Combinando lo aprendido

Acercamientos, métodos de investigación, estadísticas

Alcances

Tipos de hipótesis

- ***Hipótesis direccional***: Una hipótesis de investigación que predice una dirección particular de diferencia entre las poblaciones. Por ejemplo, una predicción de que la población, como la muestra estudiada, tiene una media más alta que la población en general.
- ***Hipótesis no direccional***: Una hipótesis de investigación que predice un efecto, pero no especifica la dirección del efecto. Ejemplo: un programa de habilidades sociales podría afectar la productividad, ya sea aumentándola o disminuyéndola.

Ejemplos de preguntas de investigación

- Describiendo las variables

  Ej. ¿Cuántas personas están a favor de las medidas de austeridad establecidas por el gobierno?

- Explorando la relación entre variables

  Ej. ¿Existe una asociación entre la educación e ingresos?

- Explorando las diferencias entre grupos o tiempos

  Ej. ¿Existen diferencias entre los procesos cognitivos de personas que juegan o no juegan videojuegos?

Elaboración de hipótesis

- Las hipótesis son las guías de una investigación o estudio. Las hipótesis indican lo que tratamos de probar y se definen como explicaciones tentativas del fenómeno investigado. Se derivan de la teoría existente y deben formularse a manera de proposiciones

Tabla 6.1 Formulación de hipótesis en estudios cuantitativos con diferentes alcances

Alcance del estudio	Formulación de hipótesis
Exploratorio	No se formulan hipótesis.
Descriptivo	Sólo se formulan hipótesis cuando se pronostica un hecho o dato.
Correlacional	Se formulan hipótesis correlacionales.
Explicativo	Se formulan hipótesis causales.

Contrastando hipótesis

- Pasos a seguir:
  1. Reformular la pregunta como una hipótesis de investigación y una hipótesis nula sobre las poblaciones.
  2. Determinar características: Identificar las características de la distribución de comparación.
  3. Determinar el punto crítico o de corte: Establecer el valor de la muestra en la distribución de comparación en el que se debe rechazar la hipótesis nula.
  4. Encontrar el puntaje de tu muestra en la distribución de comparación.
  5. Determinar si se debe rechazar la hipótesis nula.

Creando hipótesis (formen grupos de 2-3)

- Cada grupo escoja un tema de interés (e.g.: *uso de redes sociales*, *hábitos de estudio*, *salud mental*, *impactos o causas de desigualdad económica*, *etc.*) y formulen:

  - Una hipótesis descriptiva
  - Una hipótesis correlacional
  - Una hipótesis explicativa/comparativa

Modelos estadísticos avanzados

Regresión, ANOVA, Ji Cuadrada

Contraste de hipótesis

Usamos pruebas o contraste de hipótesis para examinar cuál de dos hipótesis complementarias es verdadera. Hipotetizamos sobre un parámetro poblacional, como la media poblacional, tal que

- H_0, *la hipótesis nula*, corresponda a “no hay efecto” “no hay diferencia” o “no hay relación”; y
- H_1, *la hipótesis alternativa*, corresponda a “hay efecto”, “hay diferencia” o “hay relación” entre conceptos.

❖Si fuere sobre examinar el efecto de un tratamiento (médico, política pública, etc.) nos interesa el cambio Δ.

❖Si fuere sobre comparar grupos, nos interesa saber si $\mu_1 - \mu_2 = 0$ como nula o si $\mu_1 - \mu_2 \neq 0$.

❖Y si fuere sobre la relación entre dos variables, *X* y *Y*, querremos saber si son independientes, con H_0 siendo $P(X = x \text{ y } Y = y) = P(X = x)P(Y = y)$ para todas *x*, *y*, o H_1, $P(X = x \text{ y } Y = y) \neq P(X = x)P(Y = y)$.

Ejemplo: la señora con el té

- En el libro de *Sir* Ronald Fisher, *El diseño experimental*, una colega, la Dra. Muriel Bristol, indicó que podía saber perfectamente si cuando le prepararon el té, habían echado leche en la taza primero que el té con agua caliente, o si fue al revés. Fisher no le creó y creó un experimento. Sirvióle ocho tazas, preparadas de manera distinta, y luego se las dio a probar de manera aleatoria para que ella las catalogara.
- H_0, *la hipótesis nula*, correspondía a 'la suposición de la doctora no son mejor que el azar'
- H_1, *la hipótesis alternativa*, correspondía a 'la suposición de la doctora son mejores que las que llegaría por azar'.

Esto es una prueba de independencia: sus respuestas (X) contrastarían con el contenido de la taza (Y).

Ejemplo: la señora con el té

Tabla de contingencia	Verdad: Leche primero	Verdad: Té primero
Suposición: leche primero	4	0
Suposición: té primero	0	4

Si la H_1 fuera cierta, esperaríamos ver un conteo como el que está arriba. En la práctica puede resultar en combinación unas 5 tablas (para valores 0 al 4), y la probabilidad de obtener cada una de las combinaciones bajo la H_0 es (tomando la esq. superior izquierda):

- 4 (4 tazas donde se echó leche primero, y ella respondió correctamente): 1/70,
- 3 (3 tazas donde se echó leche primero, y ella respondió correctamente excepto en una): 16/70
- 2 (2 tazas donde se echó leche primero, y ella respondió correctamente excepto en dos): 36/70
- 1 (1 tazas donde se echó leche primero, y ella respondió erradamente excepto en una): 16/70
- 0 (0 tazas donde se echó leche primero, y ella respondió erradas todas): 1/70.

Con esto en mente podemos acercarnos a los valores de probabilidad o p-valor (también valor p, p consignado, o meramente p-value).

Valor p

- El **valor p** o **p-valor** se define como la probabilidad de que un valor estadístico calculado sea posible dada la hipótesis nula ser cierta. Ayuda a diferenciar resultados productos del azar de muestreo, de resultados que sean **estadísticamente significativos.**

En formalidad esto puede ser:

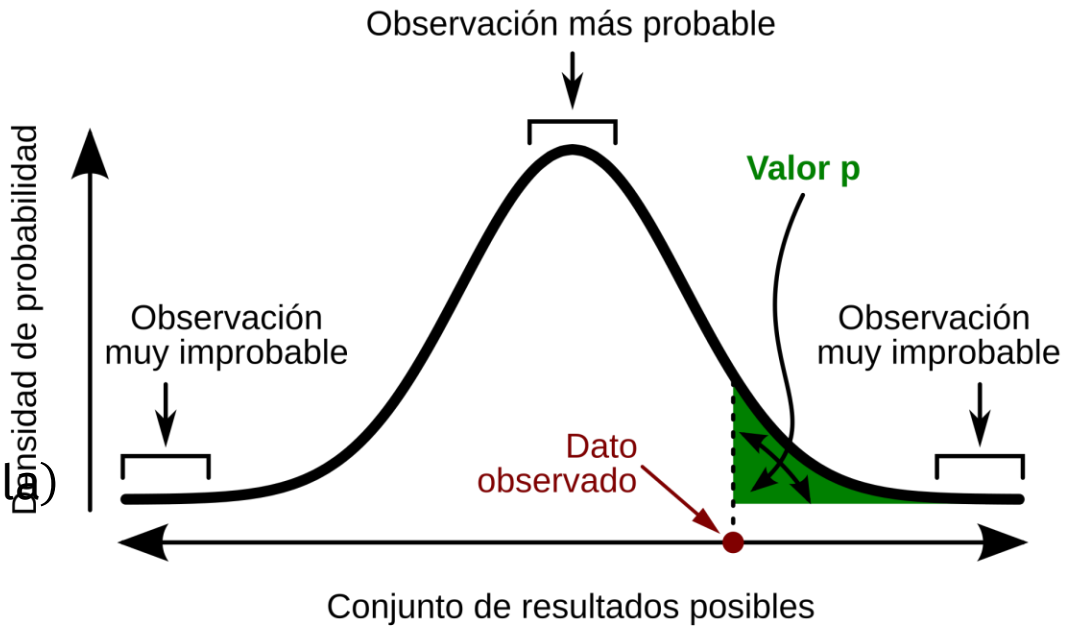
$$valor\ p = P(\text{resultado tan extremo o más}|\text{hipótesis nula})$$
$$= P(\text{resultado tan extremo o más}|H_0)$$

Importante:

Pr (observación | hipótesis) ≠ Pr (hipótesis | observación)

La probabilidad de observar un resultado dada una cierta hipótesis cierta *no es equivalente* a la probabilidad de que una hipótesis sea cierta dado un resultado observado.

Usar el valor p como un "puntaje" es cometer un error lógico inmenso: **la falacia de transposición condicional.**



El **valor p** (área de color verde) es la probabilidad de que un valor observado sea igual o más extremo que un cierto valor, asumiento que la hipótesis nula es cierta.

Cuidado con el valor-p

- Interpretación: Si el valor p es menor que el nivel de significancia (α, como 0.1, 0.05 o 0.01), se considera estadísticamente significativo y permite rechazar la hipótesis nula.
- Probabilidad: Valores p más bajos indican menor probabilidad de que el resultado se deba al azar y mayor evidencia en contra de la hipótesis nula.
- Advertencia: Rechazar la hipótesis nula no implica que la hipótesis alternativa sea verdadera, simplemente indica que hay suficiente evidencia para considerar improbable la hipótesis nula. Sin embargo, siempre existe el riesgo de cometer un error al rechazarla si el resultado es una observación atípica.

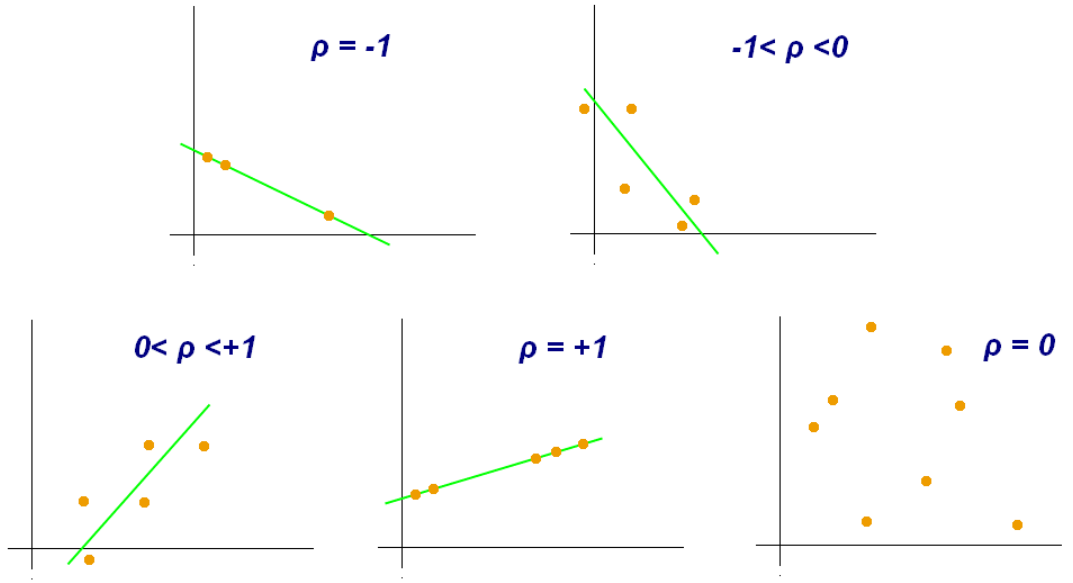
Advertencia:
no manipulen o ‘jaqueen’ el valor-p

- La manipulación o jaqueo del valor p ocurre cuando se realizan múltiples análisis o ajustes en los datos para obtener un valor p estadísticamente significativo. Esto puede generar falsos positivos y resultados que no se replican.
  - https://xkcd.com/882/ , https://www.explainxkcd.com/wiki/index.php/882:_Significant,
  - https://xkcd.com/1478/, https://www.explainxkcd.com/wiki/index.php/1478:_P-Values
- Evitar la práctica de p-hacking para mantener la integridad científica. En lugar de realizar análisis múltiples sin justificación, define de antemano tus hipótesis y el método estadístico que usarás.
- Pruebas pre-registradas.
- Uso adecuado del tamaño de muestra.
- Reportar todos los análisis realizados, no solo los que resultaron en significancia.

Correlación

- Definición: La correlación mide la fuerza y dirección de una relación lineal entre dos variables cuantitativas. Indica cómo los valores de una variable cambian sistemáticamente en relación con la otra.
- Coeficiente de correlación: El coeficiente de Pearson (notación: *r* o $\rho_{x,y}$) es el más común, variando entre -1 y 1, representando un espectro donde:
  - r = 1: Correlación positiva perfecta.
  - r = -1: Correlación negativa perfecta.
  - r = 0: No hay correlación lineal.

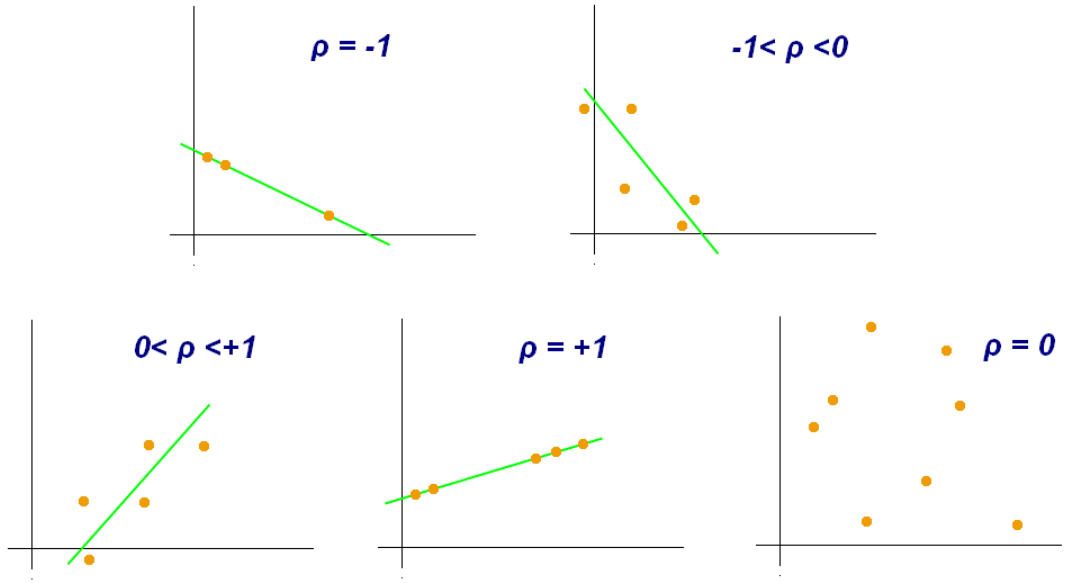
https://www.guessthecorrelation.com



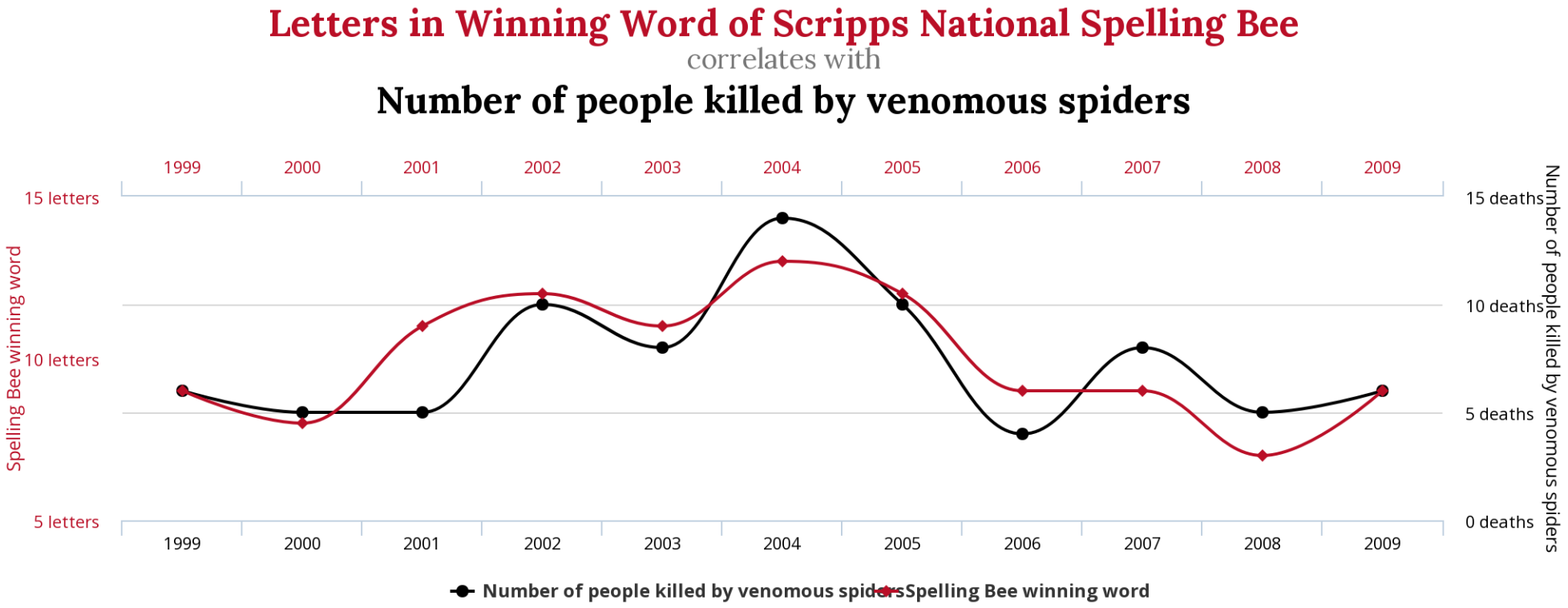
Correlación

- Definición: La correlación mide la fuerza y dirección de una relación lineal entre dos variables cuantitativas. Indica cómo los valores de una variable cambian sistemáticamente en relación con la otra.
- Coeficiente de correlación: El coeficiente de Pearson (notación: *r* o $\rho_{x,y}$) es el más común, variando entre -1 y 1, representando un espectro donde:
  - r = 1: Correlación positiva perfecta.
  - r = -1: Correlación negativa perfecta.
  - r = 0: No hay correlación lineal.

https://www.guessthecorrelation.com



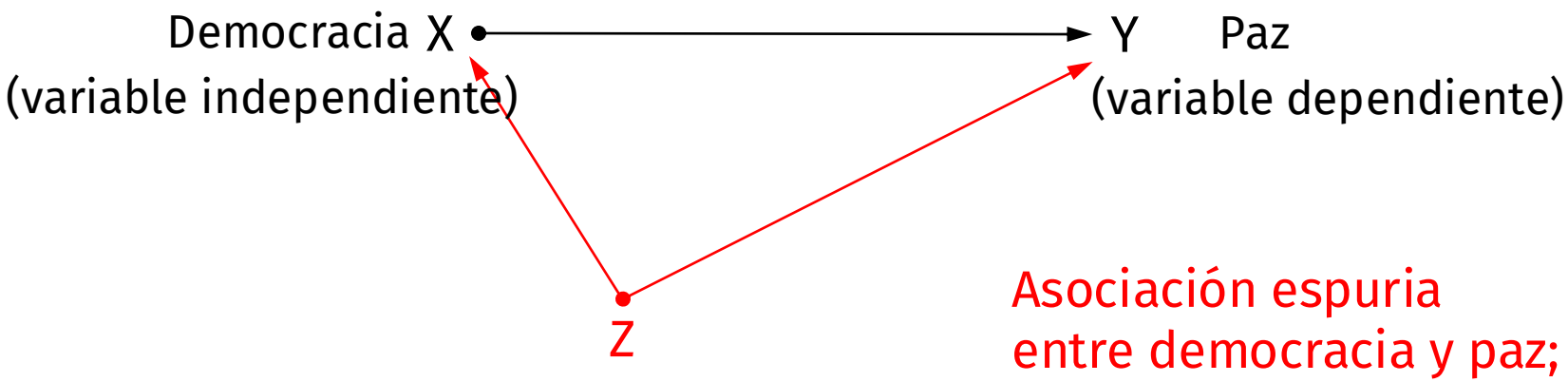
Relaciones espurias



[Spurious Correlations (tylervigen.com)]

Causalidad

Si concluimos que X causa Y, ¿estamos en lo correcto?



Asociación espuria entre democracia y paz; la asociación está causada por una tercera variable.

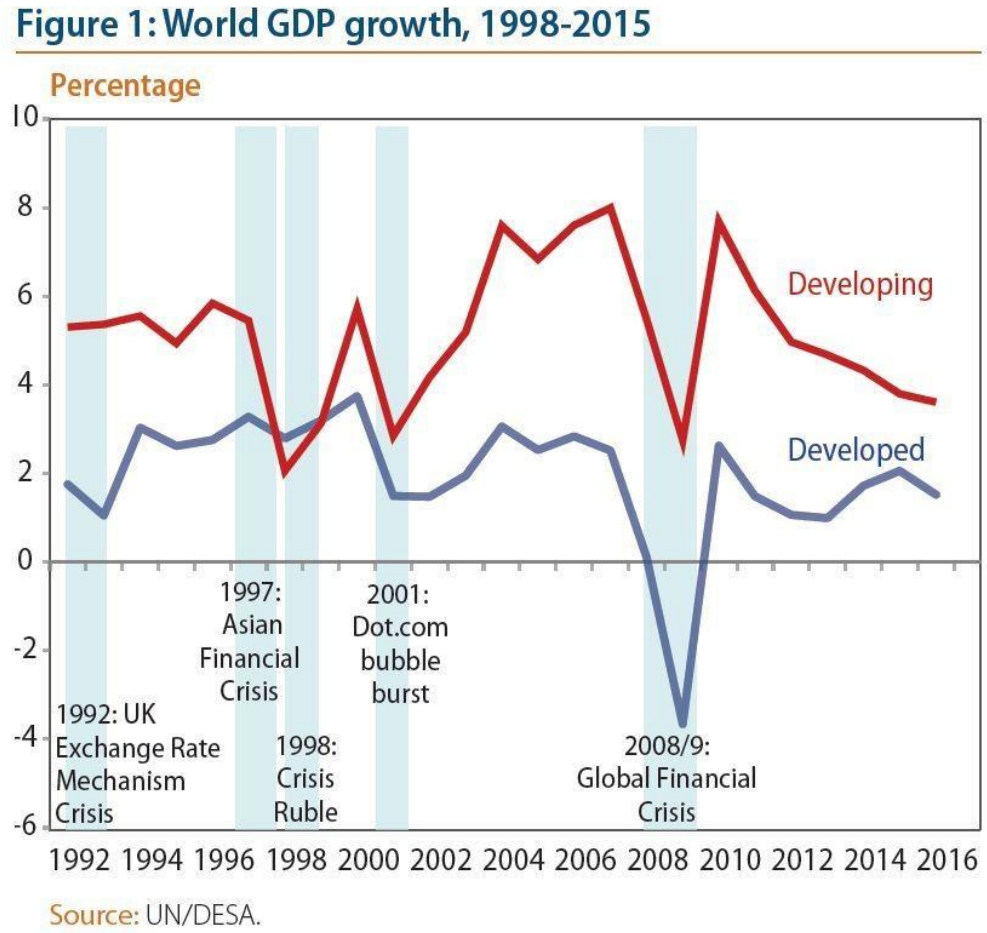
Causalidad

Si concluimos que X causa Y, ¿estamos en lo correcto?

Causalidad inversa; la dependiente en realidad influye en la que pensábamos independiente.

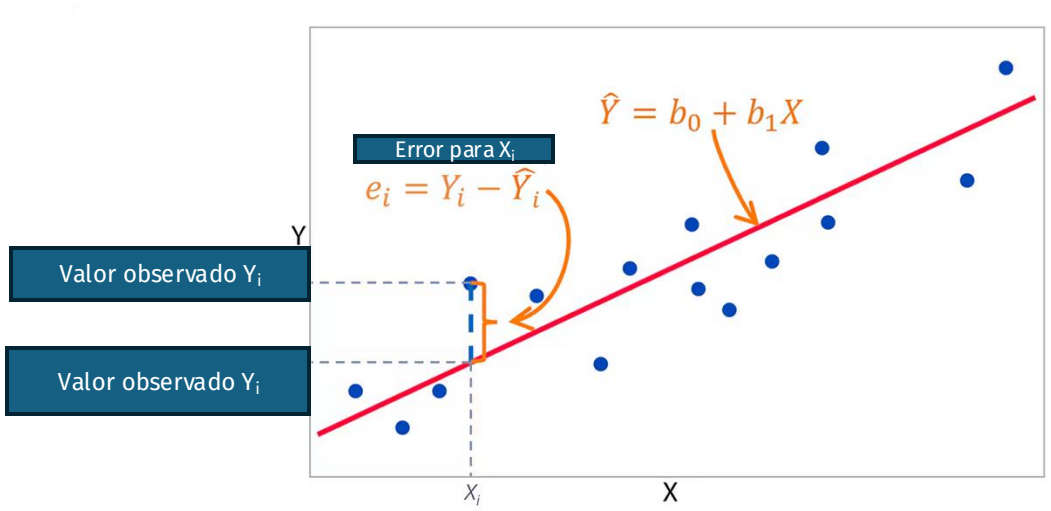
Contrafactual

- La situación que habría existido si la variable independiente NO hubiera cambiado. Lo que habría pasado si la causa no hubiera ocurrido, una realidad alternativa.

- 

Mínimos cuadrados

- Definición: Los mínimos cuadrados son una técnica que busca encontrar la mejor función que ajuste un conjunto de datos, minimizando los residuos o errores cuadráticos entre los valores observados y los predichos.
    - Objetivo: Minimizar la suma de los errores al cuadrado entre los datos reales y los valores predichos por el modelo.
    - Conexión con la Regresión: El método de mínimos cuadrados es la base para encontrar la línea de mejor ajuste en regresión lineal.
    - Ventaja: Garantiza que el modelo ajustado sea lo más preciso posible en términos de minimización del error, siempre que los errores de las observaciones sean aleatorios (no sistemáticos).



Regresión

- Definición: La regresión es una técnica estadística que estudia la relación entre una variable dependiente (resultado, lo observado) y una o más variables independientes (predictoras, explicativas). Nos ayuda a entender cómo cambia el valor promedio de una variable cuando varía otra (la expectativa condicionada).

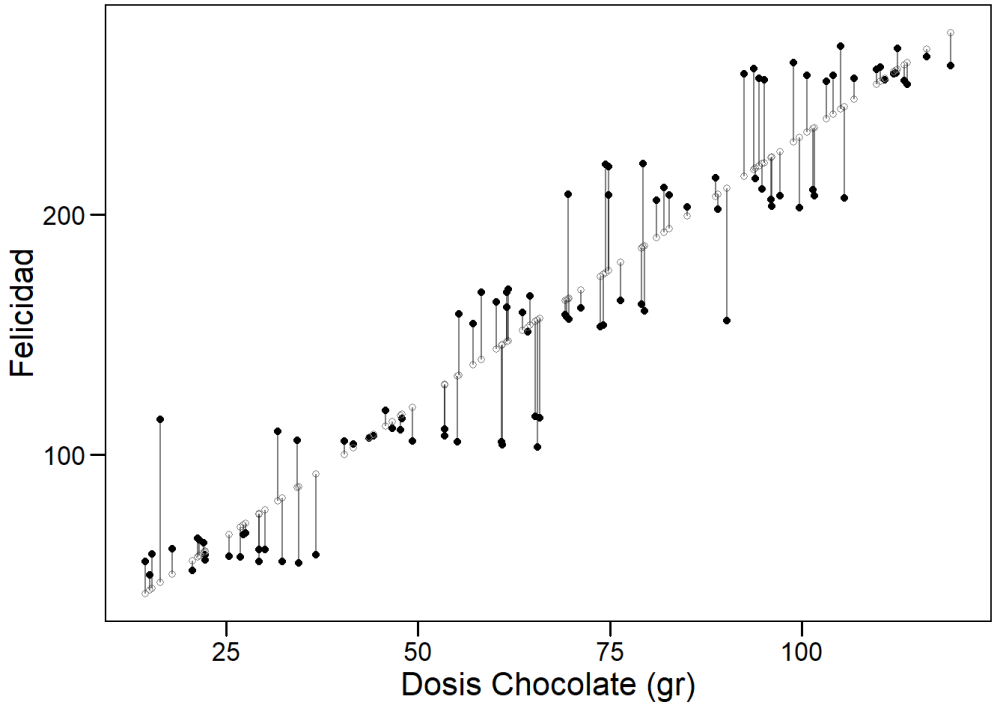
- Aplicación: Por ejemplo, podríamos intentar predecir si y cómo el ingreso y la composición racial influyen en los tiempos de espera a entrar a Unidades Electorales en las elecciones generales. O el impacto de anuncios políticos en postes y pantallas de anuncios en la movilización electoral hacia X o Y candidates.

- Cuando tenemos muchas variables predictoras, especialmente continuas, puede volverse complicado estimar la relación entre ellas y el resultado. Esto se conoce como el problema de dimensionalidad. Para resolverlo, los modelos de regresión suelen simplificar el problema al hacer suposiciones (por ejemplo, linealidad, distribución de errores) o al enfocarse en variables clave para facilitar la estimación, logrando parsimonia.

Regresión lineal simple

- Un modelo lineal es paramétrico porque asumimos que la relación entre dos variables es lineal y puede ser definida por los parámetros de una recta (el y-intercepto y la pendiente). Comenzaremos considerando un modelo lineal simple. En la siguiente figura podemos observar cómo existe una relación lineal entre la dosis de chocolate consumida y el nivel de felicidad reportado por una muestra de individuos seleccionados al azar en una población de Barcelona. Los puntos negros muestran los datos observados para cada individuo y los blancos representan a la felicidad que tendría cada individuo según la dosis de chocolate que reporta tomar.
- Un modelo lineal simple puede definirse

$$y_i = \alpha + \beta_1 x_i + \varepsilon$$



https://isglobal-brge.github.io/curso_R/modelos-de-regresión.html

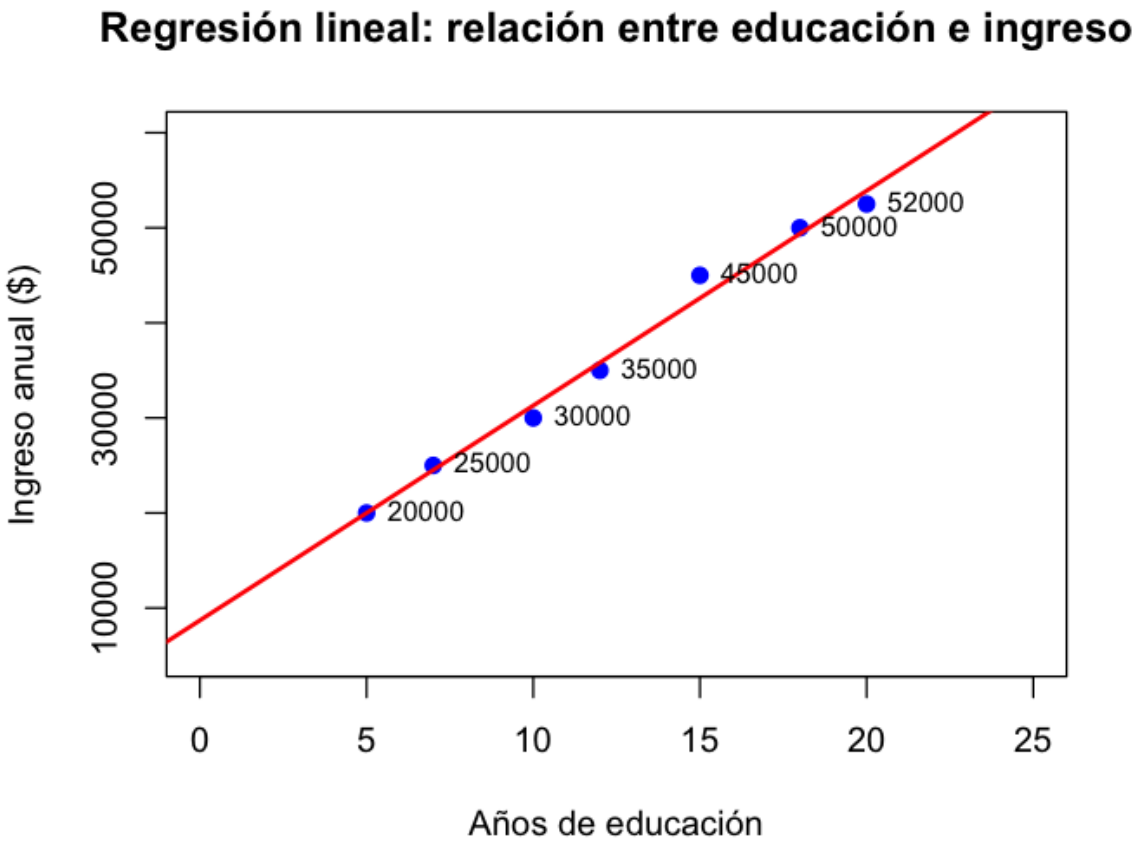
Ejemplo: ingreso y educación

	Dependent variable:
	ingreso
educación	2,259.615***
	(107.331)
Constant	8,701.923***
	(1,443.993)
Observations	7
R^2	0.989
Adjusted R^2	0.987
Residual Std. Error	1,462.677 (df =
F Statistic	443.218*** (df = 1; 5)

Note: *p<0.1; **p<0.05; ***p<0.01

La pendiente: aumenta ingreso por año de educación $2,259.62

El y-intercepto: el ingreso esperado si años de educación fuere 0 es $8,701.92. Este es el punto de partida.



Regresión lineal multivariante

- Podemos agregar predictores adicionales, p, a un modelo lineal simple, convirtiéndolo en un modelo lineal multivariante, que definimos de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- donde `i=1,…,n` y `p=1,…,p`. En esta ecuación y_i es nuevamente la variable resultado `i`-ésima, β_0 es el intercepto, β_1 es el coeficiente de la primera variable predictora, x_1, β_p es el coeficiente de la variable predictora p-ésima, x_p, y ϵ_I representa la parte estocástica del modelo, los residuos, indexados por fila. La parte determinista del modelo se puede resumir como Xβ, una matriz p x n, que llamaremos el “predictor lineal”.

ANOVA

- Definición: El Análisis de Varianza (ANOVA) es una técnica estadística que particiona la varianza en componentes relacionados con diferentes variables explicativas. Compara la varianza entre grupos con la varianza dentro de los grupos para detectar diferencias significativas.
  - Varianza entre grupos: Se calcula comparando las medias de cada grupo con la media global de los datos. Los puntos individuales no son tan relevantes como las medias de los grupos.
  - Varianza dentro de los grupos: Mide la variabilidad de cada observación respecto a la media de su grupo.
- Aplicación: ANOVA es utilizado en el diseño y análisis de experimentos para evaluar si diferentes tratamientos afectan significativamente la variabilidad de la variable respuesta.
- Origen: Desarrollado por Sir R. A. Fisher en los años 1920 y 1930, también se le conoce como “Anova de Fisher”.
- Puede ser unidireccional (una variable independiente) o bidireccional (con dos variables independientes, una cuantitativa y dos nominales, por ejemplo). También hay una variación llamada MANOVA que es para múltiples variables independientes
- Distribución F: El análisis utiliza la distribución F de Fisher para realizar pruebas de hipótesis y determinar si las diferencias observadas entre grupos son significativas.
  - Estadística F: Es el cociente entre la varianza entre grupos y la varianza dentro de los grupos. Valores grandes de F sugieren que la variabilidad entre grupos es mayor que la variabilidad dentro de los grupos.

Análisis de Varianza (ANOVA)

- Objetivo: ANOVA permite comparar si las medias de varios grupos difieren significativamente, superando la limitación de hacer comparaciones por pares.
- Modelo básico:
- $y_{\{ij\}} = \mu + \tau_i + \epsilon_{\{ij\}}$
- Donde:
  - $y_{\{ij\}}$: Valor observado.
  - μ : Media general.
  - τ_i : Efecto del tratamiento i.
  - $\epsilon_{\{ij\}}$: Error aleatorio.

Interpretando ANOVA

Modelo ANOVA: aov.model <- aov(size ~ pop) compara el tamaño (size) según los diferentes grupos de población (pop).

pop tiene 2 grados de libertad (Df = 2) y una suma de cuadrados (Sum Sq) de 34.67, con una media de cuadrados (Mean Sq) de 17.33.

El valor F es 10.4, lo que indica que la variación entre los grupos es relativamente grande en comparación con la variación dentro de los grupos (residuos).

p-valor: El p-valor es 0.00457, que es menor que 0.05, lo que indica que hay diferencias significativas entre las poblaciones con respecto al tamaño. Los asteriscos ** indican que el resultado es significativo al nivel de 0.01.

```r
aov.model <- aov(size ~ pop)
summary(aov.model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## pop          2  34.67  17.333    10.4 0.00457 **
## Residuals    9  15.00   1.667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```