Repaso Probabilidad y Estadística

2025-02-06

Repaso de Probabilidad

- Para nosotros, una *variable* es simplemente un conjunto de observaciones de la misma medida.
- Puede ser **discreta** (toma valores finitos o contables) o **continua** (toma valores en un intervalo real).

Ejemplos

- El precio promedio de la gasolina en Santiago de Chile
  - Variable continua
  - Espacio muestral: reales positivos
- Cuántos nacimientos hubo en cada comuna de Chile
  - Variable de conteo
  - Espacio amostral: naturales
- Nivel de educación: “middle school”, “high school”, “college”
  - Variable ordinal
- La raza y el sexo de una persona
  - Variable categórica

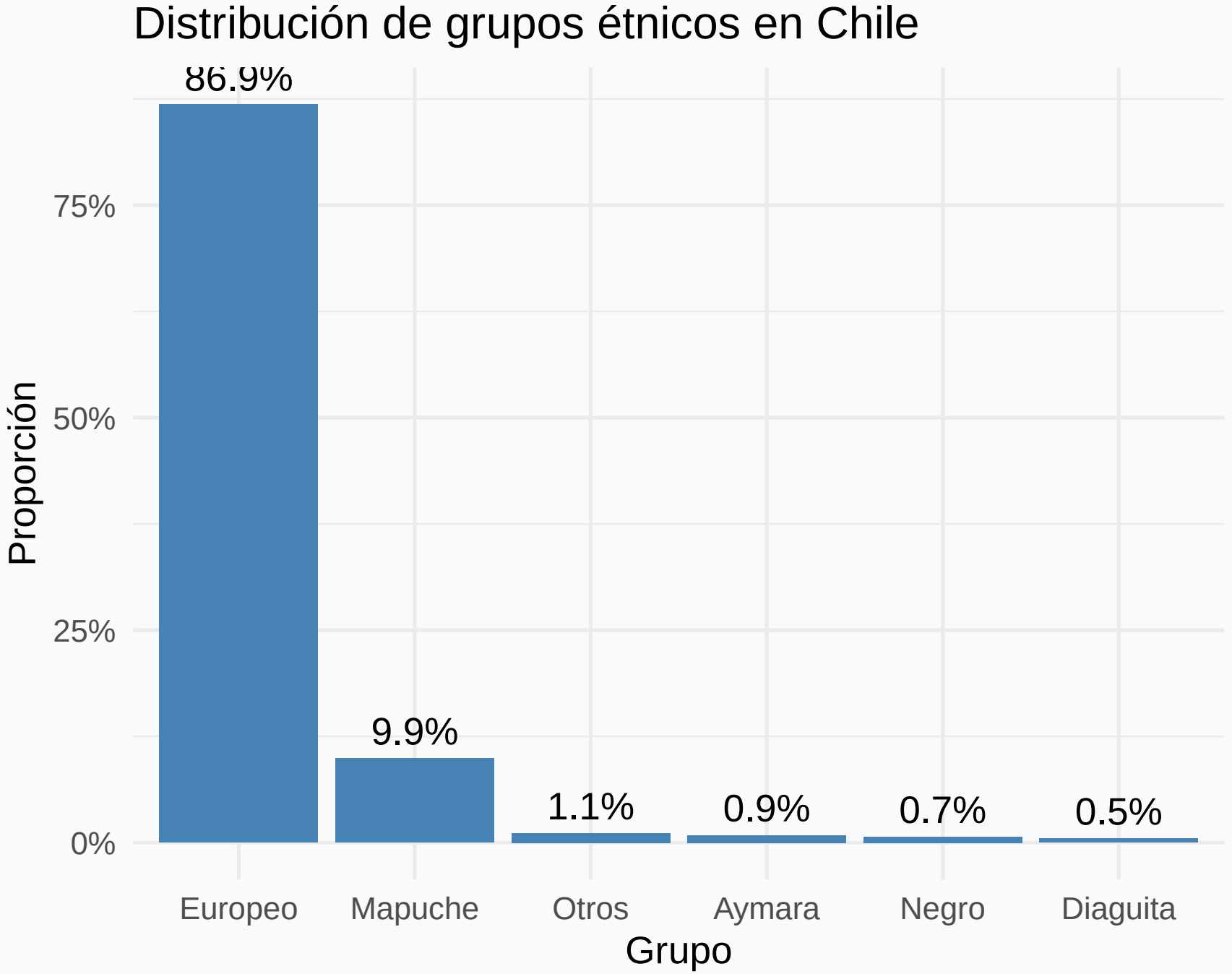
La distribución

En general, podemos describir una variable por su distribución.

Para variables discretas, eso puede ser simplemente una tabla de cada resultado y su probabilidad.

Distribución de grupos étnicos en Chile

Grupo	Proporción
Europeo	86.86%
Mapuche	9.93%
Aymara	0.89%
Diaguita	0.50%
Otros	1.12%
Negro	0.07%


Distribución de grupos étnicos en Chile
86.9%
9.9%
1.1%
0.9%
0.7%
0.5%
75%
50%
25%
0%
Proporción
Europeo
Mapuche
Otros
Aymara
Negro
Diaguita
Grupo

La Distribución

Para variables continuas, no podemos presentar una tabla de probabilidades. (¿por que?)

- Una posibilidad es transformarla en una variable discreta.
  - Histograma

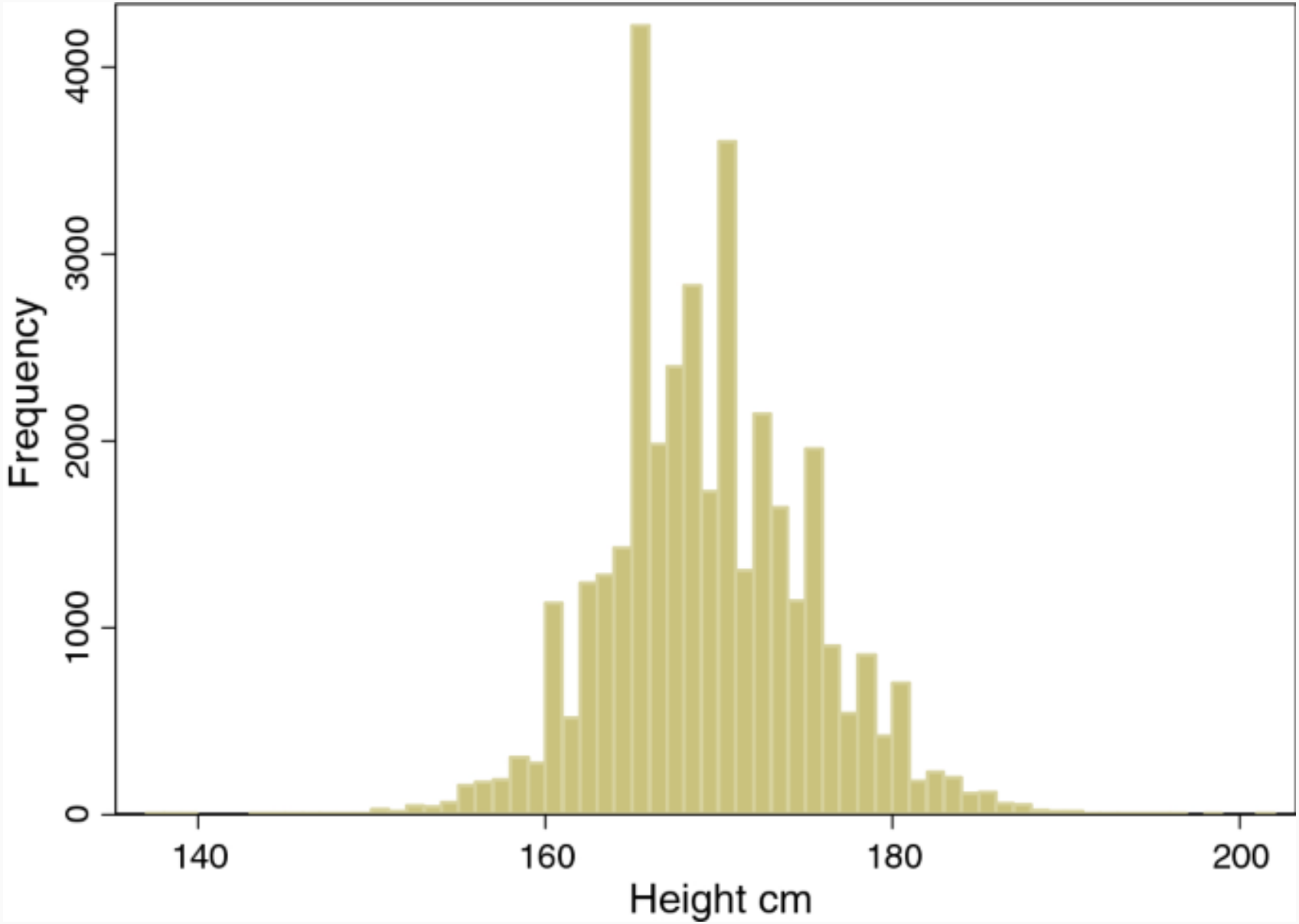
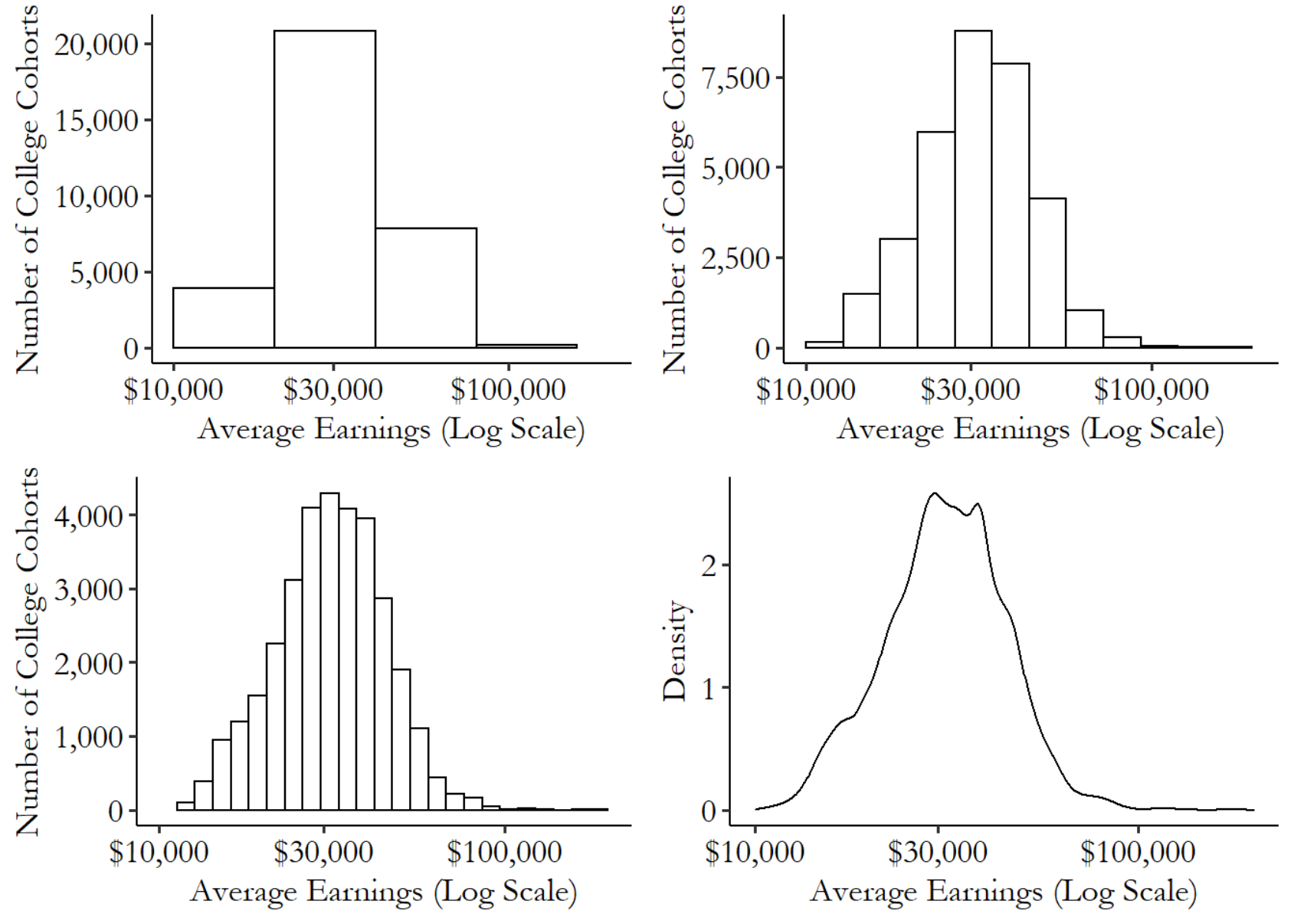


Figure 1: Altura de Hombres Chilenos

La Distribución

Otra forma de visualizar distribuciones es mediante la densidad.

Es similar a un histograma, pero con las categorias muy pequeñas.


Number of College Cohorts
20,000
15,000
10,000
5,000
0
$10,000
$30,000
$100,000
Average Earnings (Log Scale)
Number of College Cohorts
7,500
5,000
2,500
0
$10,000
$30,000
$100,000
Average Earnings (Log Scale)
Number of College Cohorts
4,000
3,000
2,000
1,000
0
$10,000
$30,000
$100,000
Average Earnings (Log Scale)
Density
2
1
0
$10,000
$30,000
$100,000
Average Earnings (Log Scale)

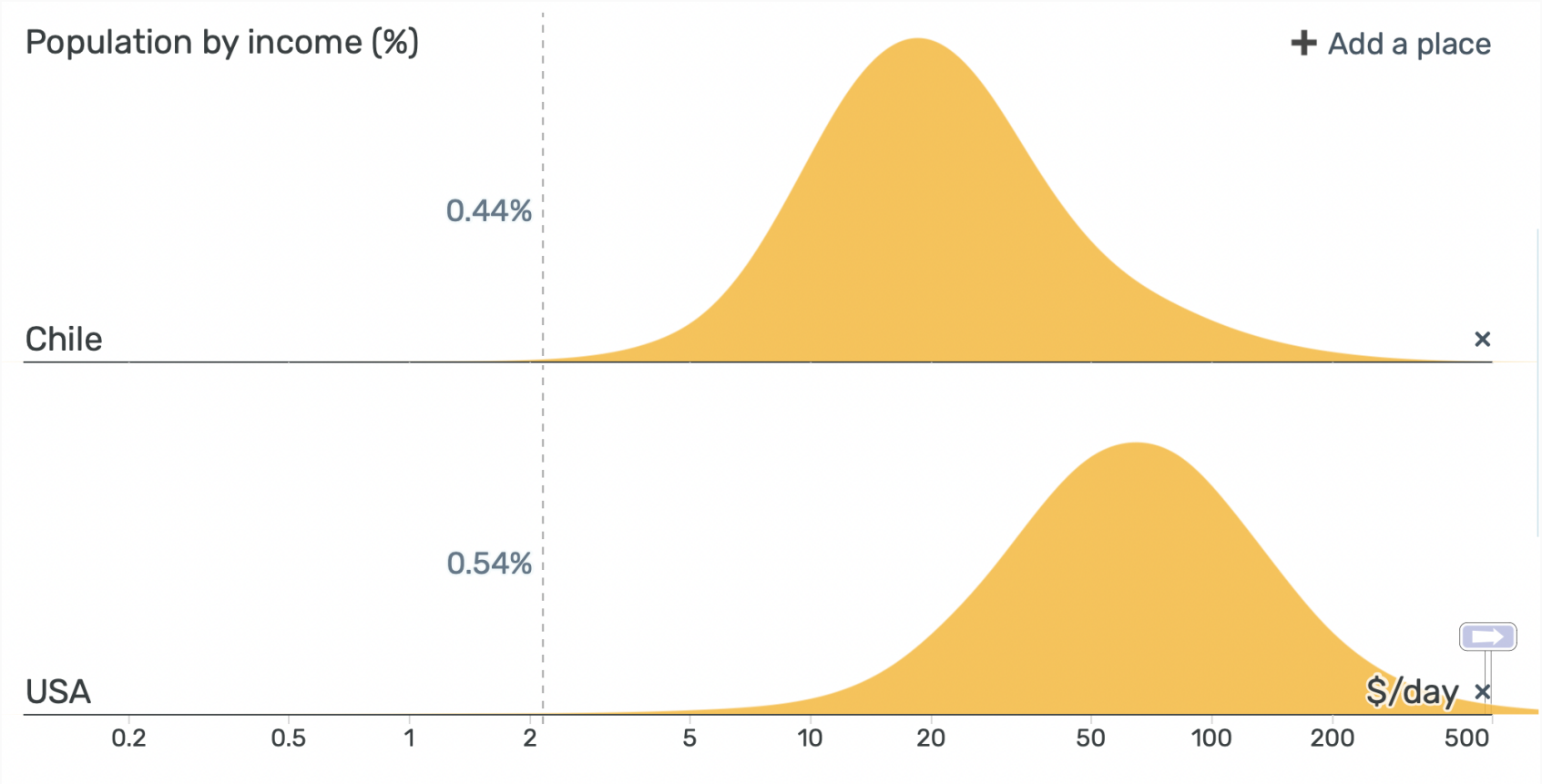


Figure 2: Distribución de Ingressos

Sumarizando la información

La distribución de una variable contiene mucha información.

Muchas veces deseamos un resumen de esa información.

La forma más usada es el valor promedio. Para variables discretas, definimos:

$$E[X] = \sum_{X} x \cdot Pr(X = x)$$

Es decir, multiplicamos cada valor por su probabilidad y sumamos todos los resultados.

Sumarizando la información

Alternativamente, para una dada muestra:

$$E[X] = \frac{1}{N} \sum x_i$$

Es decir, se suman todos los valores y se divide por el número de observaciones.

Sumarizando la información

El promedio es muy útil como medida de una observación “típica”.

Pero es muy influenciado por valores atípicos - Cando Jeff Bezos entra en una sala, el ingreso promedio aumenta muchísimo.

La **mediana** es más robusta a outliers. - Es el valor que está exactamente en el medio de la distribución: 50% arriba y 50% abajo.

Pero el promedio también tiene otras ventajas.

Sumarizando la información

Propiedades importantes del promedio:

1. Para constantes a y b, $E[a + bX] = a + bE[X]$.
2. Para N constantes a_n:
   $E[a_1X_1 + a_2X_2 + ... + a_NX_N] = a_1E[X_1] + a_2E[X_2] + ... + a_NE[X_N]$

¿Eso también es verdad para la mediana?

Sumarizando la información

También podemos resumir la dispersión de la distribución.

La medida más usada es la varianza.

$$Var(X) = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

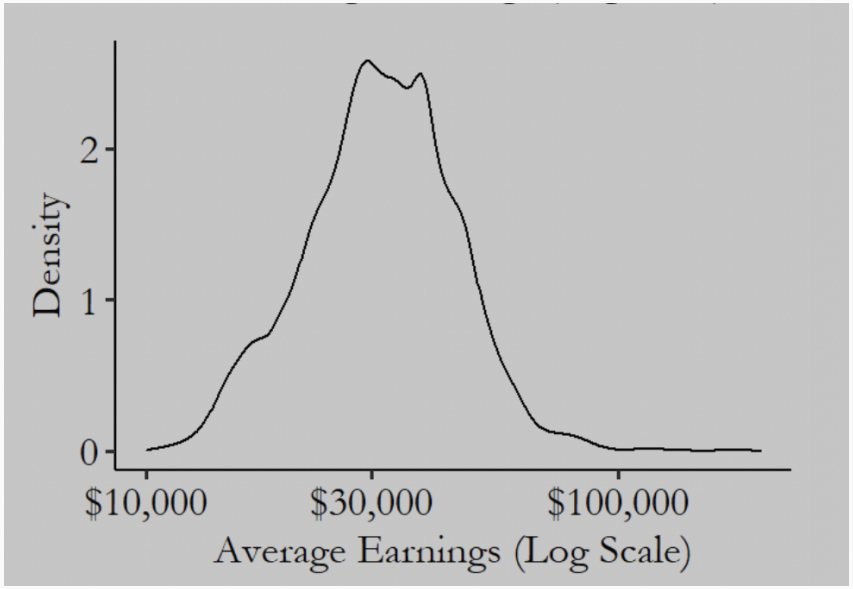
Es decir:

1. Se computa el promedio.
2. Se sustrae el promedio de cada observación.
3. Se toma el cuadrado de cada diferencia.
4. Se toma el promedio de los cuadrados.

Sumarizando la información

Como tomamos los cuadrados, la varianza se mide en unidades cuadradas.

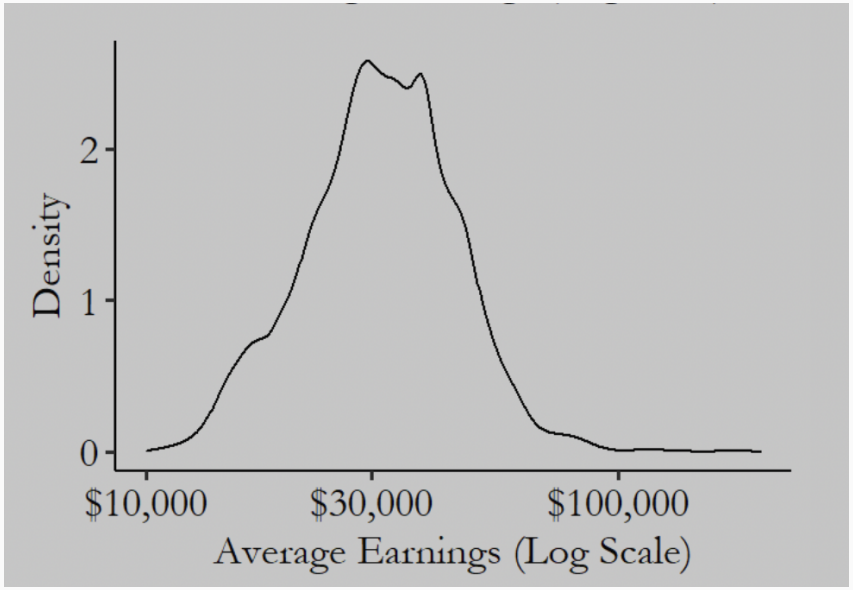
- La varianza en ingresos se mide en pesos cuadrados.



La varianza de esa distribución es 153,287,962 dólares cuadrados (?)

Sumarizando la información

La interpretación no es simple, por lo que es útil tomar la raíz cuadrada para obtener unidades naturales.



La desviación estándar de esta distribución es 12,380.95 dólares

También puede ser calculada como:

$$Var(X) = E[X^2] - E[X]^2$$

Sumarizando las relaciones entre variables

No hay mucho que pudamos decir sobre una única variable.

Nos interesan sobre todo las relaciones entre dos o más variables.

- Descripción
- Covariancia
- Distribuciones Condicionales
- Promedios Condicionales
- Independencia

Descripción

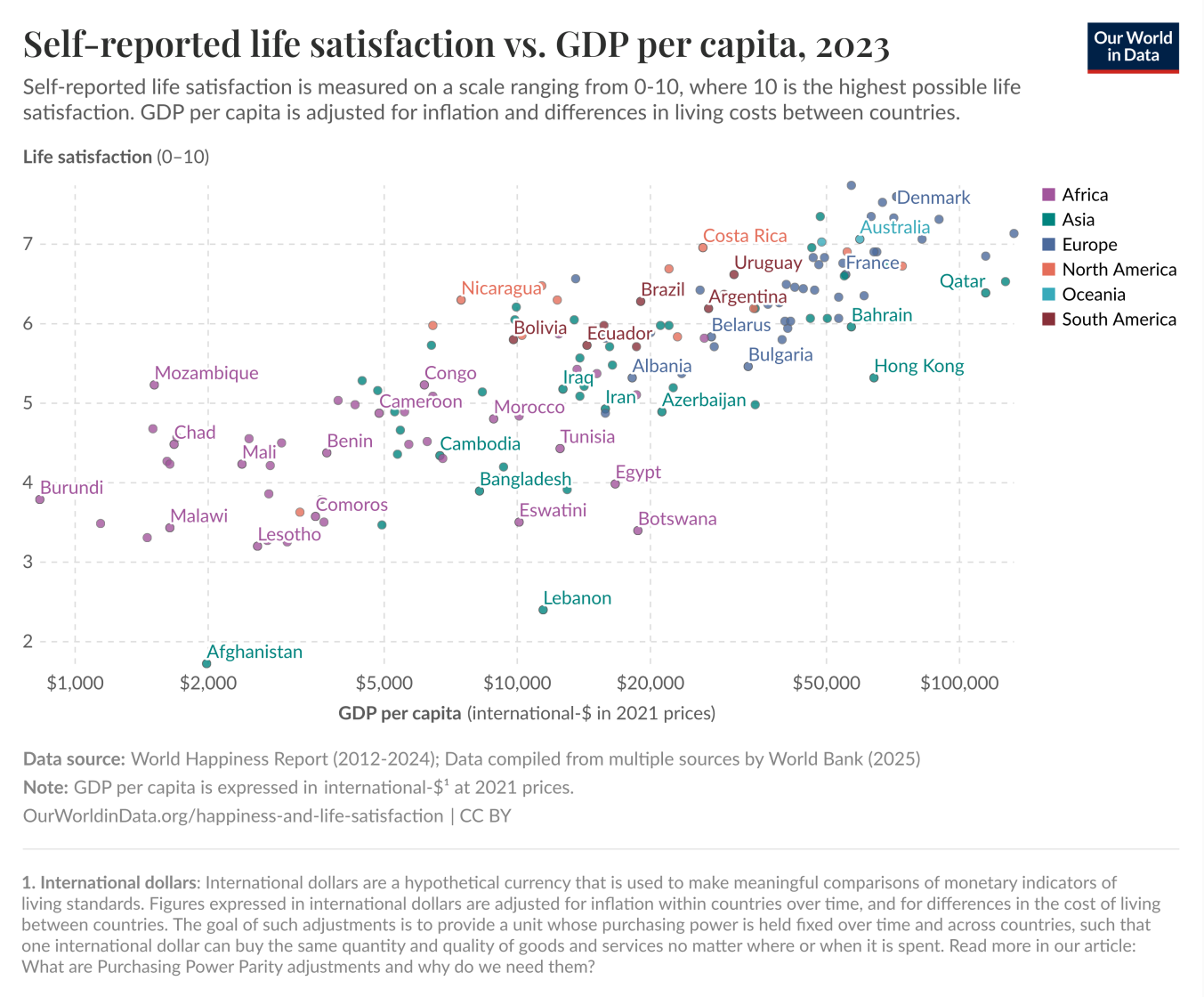
En casos más simples, podemos describir dos variables con una tabla de probabilidades.

Alfabetismo en Pakistán

Género	Sabe leer	No sabe leer
Hombres	39.85%	10.15%
Mujeres	32.6%	17.4%

Descripción

Para variables continuas, podemos usar el diagrama de dispersión.



1. **International dollars**: International dollars are a hypothetical currency that is used to make meaningful comparisons of monetary indicators of living standards. Figures expressed in international dollars are adjusted for inflation within countries over time, and for differences in the cost of living between countries. The goal of such adjustments is to provide a unit whose purchasing power is held fixed over time and across countries, such that one international dollar can buy the same quantity and quality of goods and services no matter where or when it is spent. Read more in our article: What are Purchasing Power Parity adjustments and why do we need them?

Descripción

Hay que tener cuidado con los diagramas de dispersión.

- Es muy fácil de inferir que la variable en el eje x causa la variable en el eje y. Pero es una elección arbitraria.

Covarianza

La covarianza es una medida de la dependencia lineal entre dos variables.

Se calcula como:

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

Note que Cov(X,X) = Var(X).

Y que $Cov(a_1 + b_1 X, a_2 + b_2 Y) = b_1 b_2 Cov(X, Y)$

Covarianza

La covarianza es normalmente difícil de interpretar, porque depende de la magnitud de X y Y.

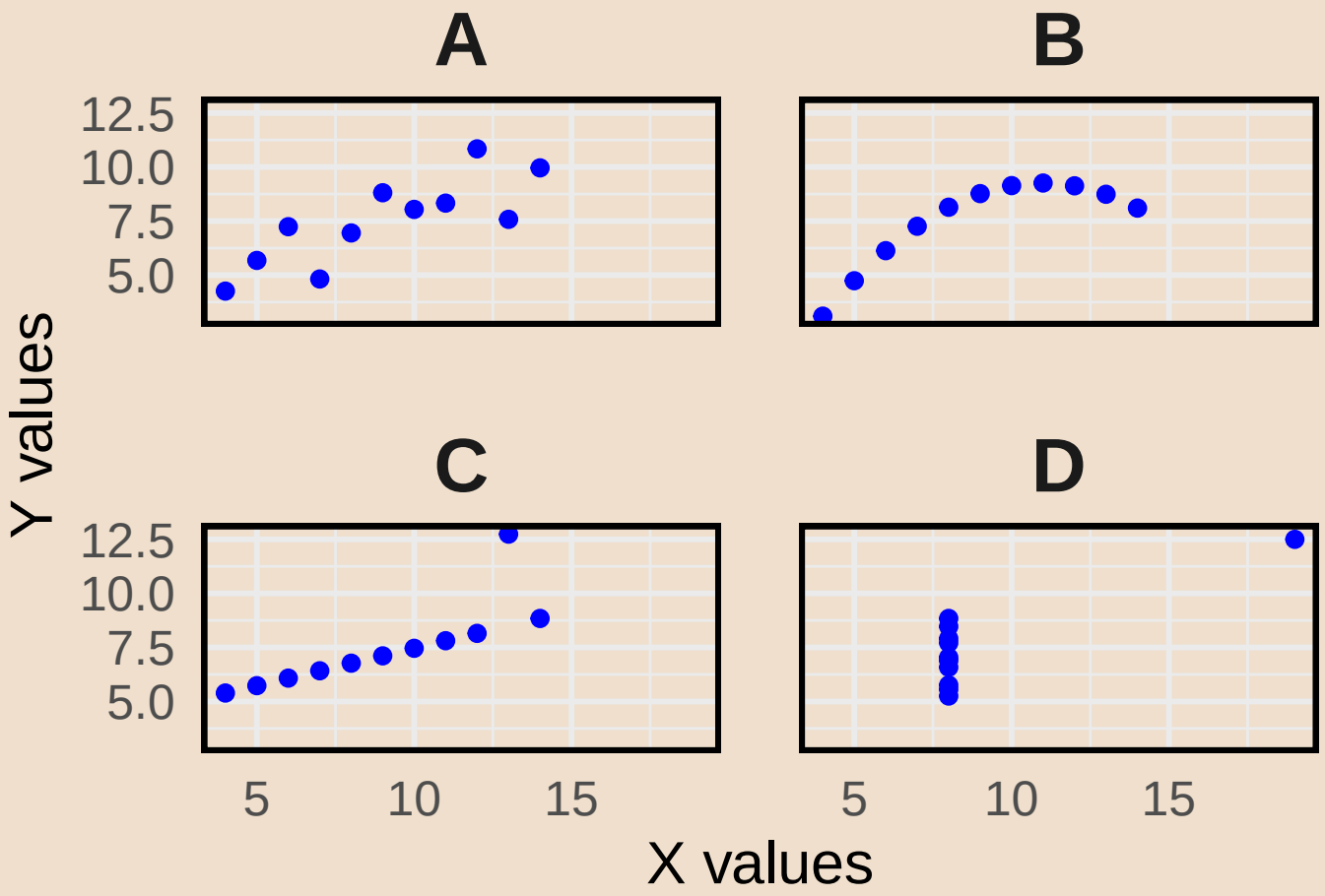
Una alternativa es transformarla para que esté entre -1 y 1.

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Refleja cuánto los datos están agrupados en una relación lineal.

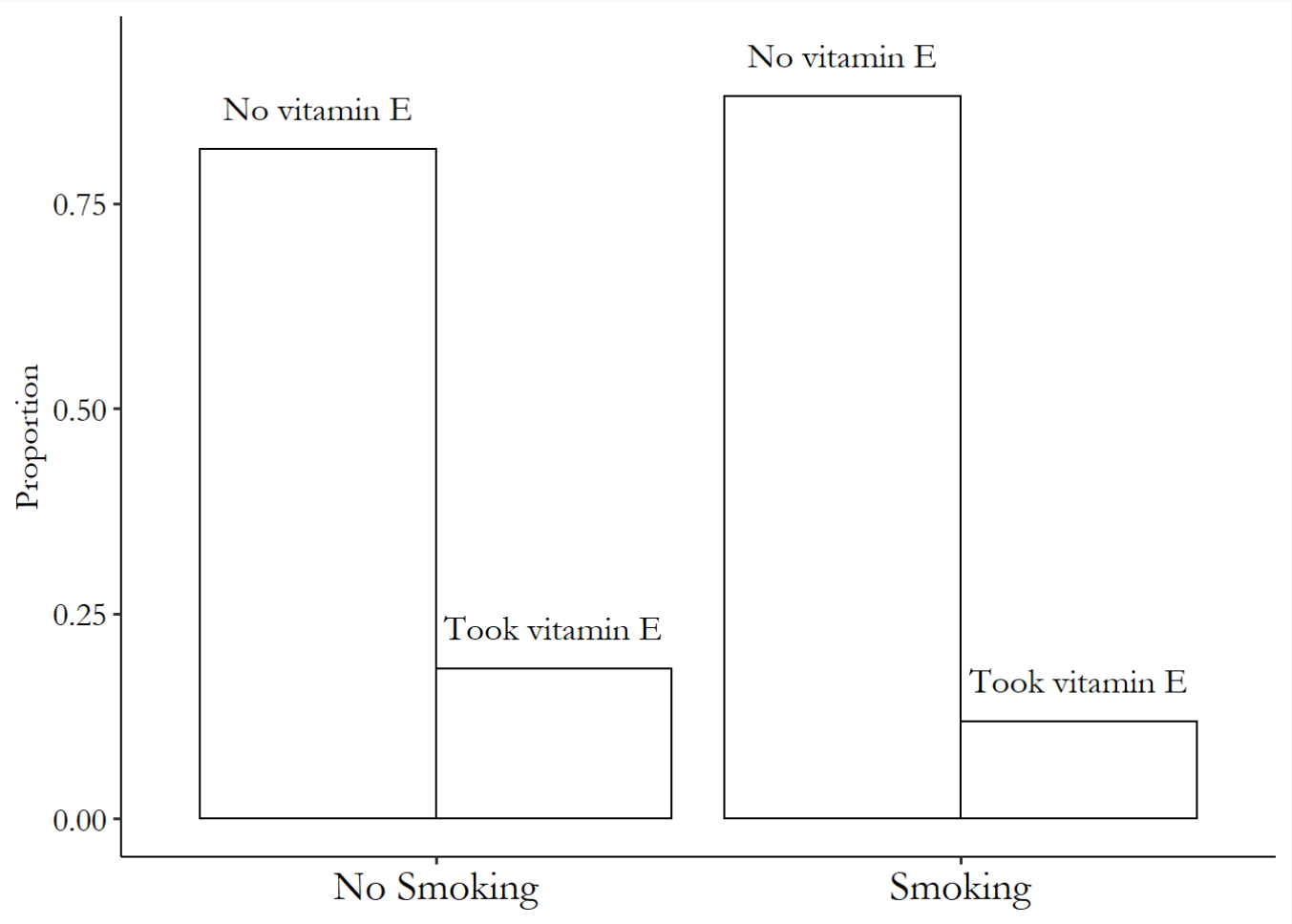
Pregunta

¿Cuál tiene el mayor coeficiente de correlación?

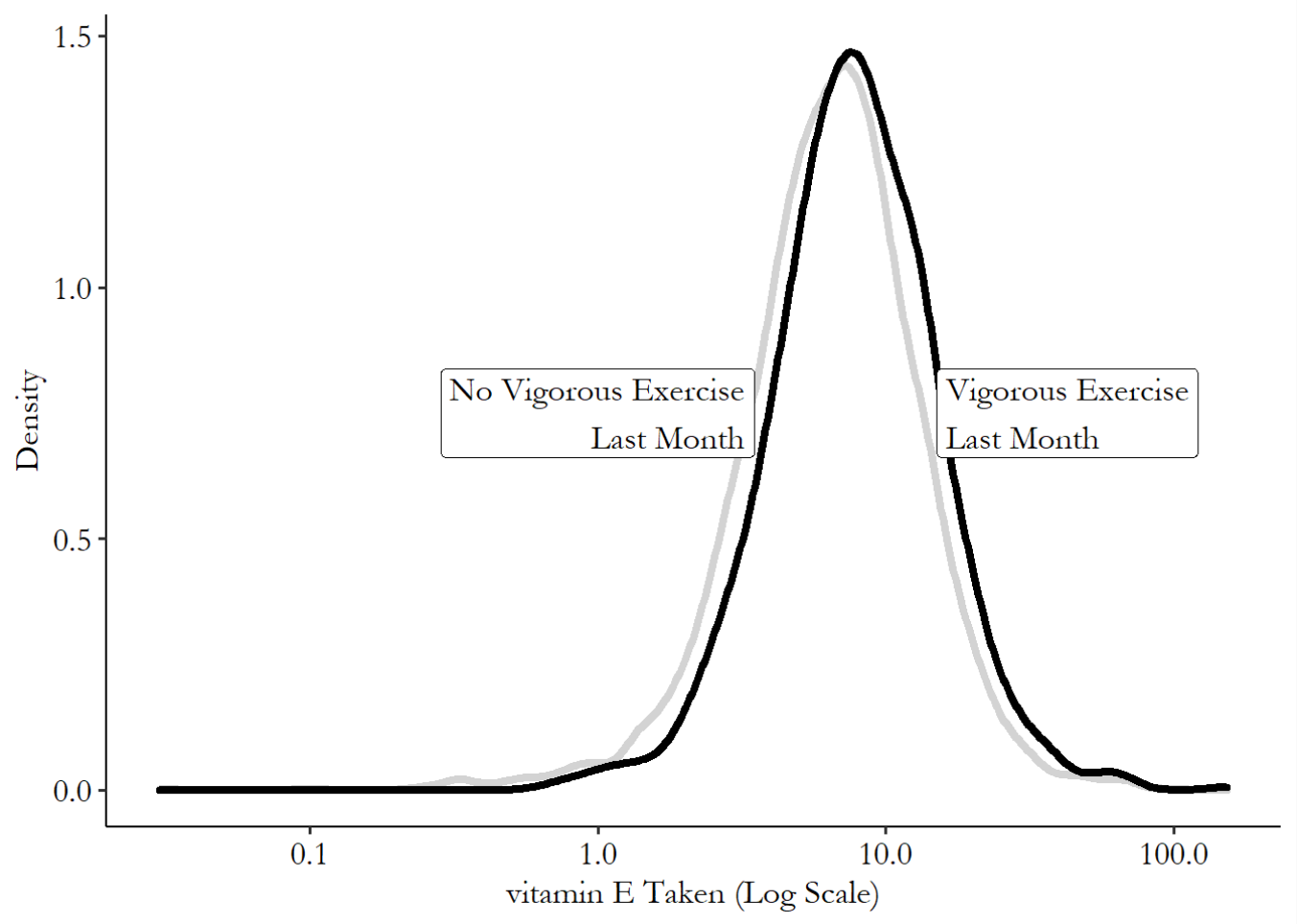


Distribuiciones Condicionales

Una distribución condicional es simplemente la distribución de una variable cuando se fija otra variable en un valor específico.



Distribuciones Condicionales



Probabilidad y Esperanza Condicional

Podemos trabajar con distribuciones condicionales de la misma manera que con distribuciones de una variable.

Escribimos $Pr(Y|X = x)$ para significar la probabilidad de Y condicional a X=x.

También escribimos $Pr(Y|X)$ para significar una *función* de x, que representa la probabilidad de Y condicional en cada valor de X.

Relacionamos la probabilidad conjunta, condicional y incondicional así:

Probabilidad Condicional

$$\Pr(X \cap Y) = \Pr(X|Y) \cdot \Pr(Y)$$

Ejercicios

Ejercicio 1

¿Cuál es la probabilidad de lanzar dos dados, y la soma ser 10 o más, dado que el dado menor es más que 2?

Ejercicio 1

¿Cuál es la probabilidad de lanzar dos dados, y la soma ser 10 o más, dado que el dado menor es más que 2?

Solución

Las formas de obtenir soma 10 son: (4,6), (5,5) y (6,4).

Los resultados posibles van de 3 a 6 para cada dado, o sea, 4 x 4 = 16.

Respuesta: 3/16

Ejercicio 2

Un hombre tiene 2 hijos/hijas. ¿Cuál es la probabilidad de que sean dos hijas, si sabemos que al menos una es una hija.

a. 1/2
b. 1/3
c. 1/4
d. 1/8
e. Other

Ejercicio 2

Un hombre tiene 2 hijos/hijas. Cual es la chance de que sean dos hijas, si sabemos que a lo menos una es una hija.

Solución

En total, existen cuatro posibilidades; (H,H), (H,M), (M,H), (M,M)

Pero sabemos que (H,H) no es el caso.

Solo una es un caso favorable.

Respuesta: *B: 1/3*

Regla de Bayes

Manipulando la ecuación de la probabilidad condicional, obtemos:

$$\Pr(X \cap Y) = \Pr(X|Y) \cdot \Pr(Y) = \Pr(Y|X) \cdot \Pr(X)$$

Regla de Bayes

$$\Pr(X|Y) = \frac{\Pr(Y|X) \cdot \Pr(X)}{\Pr(Y)}$$

Ejercício

Tres vasos contienen 10 bolas cada uno.

- Vaso 1: 4 bolas rojas y 6 bolas azules
- Vaso 2: 7 bolas rojas y 3 bolas azules
- Vaso 3: 2 bolas rojas y 8 bolas azules

Elegimos un vaso al azar y extraemos una bola. Si la bola es roja, cual es la probabilidad de que viene del vaso 2?

Ejercício

- Vaso 1: 4 bolas rojas y 6 bolas azules
- Vaso 2: 7 bolas rojas y 3 bolas azules
- Vaso 3: 2 bolas rojas y 8 bolas azules

Queremos saber $\Pr(V_2|R)$.

$$\Pr(V_2|R) = \frac{\Pr(R|V_2)\cdot\Pr(V_2)}{\Pr(R)}$$

$$\Pr(R|V_2) = \frac{7}{10}\ \Pr(V_2) = \frac{1}{3}\ \Pr(R) = \frac{13}{30}$$

$$\Pr(V_2|R) = \frac{\frac{7}{10}\cdot\frac{1}{3}}{\frac{13}{30}} = \frac{7}{13}$$

Ejercício

Una prueba médica para una condición rara tiene un 98% de chance de identificar correctamente una persona que tiene esa condición, y un 95% de chance de identificar correctamente una persona que no la tiene. Si un 1% de la población tiene la condición, ¿cuál es la probabilidad de que una persona la tiene, si un teste es positivo?

Ejercício

Queremos saber $\Pr(C|+)$.

Dados:

- $\Pr(C) = 0.01$
- $\Pr(+|C) = 0.98$
- $\Pr(+) = \Pr(+ \cap C) + \Pr(+ \cap NC) =$
  $\Pr(+|C) \cdot \Pr(C) + \Pr(+|NC) \cdot \Pr(NC)$
- $\Pr(+) = 0.98 \cdot 0.01 + 0.05 \cdot 0.99$

$$\Pr(C|+) = \frac{\Pr(+|C) \cdot \Pr(C)}{\Pr(+)}$$

$$\Pr(C|+) = \frac{0.98 \cdot 0.01}{0.98 \cdot 0.01 + 0.05 \cdot 0.99} = 16.5\%$$

Esperanza Condicional

Otra forma de mostrar como dos variables se relacionan es la esperanza condicional.

Simplemente tomamos la distribución condicional y calculamos la esperanza.

$$E[Y|X = x] = \sum y \cdot Pr(Y = y|X = x)$$

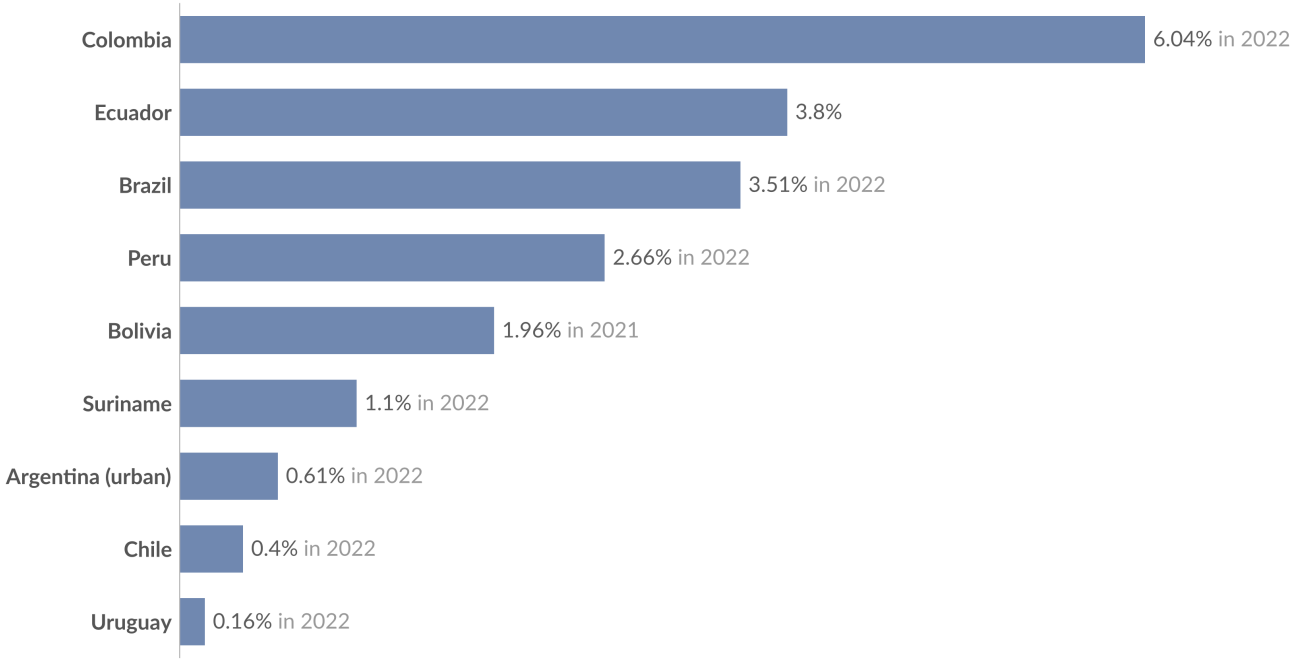
Esperanza Condicional

Es muy simple con variables discretas.

Share of population living in extreme poverty, 2023

Our World in Data

Extreme poverty is defined as living below the International Poverty Line of $2.15 per day. This data is adjusted for inflation and for differences in living costs between countries.



Data source: World Bank Poverty and Inequality Platform (2024)

Note: This data is expressed in international-$[1] at 2017 prices. Depending on the country and year, it relates to income measured after taxes and benefits, or to consumption, per capita[2].

1. **International dollars**: International dollars are a hypothetical currency that is used to make meaningful comparisons of monetary indicators of living standards. Figures expressed in international dollars are adjusted for inflation within countries over time, and for differences in the cost of living between countries. The goal of such adjustments is to provide a unit whose purchasing power is held fixed over time and across countries, such that one international dollar can buy the same quantity and quality of goods and services no matter where or when it is spent. Read more in our article: What are Purchasing Power Parity adjustments and why do we need them?

2. **Per capita (income)**: "Per capita" here means that each person (including children) is attributed an equal share of the total income received by all members of their household.

Esperanza Condicional

Supón que tienes las esperanzas condicionales para cada valor de X.

¿Cómo se puede calcular la esperanza no condicional?

Ejemplo: dadas las tasas de pobreza en cada país, ¿cómo calcular la tasa de pobreza de Latinoamérica?

¿Como se puede calcular la esperanza no-condicional?

Podemos calcular la tasa de pobreza en Latinoamérica multiplicando cada tasa por la población del país, sumando, y dividiendo por la población total.

En general:

Ley de Esperanzas Iteradas

$$E[X] = \sum_{y} E(X|Y = y) \cdot \Pr(y) = E[E[X|Y]]$$

Independencia

Un concepto central es el de independencia estadística.

Decimos que dos variables son independientes si la información sobre una no dice nada sobre la distribución de la otra.

Definición

$$X \perp Y \iff Pr(X = x|Y = y) = Pr(X = x), \forall x, y$$

Es decir, la distribución condicional de X es la misma, no importa lo que pase con Y.

Independencia

Alternativamente, si X y Y son independientes,
$Pr(x, y) = Pr(x)Pr(y)$.

Pregunta

Se lanza un dado dos veces. El evento A es “el primer resultado es impar” y el evento B es “la suma de los resultados es par”. ¿Los eventos A y B son independientes?

Independencia

El primero resultado es impar: [1,3,5]

$$\Pr(A) = \frac{1}{2}$$

La suma es par: hay 18 resultados de 36.

$$\Pr(B) = \frac{1}{2}$$

Si son independientes, la probabilidad conjunta debe ser $1/4$.

Para obtener una suma par, con el primero impar, hay 9 posibilidades.

$$\Pr(A \cap B) = \frac{9}{36} = \frac{1}{4}$$

Pregunta

Se lanza un dado dos veces. El evento A es “el primer resultado es par” y el evento B es “la suma de los resultados es 8”. ¿Los eventos A y B son independientes?

Independencia

$$\Pr(A) = \frac{1}{2}$$

Posibilidades de suma 8: (2,6),(3,5),(4,4),(5,3),(6,2).

$$\Pr(B) = \frac{5}{36}$$

Posibilidades de suma 8, con el primer número par: (2,6),(4,4),(6,2).

$$\Pr(A \cap B) = \frac{3}{36}$$

$$\Pr(A) \cdot \Pr(B) = \frac{1}{2} \cdot \frac{5}{36} = \frac{2.5}{36} \neq \frac{3}{36}$$

Independencia en promedio

Cuando Y es independiente de X, toda la distribución de Y es la misma para cualquier valor de X.

Una hipótesis más débil es la *independencia en media*. Y es independiente en media de X si su esperanza condicional es la misma para todos los valores de X.

Independencia en promedio

X/Y	0	1
-1	1/3	0
0	0	1/3
1	1/3	0

- X y Y son independientes?
- X es independiente en media de Y?
- Y es independiente en media de X?
- Cual es la correlación entre X y Y?

Conceptos de independencia

- Independencia: $Pr(x, y) = Pr(x)Pr(y)$
- Independencia en media: $E(Y|X) = E(Y)$
- Correlación cero: $E(XY) = E(X)E(Y)$

Cero correlación e independencia son simétricos, pero independencia en media no lo es.

Independencia implica independencia en media. Independencia en media implica correlación cero.