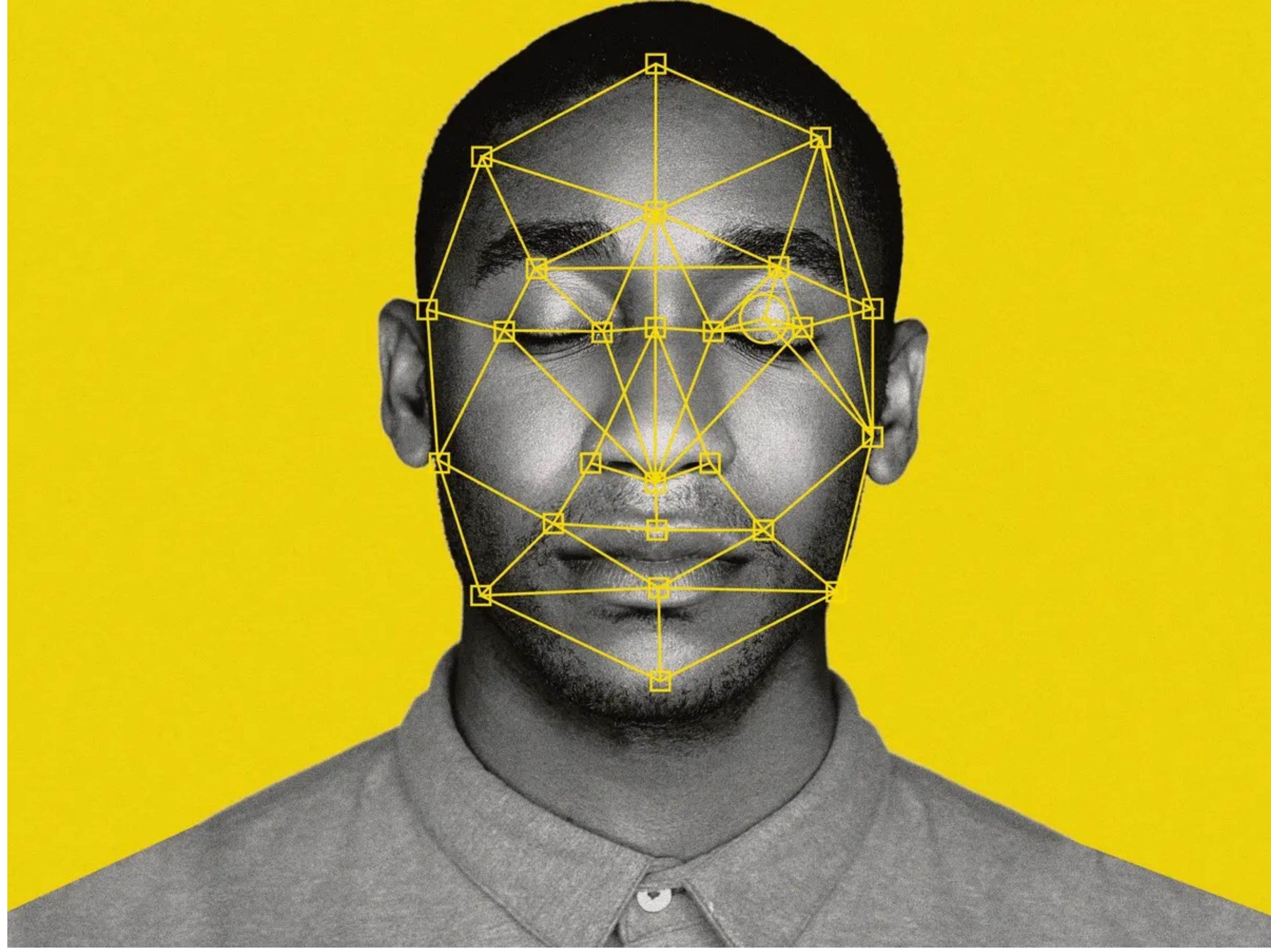
Variables aleatorias

Sebastián Ruiz-Blais, I-2024

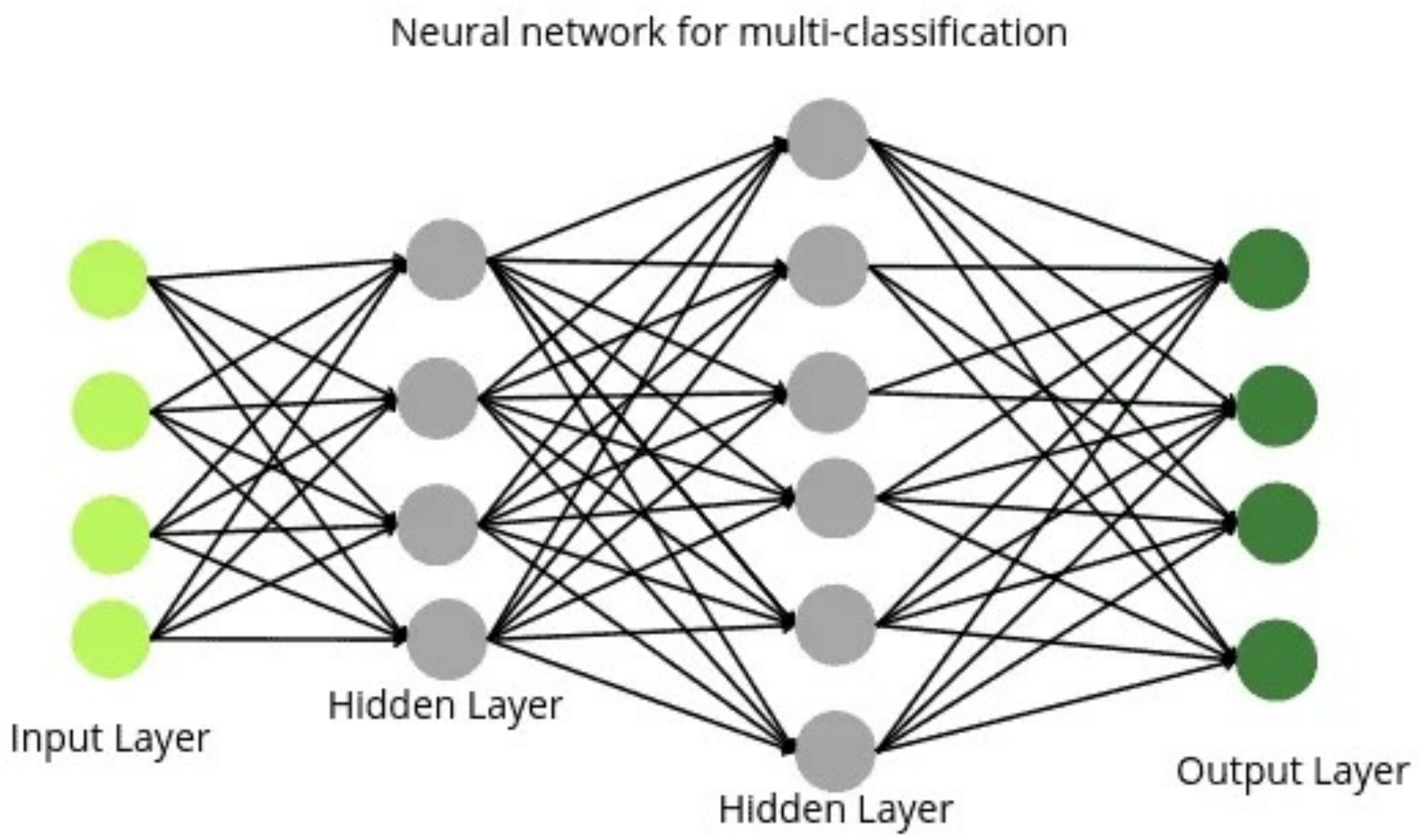
Contenidos

- Variables aleatorias
- Función de distribución de probabilidad
- Distribuciones conjuntas y marginales

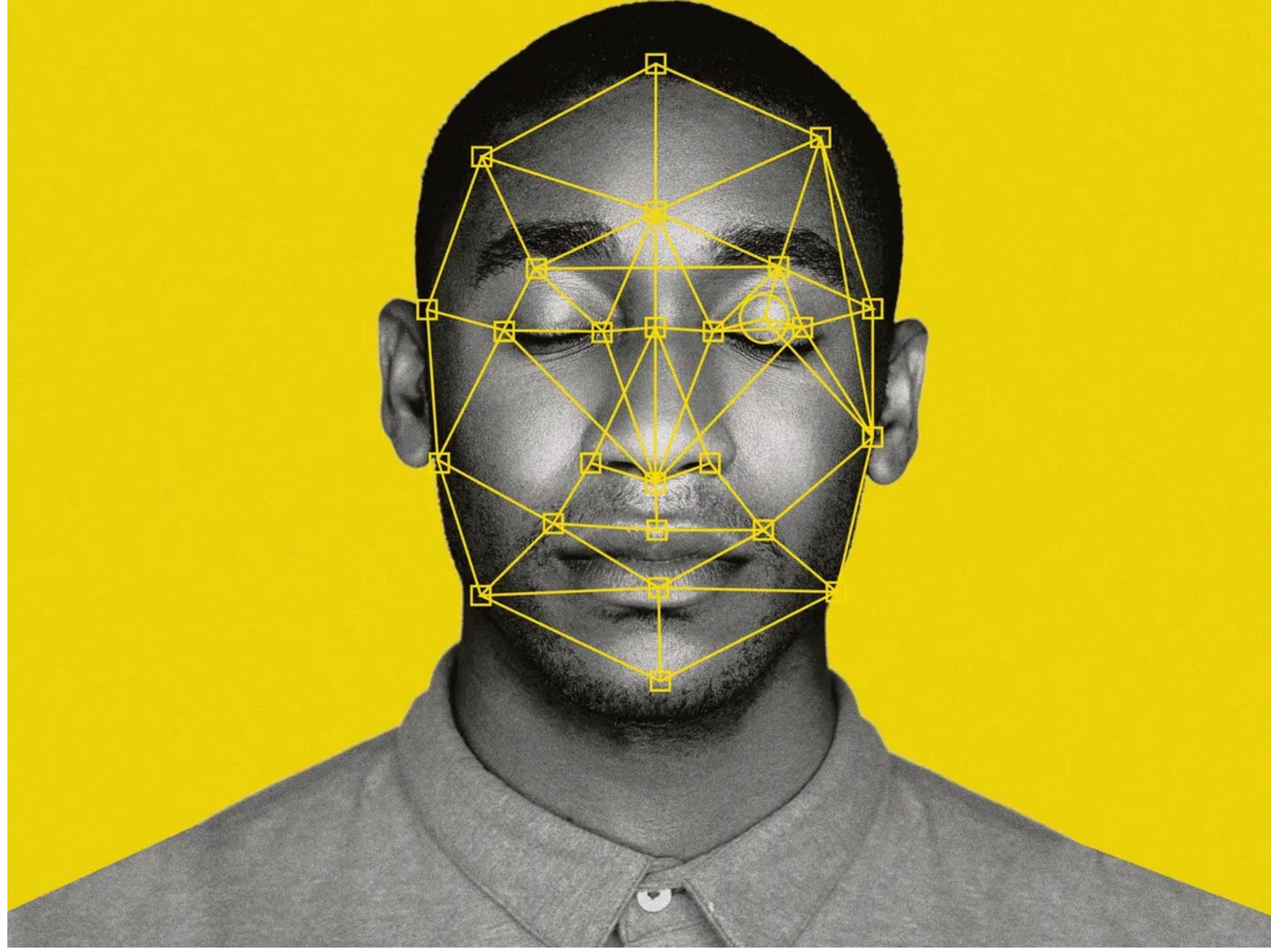
Problema



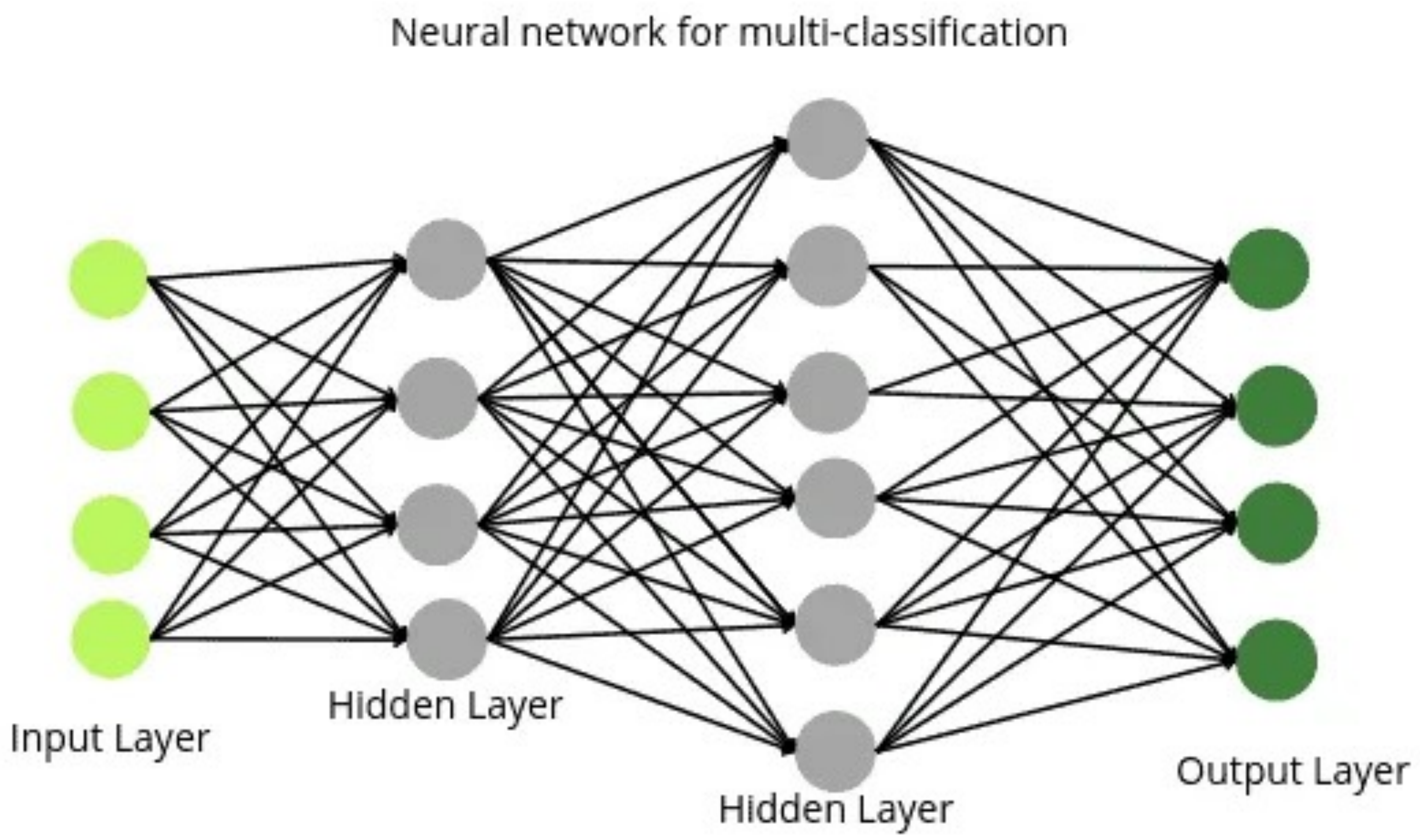
- Una empresa desarrolla un sistema de IA que detecta rostros y los asigna a un nombre. Dependiendo de las condiciones, el sistema puede dar una respuesta correcta o no.
  - Si se tiene 3 imágenes de prueba, **¿cuáles serían todos los resultados posibles**, teniendo que C:correcto e I:incorrecto?
  - Nos interesa el número de errores que realiza el sistema. **¿En cuantos casos distintos hay exactamente 2 errores**?



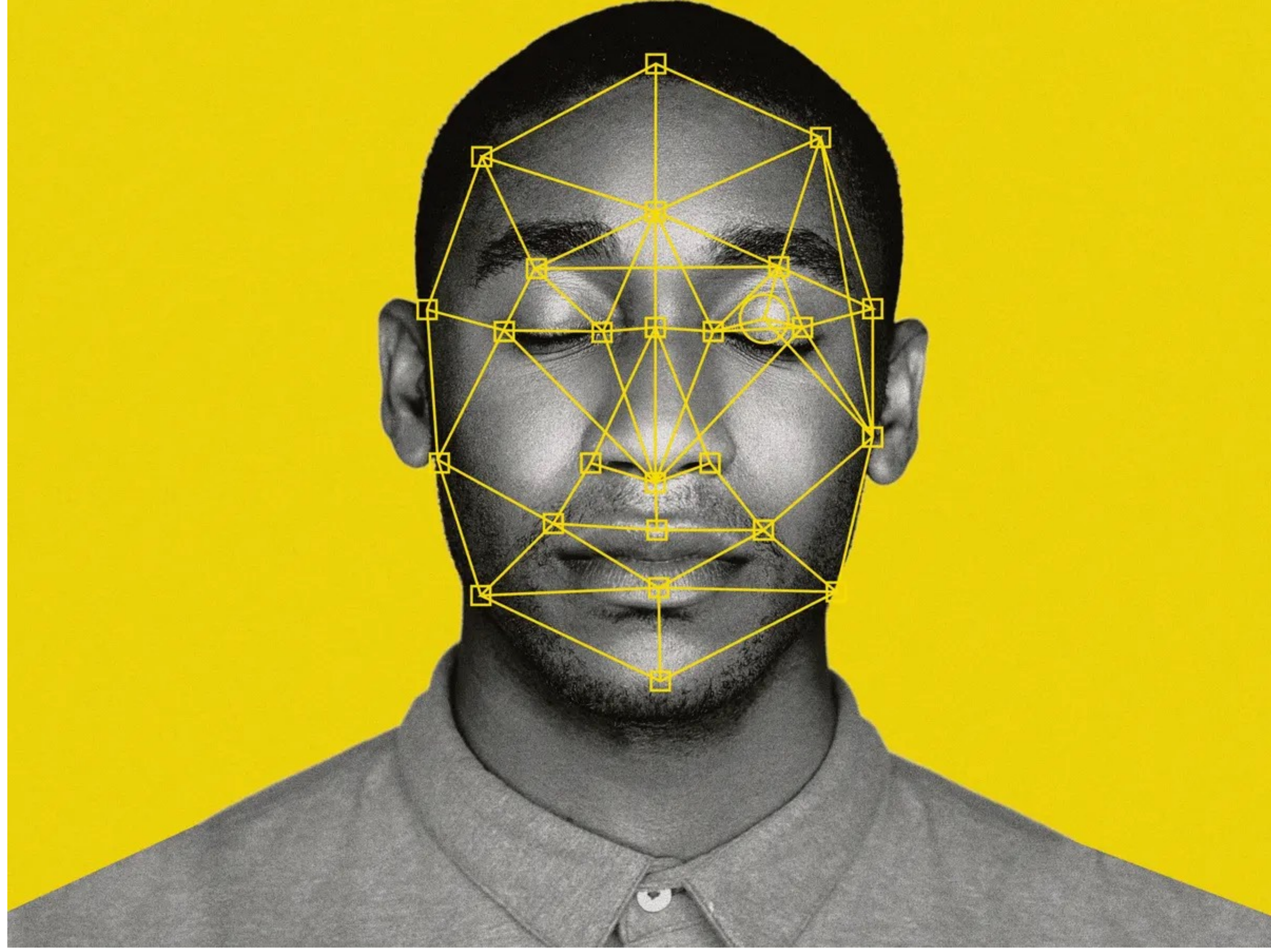
Problema



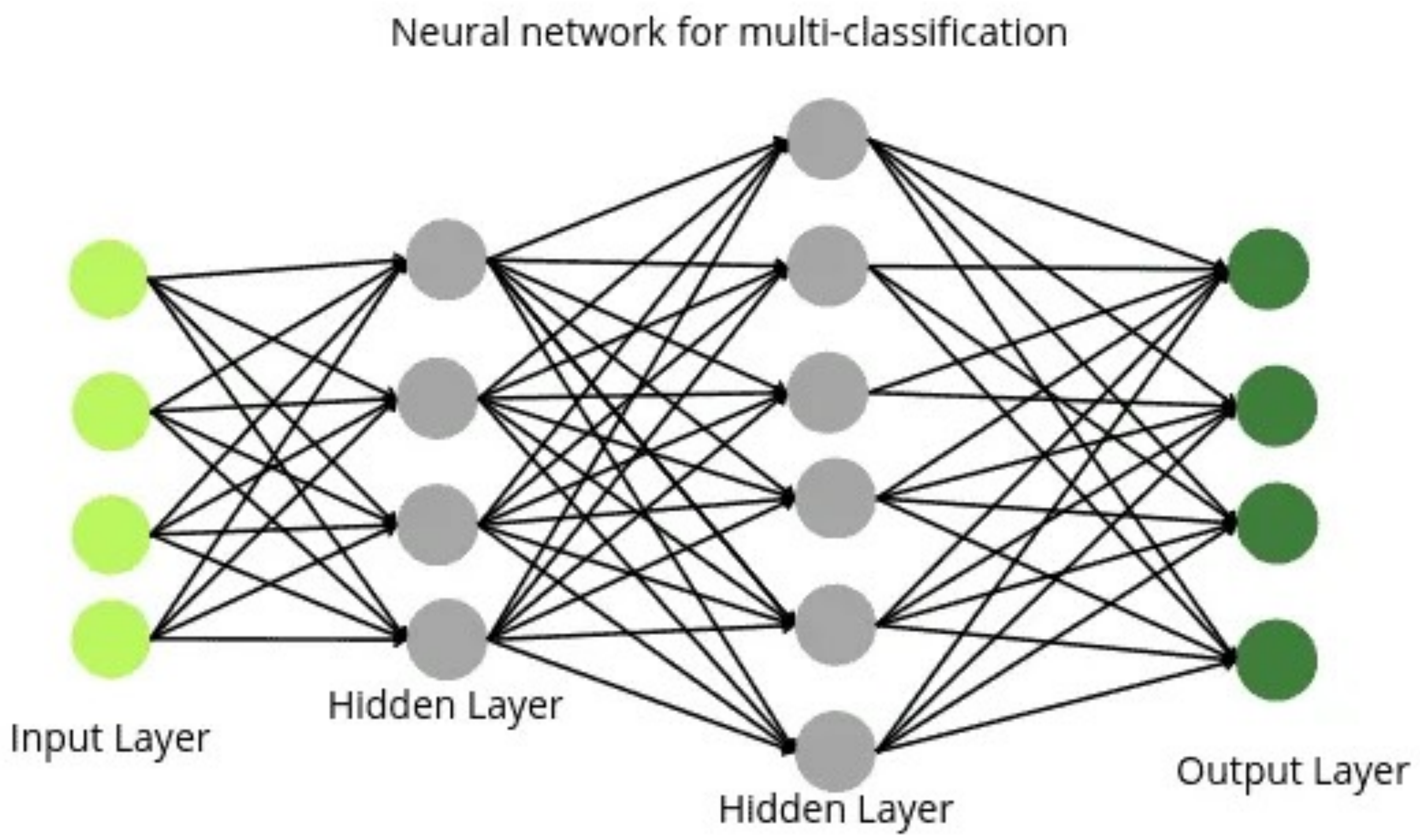
- Los resultados posibles serían:
  - CCC,
  - CCI, CIC, ICC,
  - CII, ICI, IIC,
  - III
- El número de resultados que cumplen con el evento que hay exactamente dos errores es de 3 (tercera fila)



Problema



- Una empresa desarrolla un sistema de IA que detecta rostros y los asigna a un nombre. Dependiendo de las condiciones, el sistema puede dar una respuesta correcta o no.
    - Si la **probabilidad** de que una cara sea clasificada **correctamente es de 0.90, ¿cuál es la probabilidad de que clasifique correctamente las siguientes 5 imágenes de caras?**



Variables aleatorias

- **Definición**. Una variable aleatoria es una función que *asocia* cada elemento del espacio muestral con un *número real*.
    - Por ejemplo, la **suma** de dos dados (d_1 y d_2):

+						
	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

$d_1+d_2=3$

$d_1+d_2=8$

$d_1+d_2=12$

Variables aleatorias

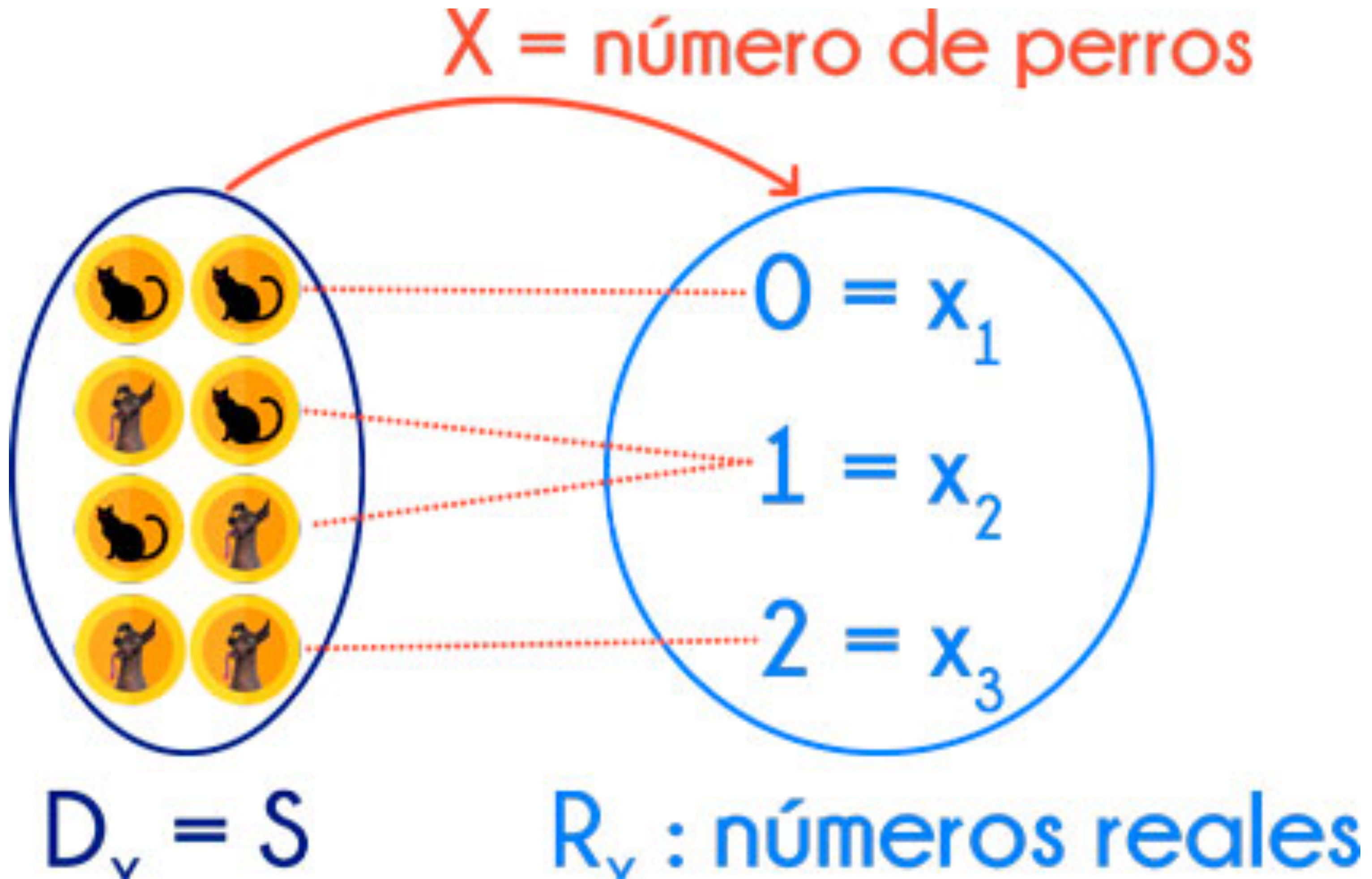
- Por ejemplo, sea **X** la **cantidad de minutos** que una persona tarda en ser atendida en atención al cliente del BNCR

  - X = 5
  - X = 15
  - X = 120
  - …

Espacio muestral discreto

- **Definición**. Si un espacio muestral contiene un número finito de posibilidades, o una serie interminable con tantos elementos como números enteros existen, se llama **espacio muestral discreto**.

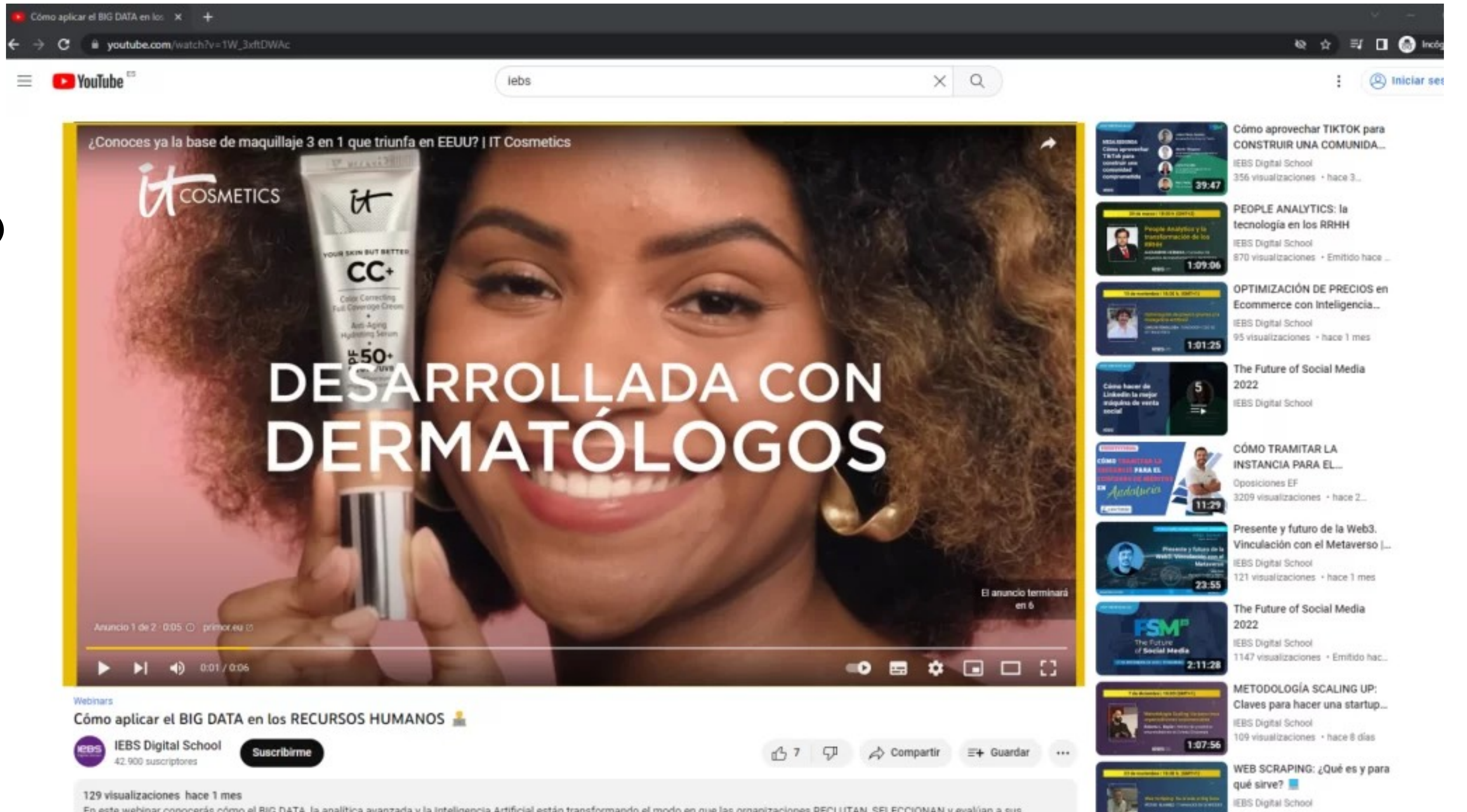
- Por ejemplo, el **número de ítems defectuosos**, el número de errores realizados por un sistema de machine learning.



Espacio muestral continuo

- **Definición.** Si un espacio muestral contiene un número infinito de posibilidades, igual al número de puntos en un segmento de recta, se le denomina **espacio muestral continuo**

- Por ejemplo, la **proporción de personas** que consumen un producto habiendo visto un anuncio en Youtube.



Función de probabilidad discreta

Definición 3.4: El conjunto de pares ordenados $(x, f(x))$ es una **función de probabilidad**, una **función de masa de probabilidad** o una **distribución de probabilidad** de la variable aleatoria discreta X si, para cada resultado x posible,

1. $f(x) \geq 0$,
2. $\sum_x f(x) = 1$,
3. $P(X = x) = f(x)$.

Distribución de probabilidad

Función de probabilidad

Walpole (2012), p.84

Función de masa de probabilidad

Problema

- Una escuela compra **dos computadoras** de una tienda con buen descuento. Sin embargo, se sabe que 3 **de las 20 computadoras** de la tienda están defectuosas. Calcule la **distribución de probabilidad** para el número de computadoras **defectuosas** comprada por la escuela.

Solución

- Sea X la variable aleatoria y *x* los números posibles de computadoras defectuosas compradas por la escuela. *x* puede ser 0, 1 ó 2.
- Buscamos:
  - $f(0) = P(X = 0)$
  - $f(1) = P(X = 1)$
  - $f(2) = P(X = 2)$

Solución

- $f(0) = P(X = 0) = \dfrac{\binom{3}{0} \times \binom{17}{2}}{\binom{20}{2}} = \dfrac{68}{95}$
- $f(1) = P(X = 1) = \dfrac{\binom{3}{1} \times \binom{17}{1}}{\binom{20}{2}} = \dfrac{51}{190}$
- $f(2) = P(X = 2) = \dfrac{\binom{3}{2} \times \binom{17}{0}}{\binom{20}{2}} = \dfrac{3}{190}$

Problema

- Se tiene una variable aleatoria X que representa los **goles que anota un equipo de fútbol durante un partido**. La probabilidad de anotar un gol es de **0.3,** de anotar dos goles: **0.15**, de anotar tres goles, **0.05**, de anotar cuatro o más goles **0.01**.
- ¿Cuál es la probabilidad de **anotar un gol o menos**?

Función de distribución acumulativa

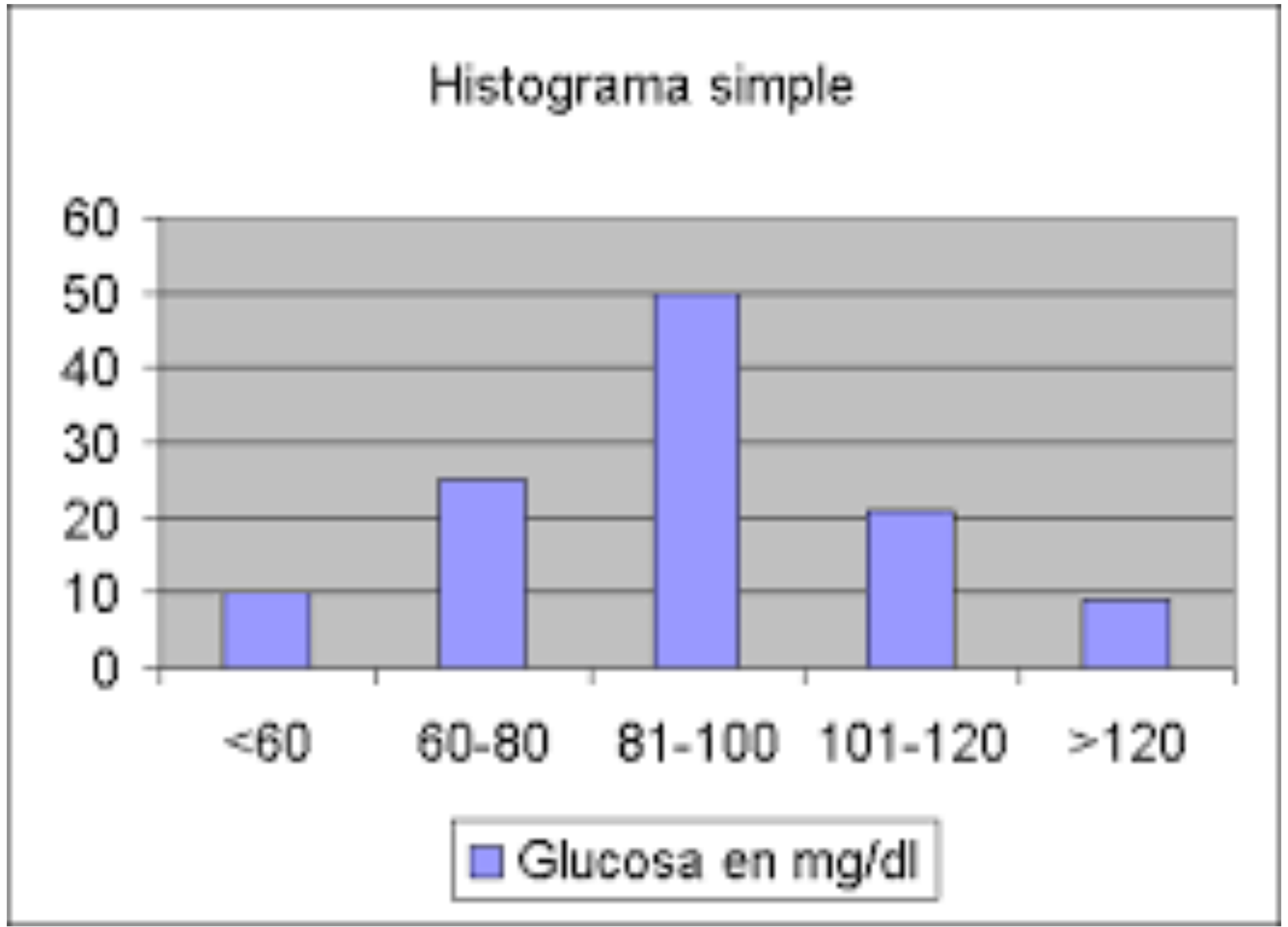
Definición 3.5: La **función de la distribución acumulativa** $F(x)$ de una variable aleatoria discreta X con distribución de probabilidad $f(x)$ es

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t), \quad \text{para} \quad -\infty < x < \infty.$$

Histograma de probabilidad

Experimento

- Se plantea el experimento de tirar la moneda 5 veces y anotar el número de caras.
- Se repite el experimento 10 veces y se anota cuántas veces aparece cada número.
- Construimos un histograma de probabilidad



Función de probabilidad continua

- Para un espacio muestral continuo, ¿cuál es la **probabilidad de un valor específico**?
  - P.ej., una distancia en metros o la temperatura de una reacción química.

Función de probabilidad continua

- Piense en una **función para generar números aleatorios computacionalmente entre 0 y 1**.

- ¿Cuál es la probabilidad de que ocurra un número en específico?

Función de probabilidad continua

- Esa probabilidad es cero.

- Por lo tanto, cuando se tienen variables aleatorias continuas, usualmente interesa la probabilidad de que ocurran **valores dentro de un rango**

Experimento: Números aleatorios

- Utilice la función numpy.random.uniform para generar 20 números en el rango [0,10]
- Calcule la probabilidad de obtener un número menor que 3
- Calcule la probabilidad de obtener un número entre 5 y 9 (inclusive)

Experimento: Números aleatorios

- Utilice números más grandes para obtener una distribución más precisa

Experimento: Números aleatorios

- Ahora repita el mismo procedimiento utilizando numpy.random.normal para generar números aleatorios con media 5 y desviación estándar de 1.5

Función de probabilidad continua

Definición 3.6: La función $f(x)$ es una **función de densidad de probabilidad** (fdp) para la variable aleatoria continua X, definida en el conjunto de números reales, si

1. $f(x) \geq 0$, para toda $x \in R$.
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
3. $P(a < X < b) = \int_a^b f(x)\, dx$.

Función de densidad de probabilidad

Lo mismo que función de probabilidad para el caso continuo

Función de probabilidad continua

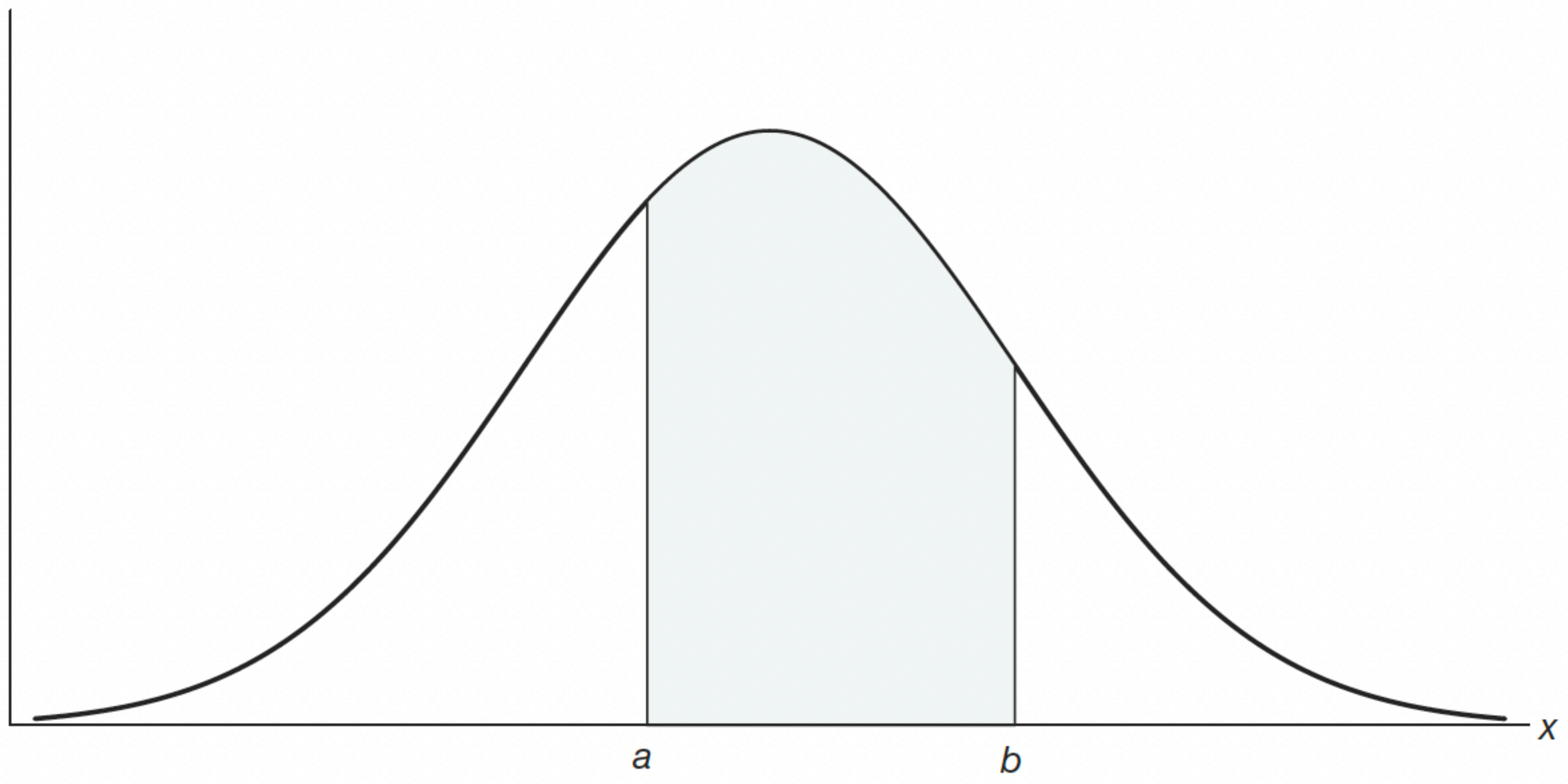


Figura 3.5: $P(a < X < b)$.

Función de distribución acumulativa

- **Ejercicio**: ¿Cuál es la función de distribución acumulativa de la función $f(x) = 0.1$, definida entre 0 y 10 (o fuera de ese rango)? Obtenga los valores para x=1, 4, 7.5 y 10

Función de distribución acumulativa

- **Ejercicio**: ¿cuál es la función de distribución acumulativa de la función $f(x) = x$, definida entre 0 y 1? Obtenga los valores 0.1, 0.5, 0.8.

Función de distribución acumulativa

Definición 3.7: La **función de distribución acumulativa** $F(x)$, de una variable aleatoria continua X con función de densidad $f(x)$, es

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt, \text{ para } -\infty < x < \infty.$$

Walpole (2012), p.90

Ejercicios

1

- Clasifique las siguientes variables aleatorias como discretas o continuas:
- X: el número de accidentes automovilísticos que ocurren al año en Cartago.
- Y: el tiempo de reposición total de un partido de fútbol.
- M: la cantidad de leche que una vaca específica produce anualmente.
- N: el número de huevos que una gallina pone mensualmente.
- P: el número de permisos para construcción que los funcionarios de una ciudad emiten cada mes.
- Q: el peso del arroz producido por hectárea.

Ejercicios

2

3.5 Determine el valor c de modo que cada una de las siguientes funciones sirva como distribución de probabilidad de la variable aleatoria discreta X:

a) $f(x) = c(x^2 + 4)$, para $x = 0, 1, 2, 3$;

b) $f(x) = c\binom{2}{x}\binom{3}{3-x}$, para $x = 0, 1, 2.$

Ejercicios

3

3.6 La vida útil, en días, para frascos de cierta medicina de prescripción es una variable aleatoria que tiene la siguiente función de densidad:

$$f(x) = \begin{cases} \frac{20,000}{(x+100)^3}, & x > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

Calcule la probabilidad de que un frasco de esta medicina tenga una vida útil de

a) al menos 200 días;
b) cualquier lapso entre 80 y 120 días.

Distribuciones de probabilidad conjunta

- En algunos casos, se considera a **dos variables aleatorias simultáneamente**.
- Esto lleva a **espacios muestrales multi-dimensionales**.
- Interesa en ese caso la **probabilidad conjunta de las dos variables**

Problema

- Se seleccionan al azar 2 canicas de una caja que contiene 3 canicas azules, 2 rojas y 3 verdes. Si X es el número de canicas azules y Y es el número de canicas rojas seleccionadas, calcule
- a) la función de probabilidad conjunta $f(x, y)$,
- b) $P[(X, Y) \in A]$, donde A es la región $\{(x, y) \mid x + y \leq 1\}$.

Solución

Pista

- Los posibles pares de valores (x, y) son $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$, $(0, 2)$ y $(2, 0)$.

Solución (a)

	$f(x, y)$	$x=0$	$x=1$	$x=2$	Totales por renglón
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Totales por columna		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Solución (b)

La probabilidad de que (X,Y) caiga en la región A es

$$P[(X,\ Y) \in A] = P(X + Y \leq 1) = f(0,0) + f(0,1) + f(1,0)$$
$$= \frac{3}{28} + \frac{3}{14} + \frac{9}{28} = \frac{9}{14}.$$

Distribuciones de probabilidad conjunta

Caso discreto

Definición 3.8: La función $f(x,y)$ es una **distribución de probabilidad conjunta** o **función de masa de probabilidad** de las variables aleatorias discretas X y Y, si

1. $f(x, y) \geq 0$ para toda (x, y),
2. $\sum_x \sum_y f(x, y) = 1$,
3. $P(X = x, Y = y) = f(x, y)$.

Para cualquier región A en el plano xy, $P[(X, Y) \in A] = \sum\sum_A f(x,y)$.

Walpole (2012), p.95

Problema

Una empresa privada opera un local que da servicio a clientes que llegan en automóvil y otro que da servicio a clientes que llegan caminando. En un día elegido al azar, sean X y Y, respectivamente, las proporciones de tiempo que ambos locales están en servicio, y suponiendo que la función de densidad conjunta de estas variables aleatorias es

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0, & \text{en otro caso.} \end{cases}$$

- a) Verifique que la doble integral de f(x,y) es igual a 1
- b) Calcule P [(X, Y) ∈A], donde A = {(x, y) | 0 < x < 0.5, 0.25< y < 0.5}.

***Solución*: *a*)** La integración de $f(x,y)$ sobre la totalidad de la región es

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = \int_0^1\int_0^1 \frac{2}{5}(2x + 3y)\, dx\, dy$$

$$= \int_0^1 \left(\frac{2x^2}{5} + \frac{6xy}{5}\right)\Bigg|_{x=0}^{x=1} dy$$

$$= \int_0^1 \left(\frac{2}{5} + \frac{6y}{5}\right) dy = \left(\frac{2y}{5} + \frac{3y^2}{5}\right)\Bigg|_0^1 = \frac{2}{5} + \frac{3}{5} = 1.$$

b) Para calcular la probabilidad utilizamos

$$P[(X, Y) \in A] = P\left(0 < X < \frac{1}{2}, \frac{1}{4} < Y < \frac{1}{2}\right)$$

$$= \int_{1/4}^{1/2}\int_0^{1/2} \frac{2}{5}(2x + 3y)\, dx\, dy$$

$$= \int_{1/4}^{1/2} \left(\frac{2x^2}{5} + \frac{6xy}{5}\right)\Bigg|_{x=0}^{x=1/2} dy = \int_{1/4}^{1/2} \left(\frac{1}{10} + \frac{3y}{5}\right) dy$$

$$= \left(\frac{y}{10} + \frac{3y^2}{10}\right)\Bigg|_{1/4}^{1/2}$$

$$= \frac{1}{10}\left[\left(\frac{1}{2} + \frac{3}{4}\right) - \left(\frac{1}{4} + \frac{3}{16}\right)\right] = \frac{13}{160}.$$

Distribuciones de probabilidad conjunta

Caso continuo

Definición 3.9: La función $f(x,y)$ es una **función de densidad conjunta** de las variables aleatorias continuas X y Y si

1. $f(x, y) \geq 0$, para toda (x, y),
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$,
3. $P[(X, Y) \in A] = \int \int_A f(x, y)\, dx\, dy$, para cualquier región A en el plano xy.

Walpole (2012), p.96

Distribuciones marginales

- Las **distribuciones marginales** de una función de densidad conjunta f(x,y) corresponden a las funciones que sólo dependen de sus variables aleatorias X y Y.
- En otras palabras, corresponden a la probabilidad de que ocurra X o Y, independientemente de la otra variable.

Ejercicio

- Verifique usando la tabla siguiente las distribuciones marginales de las variable X y Y, respectivamente.

$f(x, y)$		x = 0	x = 1	x = 2	Totales por renglón
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
Totales por columna		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Distribuciones marginales

Definición 3.10: Las **distribuciones marginales** sólo de X y sólo de Y son

$$g(x) = \sum_y f(x, y) \quad \text{y} \quad h(y) = \sum_x f(x, y)$$

para el caso discreto, y

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy \quad \text{y} \quad h(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$$

para el caso continuo.

Walpole (2012), p.97

Distribución condicional

- Al igual que con la probabilidad de eventos, las variables aleatorias también se pueden expresar en términos de **distribución condicional**.

Distribución condicional

Definición 3.11: Sean X y Y dos variables aleatorias, discretas o continuas. La **distribución condicional** de la variable aleatoria Y, dado que $X = x$, es

$$f(y|x) = \frac{f(x, y)}{g(x)}, \quad \text{siempre que } g(x) > 0.$$

De manera similar, la distribución condicional de la variable aleatoria X, dado que $Y = y$, es

$$f(x|y) = \frac{f(x, y)}{h(y)}, \quad \text{siempre que } h(y) > 0.$$

Walpole (2012), p.99

Independencia estadística

- Al igual que con la probabilidad de eventos, las variables aleatorias se pueden considerar independientes o no.
- **Definición: Son independientes dos variables X y Y si la distribución conjunta f(x,y) equivale al producto de las distribuciones marginales de X y de Y**.

Problema

- ¿Son las variables aleatorias X y Y independientes para el ejemplo de canicas?
  - Se seleccionan al azar 2 canicas de una caja que contiene 3 canicas azules, 2 rojas y 3 verdes. X es el número de canicas azules y Y es el número de canicas rojas seleccionadas.
  - **Pista**: obtenga las funciones marginales y f(0,1).

Solución

$$f(0,1) = \frac{3}{14},$$

$$g(0) = \sum_{y=0}^{2} f(0,y) = \frac{3}{28} + \frac{3}{14} + \frac{1}{28} = \frac{5}{14},$$

$$h(1) = \sum_{x=0}^{2} f(x,\ 1) = \frac{3}{14} + \frac{3}{14} + 0 = \frac{3}{7}.$$

Claramente,

$$f(0,1) \neq g(0)h(1),$$

- Por lo tanto, X y Y no son estadísticamente independientes, ya que la probabilidad conjunta es distinta del producto de ambas.

Independencia estadística

Definición 3.12: Sean X y Y dos variables aleatorias, discretas o continuas, con distribución de probabilidad conjunta $f(x, y)$ y distribuciones marginales $g(x)$ y $h(y)$, respectivamente. Se dice que las variables aleatorias X y Y son **estadísticamente independientes** si y sólo si

$$f(x, y) = g(x)h(y)$$

para toda (x,y) dentro de sus rangos.

Walpole (2012), p.102

Independencia estadística

Generalización

Definición 3.13: Sean $X_1, X_2, \ldots, X_n$, n variables aleatorias, discretas o continuas, con distribución de probabilidad conjunta $f(x_1, x_2, \ldots, x_n)$ y distribuciones marginales $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$, respectivamente. Se dice que las variables aleatorias $X_1, X_2, \ldots, X_n$ son recíproca y **estadísticamente independientes** si y sólo si

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n)$$

para toda $(x_1, x_2, \ldots, x_n)$ dentro de sus rangos.

Walpole (2012), p.103

Ejercicio

- Suponga que el tiempo de vida en almacén de cierto producto comestible perecedero empacado en cajas de cartón, en años, es una variable aleatoria cuya función de densidad de probabilidad está dada por
  - $f(x) = e^{-x}$, $x > 0$,
  - 0, en otro caso.
- Represente los tiempos de vida en almacén para tres de estas cajas seleccionadas de forma independiente con X_1, X_2 y X_3 y calcule $P(X_1 < 2, 1 < X_2 < 3, X_3 > 2)$.

Solución

***Solución*:** Como las cajas se seleccionaron de forma independiente, suponemos que las variables aleatorias X_1, X_2 y X_3 son estadísticamente independientes y que tienen la siguiente densidad de probabilidad conjunta:

$$f(x_1, x_2, x_3) = f(x_1)f(x_2)f(x_3) = e^{-x_1}e^{-x_2}e^{-x_3} = e^{-x_1-x_2-x_3},$$

para $x_1 > 0$, $x_2 > 0$, $x_3 > 0$, y $f(x_1, x_2, x_3) = 0$ en cualquier otro caso. En consecuencia,

$$P(X_1 < 2, 1 < X_2 < 3, X_3 > 2) = \int_2^\infty \int_1^3 \int_0^2 e^{-x_1-x_2-x_3}\, dx_1\, dx_2\, dx_3$$
$$= (1 - e^{-2})(e^{-1} - e^{-3})e^{-2} = 0.0372.$$