Estadística descriptiva – Práctica 6

Probabilidad y Estadística (C)

Daniela Parada

2 de junio de 2023

Plan de trabajo de hoy

Algunas definiciones iniciales

Población: conjunto total de los sujetos o unidades de análisis de interés en un estudio.

Muestra: cualquier subconjunto de sujetos o unidades de análisis de la población en estudio.

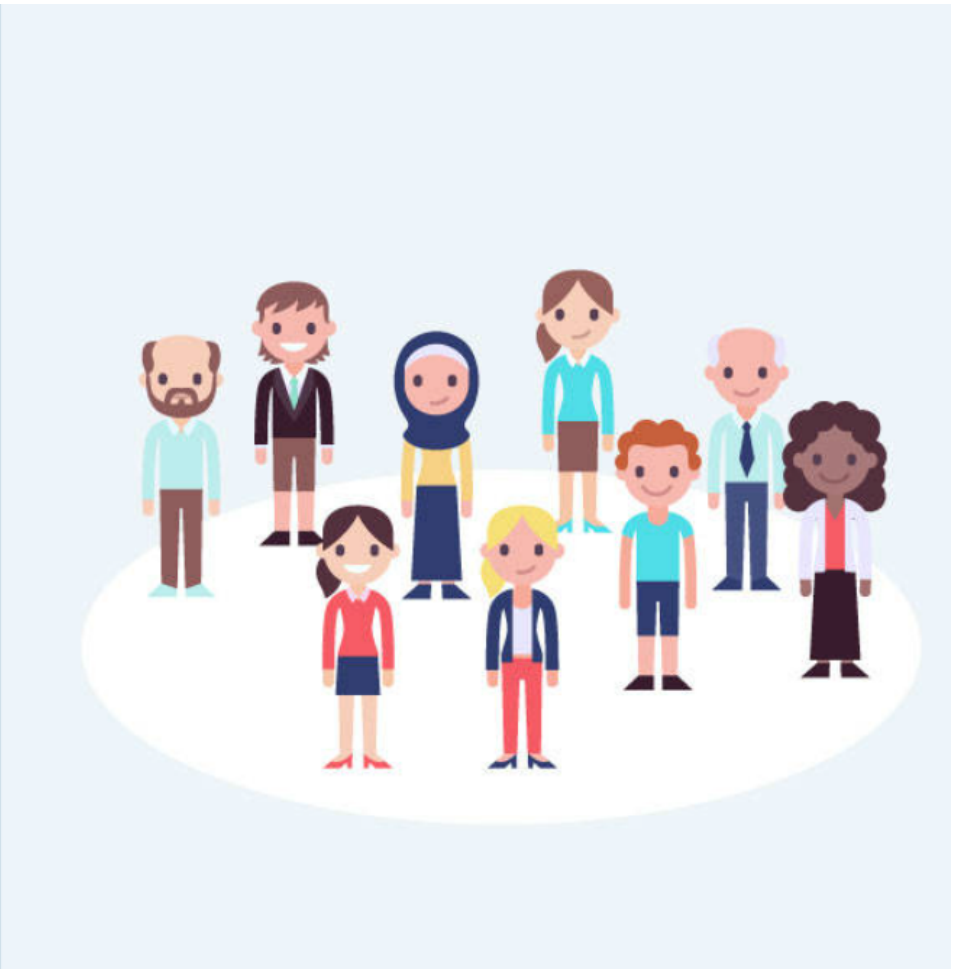
Variable: cualquier característica de la unidad de observación que interese registrar (nos restringimos a los casos en que al registrar la variable es o puede ser transformada en un número).

Valor de una variable, dato, observación, medición: número que describe a la característica de interés en el objeto de estudio.

Usamos la **estadística descriptiva** usualmente para:

- organizar, sintetizar o presentar la información,
- descubrir, visualizar sus características más relevantes, ...

Notación: $x_1, \ldots, x_n$ vs. $X_1, \ldots, X_n$



▸ Animación de TCL de Phillip Plewa.

Estadística descriptiva

Usualmente, vamos a denotar a una muestra aleatoria como $X_i, \ldots, X_n$ donde cada una de las X_i serán variables aleatorias independientes e idénticamente distribuidas a X (bajo alguna distribución conocida, o no). Y como tales, podrían tener esperanza, varianza, distribución, distribución conjunta, etc.

En cambio, notaremos $x_1, \ldots, x_n$ a los datos de una muestra, es decir, los valores observados de la realización de la muestra aleatoria.

Medidas de posición

Sean $x_1, \ldots, x_n$ las n observaciones de una variable en una muestra y sean $x^{(i)}$ las observaciones ordenadas, con $1 \leq i \leq n$. Es decir:

$$x^{(1)} \leq x^{(2)} \leq \ldots \leq x^{(n)}.$$

Media muestral

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mediana muestral

$$\tilde{x} = \begin{cases} x^{(k+1)} & \text{si } n = 2k+1 \\ \frac{x^{(k)} + x^{(k+1)}}{2} & \text{si } n = 2k \end{cases}$$

¿Cómo se define la mediana poblacional? Hallarla para $\mathcal{U}[a, b]$ y $\mathcal{N}(\mu, \sigma^2)$.

Media $\alpha-$podada:

$$\bar{x}_\alpha = \frac{x^{([n\alpha]+1)} + \ldots + x^{(n-[n\alpha])}}{n - 2[n\alpha]}$$

Ejemplo

Supongamos que tenemos este conjunto de 12 observaciones.

$$1 \quad 2 \quad 5 \quad 8 \quad 10 \quad 14 \quad 17 \quad 21 \quad 25 \quad 28 \quad 40 \quad 45$$

Calcular la media, mediana y media $0{,}10-$podada. Cambiar la última observación por 450 y repetir los cálculos.

Distribución empírica y percentiles

Percentiles

El percentil $\alpha \cdot 100\,\%$ de la muestra $(0 < \alpha < 1)$ es el valor por debajo del cual se encuentra el $\alpha \cdot 100\,\%$ de los datos en la muestra ordenada.
Algunos percentiles especiales reciben el nombre de cuartiles.

$$P_{25} = Q_1, \quad P_{50} = Q_2 = \tilde{x}, \quad P_{75} = Q_3$$

Para hallar P_α, ordenamos la muestra y buscamos el dato que ocupa la posición $[(n+1)\alpha]$.

Distribución empírica (frecuencia relativa acumulada)

$$F_n(x) = \frac{\sum_{i=1}^{n} \mathbb{I}_{\{x_i \leq x\}}}{n} = \frac{\#\{x_i \leq x\}}{n}$$

Medidas de dispersión

Rango muestral

$$R_x = \text{máx}\,(x_1, \ldots, x_n) - \text{mín}\,(x_1, \ldots, x_n)$$

Varianza muestral

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Distancia intercuartílica (o IQR) (ver con percentiles)

$$\text{IQR} = Q_3 - Q_1$$

MAD (desviación absoluta respecto de la mediana)

$$\text{MAD} = \text{med}\,(|x_i - \widetilde{x}|)$$

Ejemplo

Supongamos que tenemos este conjunto de 12 observaciones.

$$1 \quad 2 \quad 5 \quad 8 \quad 10 \quad 14 \quad 17 \quad 21 \quad 25 \quad 28 \quad 40 \quad 45$$

Calcular el rango muestral, el IQR, la varianza muestral y la MAD. Cambiar la última observación por 450 y repetir los cálculos. Dar la distribución empírica.

El histograma

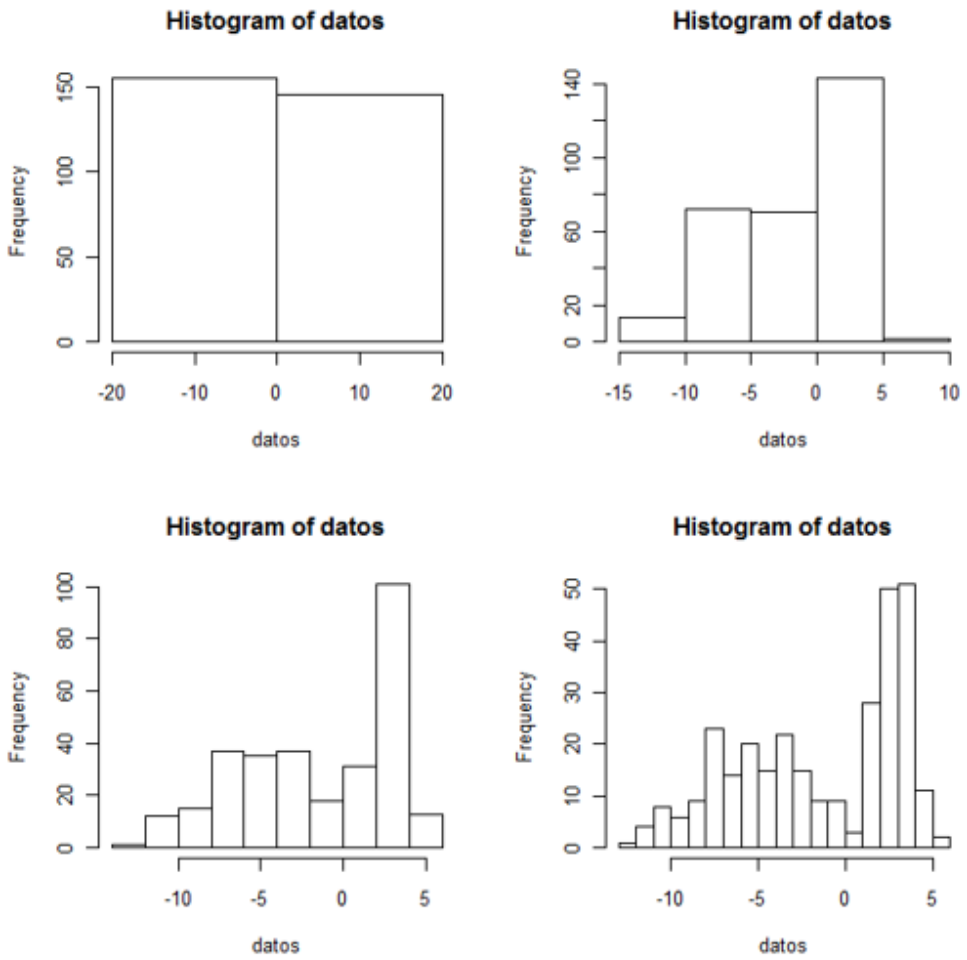


Figura 1: Diferentes histogramas para un mismo conjunto de datos.

El histograma: su construcción

- Se divide el rango de los datos en intervalos o clases excluyentes y exhaustivas.
- Se busca la frecuencia en cada intervalo o clase (o la frecuencia relativa).
- Se grafica en un par de ejes coordenados representando en las abscisas los intervalos y sobre cada uno de ellos un rectángulo cuya área sea proporcional a la frecuencia relativa de dicho intervalo. Es recomendable tomar

$$\text{altura del rectángulo} = \frac{\text{frecuencia relativa}}{\text{longitud del intervalo}}.$$

De esta manera, el área es 1 y dos histogramas son fácilmente comparables más allá de la cantidad de observaciones de cada conjunto de datos.

Ejemplo

grupo etario	cantidad de vacunas
0-18	54
19-29	251
30-39	192
40-49	191
50-59	158
60-69	118
70-79	60
80-89	22
90-99	3

Cuadro 1: Cantidad de vacunas aplicadas (miles) según grupo etario. Fuente: Datos Abiertos (19/10/21). Se desestimaron 49 casos de vacunas aplicadas a individuos de 100 o más años de edad.

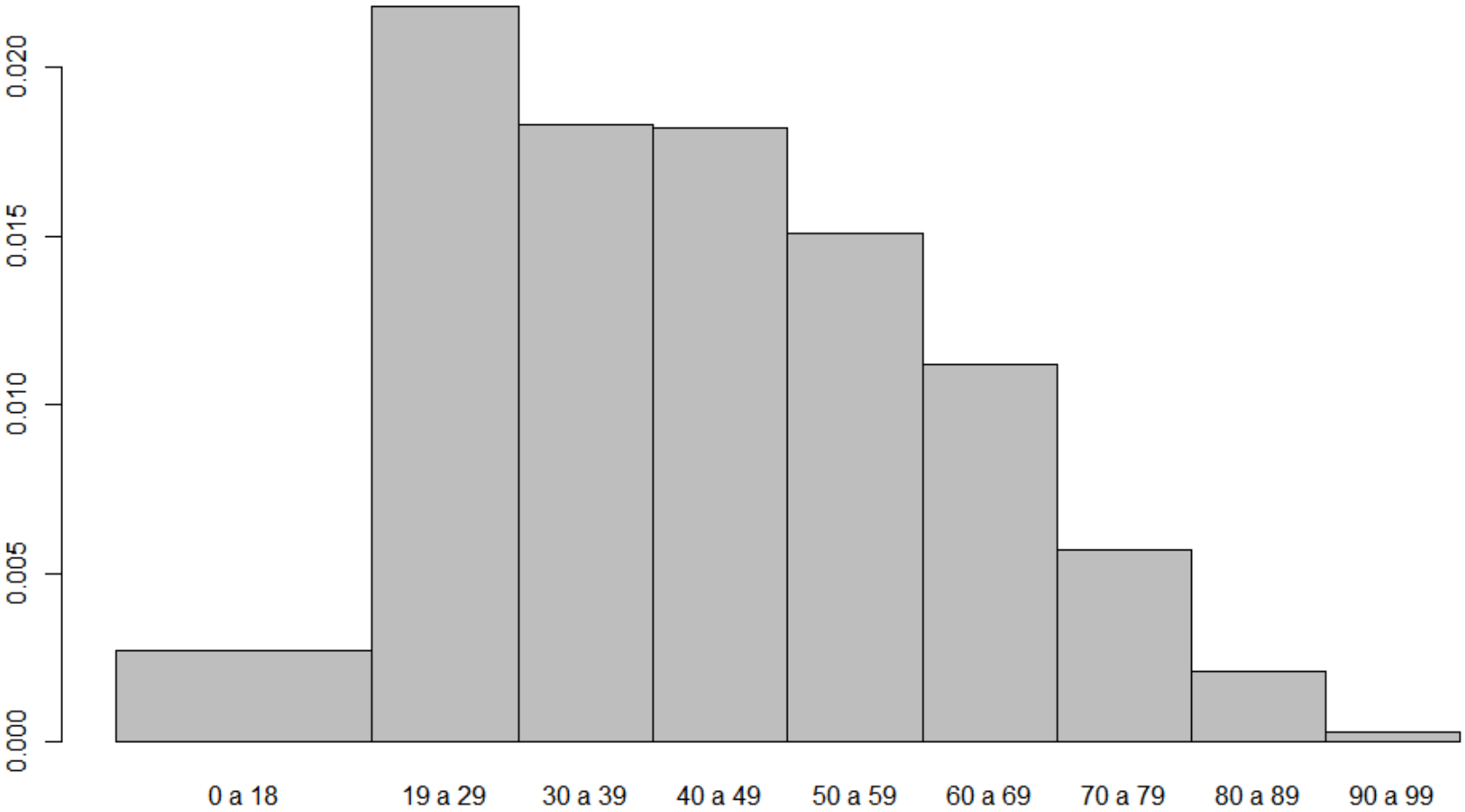
Ejemplo

intervalos	frecuencia	frec relativa	longitud	altura
[0;19)	54	0,051	19	
[19;30)	251	0,240	11	
[30;40)	192	0,183	10	
[40;50)	191	0,183	10	
[50;60)	158	0,150	10	
[60;70)	118	0,112	10	
[70;80)	60	0,057	10	
[80;90)	22	0,021	10	
[90;100)	3	0,003	10	

Ejemplo

intervalos	frecuencia	frec relativa	longitud	altura
[0;19)	54	0,051	19	**0,0027**
[19;30)	251	0,240	11	**0,0218**
[30;40)	192	0,183	10	**0,0183**
[40;50)	191	0,183	10	**0,0183**
[50;60)	158	0,150	10	**0,0150**
[60;70)	118	0,112	10	**0,0112**
[70;80)	60	0,057	10	**0,0057**
[80;90)	22	0,021	10	**0,0021**
[90;100)	3	0,003	10	**0,0003**

Ejemplo

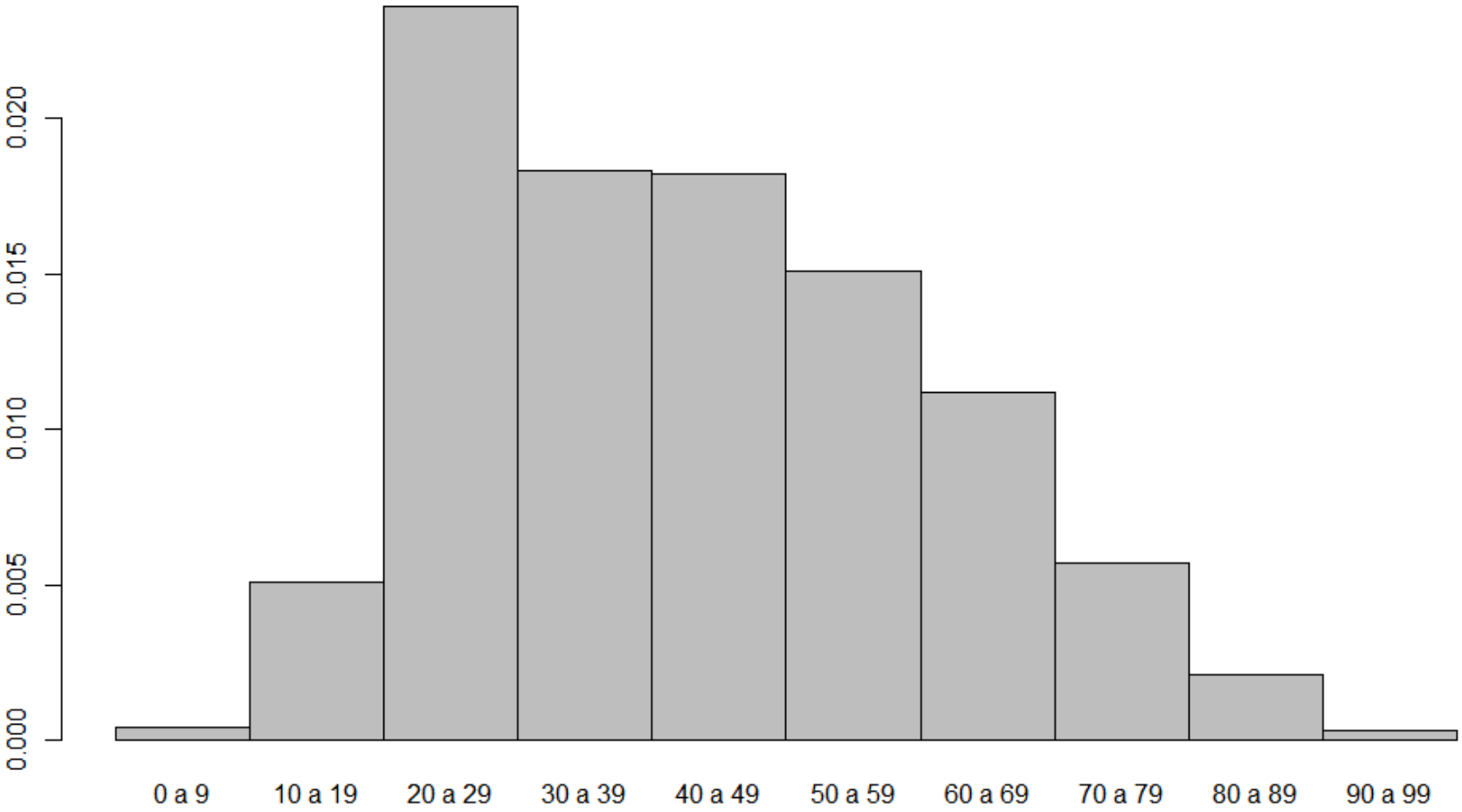

0.000
0.005
0.010
0.015
0.020
0 a 18
19 a 29
30 a 39
40 a 49
50 a 59
60 a 69
70 a 79
80 a 89
90 a 99

Ejemplo

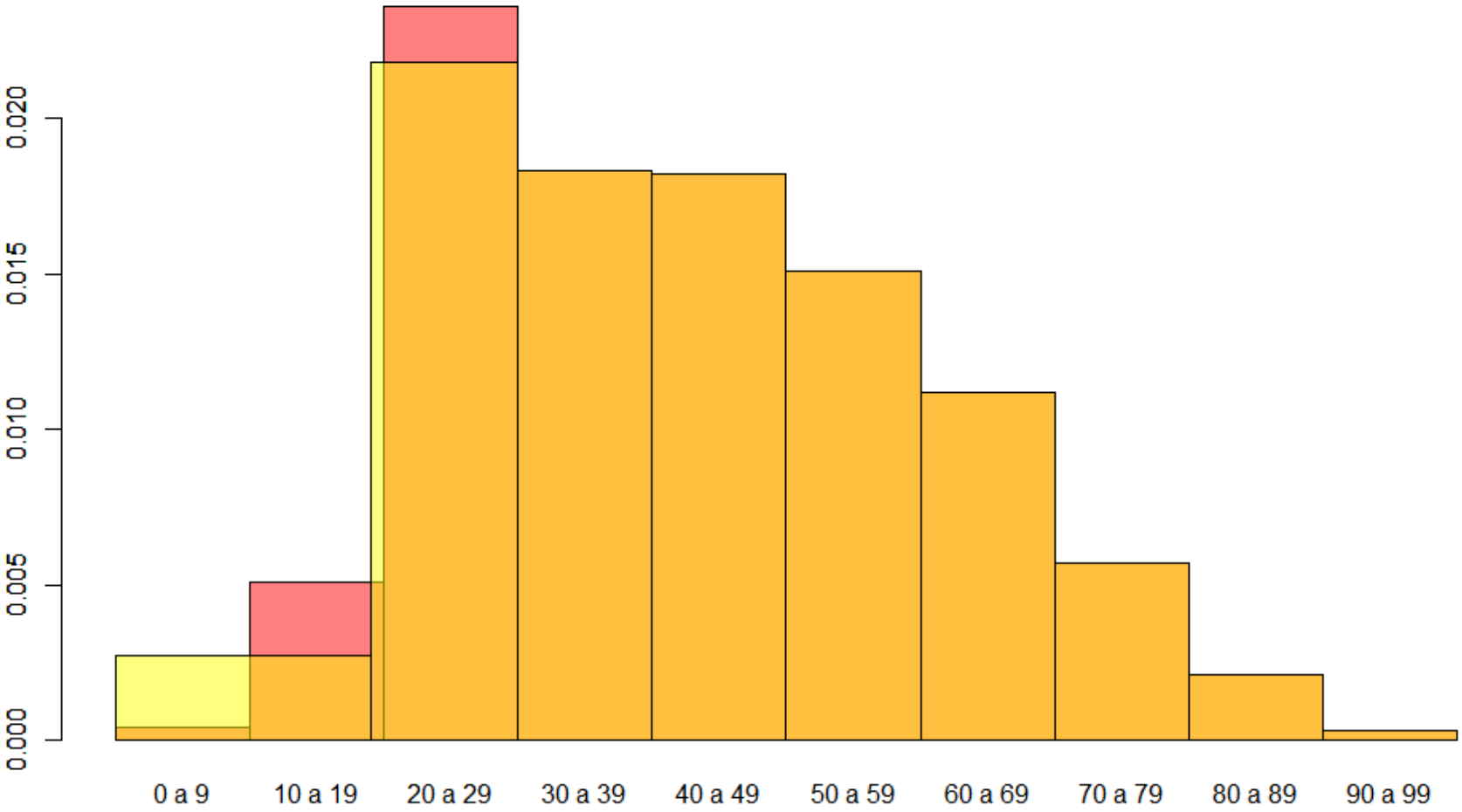
grupo etario	cantidad de vacunas
0-9	4
10-19	53
20-29	248
30-39	192
40-49	191
50-59	158
60-69	118
70-79	60
80-89	22
90-99	3

Cuadro 2: Ejemplo adaptado del anterior para tener clases de igual longitud.

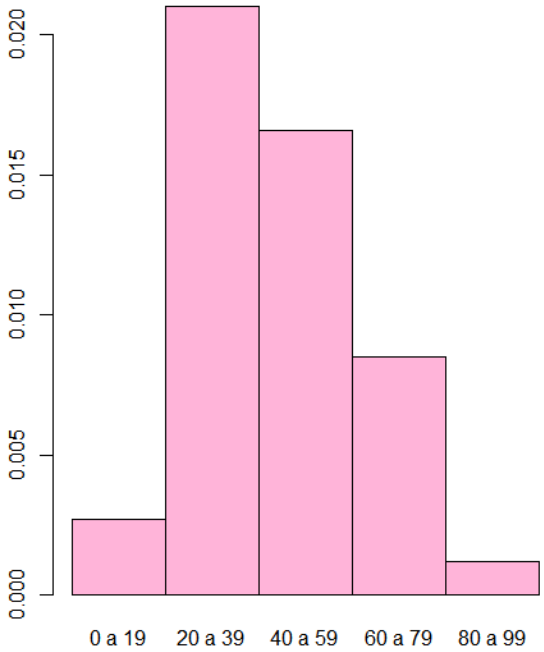
Ejemplo

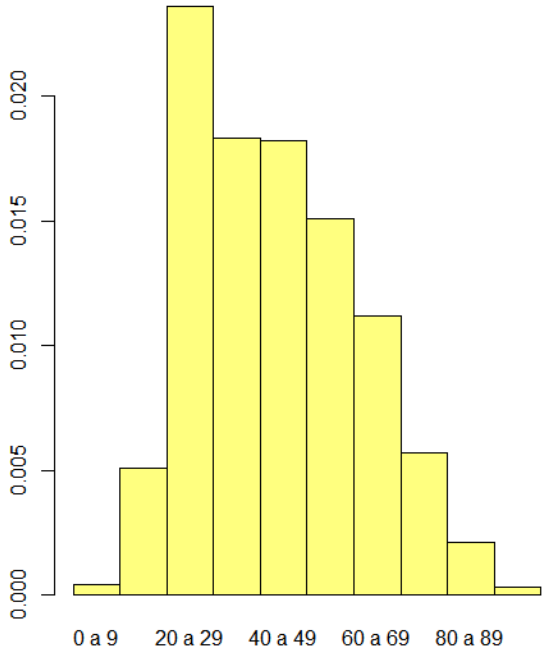

0.020
0.015
0.010
0.005
0.000
0 a 9
10 a 19
20 a 29
30 a 39
40 a 49
50 a 59
60 a 69
70 a 79
80 a 89
90 a 99

Ejemplo


0.000
0.005
0.010
0.015
0.020
0 a 9
10 a 19
20 a 29
30 a 39
40 a 49
50 a 59
60 a 69
70 a 79
80 a 89
90 a 99

La elección de la cantidad de clases/intervalos es algo no menor.

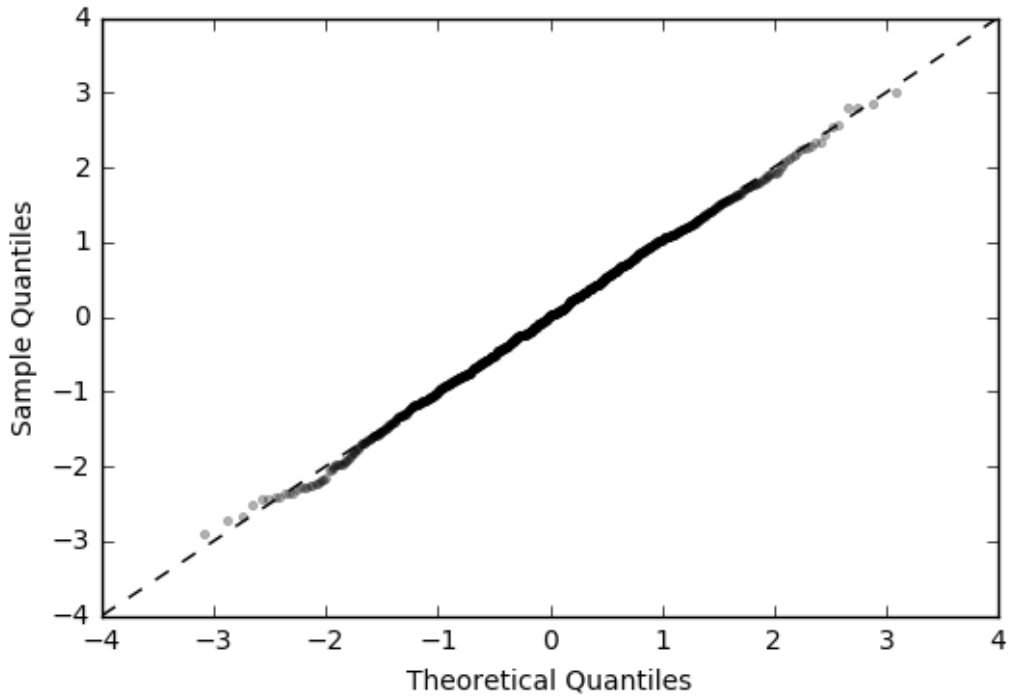




Utilizar muchos o muy pocos intervalos puede ser poco informativo.
Equilibrio entre uno muy irregular y uno demasiado suavizado.

El QQ–plot

En el QQ–plot se grafican los percentiles de una distribución teórica de interés vs. las observaciones ordenadas. Si los datos provienen de la distribución propuesta, uno esperaría que los puntos quedaran más o menos alineados sobre una recta.



El QQ–plot: su construcción

Sea una muestra de datos $x_1, x_2, \ldots, x_n$. Queremos analizar la posibilidad de que provengan de una muestra aleatoria con distribución cierta distribución, digamos $\mathcal{N}(0,1)$. Con la muestra ordenada, $x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}$ es fácil ver que

$$x^{(1)} = \text{ percentil } \frac{1}{n+1} \text{ de la muestra,}$$

$$x^{(2)} = \text{ percentil } \frac{2}{n+1} \text{ de la muestra, } \ldots$$

$$x^{(n)} = \text{ percentil } \frac{n}{n+1} \text{ de la muestra.}$$

y esto es así pues la definición de P_α que vimos era

$$x^{([\alpha\cdot(n+1)])} = x^{\left(\left[\frac{i}{n+1}(n+1)\right]\right)} = x^{(i)}.$$

Queremos comparar la distribución empírica con la de una $\mathcal{N}(0,1)$. Pero, ¿cuál sería el percentil $\frac{i}{n+1}$ teórico de una $\mathcal{N}(0,1)$? Busquémoslo así lo comparamos con el empírico $x^{(i)}$ y vemos si se parecen.

Lllamemos x_i a ese percentil teórico, que debe cumplir que $\Phi(x_i) = \frac{i}{n+1}$, es decir,

$$x_i = \Phi^{-1}\left(\frac{i}{n+1}\right).$$

Si los datos provienen de la distribución normal estándar supuesta, esos percentiles empíricos $x^{(i)}$ deberían parecerse a los percentiles teóricos x_i. Es decir, para todo $1 \leq i \leq n$, esperaríamos que

$$\Phi^{-1}\left(\frac{i}{n+1}\right) \approx x^{(i)}.$$

Una forma de evaluar cuán buen es este ajuste, es realizar un gráfico con los pares ordenados

$$\left(x_i, x^{(i)}\right) = \left(\Phi^{-1}\left(\frac{i}{n+1}\right), x^{(i)}\right).$$

En la medida en que estos puntos estén más próximos a la recta $y = x$, más evidencia a favor de la sospecha que los datos provienen de una distribución $\mathcal{N}(0, 1)$ habrá.

Adaptar este procedimiento a otras distribuciones que no sean la normal estándar tiene costo cero: basta considerar la F de la distribución supuesta, en lugar de la Φ de la normal estándar.

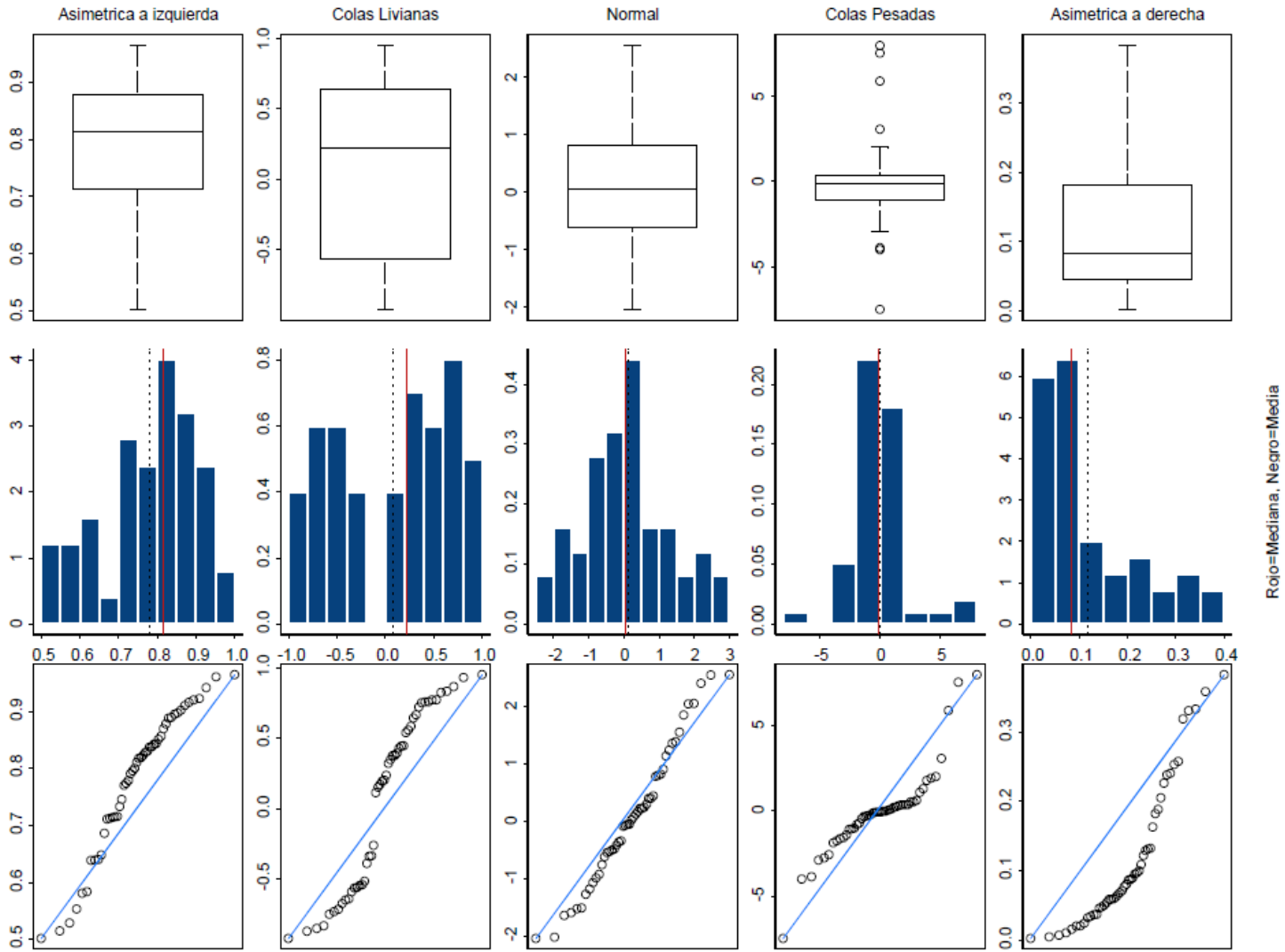


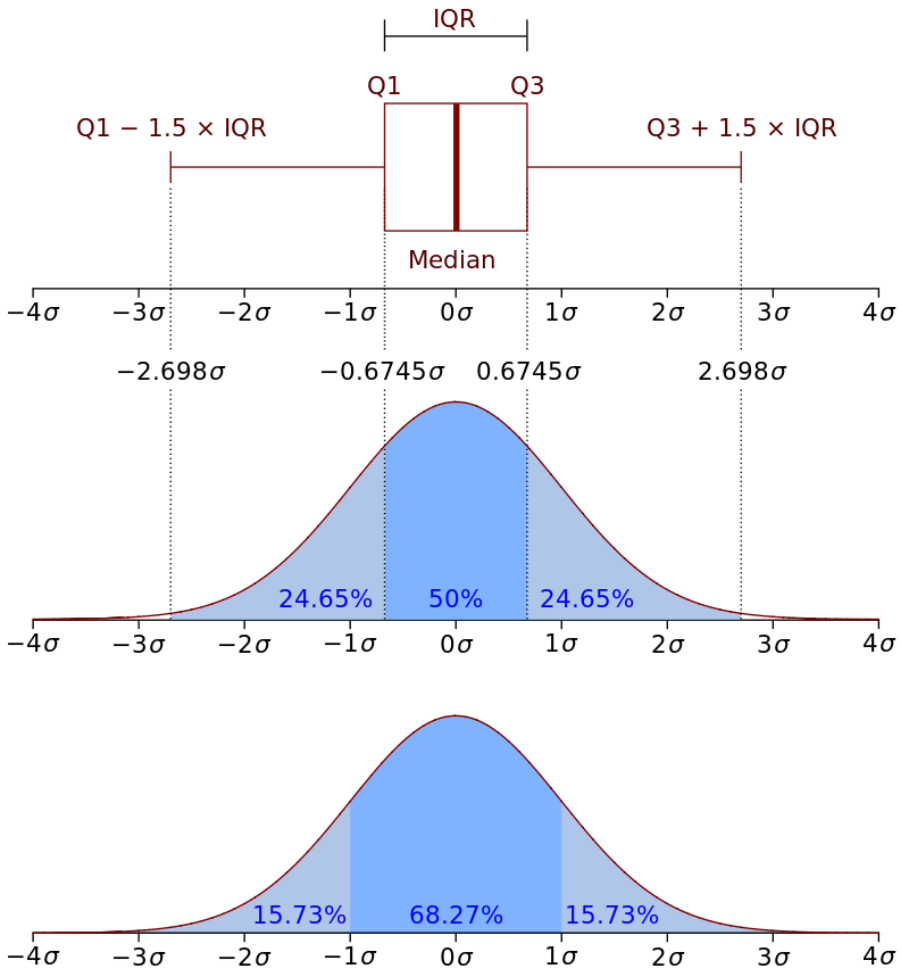
Figura 2: Gráficos para distribuciones empíricas de diferentes aspectos.

El boxplot

Su construcción (no es que lo vayamos a hacer a mano, pero...)

- Representamos una escala vertical u horizontal y bibujamos una caja cuyos extremos son Q_1 y Q_3 y dentro de ella un segmento que corresponde a la mediana Q_2.
- A partir de cada extremo, dibujamos un segmento hasta el dato más alejado que está a lo sumo a $\pm 1{,}5 d_I$ del extremo de la caja. Estos segmentos se llaman bigotes.
- Identificamos, de alguna manera, aquellos datos que están entre $1{,}5 d_I$ y $3 d_I$ de cada extremo y a aquellos que están a más de $3 d_I$ de cada extremo.
- Puede ocurrir que nos topemos con el máximo aun antes de terminar el bigote y/o no identifiquemos datos extremos. El boxplot identifica muchas observaciones extremas cuando hay asimetría y no necesariamente son, por eso, atípicos.

¿Por qué esa construcción?



El caso de datos bivariados

Recordemos que la correlación era una medida de la asociación lineal de las variables de un vector aleatorio (X, Y).

$$\rho(X,Y) = \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]}{\sqrt{V(X)V(Y)}}$$

Parecería razonable pensar que un equivalente muestral de esa correlación es la que sigue.

Correlación muestral

$$\begin{aligned}\rho(x,y) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n \sum_{i=1}^{n}(y_i - \bar{y})^2/n}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}\end{aligned}$$

El caso de datos bivariados

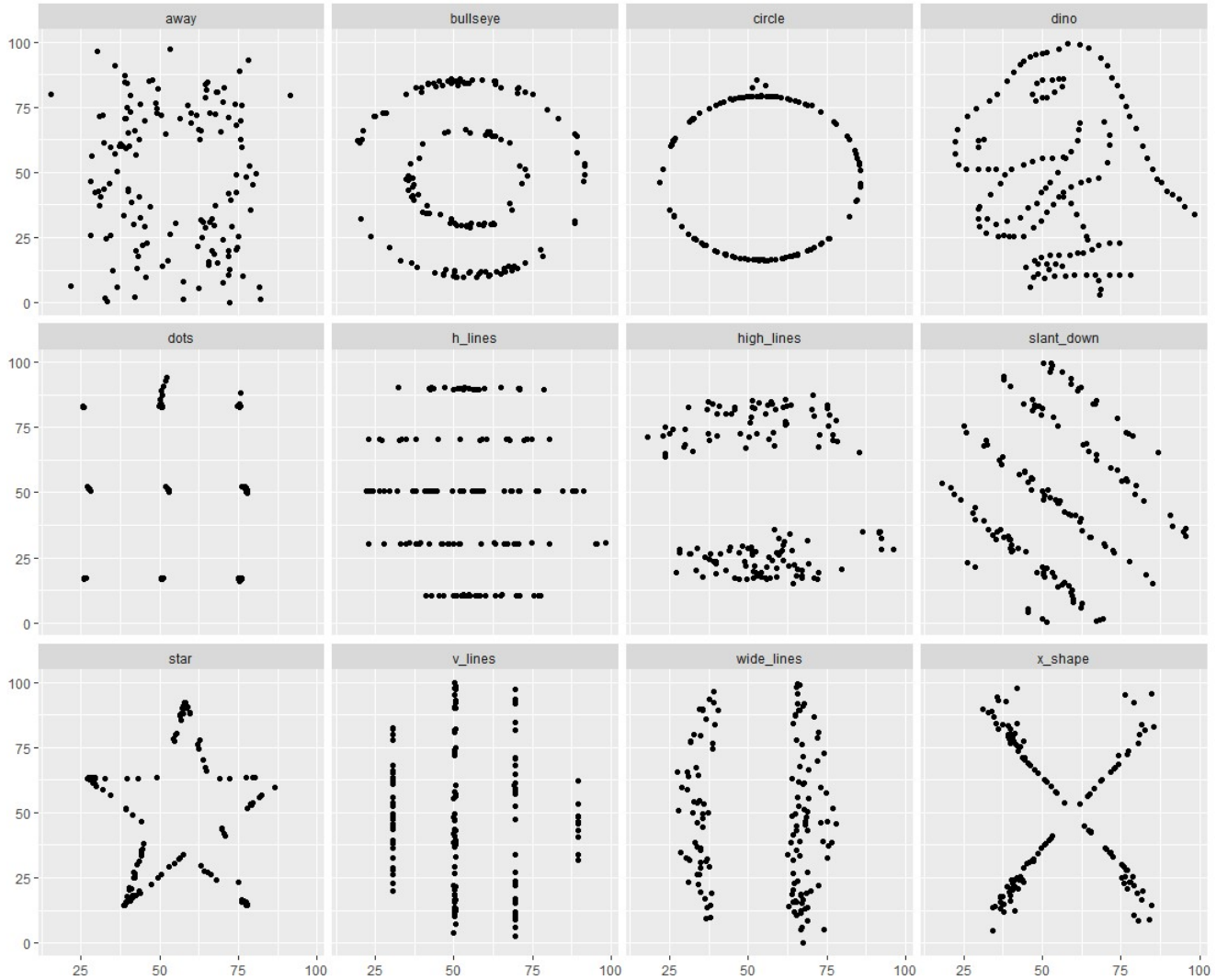
dataset	xraya	yraya	sx	sy	corr
away	54.3	47.8	16.8	26.9	-0.1
bullseye	54.3	47.8	16.8	26.9	-0.1
circle	54.3	47.8	16.8	26.9	-0.1
dino	54.3	47.8	16.8	26.9	-0.1
dots	54.3	47.8	16.8	26.9	-0.1
h_lines	54.3	47.8	16.8	26.9	-0.1
high_lines	54.3	47.8	16.8	26.9	-0.1
slant_down	54.3	47.8	16.8	26.9	-0.1
star	54.3	47.8	16.8	26.9	-0.1
v_lines	54.3	47.8	16.8	26.9	-0.1
wide_lines	54.3	47.8	16.8	26.9	-0.1
x_shape	54.3	47.8	16.8	26.9	-0.1

A veces, los estadísticos muestrales no alcanzan...

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Justin Matejka and George Fitzmaurice
Autodesk Research, Toronto Ontario Canada
{*first.last*}@autodesk.com

Visualización



Mejor predictor lineal según ECM (símil E24 P4)

El mejor predictor lineal de Y basado en X según el criterio del ECM es

$$\hat{Y} = \underbrace{\mu_Y - \frac{\sigma_Y}{\sigma_X}\rho_{XY}\mu_X}_{\alpha} + \underbrace{\frac{\sigma_Y}{\sigma_X}\rho_{XY}}_{\beta} X.$$

En la "práctica": Estimamos los parámetros α y β a partir de la muestra

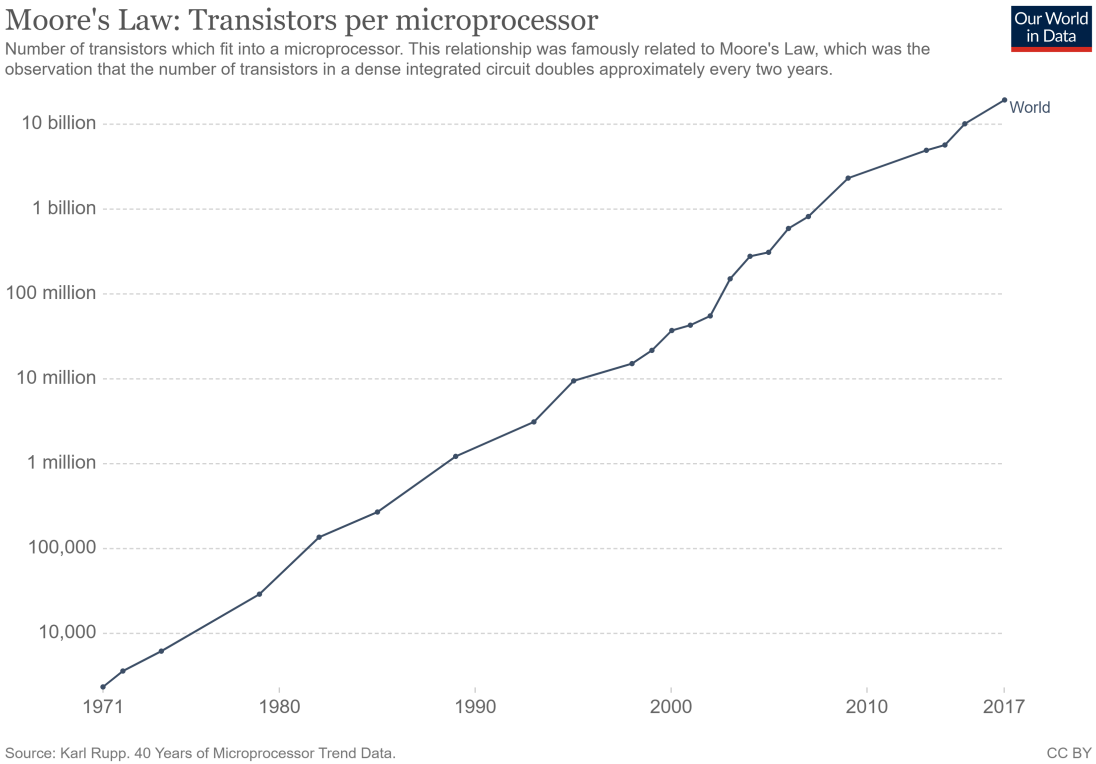
$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

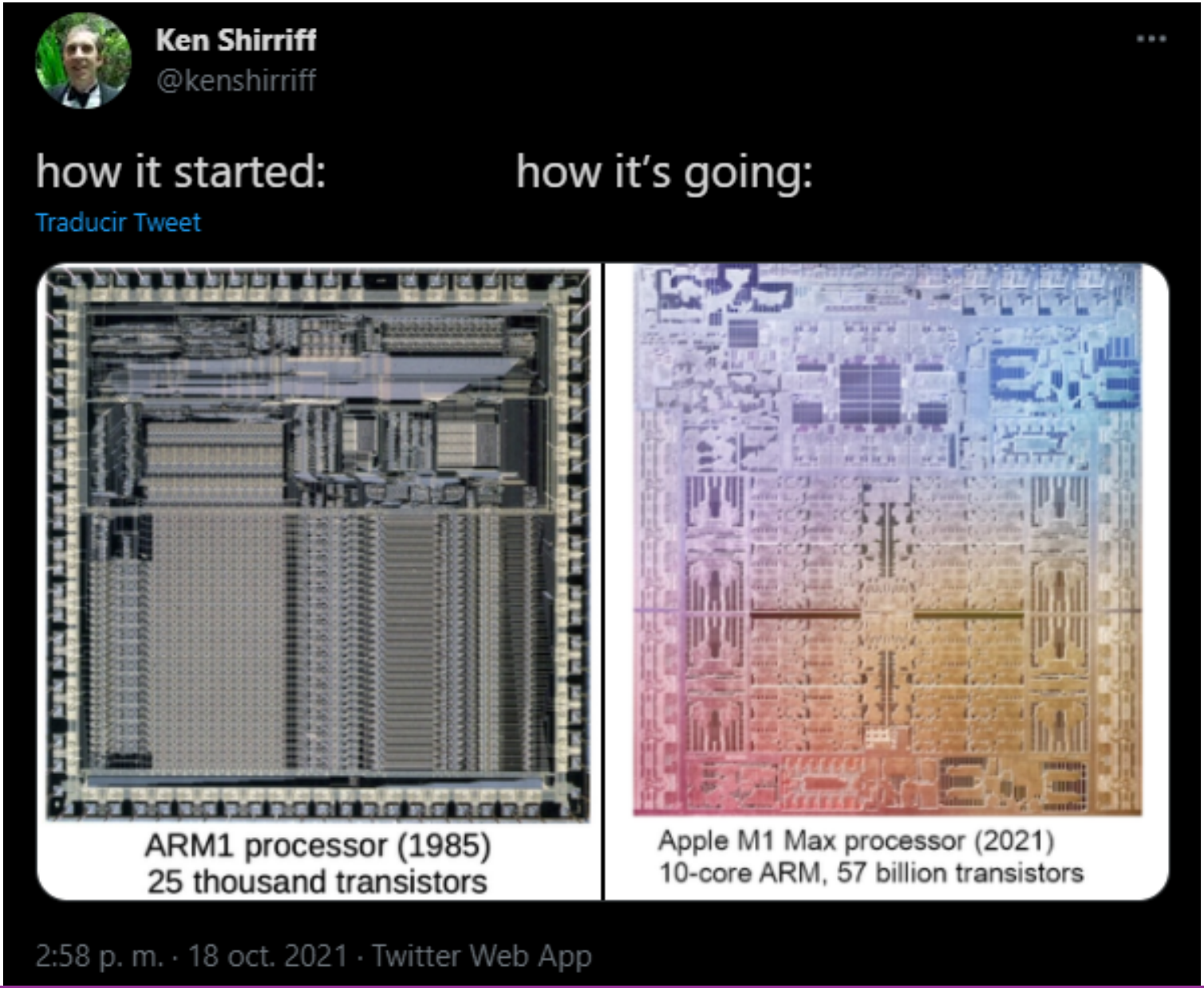
y, con esas estimaciones a partir de la muestra, obtenemos la recta de regresión estimada: $y = \hat{\alpha} + \hat{\beta}x$.

Extra: Ejemplo con la "ley de Moore"

Vamos a tratar de ver si existe un ajuste lineal (*) en la ley –empírica– de Moore. ▸ Datos acá.

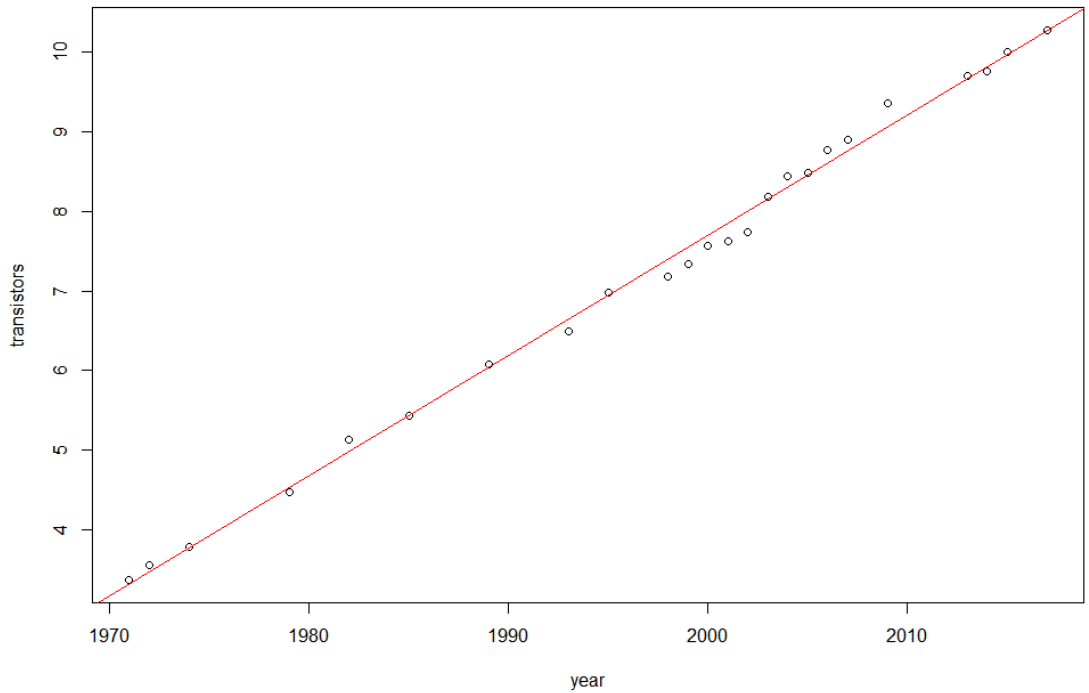


Fuente: Our World in Data


Ken Shirriff
@kenshirriff
how it started:
how it's going:
Traducir Tweet
ARM1 processor (1985)
25 thousand transistors
Apple M1 Max processor (2021)
10-core ARM, 57 billion transistors
2:58 p. m. · 18 oct. 2021 · Twitter Web App

Con los datos de la muestra, mi estimación de la recta $y = \hat{\alpha} + \hat{\beta}x$ con x el año e y el $\log$ de la cantidad de transistores, es

$$y = -294{,}7478 + 0{,}15122x.$$



Por lo tanto, una predicción para los transistores en 2021 sería

$$10^{\hat{y}_{2021}} = 10^{-294{,}7478+0{,}15122 \cdot 2021} = 74,854,017,246.$$

Sin la información que aporta X, mi mejor predicción constante de Y hubiese sido su esperanza (se puede probar, es más fácil incluso que la demo del extra).

Es decir, en términos de los datos de la muestra, si no creo que exista relación entre el año y la cantidad de transistores, mi mejor predicción para la cantidad de transistores en 2021 hubiera sido 1,854,748,968.

¡Gracias, covariables!

Extra: Mejor predictor lineal según ECM (símil E24 P4)

Sea Y una variable aleatoria que queremos aproximar por otra $\hat{Y}$. Una manera de "medir" la calidad de esa aproximación es con el error cuadrático medio:

$$\mathrm{ECM}(Y, \hat{Y}) = \mathbb{E}[(Y - \hat{Y})^2]$$

y por la misma propiedad que usamos para la varianza, se ve que

$$\mathrm{ECM}(Y, \hat{Y}) = V(Y - \hat{Y}) + \mathbb{E}[(Y - \hat{Y})]^2$$

Busquemos el mejor predictor lineal de Y (en términos del ECM). Es decir, el $\hat{Y}$ que sea de la forma $\alpha X + \beta$ con α y β tal que el ECM sea mínimo.

Deduzcamos qué forma tienen esos α y β que minimizan $\mathbb{E}[(Y - \alpha - \beta X)^2]$.

Buscamos α y β tal que $H(\alpha, \beta) = \mathbb{E}[(Y - \alpha - \beta X)^2]$ es mínimo.

$$\begin{aligned} H(\alpha, \beta) &= V(Y - \alpha - \beta X) + [\mathbb{E}(Y - \alpha - \beta X)]^2 \\ &= V(Y - \beta X) + [\mathbb{E}(Y - \alpha - \beta X)]^2 \\ &= \sigma_Y^2 + \beta^2 \sigma_X^2 - 2\beta\sigma_{XY} + (\mu_Y - \alpha - \beta\mu_X)^2 \end{aligned}$$

Buscamos el punto crítico derivando e igualando a 0

$$\frac{\partial H}{\partial \alpha} = 2\left(\mu_Y - \alpha - \beta\mu_X\right)(-1) = 0$$

$$\frac{\partial H}{\partial \beta} = 2\beta\sigma_X^2 - 2\sigma_{XY} + 2\left(\mu_Y - \alpha - \beta\mu_X\right)(-\mu_X) = 0$$

y despejando se llega a que

$$\beta = \frac{\sigma_Y}{\sigma_X}\rho_{XY}$$

$$\alpha = \mu_Y - \beta\mu_X$$

(aunque faltaría probar que allí se alcanza un mínimo...)

Extra: Mejor predictor lineal según ECM (símil E24 P4)

¿Cómo sería el mejor predictor constante de Y (en términos del ECM)? Deduzcan qué forma tiene el c que minimiza $\mathbb{E}[(Y-c)^2]$. (Mismas cuentas de antes, solo que más fáciles y nos conducen a que $\hat{Y} = c = \mathbb{E}(Y)$)

¡Hasta el jueves!