|$$

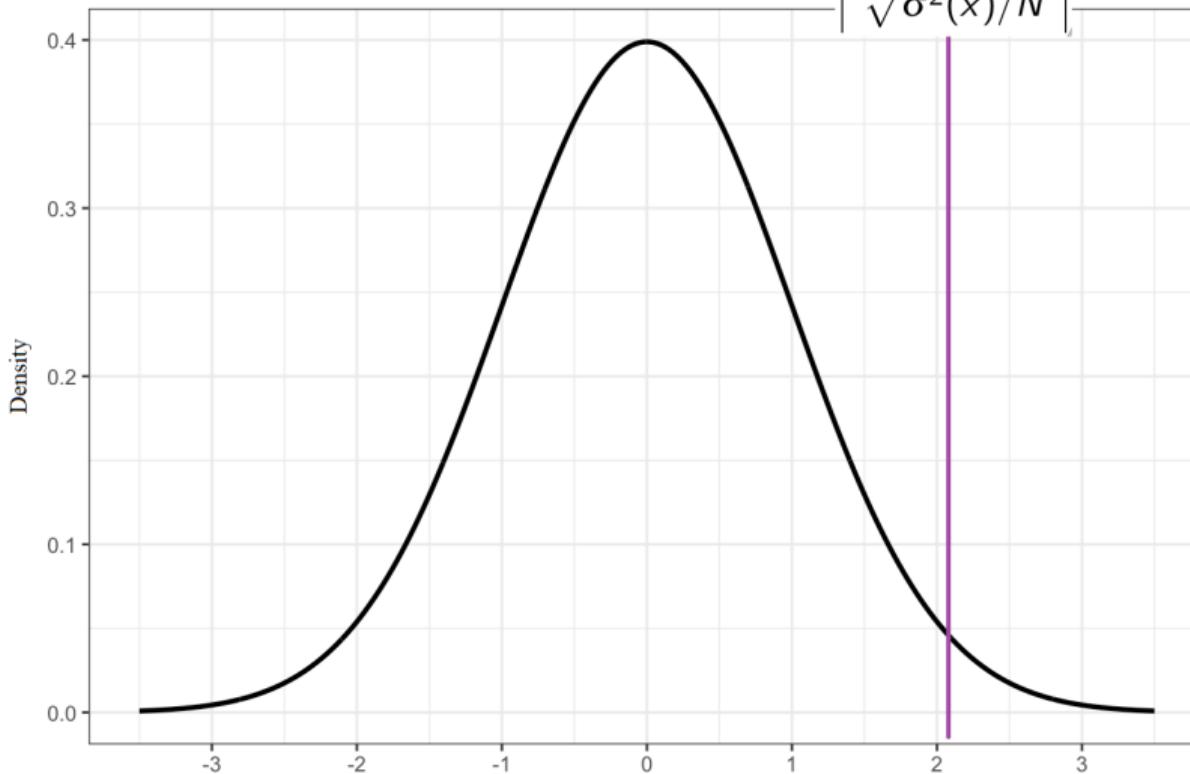
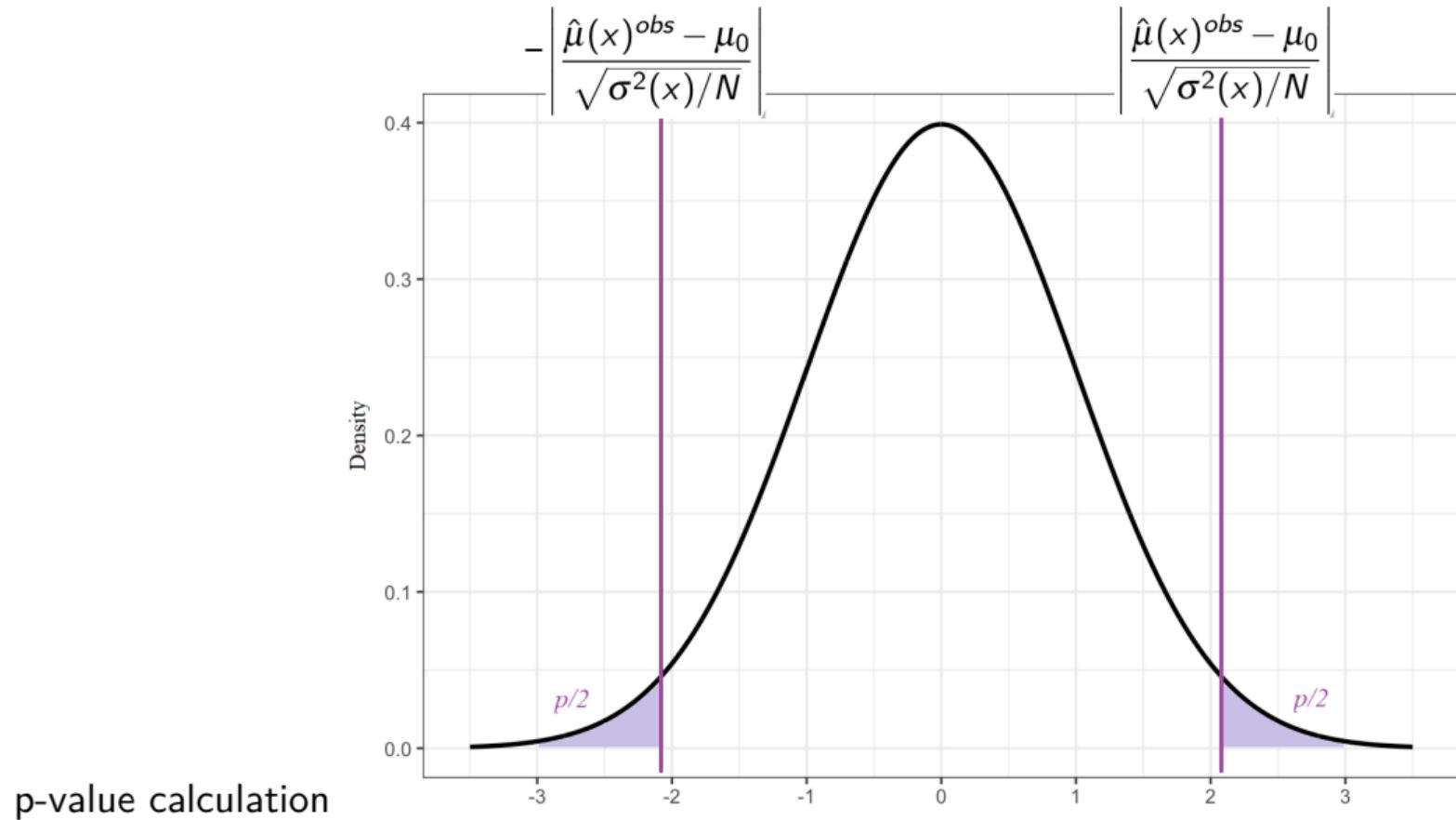


Illustration of P-Value Construction



¿Cuándo rechazar la nula?

- Recuerden que el *p*-value toma la forma

$$1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|))$$

- Resulta que $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Entonces, $p < 0.05$ si y solo sí $|\hat{t}| > 1.96$. Por lo tanto rechazamos al nivel de 5% si $|\hat{t}| > 1.96$.

¿Cuándo rechazar la nula?

- Recuerden que el *p*-value toma la forma

$$1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|))$$

- Resulta que $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Entonces, $p < 0.05$ si y solo sí $|\hat{t}| > 1.96$. Por lo tanto rechazamos al nivel de 5% si $|\hat{t}| > 1.96$.
- ¿Qué implica esto sobre el valor de μ_0 que rechazamos/no rechazamos?

¿Cuándo rechazar la nula?

- Recuerden que el p -value toma la forma

$$1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|))$$

- Resulta que $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Entonces, $p < 0.05$ si y solo sí $|\hat{t}| > 1.96$. Por lo tanto rechazamos al nivel de 5% si $|\hat{t}| > 1.96$.
- ¿Qué implica esto sobre el valor de μ_0 que rechazamos/no rechazamos?
- No rechazamos si

$$|\hat{t}| \leq 1.96$$

¿Cuándo rechazar la nula?

- Recuerden que el p -value toma la forma

$$1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|))$$

- Resulta que $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Entonces, $p < 0.05$ si y solo sí $|\hat{t}| > 1.96$. Por lo tanto rechazamos al nivel de 5% si $|\hat{t}| > 1.96$.
- ¿Qué implica esto sobre el valor de μ_0 que rechazamos/no rechazamos?
- No rechazamos si

$$|\hat{t}| \leq 1.96 \implies \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \leq 1.96$$

¿Cuándo rechazar la nula?

- Recuerden que el p -value toma la forma

$$1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|))$$

- Resulta que $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Entonces, $p < 0.05$ si y solo sí $|\hat{t}| > 1.96$. Por lo tanto rechazamos al nivel de 5% si $|\hat{t}| > 1.96$.
- ¿Qué implica esto sobre el valor de μ_0 que rechazamos/no rechazamos?
- No rechazamos si

$$|\hat{t}| \leq 1.96 \implies \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \leq 1.96 \implies \mu_0 \in [\hat{\mu} - 1.96\sigma/\sqrt{N}, \hat{\mu} + 1.96\sigma/\sqrt{N}]$$

¿Cuándo rechazar la nula?

- Recuerden que el p -value toma la forma

$$1 - (\Phi(|\hat{t}|) - \Phi(-|\hat{t}|))$$

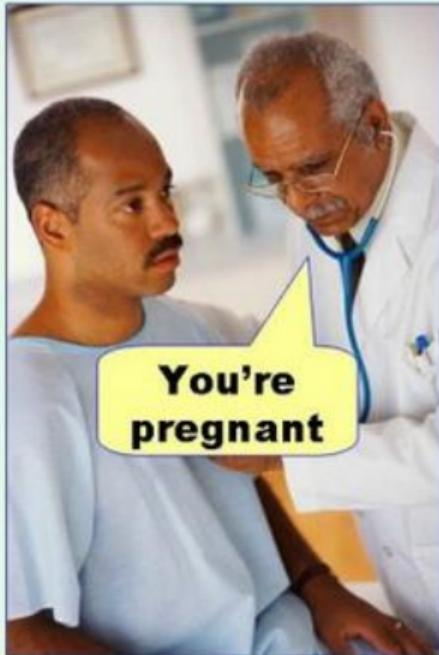
- Resulta que $\Phi(1.96) - \Phi(-1.96) \approx 0.95$. Entonces, $p < 0.05$ si y solo sí $|\hat{t}| > 1.96$. Por lo tanto rechazamos al nivel de 5% si $|\hat{t}| > 1.96$.
- ¿Qué implica esto sobre el valor de μ_0 que rechazamos/no rechazamos?
- No rechazamos si

$$|\hat{t}| \leq 1.96 \implies \frac{|\hat{\mu} - \mu_0|}{\sigma/\sqrt{N}} \leq 1.96 \implies \mu_0 \in [\hat{\mu} - 1.96\sigma/\sqrt{N}, \hat{\mu} + 1.96\sigma/\sqrt{N}]$$

- El intervalo $\hat{\mu} \pm 1.96\sigma/\sqrt{N}$ es por tanto el 95% intervalo de confianza (CI)
 - Tiene la propiedad de que $Pr(\mu_0 \in CI) = 0.95$ cuando $H_0 : \mu = \mu_0$ es verdadero

Dos Formas de Fregarla

Type I error
(false positive)



Type II error
(false negative)



Significancia y Poder Estadístico

- El *nivel de significancia* (suele decirsele solo *size*) de un test es la probabilidad pre-especificada de incorrectamente rechazar la nula cuando es verdad (% error tipo I)

Significancia y Poder Estadístico

- El *nivel de significancia* (suele decirsele solo *size*) de un test es la probabilidad pre-especificada de incorrectamente rechazar la nula cuando es verdad (% error tipo I)
 - Por ej. un test al 5% rechaza cuando $p < 0.05$.

Significancia y Poder Estadístico

- El *nivel de significancia* (suele decirsele solo *size*) de un test es la probabilidad pre-especificada de incorrectamente rechazar la nula cuando es verdadera (% error tipo I)
 - Por ej. un test al 5% rechaza cuando $p < 0.05$.
- El *poder* de un test es la probabilidad de correctamente rechazar la nula cuando es falsa (1- % error tipo II)
 - El poder es una función de la hipótesis *alternative*.
 - Ejemplo 1: la probabilidad de rechazar $H_0 : \mu = \mu_0$ cuando en realidad $\mu = \mu_A$
 - Ejemplo 2: es más probable rechazar que yo soy Brad Pitt a que soy Pedro Castillo (mayor poder)

Precaución sobre interpretar *P*-Value

- A veces los p-values se interpretan como la “probabilidad de que H_0 es verdad.” ¿Es esto correcto?

Precaución sobre interpretar *P*-Value

- A veces los p-values se interpretan como la “probabilidad de que H_0 es verdad.” ¿Es esto correcto? ¡No!
 - *p*-value dice la probabilidad de obtener la data observada *asumiendo* que la nula es verdad
 - Es decir, *p*-value nos dice sobre $P(\text{data}|H_0)$, no $P(H_0|\text{data})$.
 - Por la regla de Bayes, $P(H_0|\text{Data}) = P(\text{Data}|H_0) * P(H_0)/P(\text{Data})$. Para formalizar esto, debemos tomar una postura sobre cual es nuestra creencia previa sobre si H_0 es verdadera, $P(H_0)$.

Precaución sobre interpretar *P*-Value

- A veces los p-values se interpretan como la “probabilidad de que H_0 es verdad.” ¿Es esto correcto? ¡No!
 - *p*-value dice la probabilidad de obtener la data observada *asumiendo* que la nula es verdad
 - Es decir, *p*-value nos dice sobre $P(\text{data}|H_0)$, no $P(H_0|\text{data})$.
 - Por la regla de Bayes, $P(H_0|\text{Data}) = P(\text{Data}|H_0) * P(H_0)/P(\text{Data})$. Para formalizar esto, debemos tomar una postura sobre cual es nuestra creencia previa sobre si H_0 es verdadera, $P(H_0)$.
- La gente interpreta $p < 0.05$ como fuerte evidencia de que hay un efecto, y $p \geq 0.05$ como evidencia de no efecto.
 - Pero $p = 0.05$ es un umbral arbitrario.
 - Mejor vean el *p*-value como un espectro indicando que tan probable es que veamos la data observada dada la hip. nula.

Precaución sobre interpretar *P*-Value

- A veces los p-values se interpretan como la “probabilidad de que H_0 es verdad.” ¿Es esto correcto? ¡No!
 - *p*-value dice la probabilidad de obtener la data observada *asumiendo* que la nula es verdad
 - Es decir, *p*-value nos dice sobre $P(\text{data}|H_0)$, no $P(H_0|\text{data})$.
 - Por la regla de Bayes, $P(H_0|\text{Data}) = P(\text{Data}|H_0) * P(H_0)/P(\text{Data})$. Para formalizar esto, debemos tomar una postura sobre cual es nuestra creencia previa sobre si H_0 es verdadera, $P(H_0)$.
- La gente interpreta $p < 0.05$ como fuerte evidencia de que hay un efecto, y $p \geq 0.05$ como evidencia de no efecto.
 - Pero $p = 0.05$ es un umbral arbitrario.
 - Mejor vean el *p*-value como un espectro indicando que tan probable es que veamos la data observada dada la hip. nula.
 - Además, *p*-values pueden ser super altos cuando la nula es verdadera (¡bajo poder estadístico!)

¿Qué hacemos con data no-normal?

- Ok, chévere ... ¿y ahora que hacemos con la data de la vida real?

¿Qué hacemos con data no-normal?

- Ok, chévere ... ¿y ahora que hacemos con la data de la vida real?
 - ¡Casi todas las variables no tienen esta distribución!

¿Qué hacemos con data no-normal?

- Ok, chévere ... ¿y ahora que hacemos con la data de la vida real?
 - ¡Casi todas las variables no tienen esta distribución!
 - ¡¿Incluso si fueran normales, cómo saber su varianza verdadera?!

¿Qué hacemos con data no-normal?

- Ok, chévere ... ¿y ahora que hacemos con la data de la vida real?
 - ¡Casi todas las variables no tienen esta distribución!
 - ¡¿Incluso si fueran normales, cómo saber su varianza verdadera?!
 - ¡¿Les acabo de hacer perder un buen par de horas!?!

¿Qué hacemos con data no-normal?

- Ok, chévere ... ¿y ahora que hacemos con la data de la vida real?
 - ¡Casi todas las variables no tienen esta distribución!
 - ¡¿Incluso si fueran normales, cómo saber su varianza verdadera?!
 - ¡¿Les acabo de hacer perder un buen par de horas!?! No ... espero :)

¿Qué hacemos con data no-normal?

- Ok, chévere ... ¿y ahora que hacemos con la data de la vida real?
 - ¡Casi todas las variables no tienen esta distribución!
 - ¡¿Incluso si fueran normales, cómo saber su varianza verdadera?!
 - ¡¿Les acabo de hacer perder un buen par de horas!?! No ... espero :)
- Ahora veremos que pasa en el mundo de las muestras grandes: **asymptopia**. Esto nos permitirá aplicar inferencia similar incluso cuando la muestra Y_i no sea normalmente distribuida.

Outline

1. Media Condicional ✓
2. Muestreo Aleatorio ✓
3. Test de Hipótesis e Inferencia ✓
4. Teoría Asintótica

Motivación

- Hemos visto el caso de hacer test de hipótesis sobre media poblacional usando $\hat{\mu}$ cuando es **normalmente distribuida** con una varianza conocida
- Esto surge del caso en que $Y_i \sim N(\mu, \sigma^2)$ cuando σ es conocido
- Esta situación es bastante rara... ¿cómo hacemos inferencia de manera general?

Motivación

- Hemos visto el caso de hacer test de hipótesis sobre media poblacional usando $\hat{\mu}$ cuando es **normalmente distribuida** con una varianza conocida
- Esto surge del caso en que $Y_i \sim N(\mu, \sigma^2)$ cuando σ es conocido
- Esta situación es bastante rara... ¿cómo hacemos inferencia de manera general?
- Afortunadamente, el supuesto de que el promedio muestral está normalmente distribuido resulta ser una buena **aproximación** cuando la muestra es grande

Motivación

- Hemos visto el caso de hacer test de hipótesis sobre media poblacional usando $\hat{\mu}$ cuando es **normalmente distribuida** con una varianza conocida
- Esto surge del caso en que $Y_i \sim N(\mu, \sigma^2)$ cuando σ es conocido
- Esta situación es bastante rara... ¿cómo hacemos inferencia de manera general?
- Afortunadamente, el supuesto de que el promedio muestral está normalmente distribuido resulta ser una buena **aproximación** cuando la muestra es grande
- A lo que nos referimos con “buena aproximación” es algo formalizado por la teoría asintótica, que considera la distribución de $\hat{\mu}$ en el límite cuando $N \rightarrow \infty$

Resumen de Resultados Importantes

- La **Ley de los Grandes Números** (LLN) dice que cuando N es grande, $\hat{\mu}$ está cerca de μ con probabilidad muy alta

Resumen de Resultados Importantes

- La **Ley de los Grandes Números** (LLN) dice que cuando N es grande, $\hat{\mu}$ está cerca de μ con probabilidad muy alta
- El **Teorema del Límite Central** (CLT) dice que cuando N es grande, la distribución de $\hat{\mu}$ es aproximadamente normal con media μ y varianza σ^2/n

Resumen de Resultados Importantes

- La **Ley de los Grandes Números** (LLN) dice que cuando N es grande, $\hat{\mu}$ está cerca de μ con probabilidad muy alta
- El **Teorema del Límite Central** (CLT) dice que cuando N es grande, la distribución de $\hat{\mu}$ es aproximadamente normal con media μ y varianza σ^2/n
- El **Continuous Mapping Theorem** dice que cuando N es grande, las funciones continuas de $\hat{\mu}$, digamos $g(\hat{\mu})$, también son cercanas a $g(\mu)$

Convergencia en Probabilidad

- Intuitivamente, una variable aleatoria X_N **converge en probabilidad** a x si la probabilidad de que X_N está “cerca” a x es casi 1 cuando N es grande

Convergencia en Probabilidad

- Intuitivamente, una variable aleatoria X_N **converge en probabilidad** a x si la probabilidad de que X_N está “cerca” a x es casi 1 cuando N es grande
- Formalmente, decimos que X_N converge en probabilidad a x , $X_n \rightarrow_p x$, si para todos los $\varepsilon > 0$,

$$\lim_{N \rightarrow \infty} P(|X_N - x| > \varepsilon) \rightarrow 0$$

Convergencia en Probabilidad

- Intuitivamente, una variable aleatoria X_N **converge en probabilidad** a x si la probabilidad de que X_N está “cerca” a x es casi 1 cuando N es grande
- Formalmente, decimos que X_N converge en probabilidad a x , $X_n \rightarrow_p x$, si para todos los $\varepsilon > 0$,

$$\lim_{N \rightarrow \infty} P(|X_N - x| > \varepsilon) \rightarrow 0$$

- Si $X_n \rightarrow_p x$ para una constante x , decimos que X_n es *consistente* para x

Convergencia en Probabilidad (Cont.)

- Hecho útil: si $E[(X_N - x)^2] \rightarrow 0$, entonces $X_N \rightarrow_p x$

Convergencia en Probabilidad (Cont.)

- Hecho útil: si $E[(X_N - x)^2] \rightarrow 0$, entonces $X_N \rightarrow_p x$
- **Prueba** (recuerden que en este curso no se les pedirá derivar estas cosas):
Por ley de esperanzas iteradas (LIE),

$$\begin{aligned} E[(X_N - x)^2] &= P(|X_n - x| > \varepsilon)E[(X_N - x)^2 | |X_n - x| > \varepsilon] + \\ &\quad P(|X_n - x| \leq \varepsilon)E[(X_N - x)^2 | |X_n - x| \leq \varepsilon] \end{aligned}$$

Convergencia en Probabilidad (Cont.)

- Hecho útil: si $E[(X_N - x)^2] \rightarrow 0$, entonces $X_N \rightarrow_p x$
- **Prueba** (recuerden que en este curso no se les pedirá derivar estas cosas):
Por ley de esperanzas iteradas (LIE),

$$\begin{aligned} E[(X_N - x)^2] &= P(|X_n - x| > \varepsilon)E[(X_N - x)^2 | |X_n - x| > \varepsilon] + \\ &\quad P(|X_n - x| \leq \varepsilon)E[(X_N - x)^2 | |X_n - x| \leq \varepsilon] \\ &\geq P(|X_n - x| > \varepsilon)\varepsilon^2 + 0 \end{aligned}$$

Convergencia en Probabilidad (Cont.)

- Hecho útil: si $E[(X_N - x)^2] \rightarrow 0$, entonces $X_N \rightarrow_p x$
- **Prueba** (recuerden que en este curso no se les pedirá derivar estas cosas):
Por ley de esperanzas iteradas (LIE),

$$\begin{aligned} E[(X_N - x)^2] &= P(|X_n - x| > \varepsilon)E[(X_N - x)^2 | |X_n - x| > \varepsilon] + \\ &\quad P(|X_n - x| \leq \varepsilon)E[(X_N - x)^2 | |X_n - x| \leq \varepsilon] \\ &\geq P(|X_n - x| > \varepsilon)\varepsilon^2 + 0 \end{aligned}$$

Esto implica que

$$P(|X_N - x| > \varepsilon) \leq E[(X_N - x)^2]/\varepsilon^2 \text{ (Chebychev's Inequality)}$$

Convergencia en Probabilidad (Cont.)

- Hecho útil: si $E[(X_N - x)^2] \rightarrow 0$, entonces $X_N \rightarrow_p x$
- **Prueba** (recuerden que en este curso no se les pedirá derivar estas cosas):
Por ley de esperanzas iteradas (LIE),

$$\begin{aligned} E[(X_N - x)^2] &= P(|X_n - x| > \varepsilon)E[(X_N - x)^2 | |X_n - x| > \varepsilon] + \\ &\quad P(|X_n - x| \leq \varepsilon)E[(X_N - x)^2 | |X_n - x| \leq \varepsilon] \\ &\geq P(|X_n - x| > \varepsilon)\varepsilon^2 + 0 \end{aligned}$$

Esto implica que

$$P(|X_N - x| > \varepsilon) \leq E[(X_N - x)^2]/\varepsilon^2 \text{ (Chebychev's Inequality)}$$

Por tanto, $E[(X_N - x)^2] \rightarrow 0$ implica $P(|X_N - x| > \varepsilon) \rightarrow 0$

Ley de los Grandes Números

- **Ley de los Grandes Números.** Suponga que Y_1, \dots, Y_N son sacados *iid* de una distribución con $\text{Var}(Y_i) = \sigma^2 < \infty$. Entonces

$$\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^N Y_i \rightarrow_p \mu = E[Y_i]$$

- En cristiano: si la muestra se hace grande, la media muestral se volverá cercana a la media poblacional con alta probabilidad.

Ley de los Grandes Números

- **Ley de los Grandes Números.** Suponga que Y_1, \dots, Y_N son sacados *iid* de una distribución con $\text{Var}(Y_i) = \sigma^2 < \infty$. Entonces

$$\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^N Y_i \rightarrow_p \mu = E[Y_i]$$

- En cristiano: si la muestra se hace grande, la media muestral se volverá cercana a la media poblacional con alta probabilidad.
- **Prueba:** Como saben $E[\hat{\mu}_N] = \mu$ y $\text{Var}(\hat{\mu}_N) = \sigma^2/N$. Entonces,

$$\text{Var}(\hat{\mu}_N) = E[(\hat{\mu}_N - \mu)^2] = \sigma^2/N \rightarrow 0$$

Por tanto, $\hat{\mu}_N \rightarrow_p \mu$ por nuestro “hecho útil”.

Ilustración de LLN

Distribución y media de $\frac{1}{N} \sum_i Z_i$ cuando $Z_i \sim U(0,1)$, $N = 1$

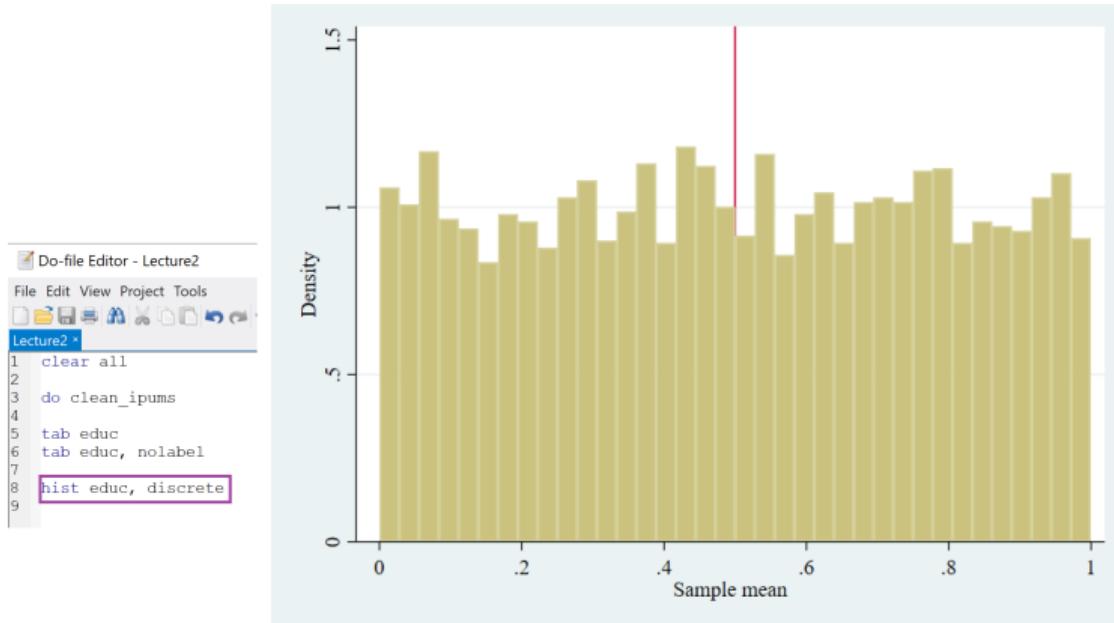


Ilustración de LLN

Distribución y media de $\frac{1}{N} \sum_i Z_i$ cuando $Z_i \sim U(0,1)$, $N = 10$

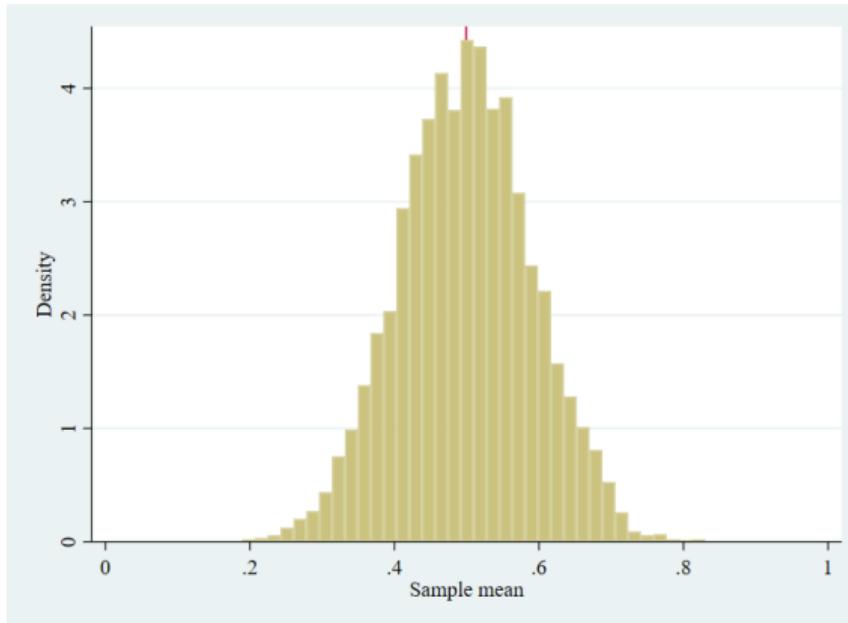
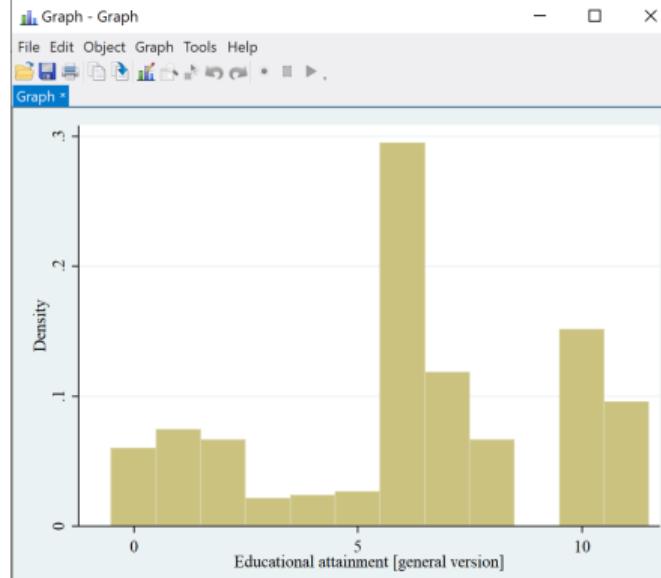
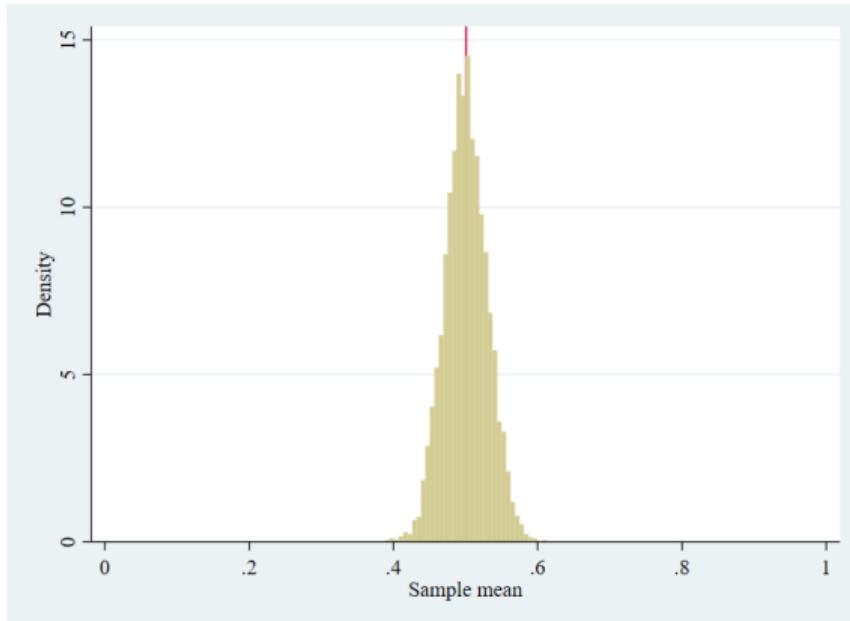


Ilustración de LLN

Distribución y media de $\frac{1}{N} \sum_i Z_i$ cuando $Z_i \sim U(0,1)$, $N = 100$

```
Do-file Editor - Lecture2
File Edit View Project Tools
Lecture2*
1 clear all
2
3 do clean_ipums
4
5 replace incwage = . ///
6     if incwage==0 | incwage==999999
7
8 cumul incwage, gen(c_incwage)
9 sort c_incwage
10 line c_incwage incwage
11
12 kdensity incwage
13
```

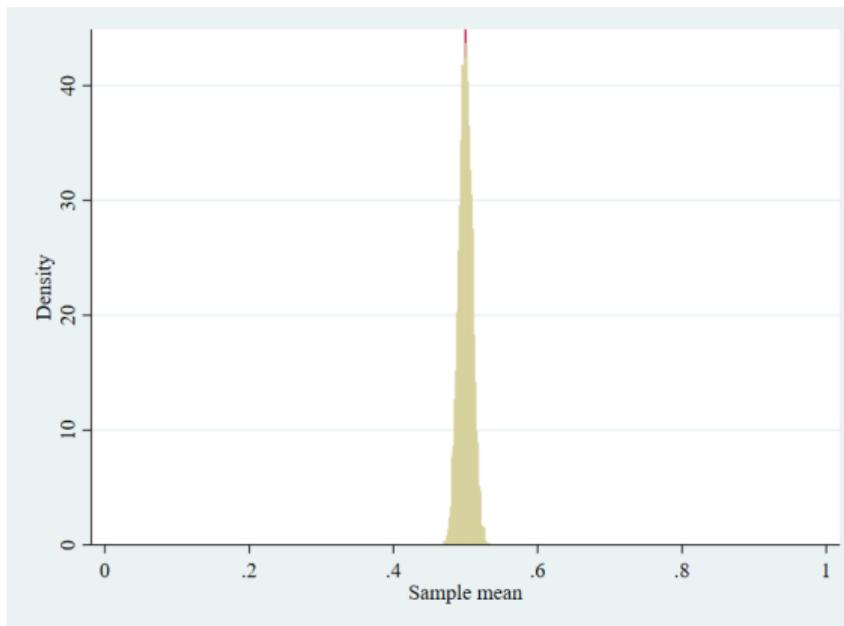


Laws of Large Numbers Illustration

Distribution and mean of $\frac{1}{N} \sum_i Z_i$ when $Z_i \sim U(0, 1)$, $N = 1000$

Do-file Editor - Lecture2

```
File Edit View Project Tools
Lecture2.x
1 clear all
2
3 do clean_ipums
4
5 replace incwage = . ///
6     if incwage==0 | incwage==999999
7
8 cumul incwage, gen(c_incwage)
9 sort c_incwage
10 line c_incwage incwage
11
12 kdensity incwage
13
```



Convergencia en Distribución

- Se habrán dado cuenta que la distribución de $\hat{\mu}$ en la simulación se aproxima a una normal cuando N se vuelve grande
- La noción de **convergencia en distribución** formaliza lo que significa que dos distribuciones se aproximen entre ellas

Convergencia en Distribución

- Se habrán dado cuenta que la distribución de $\hat{\mu}$ en la simulación se aproxima a una normal cuando N se vuelve grande
- La noción de **convergencia en distribución** formaliza lo que significa que dos distribuciones se aproximen entre ellas
- Definición: Decimos que X_N converge en distribución a una variable continuamente distribuida X , denotado por $X_n \rightarrow_d X$ o $X_n \Rightarrow X$, si el CDF de X_N converge (punto por punto - pointwise) al CDF de X ,

$$F_{X_N}(x) \rightarrow F_X(x) \text{ para cada punto } x$$

Teorema del Límite Central

- **The Central Limit Theorem (CLT)** formalizes the sense in which sample means are approximately normally distributed in large samples

Teorema del Límite Central

- **The Central Limit Theorem (CLT)** formalizes the sense in which sample means are approximately normally distributed in large samples
- Theorem: Suppose that Y_1, \dots, Y_N are drawn *iid* from a distribution with mean $\mu = E[Y_i]$ and variance $Var(Y_i) = \sigma^2 < \infty$. Then the sample mean $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N Y_i$ satisfies

$$\sqrt{N}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Teorema del Límite Central

- **The Central Limit Theorem (CLT)** formalizes the sense in which sample means are approximately normally distributed in large samples
- Theorem: Suppose that Y_1, \dots, Y_N are drawn *iid* from a distribution with mean $\mu = E[Y_i]$ and variance $Var(Y_i) = \sigma^2 < \infty$. Then the sample mean $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N Y_i$ satisfies

$$\sqrt{N}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

- In words, the theorem says the following:
 - ① We can start with any distribution Y_i , possibly non-normal
 - ② If we take the average of the Y_1, \dots, Y_N in a sample sufficiently large, the distribution of $\hat{\mu} = \frac{1}{N} \sum_i Y_i$ is (approximately) normal!

Ilustración de CLT

Distributions of $\hat{\mu} = \frac{1}{N} \sum_i X_i$ vs. $N(E[\hat{\mu}], Var(\hat{\mu}))$: $X_i \sim U(0, 1)$, $N = 1$

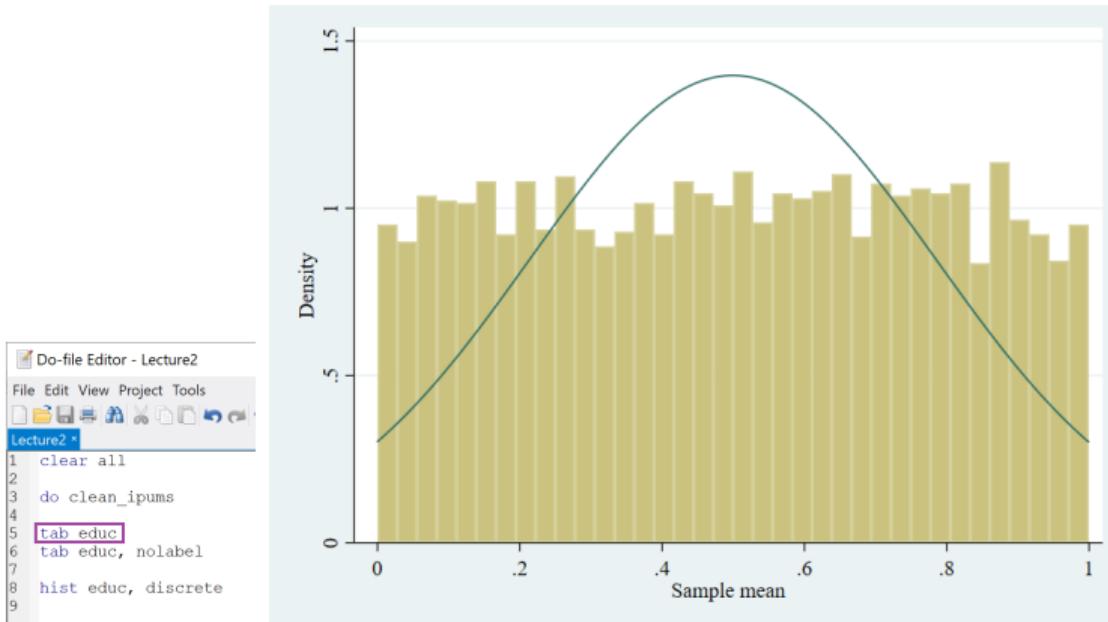


Ilustración de CLT

Distributions of $\hat{\mu} = \frac{1}{N} \sum_i X_i$ vs. $N(E[\hat{\mu}], Var(\hat{\mu}))$: $X_i \sim U(0, 1)$, $N = 2$

Educational attainment [general version]	Freq.	Percent	Cum.
N/A or no schooling	192,718	6.00	6.00
Nursery school to grade 4	239,568	7.45	13.45
Grade 5, 6, 7, or 8	213,469	6.64	20.09
Grade 9	69,213	2.15	22.24
Grade 10	76,684	2.39	24.63
Grade 11	85,612	2.66	27.29
Grade 12	948,049	29.49	56.78
1 year of college	381,192	11.86	68.64
2 years of college	213,684	6.65	75.29
4 years of college	486,531	15.14	90.42
5+ years of college	307,819	9.58	100.00
Total	3,214,539	100.00	

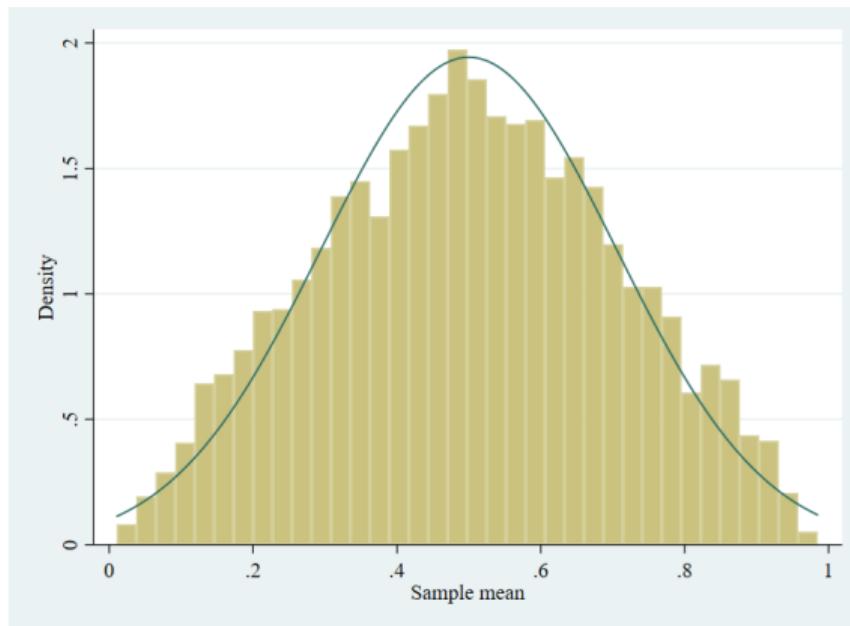


Ilustración de CLT

Distributions of $\hat{\mu} = \frac{1}{N} \sum_i X_i$ vs. $N(E[\hat{\mu}], Var(\hat{\mu}))$: $X_i \sim U(0, 1)$, $N = 5$

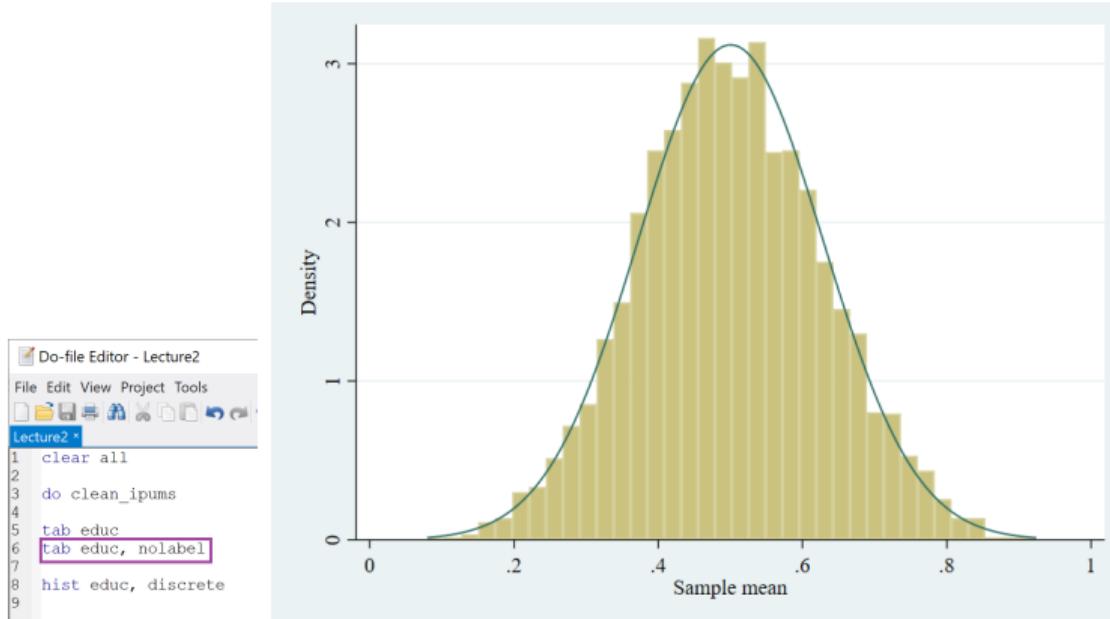
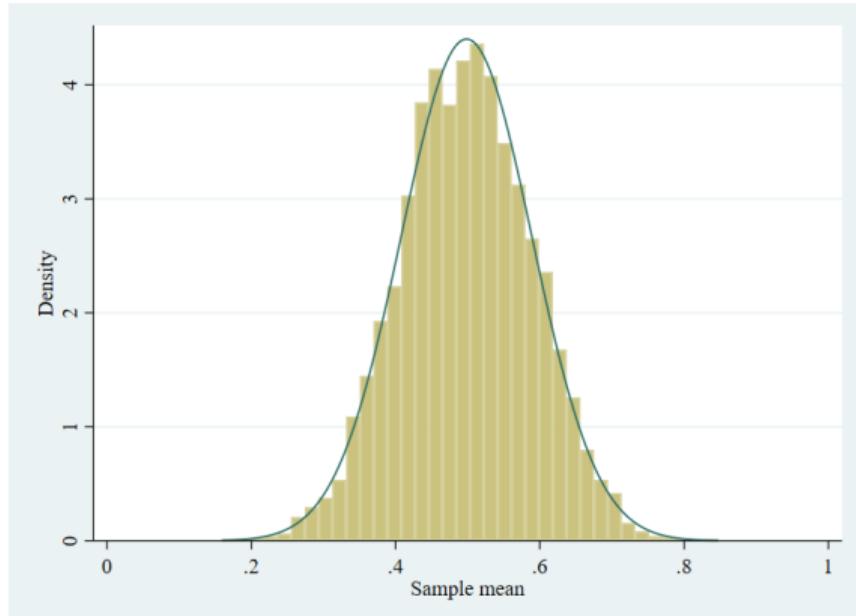


Ilustración de CLT

Distributions of $\hat{\mu} = \frac{1}{N} \sum_i X_i$ vs. $N(E[\hat{\mu}], Var(\hat{\mu}))$: $X_i \sim U(0, 1)$, $N = 10$

Educational attainment (general version)	Freq.	Percent	Cum.
0	192,718	6.00	6.00
1	239,568	7.45	13.45
2	213,469	6.64	20.09
3	69,213	2.15	22.24
4	76,684	2.39	24.63
5	85,612	2.66	27.29
6	948,049	29.49	56.78
7	381,192	11.86	68.64
8	213,684	6.65	75.29
10	486,531	15.14	90.42
11	307,819	9.58	100.00
Total	3,214,539	100.00	



No puede ser :o



Continuous Mapping Theorem

- A veces estamos interesado en funciones de la media muestral (por ej., el estadístico t es una función de $\hat{\mu}$ and σ).

Continuous Mapping Theorem

- A veces estamos interesado en funciones de la media muestral (por ej., el estadístico t es una función de $\hat{\mu}$ and σ).
- El **continuous mapping theorem** (CMT) nos dice que pasa con funciones continuas de variables aleatorias que convergen en distribución/probabilidad

Continuous Mapping Theorem

- A veces estamos interesado en funciones de la media muestral (por ej., el estadístico t es una función de $\hat{\mu}$ and σ).
- El **continuous mapping theorem** (CMT) nos dice que pasa con funciones continuas de variables aleatorias que convergen en distribución/probabilidad
- Teorema: supongamos que $g(\cdot)$ es una función continua
Si $X_N \rightarrow_p X$, entonces $g(X_N) \rightarrow_p g(X)$

Continuous Mapping Theorem

- A veces estamos interesado en funciones de la media muestral (por ej., el estadístico t es una función de $\hat{\mu}$ and σ).
- El **continuous mapping theorem** (CMT) nos dice que pasa con funciones continuas de variables aleatorias que convergen en distribución/probabilidad
- Teorema: supongamos que $g(\cdot)$ es una función continua

Si $X_N \rightarrow_p X$, entonces $g(X_N) \rightarrow_p g(X)$

Si $X_N \rightarrow_d X$, entonces $g(X_N) \rightarrow_d g(X)$

Continuous Mapping Theorem

- A veces estamos interesado en funciones de la media muestral (por ej., el estadístico t es una función de $\hat{\mu}$ and σ).
- El **continuous mapping theorem** (CMT) nos dice que pasa con funciones continuas de variables aleatorias que convergen en distribución/probabilidad
- Teorema: supongamos que $g(\cdot)$ es una función continua

Si $X_N \rightarrow_p X$, entonces $g(X_N) \rightarrow_p g(X)$

Si $X_N \rightarrow_d X$, entonces $g(X_N) \rightarrow_d g(X)$

En el caso multivariado es igual: Si $\mathbf{X}_N \rightarrow_p \mathbf{X}$, then $g(\mathbf{X}_N) \rightarrow_p g(\mathbf{X})$ y si $\mathbf{X}_N \rightarrow_d \mathbf{X}$, entonces $g(\mathbf{X}_N) \rightarrow_d g(\mathbf{X})$

Convergencia de Varianza Muestral

- Una aplicación útil del CMT es probar la convergencia en probabilidad de la varianza muestral

Convergencia de Varianza Muestral

- Una aplicación útil del CMT es probar la convergencia en probabilidad de la varianza muestral
- Escribimos $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\mu})^2$ como la varianza muestral de Y_i .

Convergencia de Varianza Muestral

- Una aplicación útil del CMT es probar la convergencia en probabilidad de la varianza muestral
- Escribimos $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\mu})^2$ como la varianza muestral de Y_i .
- Claim: si Y_1, \dots, Y_N son *iid* y $Var(Y_i^2)$ es finito, entonces $\hat{\sigma}^2 \rightarrow_p \sigma^2 = Var(Y_i)$.

Convergencia de Varianza Muestral

- Una aplicación útil del CMT es probar la convergencia en probabilidad de la varianza muestral
- Escribimos $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\mu})^2$ como la varianza muestral de Y_i .
- Claim: si Y_1, \dots, Y_N son *iid* y $Var(Y_i^2)$ es finito, entonces $\hat{\sigma}^2 \rightarrow_p \sigma^2 = Var(Y_i)$.
- Prueba:
Podemos escribir la varianza muestral como $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \hat{\mu}^2$.

Convergencia de Varianza Muestral

- Una aplicación útil del CMT es probar la convergencia en probabilidad de la varianza muestral
- Escribimos $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\mu})^2$ como la varianza muestral de Y_i .
- Claim: si Y_1, \dots, Y_N son *iid* y $Var(Y_i^2)$ es finito, entonces $\hat{\sigma}^2 \rightarrow_p \sigma^2 = Var(Y_i)$.
- Prueba:
Podemos escribir la varianza muestral como $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \hat{\mu}^2$.
Primer término: por LLN, $\frac{1}{N} \sum_{i=1}^N Y_i^2 \rightarrow_p E[Y_i^2]$.

Convergencia de Varianza Muestral

- Una aplicación útil del CMT es probar la convergencia en probabilidad de la varianza muestral
- Escribimos $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\mu})^2$ como la varianza muestral de Y_i .
- Claim: si Y_1, \dots, Y_N son *iid* y $Var(Y_i^2)$ es finito, entonces $\hat{\sigma}^2 \rightarrow_p \sigma^2 = Var(Y_i)$.
- Prueba:
Podemos escribir la varianza muestral como $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \hat{\mu}^2$.
Primer término: por LLN, $\frac{1}{N} \sum_{i=1}^N Y_i^2 \rightarrow_p E[Y_i^2]$.
Segundo término: por LLN, $\hat{\mu} \rightarrow_p \mu = E[Y_i]$. Entonces, por el CMT, $\hat{\mu}^2 \rightarrow_p E[Y_i]^2$.

Convergencia de Varianza Muestral

- Una aplicación útil del CMT es probar la convergencia en probabilidad de la varianza muestral
- Escribimos $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{\mu})^2$ como la varianza muestral de Y_i .
- Claim: si Y_1, \dots, Y_N son *iid* y $Var(Y_i^2)$ es finito, entonces $\hat{\sigma}^2 \rightarrow_p \sigma^2 = Var(Y_i)$.
- Prueba:

Podemos escribir la varianza muestral como $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \hat{\mu}^2$.

Primer término: por LLN, $\frac{1}{N} \sum_{i=1}^N Y_i^2 \rightarrow_p E[Y_i^2]$.

Segundo término: por LLN, $\hat{\mu} \rightarrow_p \mu = E[Y_i]$. Entonces, por el CMT, $\hat{\mu}^2 \rightarrow_p E[Y_i]^2$.

Aplicando nuevamente el CMT, $\frac{1}{N} \sum_{i=1}^N Y_i^2 - \hat{\mu}^2 \rightarrow_p E[Y_i^2] - E[Y_i]^2 = \sigma^2$.

Lemma de Slutsky

- **Lemma de Slutsky** (a veces también llamado Teorema) resume *casos especiales* del CMT que son útiles

Lemma de Slutsky

- **Lemma de Slutsky** (a veces también llamado Teorema) resume *casos especiales* del CMT que son útiles
- Supongamos que $X_N \rightarrow_p c$ para una constante c , y $Y_N \rightarrow_d Y$. Entonces:
- $X_N + Y_N \rightarrow_d c + Y$.

Lemma de Slutsky

- **Lemma de Slutsky** (a veces también llamado Teorema) resume *casos especiales* del CMT que son útiles
- Supongamos que $X_N \rightarrow_p c$ para una constante c , y $Y_N \rightarrow_d Y$. Entonces:
 - $X_N + Y_N \rightarrow_d c + Y$.
 - $X_n Y_n \rightarrow_d cY$.

Lemma de Slutsky

- **Lemma de Slutsky** (a veces también llamado Teorema) resume *casos especiales* del CMT que son útiles
- Supongamos que $X_N \rightarrow_p c$ para una constante c , y $Y_N \rightarrow_d Y$. Entonces:
 - $X_N + Y_N \rightarrow_d c + Y$.
 - $X_n Y_n \rightarrow_d cY$.
 - If $c \neq 0$, then $Y_n/X_n \rightarrow_d Y/c$.

Lemma de Slutsky

- **Lemma de Slutsky** (a veces también llamado Teorema) resume *casos especiales* del CMT que son útiles
- Supongamos que $X_N \rightarrow_p c$ para una constante c , y $Y_N \rightarrow_d Y$. Entonces:
 - $X_N + Y_N \rightarrow_d c + Y$.
 - $X_n Y_n \rightarrow_d cY$.
 - If $c \neq 0$, then $Y_n/X_n \rightarrow_d Y/c$.
 - Versiones análogas a esto aplican para vectores de variables aleatorias.

Test de Hipótesis Asintótico

- Recuerden que cuando $Y_i \sim N(\mu, \sigma^2)$, mostramos que el t -statistic $\hat{t} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ bajo $H_0 : \mu = \mu_0$.

Test de Hipótesis Asintótico

- Recuerden que cuando $Y_i \sim N(\mu, \sigma^2)$, mostramos que el t -statistic $\hat{t} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ bajo $H_0 : \mu = \mu_0$.
- Entonces, cuando $Y_i \sim N(\mu, \sigma^2)$, teniamos que $Pr(|\hat{t}| > 1.96) = 0.05$ bajo la nula.

Test de Hipótesis Asintótico

- Recuerden que cuando $Y_i \sim N(\mu, \sigma^2)$, mostramos que el t -statistic $\hat{t} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ bajo $H_0 : \mu = \mu_0$.
- Entonces, cuando $Y_i \sim N(\mu, \sigma^2)$, teniamos que $Pr(|\hat{t}| > 1.96) = 0.05$ bajo la nula.
- Ahora, suponemos que Y_i no es normalmente distribuida y no sabemos su varianza.

Test de Hipótesis Asintótico

- Recuerden que cuando $Y_i \sim N(\mu, \sigma^2)$, mostramos que el t -statistic $\hat{t} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ bajo $H_0 : \mu = \mu_0$.
- Entonces, cuando $Y_i \sim N(\mu, \sigma^2)$, teniamos que $Pr(|\hat{t}| > 1.96) = 0.05$ bajo la nula.
- Ahora, suponemos que Y_i no es normalmente distribuida y no sabemos su varianza.
- Por CLT, $\sqrt{N}(\hat{\mu} - \mu_0) \rightarrow_d N(0, \sigma^2)$.
Por CMT y LLN (como se mostró arriba), $\hat{\sigma} \rightarrow_p \sigma$.

Test de Hipótesis Asintótico

- Recuerden que cuando $Y_i \sim N(\mu, \sigma^2)$, mostramos que el t -statistic $\hat{t} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ bajo $H_0 : \mu = \mu_0$.
- Entonces, cuando $Y_i \sim N(\mu, \sigma^2)$, teniamos que $Pr(|\hat{t}| > 1.96) = 0.05$ bajo la nula.
- Ahora, suponemos que Y_i no es normalmente distribuida y no sabemos su varianza.
- Por CLT, $\sqrt{N}(\hat{\mu} - \mu_0) \rightarrow_d N(0, \sigma^2)$.
Por CMT y LLN (como se mostró arriba), $\hat{\sigma} \rightarrow_p \sigma$.
- Entonces, por Slutsky's, $\hat{t} = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} \rightarrow_d N(0, 1)$.

Test de Hipótesis Asintótico

- Recuerden que cuando $Y_i \sim N(\mu, \sigma^2)$, mostramos que el t -statistic $\hat{t} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$ bajo $H_0 : \mu = \mu_0$.
- Entonces, cuando $Y_i \sim N(\mu, \sigma^2)$, teniamos que $Pr(|\hat{t}| > 1.96) = 0.05$ bajo la nula.
- Ahora, suponemos que Y_i no es normalmente distribuida y no sabemos su varianza.
- Por CLT, $\sqrt{N}(\hat{\mu} - \mu_0) \rightarrow_d N(0, \sigma^2)$.
Por CMT y LLN (como se mostró arriba), $\hat{\sigma} \rightarrow_p \sigma$.
- Entonces, por Slutsky's, $\hat{t} = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}} \rightarrow_d N(0, 1)$.
- Por tanto, asintóticamente $Pr(|\hat{t}| > 1.96) \rightarrow 0.05$, incluso cuando Y_i no es normal y $\hat{\sigma}$ is estimado! Podemos hacer test de hipótesis igual que antes.

Test de Hipótesis Asintótico

- Similarmente, cuando Y_i era normal con σ conocido, mostramos que el intervalo de confianza $\hat{\mu} \pm 1.96\sigma/\sqrt{N}$ contiene el verdadero μ 95% de las veces (asumiendo como cierta la nula)

Test de Hipótesis Asintótico

- Similarmente, cuando Y_i era normal con σ conocido, mostramos que el intervalo de confianza $\hat{\mu} \pm 1.96\sigma/\sqrt{N}$ contiene el verdadero μ 95% de las veces (asumiendo como cierta la nula)
- De manera análoga, cuando Y_i es no-normal con varianza desconocida, $\hat{\mu} \pm 1.96\hat{\sigma}/\sqrt{N}$ contiene el verdadero μ con probabilidad aproximando 95% a medida que N se vuelve grande.

Example – Oregon Health Insurance Experiment

In 2008, a group of uninsured low-income adults in Oregon was selected by lottery to be given the chance to apply for Medicaid. This lottery provides an opportunity to gauge the effects of expanding access to public health insurance on the health care use, financial strain, and health of low-income adults using a randomized controlled design. In the year after random assignment, the treatment group selected by the lottery was about 25 percentage points more likely to have insurance than the control group that was not selected. We find that in this first year, the treatment group had substantively and statistically significantly higher health care utilization (including primary and preventive care as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health than the control group. *JEL* Codes: H51, H75, I1.

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

- Imaginen que queremos un CI para la media poblacional en el grupo de control

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

- Imaginen que queremos un CI para la media poblacional en el grupo de control
- Tenemos

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} =$$

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

- Imaginen que queremos un CI para la media poblacional en el grupo de control
- Tenemos

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} = 0.329 \pm 1.96 \times 0.470 / \sqrt{10426} =$$

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

- Imaginen que queremos un CI para la media poblacional en el grupo de control
- Tenemos

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} = 0.329 \pm 1.96 \times 0.470 / \sqrt{10426} = [0.319, 0.338]$$

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

- Imaginen que queremos un CI para la media poblacional en el grupo de control
- Tenemos

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} = 0.329 \pm 1.96 \times 0.470 / \sqrt{10426} = [0.319, 0.338]$$

- ¿Y para el grupo de tratamiento?

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

- Imaginen que queremos un CI para la media poblacional en el grupo de control
- Tenemos

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} = 0.329 \pm 1.96 \times 0.470 / \sqrt{10426} = [0.319, 0.338]$$

- ¿Y para el grupo de tratamiento?

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} =$$

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

- Imaginen que queremos un CI para la media poblacional en el grupo de control
- Tenemos

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} = 0.329 \pm 1.96 \times 0.470 / \sqrt{10426} = [0.319, 0.338]$$

- ¿Y para el grupo de tratamiento?

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} = 0.306 \pm 1.96 \times 0.461 / \sqrt{13315} =$$

Ejemplo: acceso a salud y depresión

	Grupo Control	Grupo Tratamiento
Media	0.329	0.306
DesvEst	0.470	0.461
N	10426	13315

- Imaginen que queremos un CI para la media poblacional en el grupo de control
- Tenemos

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} = 0.329 \pm 1.96 \times 0.470 / \sqrt{10426} = [0.319, 0.338]$$

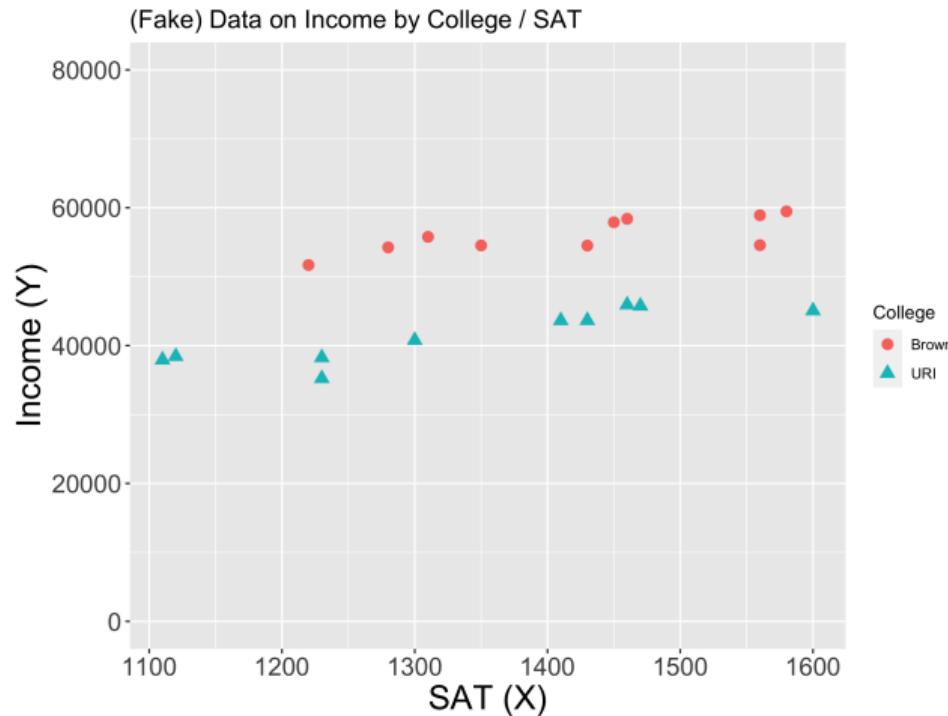
- ¿Y para el grupo de tratamiento?

$$\hat{\mu} \pm 1.96 \times \hat{\sigma} / \sqrt{N} = 0.306 \pm 1.96 \times 0.461 / \sqrt{13315} = [0.298, 0.313]$$

Outline

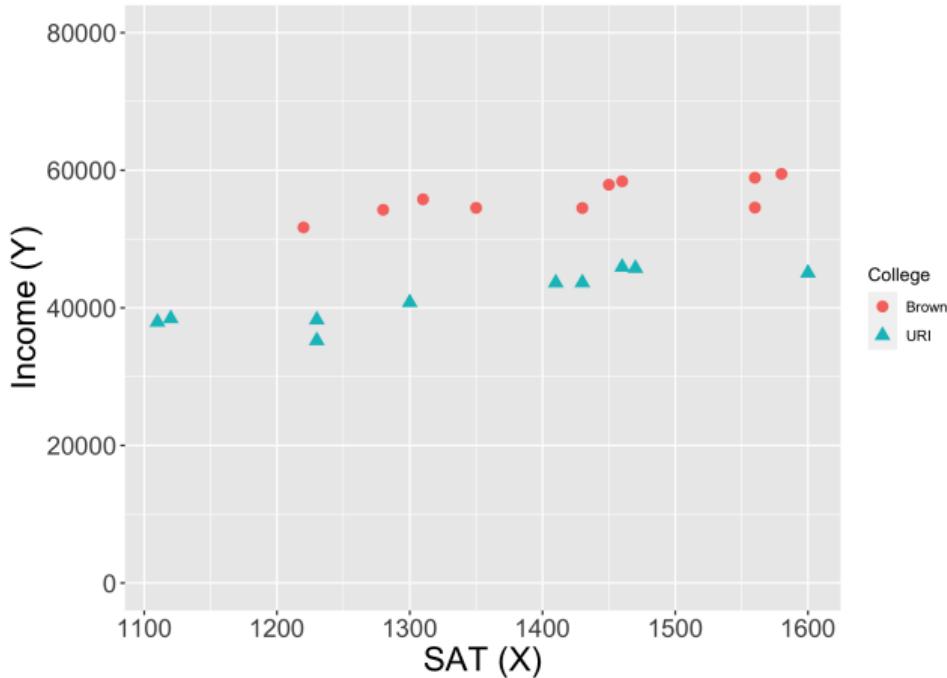
5. Regresión Lineal

6. Regularización

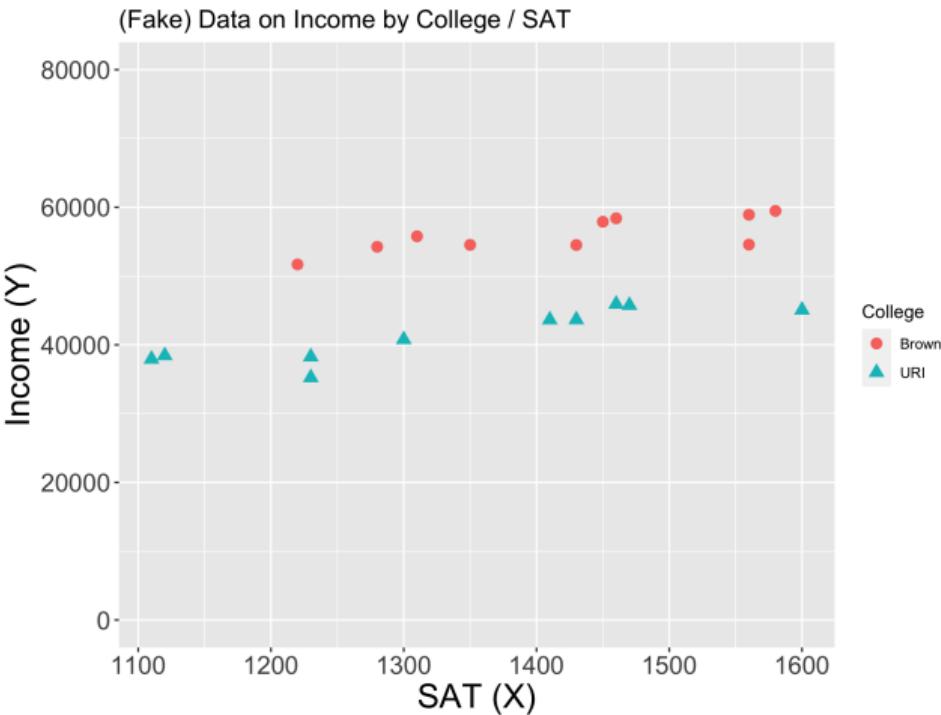


- Imagina que es nuestra data

(Fake) Data on Income by College / SAT

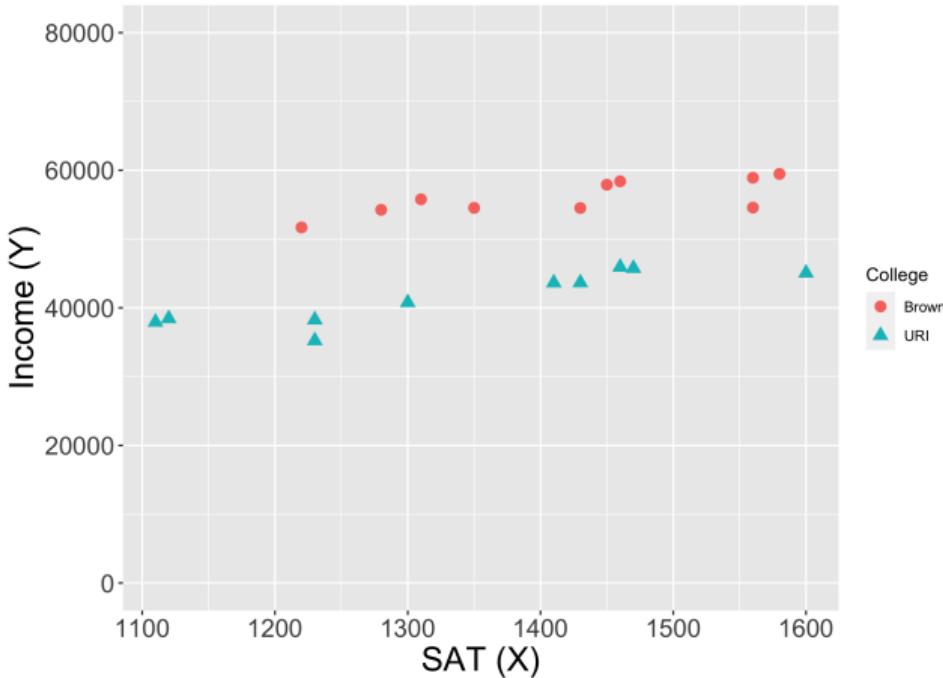


- Estamos dispuestos a asumir que el puntaje del examen SAT es casi tan bueno como aleatorio

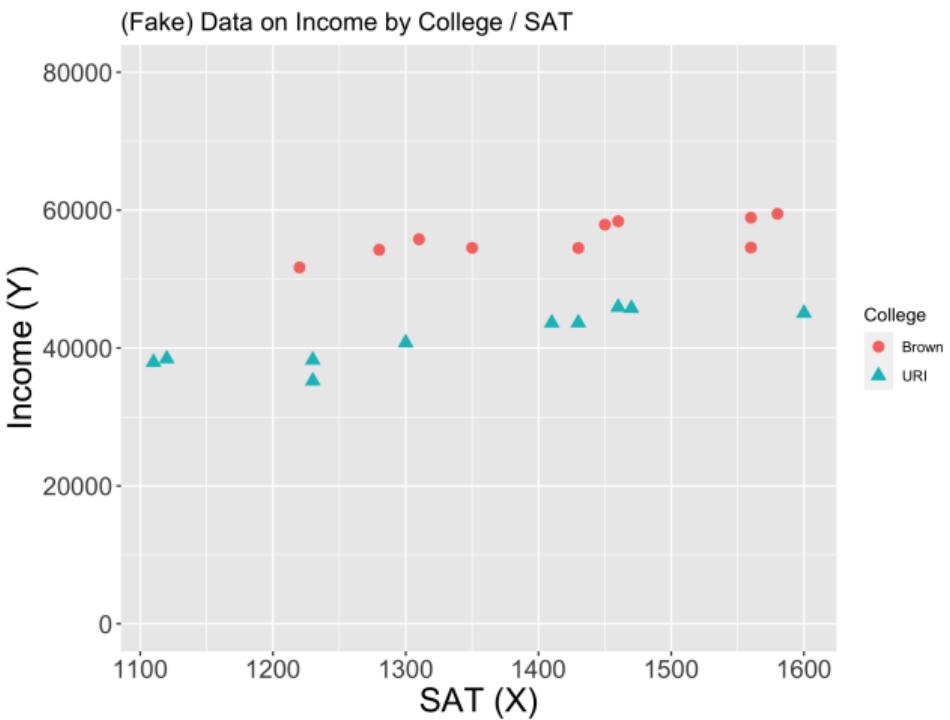


- Imagina que estamos interesados en el efecto (condicional) en $X = 1350$. La teoría nos diría que comparemos colegios con puntaje $X = 1350$.

(Fake) Data on Income by College / SAT

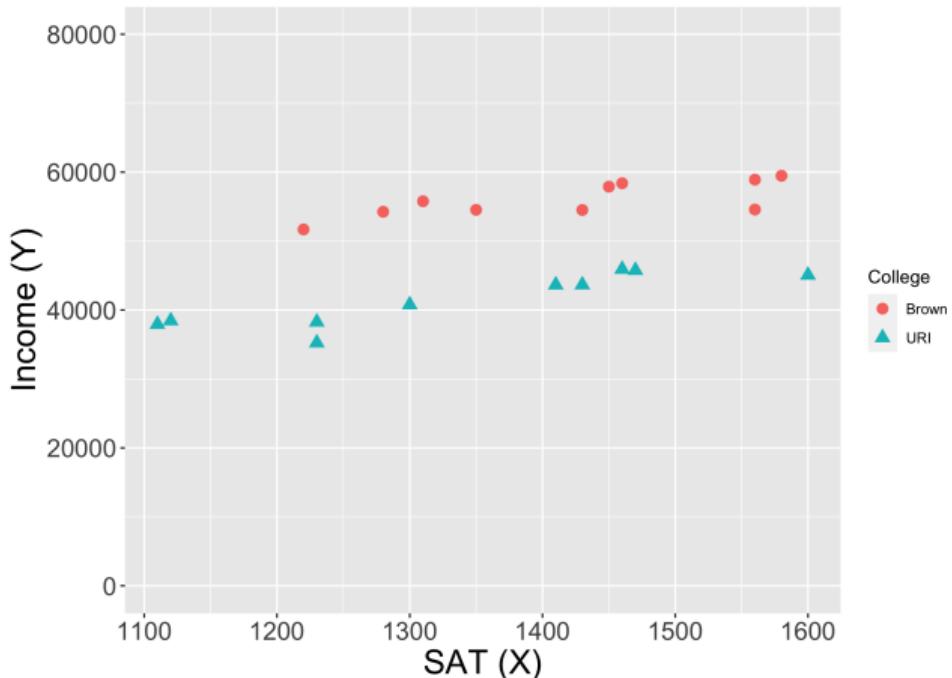


- Podemos estimar el efecto promedio usando la única observación que tenemos para Brown en $X = 1350$. Este estimado es muy ruidoso, & no podemos usar CLT



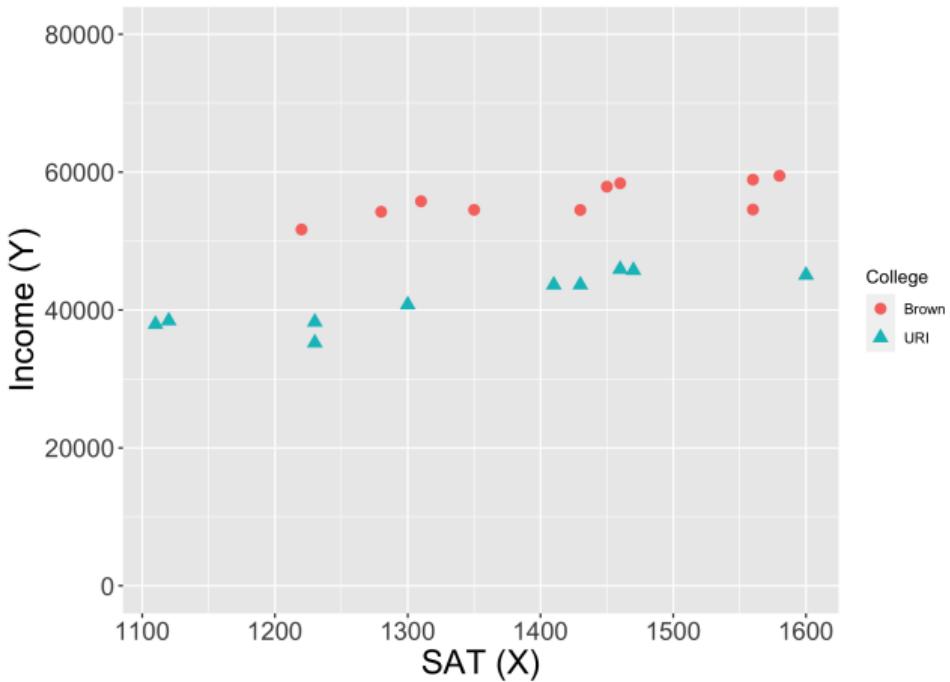
- Podemos estimar el efecto promedio usando la única observación que tenemos para Brown en $X = 1350$. Este estimado es muy ruidoso, & no podemos usar CLT
- Peor aún, no tenemos alumnos de URI con $X = 1350!$

(Fake) Data on Income by College / SAT



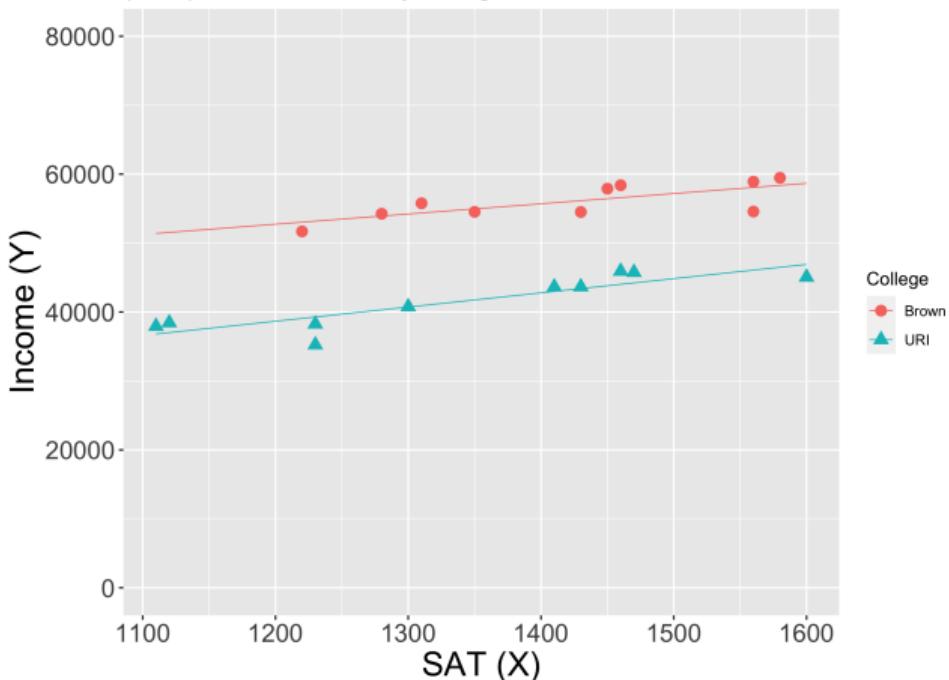
- ¡¡Necesitamos extrapolar!!

(Fake) Data on Income by College / SAT

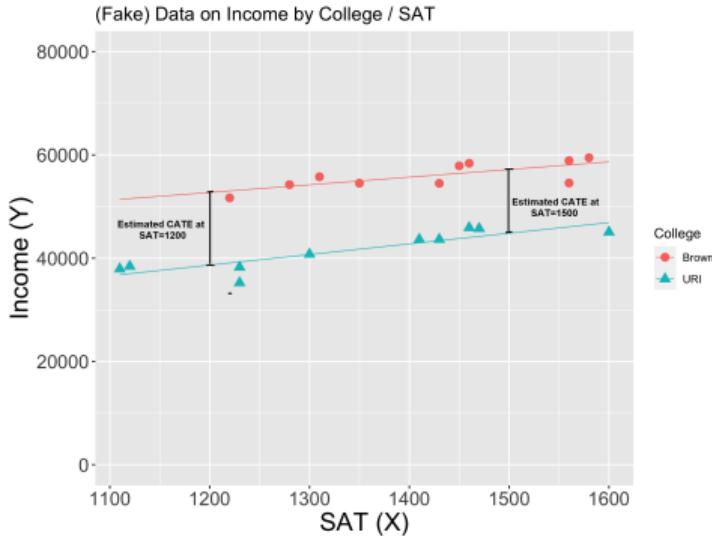


- ¡¡Necesitamos extrapolar!!
- ¿Qué haríamos más o menos al ojo?

(Fake) Data on Income by College / SAT



- ¡¡Necesitamos extrapolar!!
- ¿Qué haríamos más o menos al ojo?
- Dibujar quizás una linea que represente bien en promedio



- Con esto podemos estimar el efecto promedio $CATE(x)$ en cualquier x

Introducción a la Regresión

- La idea de **regresión** es formalizar el proceso de estimar la función de esperanza condicional (CEF) al extrapolar a través de unidades usando una forma de función específica (por ej. linear, quadratic, etc.)

Introducción a la Regresión

- La idea de **regresión** es formalizar el proceso de estimar la función de esperanza condicional (CEF) al extrapolar a través de unidades usando una forma de función específica (por ej. linear, quadratic, etc.)
- Falta responder:

Introducción a la Regresión

- La idea de **regresión** es formalizar el proceso de estimar la función de esperanza condicional (CEF) al extrapolar a través de unidades usando una forma de función específica (por ej. linear, quadratic, etc.)
- Falta responder:
- ¿Cómo aproximar el CEF en la muestra que tenemos? (I.e. como hiciste ese dibujito!)

Introducción a la Regresión

- La idea de **regresión** es formalizar el proceso de estimar la función de esperanza condicional (CEF) al extrapolar a través de unidades usando una forma de función específica (por ej. linear, quadratic, etc.)
- Falta responder:
- ¿Cómo aproximar el CEF en la muestra que tenemos? (I.e. como hiciste ese dibujito!)
- ¿Cómo construir intervalos de confianza/hipótesis test para estimados del CEF?

Introducción a la Regresión

- La idea de **regresión** es formalizar el proceso de estimar la función de esperanza condicional (CEF) al extrapolar a través de unidades usando una forma de función específica (por ej. linear, quadratic, etc.)
- Falta responder:
- ¿Cómo aproximar el CEF en la muestra que tenemos? (I.e. como hiciste ese dibujito!)
- ¿Cómo construir intervalos de confianza/hipótesis test para estimados del CEF?
- ¿Qué pasa si el CEF no es lineal?

Introducción a la Regresión

- La idea de **regresión** es formalizar el proceso de estimar la función de esperanza condicional (CEF) al extrapolar a través de unidades usando una forma de función específica (por ej. linear, quadratic, etc.)
- Falta responder:
- ¿Cómo aproximar el CEF en la muestra que tenemos? (I.e. como hiciste ese dibujito!)
- ¿Cómo construir intervalos de confianza/hipótesis test para estimados del CEF?
- ¿Qué pasa si el CEF no es lineal?
- Ok, trataremos de contestar estas cosas

Camino

- **Lo que sabemos hacer:** Estimar y hacer test de hipótesis de medias poblacionales usando promedios muestrales
- **Queremos hacer:** Estimar aproximación de CEF y hacer test de hipótesis

Camino

- **Lo que sabemos hacer:** Estimar y hacer test de hipótesis de medias poblacionales usando promedios muestrales
- **Queremos hacer:** Estimar aproximación de CEF y hacer test de hipótesis

¿Cómo empezar?

Camino

- **Lo que sabemos hacer:** Estimar y hacer test de hipótesis de medias poblacionales usando promedios muestrales
- **Queremos hacer:** Estimar aproximación de CEF y hacer test de hipótesis

¿Cómo empezar?

- 1) Asumir CEF toma cierta forma, por ej. lineal:

$$E[Y_i|X_i = x] = \alpha + x\beta$$

Camino

- **Lo que sabemos hacer:** Estimar y hacer test de hipótesis de medias poblacionales usando promedios muestrales
- **Queremos hacer:** Estimar aproximación de CEF y hacer test de hipótesis

¿Cómo empezar?

- 1) Asumir CEF toma cierta forma, por ej. lineal:

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- 2) Mostrar que, α and β can se pueden representar como medias poblacionales

Camino

- **Lo que sabemos hacer:** Estimar y hacer test de hipótesis de medias poblacionales usando promedios muestrales
- **Queremos hacer:** Estimar aproximación de CEF y hacer test de hipótesis

¿Cómo empezar?

- 1) Asumir CEF toma cierta forma, por ej. lineal:

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- 2) Mostrar que, α and β can se pueden representar como medias poblacionales
- 3) Estimar eso con promedio muestral $\hat{\alpha}, \hat{\beta}$ y hacer test de hipótesis

Camino

- **Lo que sabemos hacer:** Estimar y hacer test de hipótesis de medias poblacionales usando promedios muestrales
- **Queremos hacer:** Estimar aproximación de CEF y hacer test de hipótesis

¿Cómo empezar?

- 1) Asumir CEF toma cierta forma, por ej. lineal:

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- 2) Mostrar que, α and β can se pueden representar como medias poblacionales
- 3) Estimar eso con promedio muestral $\hat{\alpha}, \hat{\beta}$ y hacer test de hipótesis
- 4) Argumentar que incluso si CEF no es lineal, α y β proveen una buena aproximación

El “Least Squares” Problem

- Suponer que X_i es escalar y CEF lineal (se puede relajar luego):

$$E[Y_i|X_i = x] = \alpha + x\beta$$

El “Least Squares” Problem

- Suponer que X_i es escalar y CEF lineal (se puede relajar luego):

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- Algo útil que se puede explotar es que si la ecuación de arriba es cierta, entonces (α, β) solve the “least squares” problem:

$$(\alpha, \beta) = \arg \min_{a,b} E[(Y_i - (a + bX_i))^2]$$

El “Least Squares” Problem

- Suponer que X_i es escalar y CEF lineal (se puede relajar luego):

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- Algo útil que se puede explotar es que si la ecuación de arriba es cierta, entonces (α, β) solve the “least squares” problem:

$$(\alpha, \beta) = \arg \min_{a,b} E[(Y_i - (a + bX_i))^2]$$

- ¿De dónde viene esto?

Un problema más simple

- Antes de mostrar que (α, β) resuelve el “least-squares” problem, consideremos un problema más sencillo:

Un problema más simple

- Antes de mostrar que (α, β) resuelve el “least-squares” problem, consideremos un problema más sencillo:
- Imaginen que queremos encontrar una constante u para minimizar

$$\min_u E[(Y_i - u)^2]$$

Un problema más simple

- Antes de mostrar que (α, β) resuelve el “least-squares” problem, consideremos un problema más sencillo:
- Imaginen que queremos encontrar una constante u para minimizar

$$\min_u E[(Y_i - u)^2]$$

- ¿Qué constante u deberíamos elegir?

Un problema más simple

- Antes de mostrar que (α, β) resuelve el “least-squares” problem, consideremos un problema más sencillo:
- Imaginen que queremos encontrar una constante u para minimizar

$$\min_u E[(Y_i - u)^2]$$

- ¿Qué constante u deberíamos elegir? La media poblacional $\mu = E[Y_i]$!

Un problema más simple

- Antes de mostrar que (α, β) resuelve el “least-squares” problem, consideremos un problema más sencillo:
- Imaginen que queremos encontrar una constante u para minimizar

$$\min_u E[(Y_i - u)^2]$$

- ¿Qué constante u deberíamos elegir? La media poblacional $\mu = E[Y_i]$!
- Prueba:
La derivada de $E[(Y_i - u)^2]$ con respecto a u es $E[2(Y_i - u)]$.

Un problema más simple

- Antes de mostrar que (α, β) resuelve el “least-squares” problem, consideremos un problema más sencillo:
- Imaginen que queremos encontrar una constante u para minimizar

$$\min_u E[(Y_i - u)^2]$$

- ¿Qué constante u deberíamos elegir? La media poblacional $\mu = E[Y_i]$!
- Prueba:
La derivada de $E[(Y_i - u)^2]$ con respecto a u es $E[2(Y_i - u)]$.
Igualando la derivada a cero nos da

$$E[2(Y_i - \mu)] = 0 \Rightarrow 2E[Y_i] = 2\mu \Rightarrow u = E[Y_i].$$

Un problema ligeramente más difícil

- Ahora queremos una función de predicción que dependa de x , $u(x)$, que minimice

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

Un problema ligeramente más difícil

- Ahora queremos una función de predicción que dependa de x , $u(x)$, que minimice

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- ¿Qué función $u(x)$ usar?

Un problema ligeramente más difícil

- Ahora queremos una función de predicción que dependa de x , $u(x)$, que minimice

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- ¿Qué función $u(x)$ usar? La esperanza condicional $u(x) = E[Y|X = x]$.

Un problema ligeramente más difícil

- Ahora queremos una función de predicción que dependa de x , $u(x)$, que minimice

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- ¿Qué función $u(x)$ usar? La esperanza condicional $u(x) = E[Y|X = x]$.
- Prueba:
Por LIE,

$$E[(Y_i - u(X_i))^2] = E[E[(Y_i - u(X_i))^2 | X_i]].$$

Un problema ligeramente más difícil

- Ahora queremos una función de predicción que dependa de x , $u(x)$, que minimice

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- ¿Qué función $u(x)$ usar? La esperanza condicional $u(x) = E[Y|X = x]$.
- Prueba:
Por LIE,

$$E[(Y_i - u(X_i))^2] = E[E[(Y_i - u(X_i))^2 | X_i]].$$

Entonces, para cada valor x , queremos elegir $u(x)$ tal que minimice

$$E[(Y_i - u(x))^2 | X_i = x].$$

Un problema ligeramente más difícil

- Ahora queremos una función de predicción que dependa de x , $u(x)$, que minimice

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- ¿Qué función $u(x)$ usar? La esperanza condicional $u(x) = E[Y|X = x]$.
- Prueba:
Por LIE,

$$E[(Y_i - u(X_i))^2] = E[E[(Y_i - u(X_i))^2 | X_i]].$$

Entonces, para cada valor x , queremos elegir $u(x)$ tal que minimice

$$E[(Y_i - u(x))^2 | X_i = x].$$

Sin embargo, nuestro argumento en la slide anterior implica que la solución es dada por $u(x) = E[Y_i | X_i = x]$.

Regresando...

- Hemos mostrado que $u(x) = E[Y_i|X_i = x]$ resuelve

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

Regresando...

- Hemos mostrado que $u(x) = E[Y_i|X_i = x]$ resuelve

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- Entonces, si $E[Y_i|X_i = x] = \alpha + \beta x$, entonces $u(x) = \alpha + \beta x$ minimiza

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

Regresando...

- Hemos mostrado que $u(x) = E[Y_i|X_i = x]$ resuelve

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- Entonces, si $E[Y_i|X_i = x] = \alpha + \beta x$, entonces $u(x) = \alpha + \beta x$ minimiza

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- La minimización de arriba es sobre *todas* las funciones $u(\cdot)$, incluyendo las lineales $a + bx$.
Por tanto,

$$E[(Y_i - (\alpha + \beta X_i))^2] \leq E[(Y_i - (a + bX_i))^2] \text{ para todo } a, b.$$

Regresando...

- Hemos mostrado que $u(x) = E[Y_i|X_i = x]$ resuelve

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- Entonces, si $E[Y_i|X_i = x] = \alpha + \beta x$, entonces $u(x) = \alpha + \beta x$ minimiza

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- La minimización de arriba es sobre *todas* las funciones $u(\cdot)$, incluyendo las lineales $a + bx$.
Por tanto,

$$E[(Y_i - (\alpha + \beta X_i))^2] \leq E[(Y_i - (a + b X_i))^2] \text{ para todo } a, b.$$

- Esto implica que (α, β) resuelve

$$\min_{a,b} E[(Y_i - (a + b X_i))^2],$$

tal como queríamos mostrar

¿Por qué es útil esto?

- Hemos mostrado que α, β solucionan

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- ¿En qué ayuda esto?

¿Por qué es útil esto?

- Hemos mostrado que α, β solucionan

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- ¿En qué ayuda esto? Resolviendo esto podemos expresar α, β como funciones de la población

¿Por qué es útil esto?

- Hemos mostrado que α, β solucionan

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- ¿En qué ayuda esto? Resolviendo esto podemos expresar α, β como funciones de la población
- Derivemos con respecto a a y b e igualamos a cero en ambos (α, β):

¿Por qué es útil esto?

- Hemos mostrado que α, β solucionan

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- ¿En qué ayuda esto? Resolviendo esto podemos expresar α, β como funciones de la población
- Derivemos con respecto a a y b e igualamos a cero en ambos (α, β):

$$E[-2(Y_i - (\alpha + \beta X_i))] = 0$$

¿Por qué es útil esto?

- Hemos mostrado que α, β solucionan

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- ¿En qué ayuda esto? Resolviendo esto podemos expresar α, β como funciones de la población
- Derivemos con respecto a a y b e igualamos a cero en ambos (α, β):

$$E[-2(Y_i - (\alpha + \beta X_i))] = 0$$

$$E[-2X_i(Y_i - (\alpha + \beta X_i))] = 0$$

¿Por qué es útil esto?

- Hemos mostrado que α, β solucionan

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- ¿En qué ayuda esto? Resolviendo esto podemos expresar α, β como funciones de la población
- Derivemos con respecto a a y b e igualamos a cero en ambos (α, β) :

$$E[-2(Y_i - (\alpha + \beta X_i))] = 0$$

$$E[-2X_i(Y_i - (\alpha + \beta X_i))] = 0$$

- Tenemos 2 ecuaciones, 2 incógnitas, podemos resolver para (α, β)

La solución Least Squares

- La solución al sistema es

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)}$$

$$\alpha = E[Y_i] - E[X_i]\beta$$

La solución Least Squares

- La solución al sistema es

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)}$$

$$\alpha = E[Y_i] - E[X_i]\beta$$

- Son funciones continuas de medias poblacionales!

La solución Least Squares

- La solución al sistema es

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)}$$
$$\alpha = E[Y_i] - E[X_i]\beta$$

- Son funciones continuas de medias poblacionales!
- Podemos usar las técnicas anteriores para estimar y hacer test de hipótesis sobre el CEF!

Estimando Coeficientes de Regresión

- Mostramos que $E[Y_i | X_i = x] = \alpha + \beta x$

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)}$$
$$\alpha = E[Y_i] - E[X_i]\beta$$

- ¿Cómo estimar α, β ?

Estimando Coeficientes de Regresión

- Mostramos que $E[Y_i | X_i = x] = \alpha + \beta x$

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)}$$
$$\alpha = E[Y_i] - E[X_i]\beta$$

- ¿Cómo estimar α, β ?
Medias poblacionales por promedios muestrales

Estimando Coeficientes de Regresión

- Mostramos que $E[Y_i | X_i = x] = \alpha + \beta x$

$$\begin{aligned}\beta &= \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])] }{E[(X_i - E[X_i])^2]} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} \\ \alpha &= E[Y_i] - E[X_i]\beta\end{aligned}$$

- ¿Cómo estimar α, β ?

Medias poblacionales por promedios muestrales

$$\hat{\beta} = \frac{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2} = \frac{\widehat{\text{Cov}}(X_i, Y_i)}{\widehat{\text{Var}}(X_i)}$$

Estimando Coeficientes de Regresión

- Mostramos que $E[Y_i | X_i = x] = \alpha + \beta x$

$$\begin{aligned}\beta &= \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])] }{E[(X_i - E[X_i])^2]} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} \\ \alpha &= E[Y_i] - E[X_i]\beta\end{aligned}$$

- ¿Cómo estimar α, β ?

Medias poblacionales por promedios muestrales

$$\begin{aligned}\hat{\beta} &= \frac{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2} = \frac{\widehat{\text{Cov}}(X_i, Y_i)}{\widehat{\text{Var}}(X_i)} \\ \hat{\alpha} &= \bar{Y} - \bar{X}\hat{\beta}\end{aligned}$$

Estimando Coeficientes de Regresión

- Mostramos que $E[Y_i | X_i = x] = \alpha + \beta x$

$$\begin{aligned}\beta &= \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])] }{E[(X_i - E[X_i])^2]} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} \\ \alpha &= E[Y_i] - E[X_i]\beta\end{aligned}$$

- ¿Cómo estimar α, β ?

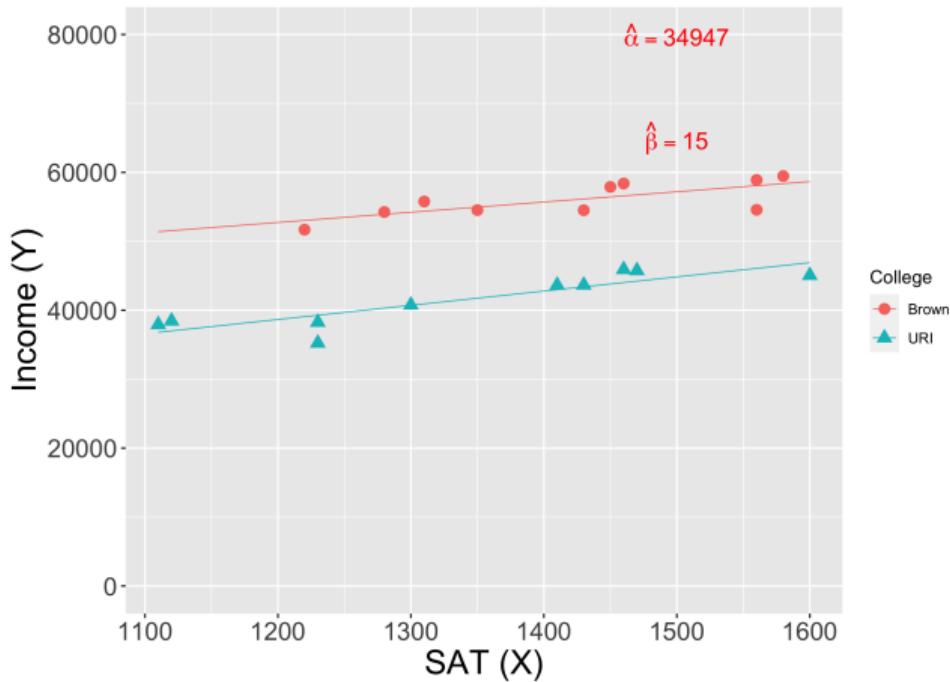
Medias poblacionales por promedios muestrales

$$\begin{aligned}\hat{\beta} &= \frac{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2} = \frac{\widehat{\text{Cov}}(X_i, Y_i)}{\widehat{\text{Var}}(X_i)} \\ \hat{\alpha} &= \bar{Y} - \bar{X}\hat{\beta}\end{aligned}$$

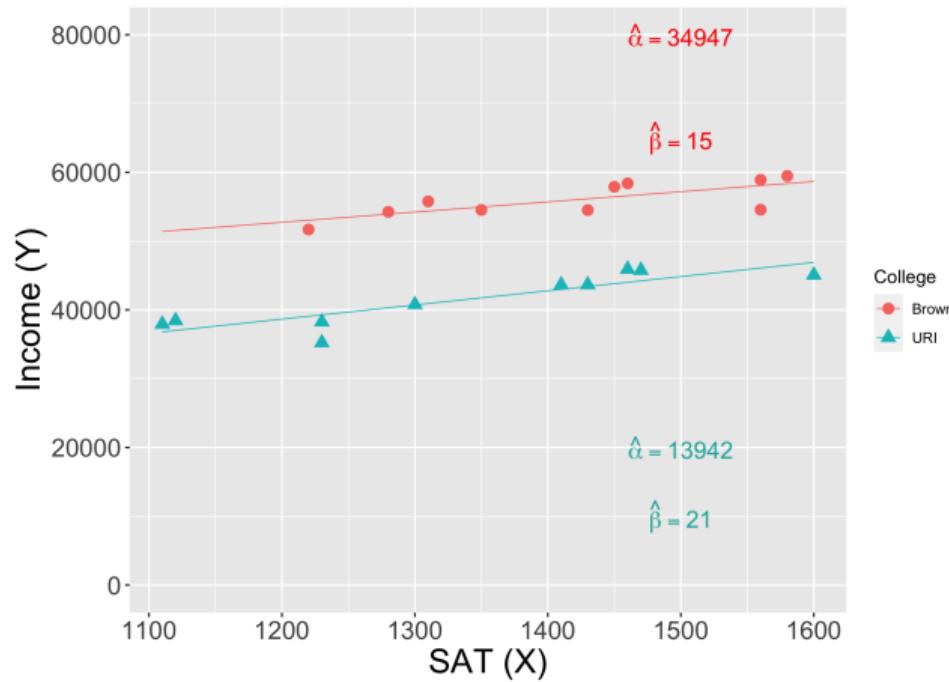
- Estos $\hat{\alpha}, \hat{\beta}$ se llaman coeficientes de *ordinary least squares* (OLS)

- Resuelven el problema análogo en la muestra, $\min_{a,b} \frac{1}{N} \sum_i (Y_i - (a + bX_i))^2$

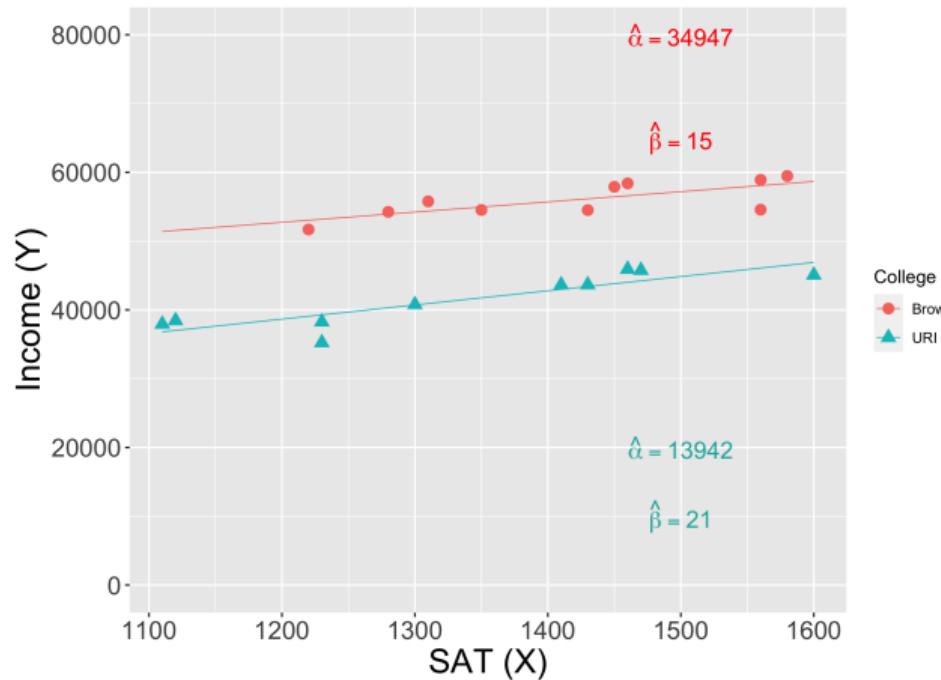
(Fake) Data on Income by College / SAT



(Fake) Data on Income by College / SAT

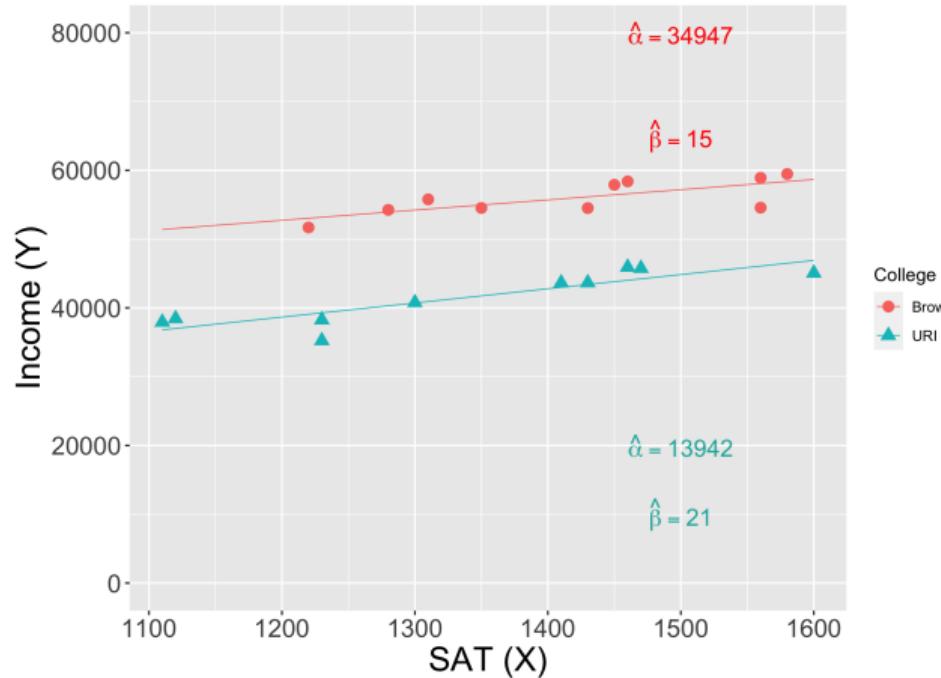


(Fake) Data on Income by College / SAT



- ¿Cuál es el valor estimado de $E[Y_i|D_i = 1, X_i = 1350]$?

(Fake) Data on Income by College / SAT



- ¿Cuál es el valor estimado de $E[Y_i|D_i = 1, X_i = 1350]$?
 $\hat{\alpha} + \hat{\beta} \cdot 1350 = 34947 + 15 \cdot 1350 = 55197.$

Consistencia de OLS

- Podemos usar los resultados que sabemos anteriormente para mostrar que $\hat{\beta}$ es consistente para β , i.e. $\hat{\beta} \xrightarrow{p} \beta$.

Consistencia de OLS

- Podemos usar los resultados que sabemos anteriormente para mostrar que $\hat{\beta}$ es consistente para β , i.e. $\hat{\beta} \xrightarrow{p} \beta$.
- Tenemos que

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

Consistencia de OLS

- Podemos usar los resultados que sabemos anteriormente para mostrar que $\hat{\beta}$ es consistente para β , i.e. $\hat{\beta} \rightarrow_p \beta$.
- Tenemos que

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N X_i^2 - \bar{X}^2 \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N X_i Y_i - \bar{X} \bar{Y} \right)\end{aligned}$$

Consistencia de OLS

- Podemos usar los resultados que sabemos anteriormente para mostrar que $\hat{\beta}$ es consistente para β , i.e. $\hat{\beta} \xrightarrow{p} \beta$.
- Tenemos que

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N X_i^2 - \bar{X}^2 \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N X_i Y_i - \bar{X} \bar{Y} \right) \\ &\xrightarrow{p} (E[X_i^2] - E[X_i]^2)^{-1} (E[X_i Y_i] - E[X_i] E[Y_i])\end{aligned}$$

Consistencia de OLS

- Podemos usar los resultados que sabemos anteriormente para mostrar que $\hat{\beta}$ es consistente para β , i.e. $\hat{\beta} \xrightarrow{p} \beta$.
- Tenemos que

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N X_i^2 - \bar{X}^2 \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N X_i Y_i - \bar{X} \bar{Y} \right) \\ &\xrightarrow{p} (E[X_i^2] - E[X_i]^2)^{-1} (E[X_i Y_i] - E[X_i] E[Y_i]) \\ &= Var(X_i)^{-1} Cov(X_i, Y_i)\end{aligned}$$

Consistencia de OLS

- Podemos usar los resultados que sabemos anteriormente para mostrar que $\hat{\beta}$ es consistente para β , i.e. $\hat{\beta} \xrightarrow{p} \beta$.
- Tenemos que

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N X_i^2 - \bar{X}^2 \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N X_i Y_i - \bar{X} \bar{Y} \right) \\ &\xrightarrow{p} (E[X_i^2] - E[X_i]^2)^{-1} (E[X_i Y_i] - E[X_i] E[Y_i]) \\ &= Var(X_i)^{-1} Cov(X_i, Y_i) = \beta\end{aligned}$$

Consistencia de OLS

- Podemos usar los resultados que sabemos anteriormente para mostrar que $\hat{\beta}$ es consistente para β , i.e. $\hat{\beta} \rightarrow_p \beta$.
- Tenemos que

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N X_i^2 - \bar{X}^2 \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^N X_i Y_i - \bar{X} \bar{Y} \right) \\ &\rightarrow_p (E[X_i^2] - E[X_i]^2)^{-1} (E[X_i Y_i] - E[X_i] E[Y_i]) \\ &= Var(X_i)^{-1} Cov(X_i, Y_i) = \beta\end{aligned}$$

- De la misma manera, podemos mostrar que $\hat{\alpha} \rightarrow_p \alpha$.

Distribución Asintótica de OLS

- Los estimados $\hat{\alpha}, \hat{\beta}$ son funciones continuas de medias muestrales

Distribución Asintótica de OLS

- Los estimados $\hat{\alpha}, \hat{\beta}$ son funciones continuas de medias muestrales
- Podemos usar el CLT y CMT para mostrar que están asintóticamente normalmente distribuidas

Distribución Asintótica de OLS

- Los estimados $\hat{\alpha}, \hat{\beta}$ son funciones continuas de medias muestrales
- Podemos usar el CLT y CMT para mostrar que están asintóticamente normalmente distribuidas
- En particular, podemos mostrar que

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2),$$

donde

$$\sigma^2 = \frac{Var((X_i - E[X_i])\varepsilon_i)}{Var(X_i)^2}$$

Distribución Asintótica de OLS

- Los estimados $\hat{\alpha}, \hat{\beta}$ son funciones continuas de medias muestrales
- Podemos usar el CLT y CMT para mostrar que están asintóticamente normalmente distribuidas
- En particular, podemos mostrar que

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2),$$

donde

$$\sigma^2 = \frac{\text{Var}((X_i - E[X_i])\varepsilon_i)}{\text{Var}(X_i)^2}$$

- Esto es útil porque podemos crear CIs para β de la forma $\hat{\beta} \pm 1.96\hat{\sigma}/\sqrt{N}$, donde $\hat{\sigma}$ es nuestro valor estimado de σ .

Derivando Distribución Asintótica de OLS

- Define el **residual** $\varepsilon_i = Y_i - (\alpha + X_i\beta)$, implicando

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

Derivando Distribución Asintótica de OLS

- Define el **residual** $\varepsilon_i = Y_i - (\alpha + X_i\beta)$, implicando

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

- Las condiciones de primer orden de (α, β) implican que el residual tiene media cero, y es **ortogonal to the regressor**: $E[\varepsilon_i] = E[X_i\varepsilon_i] = 0$

Derivando Distribución Asintótica de OLS

- Define el **residual** $\varepsilon_i = Y_i - (\alpha + X_i\beta)$, implicando

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

- Las condiciones de primer orden de (α, β) implican que el residual tiene media cero, y es **ortogonal to the regressor**: $E[\varepsilon_i] = E[X_i\varepsilon_i] = 0$
- Tomando promedios, $\bar{Y} = \alpha + \bar{X}\beta + \bar{\varepsilon}$.

Derivando Distribución Asintótica de OLS

- Define el **residual** $\varepsilon_i = Y_i - (\alpha + X_i\beta)$, implicando

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

- Las condiciones de primer orden de (α, β) implican que el residual tiene media cero, y es **ortogonal to the regressor**: $E[\varepsilon_i] = E[X_i\varepsilon_i] = 0$
- Tomando promedios, $\bar{Y} = \alpha + \bar{X}\beta + \bar{\varepsilon}$. Entonces $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$

Derivando Distribución Asintótica de OLS (cont.)

- Acabamos de derivar que $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$.
- Entonces,

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

=

Derivando Distribución Asintótica de OLS (cont.)

- Acabamos de derivar que $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$.
- Entonces,

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})((X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon}))\end{aligned}$$

Derivando Distribución Asintótica de OLS (cont.)

- Acabamos de derivar que $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$.
- Entonces,

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})((X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})) \\ &= \beta + \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})\end{aligned}$$

Derivando Distribución Asintótica de OLS (cont.)

- Por tanto,

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})$$

Derivando Distribución Asintótica de OLS (cont.)

- Por tanto,

$$\begin{aligned}\sqrt{N}(\hat{\beta} - \beta) &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - E[X_i])\varepsilon_i \\ &\quad - \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} (\bar{\varepsilon} \sqrt{N}(\bar{X} - E[X_i]))\end{aligned}$$

- Por LLN y CMT, $\left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \rightarrow_p Var(X_i)^{-1}$

Derivando Distribución Asintótica de OLS (cont.)

- Por tanto,

$$\begin{aligned}\sqrt{N}(\hat{\beta} - \beta) &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - E[X_i])\varepsilon_i \\ &\quad - \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} (\bar{\varepsilon} \sqrt{N}(\bar{X} - E[X_i]))\end{aligned}$$

- Por LLN y CMT, $\left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \rightarrow_p Var(X_i)^{-1}$
- Por CLT, $\sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - E[X_i])\varepsilon_i \rightarrow_d N(0, Var((X_i - E[X_i])\varepsilon_i))$.

Derivando Distribución Asintótica de OLS (cont.)

- Por tanto,

$$\begin{aligned}\sqrt{N}(\hat{\beta} - \beta) &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon}) \\ &= \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - E[X_i])\varepsilon_i \\ &\quad - \left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} (\bar{\varepsilon} \sqrt{N}(\bar{X} - E[X_i]))\end{aligned}$$

- Por LLN y CMT, $\left(\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 \right)^{-1} \rightarrow_p Var(X_i)^{-1}$
- Por CLT, $\sqrt{N} \frac{1}{N} \sum_{i=1}^N (X_i - E[X_i])\varepsilon_i \rightarrow_d N(0, Var((X_i - E[X_i])\varepsilon_i))$.
- Por LLN, CLT, y Slutsky, $\bar{\varepsilon} \sqrt{N}(\bar{X} - E[X_i]) \rightarrow_d 0 \times N(0, Var(X_i)) = 0$

Terminando (!)

- Poniendo las piezas juntas, tenemos que

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2),$$

donde

$$\sigma^2 = \frac{Var((X_i - E[X_i])\varepsilon_i)}{Var(X_i)^2}$$

- Como antes, podemos estimar la varianza σ^2 usando promedios muestrales,

$$\hat{\sigma}^2 = \frac{\frac{1}{N} \sum_i ((X_i - \bar{X})\hat{\varepsilon}_i)^2}{\left(\frac{1}{N} \sum_i (X_i - \bar{X})^2\right)^2}, \text{ where } \hat{\varepsilon}_i = Y_i - (\hat{\alpha} + X_i \hat{\beta})$$

Terminando (!)

- Poniendo las piezas juntas, tenemos que

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2),$$

donde

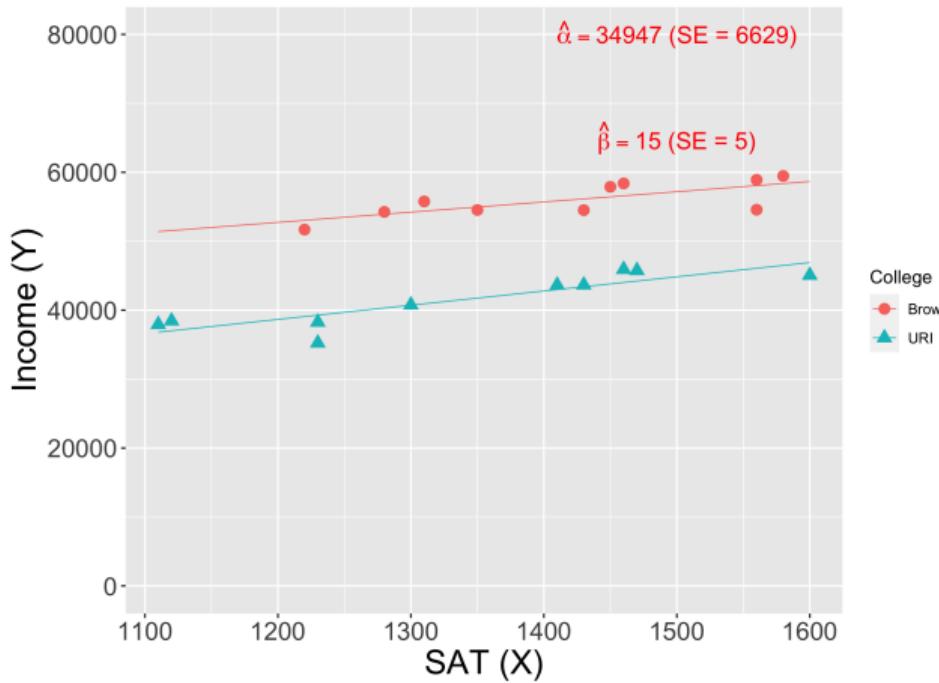
$$\sigma^2 = \frac{Var((X_i - E[X_i])\varepsilon_i)}{Var(X_i)^2}$$

- Como antes, podemos estimar la varianza σ^2 usando promedios muestrales,

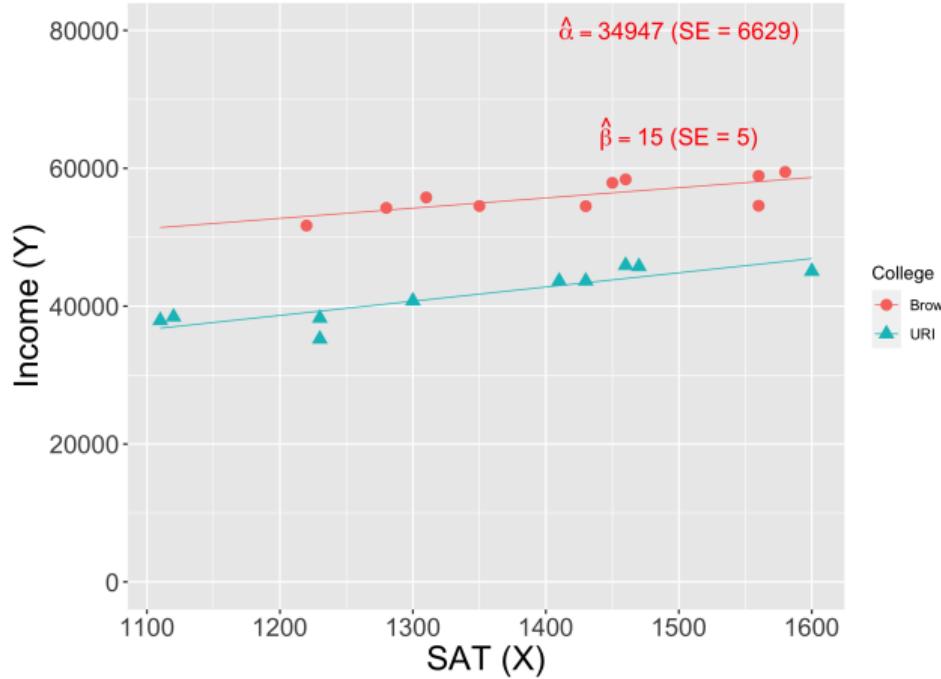
$$\hat{\sigma}^2 = \frac{\frac{1}{N} \sum_i ((X_i - \bar{X})\hat{\varepsilon}_i)^2}{\left(\frac{1}{N} \sum_i (X_i - \bar{X})^2\right)^2}, \text{ where } \hat{\varepsilon}_i = Y_i - (\hat{\alpha} + X_i \hat{\beta})$$

- Podemos hacer los mismos pasos para mostrar que $\hat{\alpha}$ también se distribuye de manera normal asintóticamente.

(Fake) Data on Income by College / SAT

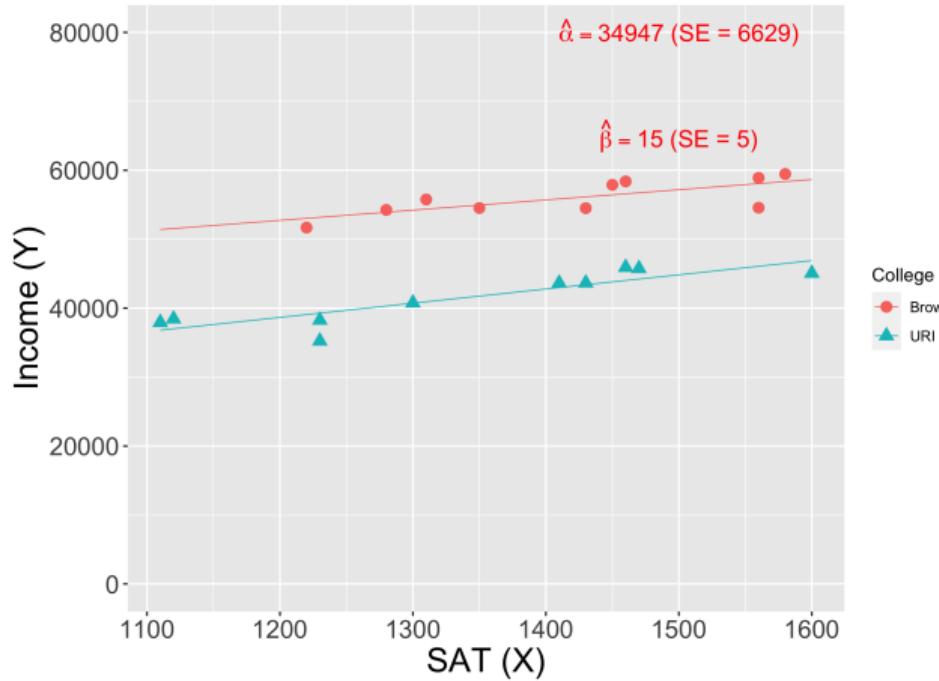


(Fake) Data on Income by College / SAT



- Un CI para β es $\hat{\beta} \pm 1.96 \times SE$

(Fake) Data on Income by College / SAT



- Un CI para β es $\hat{\beta} \pm 1.96 \times SE \approx [5, 25]$

Regression como aproximación al CEF

- Hasta ahora asumimos que el CEF es lineal: $E[Y_i|X_i = x] = \alpha + \beta x$
- ¡¿Qué pasa si no lo es?!

Regression como aproximación al CEF

- Hasta ahora asumimos que el CEF es lineal: $E[Y_i|X_i = x] = \alpha + \beta x$
- ¡¿Qué pasa si no lo es?!
- Claim: si CEF no es lineal, entonces OLS nos da la “mejor aproximación lineal” que hay

Regression como aproximación al CEF

- Hasta ahora asumimos que el CEF es lineal: $E[Y_i|X_i = x] = \alpha + \beta x$
- ¡¿Qué pasa si no lo es?!
- Claim: si CEF no es lineal, entonces OLS nos da la “mejor aproximación lineal” que hay
- Con eso me refiero a que α, β de OLS minimiza

$$\min_{\alpha, \beta} E[(E[Y|X] - (\alpha + \beta X))^2]$$

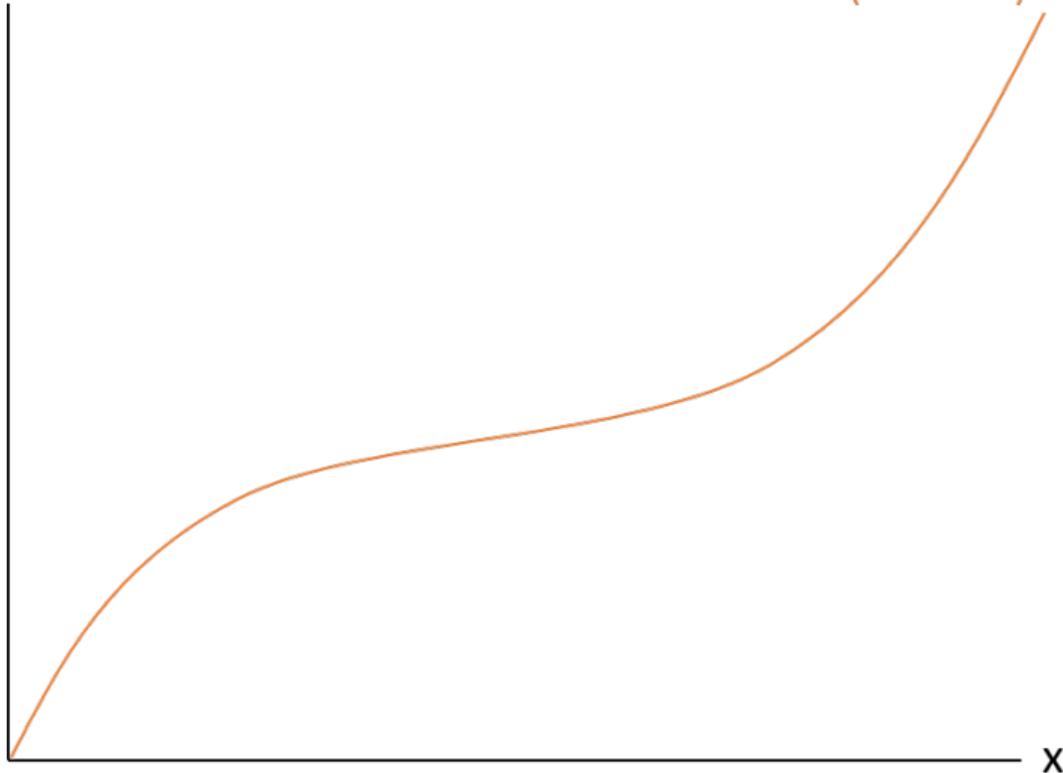
Regression como aproximación al CEF

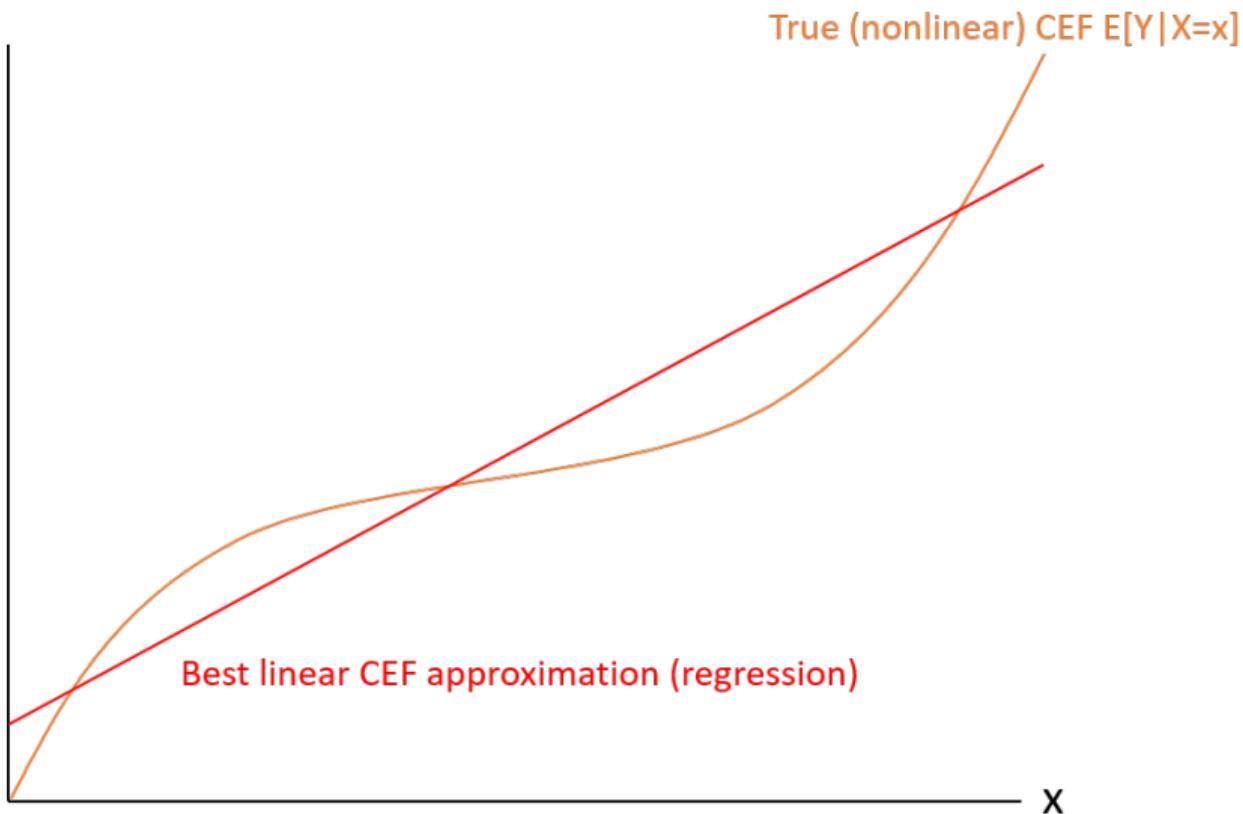
- Hasta ahora asumimos que el CEF es lineal: $E[Y_i|X_i = x] = \alpha + \beta x$
- ¡¿Qué pasa si no lo es?!
- Claim: si CEF no es lineal, entonces OLS nos da la “mejor aproximación lineal” que hay
- Con eso me refiero a que α, β de OLS minimiza

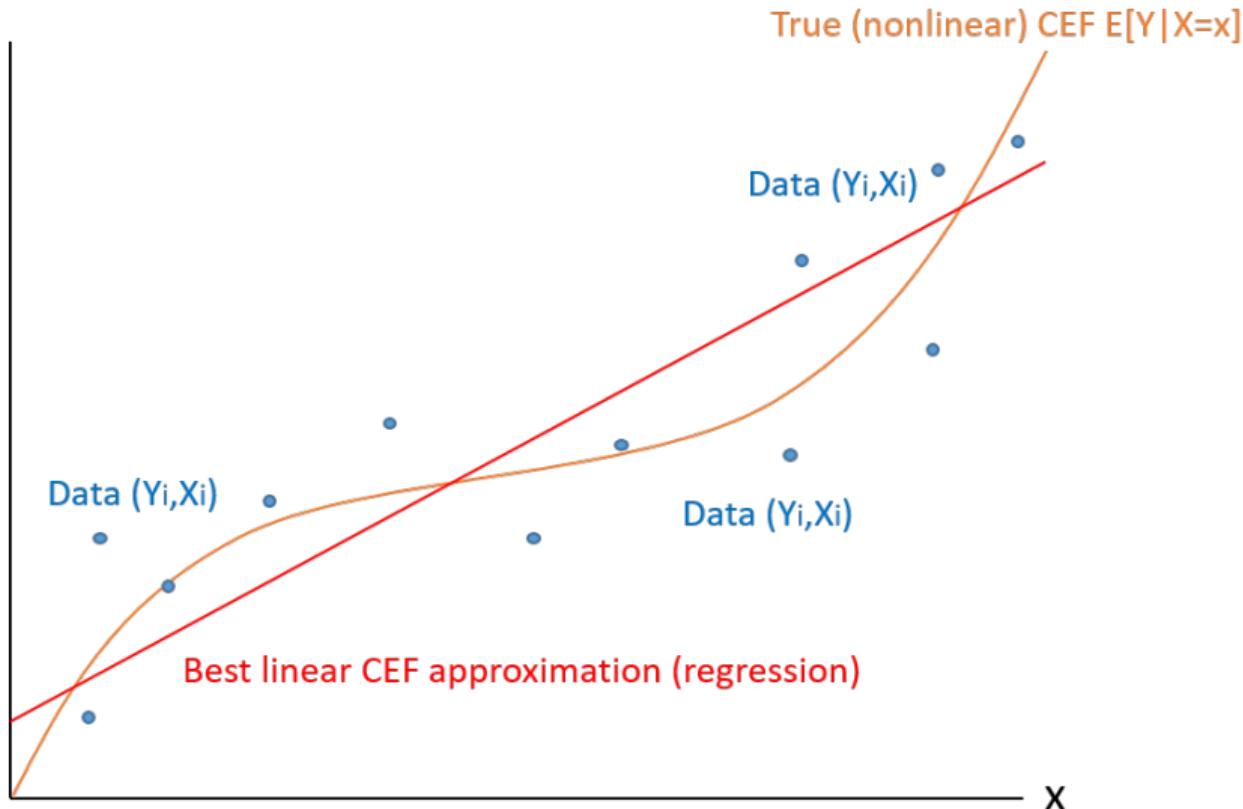
$$\min_{\alpha, \beta} E[(E[Y|X] - (\alpha + \beta X))^2]$$

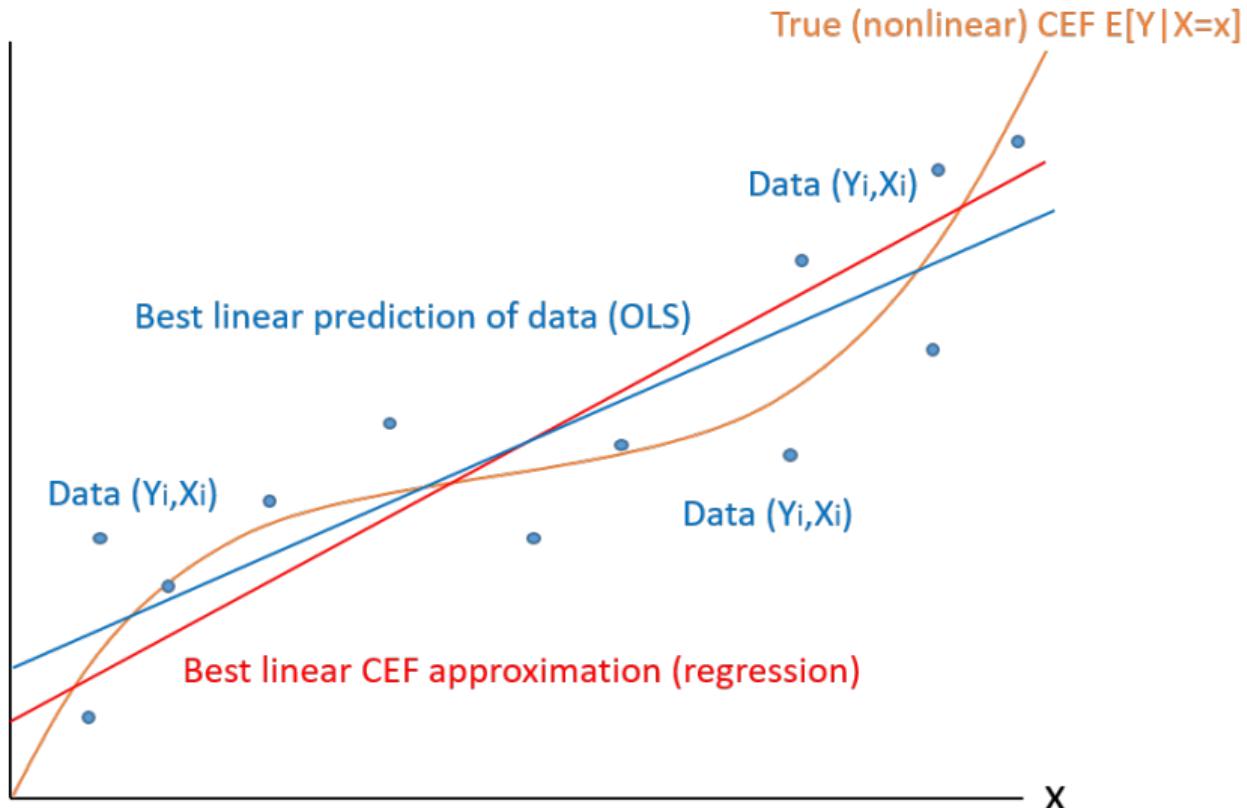
Es decir, tenemos la función lineal “más cercana” al CEF en términos de error cuadrático medio

True (nonlinear) CEF $E[Y|X=x]$









Fin de la parte I

Regresión Lineal Múltiple

La clase pasada vimos como podemos aproximar la función de esperanza condicional (en variables que observamos X_i) en una forma lineal

$$E[Y_i | X_i = x] \approx \alpha + x\beta,$$

cuando X_i es un escalar (es decir, $X_i \in \mathbb{R}$ en lugar de ser un vector en \mathbb{R}^k).

- Y mostramos como el estimando (α, β) se puede estimar por OLS

Esto se puede generalizar al caso donde observas varias características $X_{i2}, X_{i3}, \dots, X_{id}$. Simplemente podemos escribirlo en forma de vectores.

$$E[Y_i | \mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta} \text{ dado un vector } \mathbf{X}_i = (1, X_{i2}, \dots, X_{id})'$$

Visualmente

Por convención*, definimos la primera columna como el intercepto: vector que toma el valor de 1 en todos sus elementos.

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1d} \\ X_{21} & X_{22} & \cdots & X_{2d} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nd} \end{bmatrix} = \begin{bmatrix} \cdots & X_1^T & \cdots \\ \cdots & X_2^T & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & X_n^T & \cdots \end{bmatrix}$$

*: el intercepto hace que el error de predicción en la muestra sea 0 en promedio. Esta es una propiedad bastante deseable

Plan para hoy

1. Introducción a Machine Learning
2. Comandos de Stata para Regresión Lineal con ML: *lasso*, *elasticnet*

¿Qué es Machine Learning (ML)?

- Le llamamos así al uso de computadoras para **detectar patrones en la data y usarlos para hacer predicciones** o decisiones
- Suele ser útil cuando:
 - Automatizar algo que un humano no puede hacer
 - Procesar una cantidad de información que un humano no puede hacerlo de manera rápida (ej. Terabytes de data)

Aplicaciones

¡¡Está en todos lados!!

- Filtra correo spam
- Recomendación de videos en Youtube
- Google Translate
- Analizar bases de datos ("automatizar" a un estadista)
- Videojuegos
- ChatBots? (ej. ChatGPT, Brok)

Combinar matemática/estadísticas con MUCHA data o poder computacional abre un abanico de herramientas útiles.

Conocer las limitaciones

Como toda herramienta, deben conocer sus **limitaciones** a la hora de usarla.

Muchas personas que usan estas herramientas indiscriminadamente están cayendo en una de las trampas más grandes de la estadística: **overfitting**

- Método no funciona bien fuera de la muestra estadística en la que se analizó originalmente (*out-of-sample prediction*)

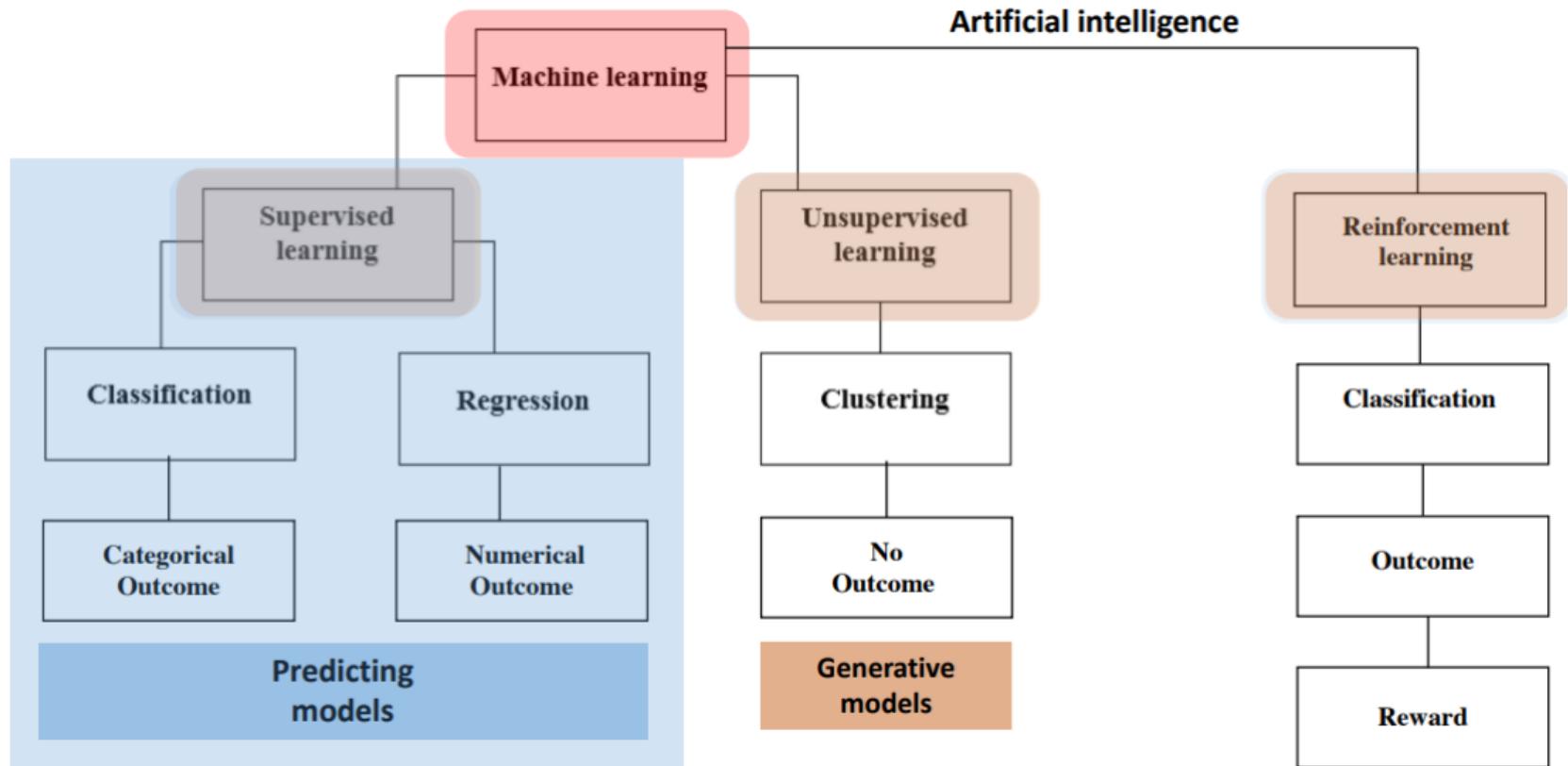


horse → zebra

Todo es un tema de marketing...

STATISTICS	MACHINE LEARNING
<i>Statistical model</i>	<i>Learner</i>
<i>Estimation sample</i>	<i>Training dataset</i>
<i>Out-of-sample observations</i>	<i>Test dataset</i>
<i>Estimation method</i>	<i>Algorithm</i>
<i>Observation</i>	<i>Instance</i>
<i>Predictor</i>	<i>Feature</i>
<i>Dependent variable</i>	<i>Target</i>

Supervised, Unsupervised and Artificial Intelligence



Supervised Learning

Supervised Learning es el caso que más nos sirve como economistas, pues discute cómo realizar predicciones buenas cuando observamos características X y resultados y .

- Cuando y es categórico (ejemplo, toma valores 0 o 1), se le llama **Clasificación**
- Cuando y es continuo (ejemplo, salarios), se le llama **Regresión**

Egg	Milk	Fish	Wheat	Shellfish	Peanuts	...	Sick?
0	0.7	0	0.3	0	0		1
0.3	0.7	0	0.6	0	0.01		1
0	0	0	0.8	0	0		0
0.3	0.7	1.2	0	0.10	0.01		1
0.3	0	1.2	0.3	0.10	0.01		1

Train vs Test Data

$$X = \begin{bmatrix} X_{\text{train}} \\ \cdots \\ X_{\text{validate}} \end{bmatrix} \quad Y = \begin{bmatrix} y_{\text{train}} \\ \cdots \\ y_{\text{validate}} \end{bmatrix}$$

The diagram illustrates the decomposition of training and validation data into feature matrices X and target vectors y . The matrix X contains two distinct sections: X_{train} (top) and X_{validate} (bottom), separated by a dashed line. Similarly, the vector y contains two sections: y_{train} (top) and y_{validate} (bottom), also separated by a dashed line. Braces on the right side group the "train" and "validation" sections respectively, indicating that the validation set is not included in the training set.

Train vs Test Data

Usamos una parte de la data para entrenar (*train*)

$$X = \\ n \times d$$

Egg	Milk	Fish	Wheat	Shellfish	Peanuts	...	Sick?
0	0.7	0	0.3	0	0		1
0.3	0.7	0	0.6	0	0.01		1
0	0	0	0.8	0	0		0
0.3	0.7	1.2	0	0.10	0.01		1
0.3	0	1.2	0.3	0.10	0.01		1

$$y = \\ n \times 1$$

Usamos otra parte de la data para validar (*test*)

$$\tilde{X} = \\ t \times d$$

Egg	Milk	Fish	Wheat	Shellfish	Peanuts	...	Sick?
0.5	0	1	0.6	2	1		?
0	0.7	0	1	0	0		?
3	1	0	0.5	0	0		?

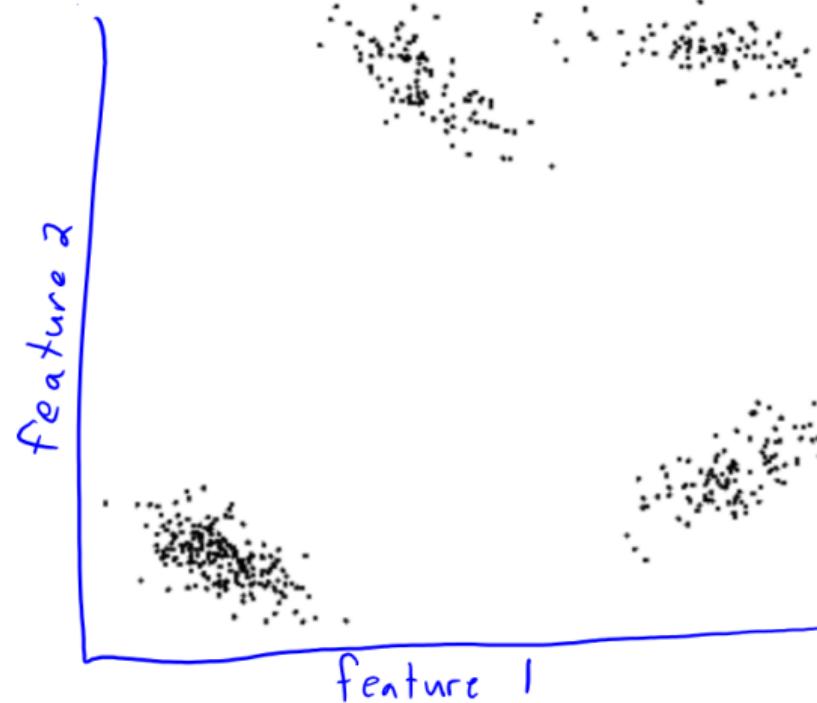
$$\tilde{y} = \\ t \times 1$$

Unsupervised Learning

Unsupervised learning es el caso donde observamos características X pero no los resultados y , y buscamos obtener algún patrón.

Input: data matrix 'X'.

$$X = \begin{bmatrix} -9.0 & -7.3 \\ -10.9 & -9.0 \\ 13.7 & 19.3 \\ 13.8 & 20.4 \\ 12.8 & 20.6 \\ \vdots & \vdots \end{bmatrix}$$

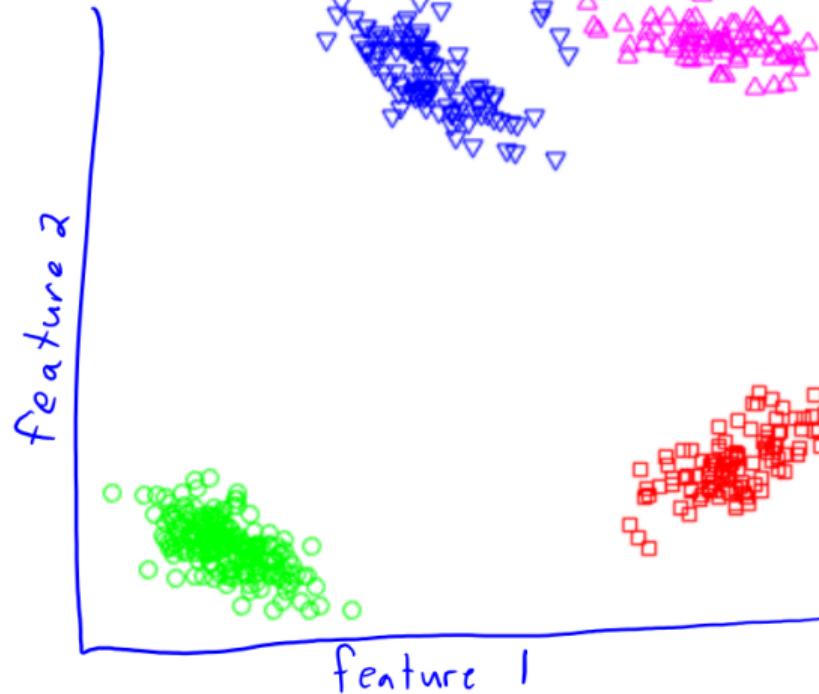


Unsupervised Learning

Unsupervised learning es el caso donde observamos características X pero no los resultados y , y buscamos obtener algún patrón.

Input: data matrix 'X'.

$$X = \begin{bmatrix} -9.0 & -7.3 \\ -10.9 & -9.0 \\ 13.7 & 19.3 \\ 13.8 & 20.4 \\ 12.8 & 20.6 \\ \vdots & \vdots \end{bmatrix}$$



Output: clusters \hat{y} .

$$\hat{y} = \begin{bmatrix} 2 \\ 2 \\ 3 \\ 3 \\ 1 \\ \vdots \end{bmatrix}$$

Regresión

Hoy nos centaremos en la regresión lineal. La manera en que suelen definir el problema es así

$$\hat{y}_i = x_i' w,$$

donde w es el peso (weight), que corresponde a lo que nosotros conocemos como el coeficiente de regresión β .

Que recordemos que busca minimizar

$$\sum_{i=1}^n (\hat{y}_i - y_i)^2 = \sum_{i=1}^n (w' x_i - y_i)^2$$

Trade-off Fundamental

Training dataset

N in-sample available observations



$$\mathcal{T}_r = \{x_i, y_i\}_1^N$$



$$\text{MSE}_{\mathcal{T}_r} = \text{Ave}_{i \in \mathcal{T}_r} [y_i - \hat{f}(x_i)]^2$$



Overfitting as flexibility increases

Testing dataset

M out-of-sample observations



$$\mathcal{T}_e = \{x_i, y_i\}_1^M$$



$$\text{MSE}_{\mathcal{T}_e} = \text{Ave}_{i \in \mathcal{T}_e} [y_i - \hat{f}(x_i)]^2$$



True fitting accuracy

Trade-off Fundamental

Algunos llaman MSE_{tr} el **error de train** y a MSE_{te} el **error de test**.

$$E_{test} = (E_{test} - E_{train}) + E_{train}$$

"test error" *"generalization gap"* *"training error"*
E_{gap}

Predecir mal fuera de muestra (E_{test} alto) puede ser a causa de

- E_{gap} alto: tu modelo no generaliza bien fuera de la muestra train
- E_{train} alto: tu modelo no es bueno prediciendo ni siquiera dentro de la muestra train

0 ¿De qué depende el error de generalización?

- Se reduce cuando la muestra n aumenta
- Aumenta cuando la complejidad del modelo aumenta (por ej. número de características X a considerar)

Trade-off Fundamental

Algunos llaman MSE_{tr} el **error de train** y a MSE_{te} el **error de test**.

$$E_{test} = (E_{test} - E_{train}) + E_{train}$$

"test error" *"generalization gap"* *"training error"*
E_{gap}

Predecir mal fuera de muestra (E_{test} alto) puede ser a causa de

- E_{gap} alto: tu modelo no generaliza bien fuera de la muestra train
- E_{train} alto: tu modelo no es bueno prediciendo ni siquiera dentro de la muestra train

¿De qué depende el error de generalización?

- Se reduce cuando la muestra n aumenta → **no controlable**
- Aumenta cuando la complejidad del modelo aumenta (por ej. número de características X a considerar)

Trade-off Fundamental

El caso cuando E_{train} es alto, se le llama **overfitting**



**THE BEST WAY TO
EXPLAIN OVERTFITTING**

Regularización

Una manera de controlar la complejidad del modelo es tratar de hacer que hayan menos X (o cercano a ello). A esto se le conoce como **regularización**. La manera en que se aplica en el

contexto de la regresión lineal es agregando un término de penalización (a lo multiplicador de Lagrange) para cuando los coeficientes w son muy altos.

Es decir, la estimación/goritmo priorizará tener un modelo menos complejo.

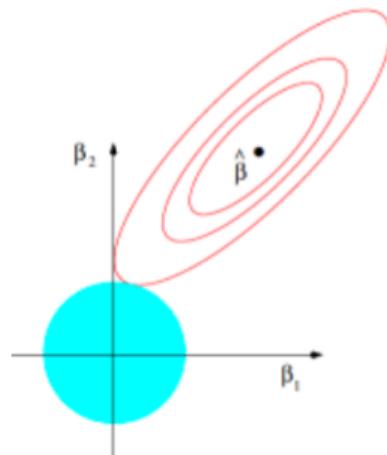
Denotamos $1/n \sum_{i=1}^n w_i^2 := \|w\|_2$ a la norma L_2 y $1/n \sum_{i=1}^n |w_i| := \|w\|_1$ a la norma L_1 .

Ridge Regression

La regresión Ridge pone el siguiente término de penalización

$$f(w) = \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$$

donde λ es el parámetro de penalización elegido por el economista.
Gráficamente se ve así



Ridge Regression

Pros

- Tiene una solución única incluso cuando OLS no lo tiene
- Robusto a valores extremos en la data

Cons

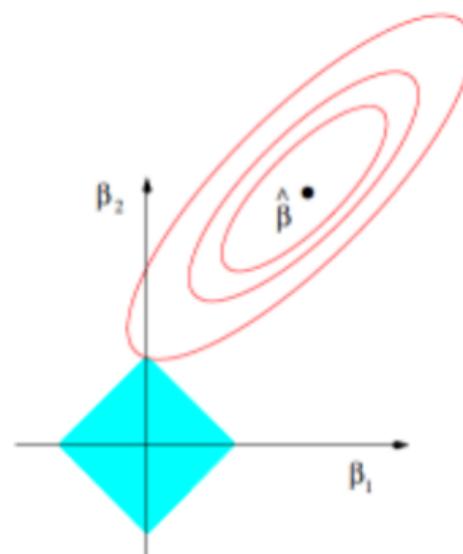
- Retiene todos los coeficientes potencialmente
- Introducimos sesgo (hacia 0) a los coeficientes
- Hay que elegir λ

Lasso Regression

La regresión Lasso pone el siguiente término de penalización

$$f(w) = \frac{1}{2} \|Xw - y\|^2 + \lambda \|w\|_1$$

donde λ es el parámetro de penalización elegido por el economista.
Gráficamente se ve así



Lasso Regression

Pros

- Reduce mucho más la complejidad del modelo al hacer coeficientes exactamente cero
- Computacionalmente rápido

Cons

- No es muy robusto cuando hay fuerte correlacion entre distintos X
- Introducimos sesgo (hacia 0) a los coeficientes
- Hay que elegir λ

Elastic Net Regression

¡Podemos combinar el beneficio de ambos!

$$\frac{\sum_{i=1}^n (y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^m \hat{\beta}_j^2 + \alpha \sum_{j=1}^m |\hat{\beta}_j| \right)$$

donde λ es el parámetro de penalización elegido por el economista y $\alpha \in [0, 1]$ es la importancia relativa que le damos a Lasso con respecto a Ridge.

¡¡Ahora tenemos otro hiperparámetro más!!



¿Cómo elegir hiperparámetros?

Una propuesta de pasos

- ① Partir la data en distintas partes (por ahora pensemos solo en 2), y una se usa para training y otra para test
- ② Repetimos el paso 1 con distintos valores de λ que elegimos (ej. probemos $\lambda = 0, \lambda = 1, \lambda = 2$)
- ③ Nos quedamos con el λ que tuvo el menor error de predicción fuera de la muestra

A este procedimiento se le conoce como **Cross-Validation**.

¡A la computadora!