Conceptos de probabilidad y estadística descriptiva

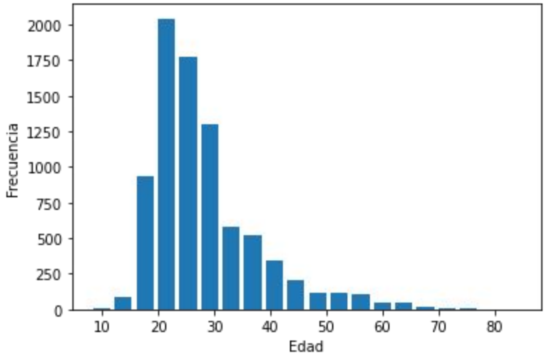
Laboratorio de Datos 1°C 2021

¿Por qué estadística?

	1_extraversion	1_agreeableness	1_conscientiousness	1_neuroticism	1_openness	2_edad	2_genero
0	27	31	31	37	32	21	
1	16	37	26	31	45	33	
2	23	33	38	22	44	29	
3	20	25	25	20	18	24	
4	22	34	24	35	42	35	
	...	...	...	...	...	...	...
	21	35	44	13	35	21	2
	32	30	36	12	33	21	2
	20	37	30	13	37	18	1

Ver los datos crudos puede ser poco informativo

Media: 29.95
Desviación: 6.69
90% central: 19 - 41
Mediana: 30
Moda: 31



Con unos **pocos observables y una visualización de los datos** podemos tener una noción de dónde están la mayoría de los datos, entre qué valores fluctúan, cuál es el más probable, etc.

¿Por qué probabilidad?

- Muchas veces podemos interpretar nuestros datos como el resultado de una secuencia de eventos al azar, por lo tanto es bueno para el análisis pensar en el proceso que los generó.

- Muchos algoritmos que usan un lenguaje probabilístico (PCA, Naive Bayes, Latent Dirichlet Allocation), por lo tanto es bueno saber conceptos tales como la covarianza, probabilidad condicional, variables independientes.

Esquema de la clase

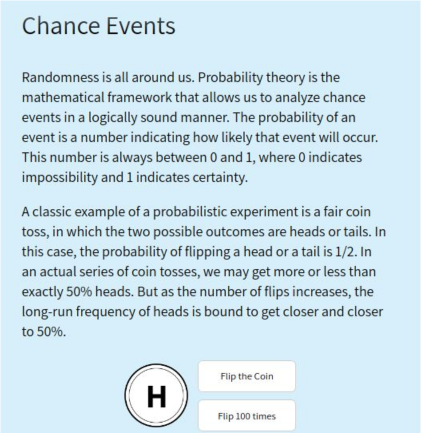
- Hacer un breve repaso de conceptos de probabilidad ¿qué es una variable aleatoria? ¿qué es una distribución de probabilidad? Ejemplos de distribución, definición de valores medios y algunos resultados teóricos útiles.

- Definir observables y medidas análogos a los definidos para distribuciones de probabilidad para describir nuestra distribución de datos.

Conceptos de probabilidad

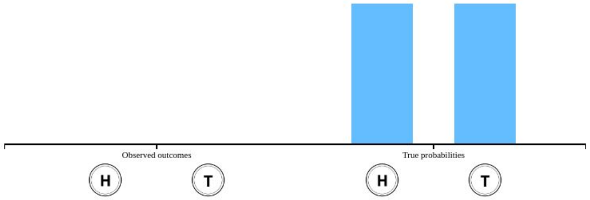
La probabilidad es una medida de la incerteza que tenemos del resultado de un experimento.

Podemos pensar en la probabilidad como la cantidad relativa de veces que obtenemos dicho resultado al realizar muchas veces el experimento.


Chance Events

Randomness is all around us. Probability theory is the mathematical framework that allows us to analyze chance events in a logically sound manner. The probability of an event is a number indicating how likely that event will occur. This number is always between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.

A classic example of a probabilistic experiment is a fair coin toss, in which the two possible outcomes are heads or tails. In this case, the probability of flipping a head or a tail is 1/2. In an actual series of coin tosses, we may get more or less than exactly 50% heads. But as the number of flips increases, the long-run frequency of heads is bound to get closer and closer to 50%.



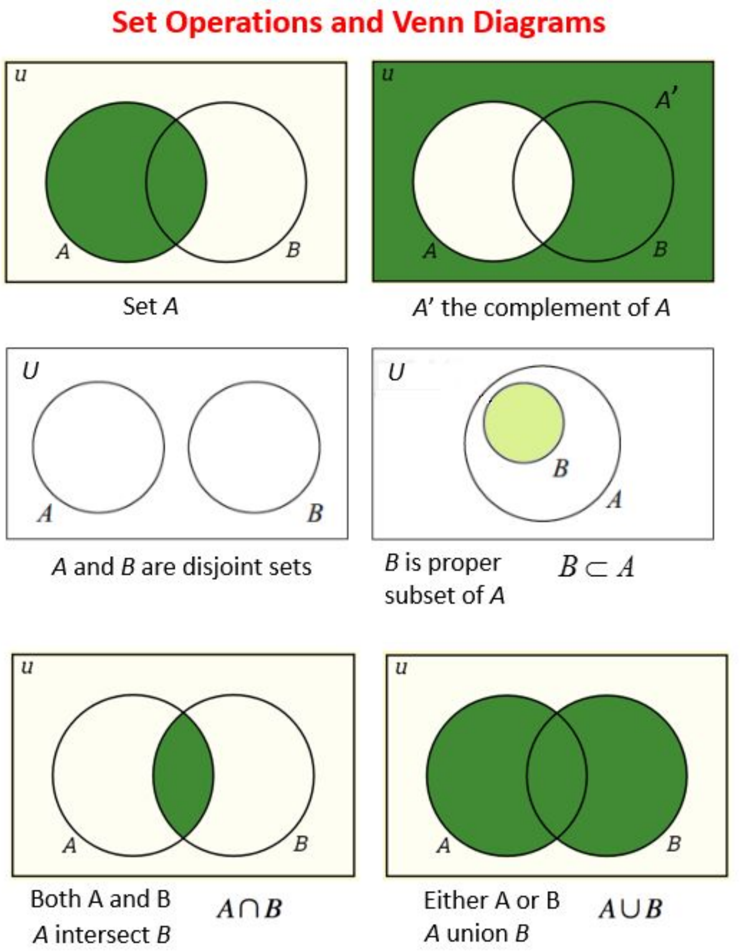
https://seeing-theory.brown.edu/

Conceptos de probabilidad

El lenguaje de la teoría de probabilidad es prácticamente el mismo que el de teoría de conjuntos.

Alerta! un poco de teoría

https://medium.com/@sukhrobgolibboev/understanding-set-theory-de2532f746ac



Conceptos de probabilidad

Espacio muestral: conjunto con todos los posibles resultados de un experimento. Los subconjuntos de este espacio se denominan eventos.

Sea Ω el espacio muestral y *A* un evento, se define a la probabilidad como una función *P* que cumple los siguientes axiomas:

- $P(A) \geq 0$ para todo A.
- $P(\Omega) = 1$.
- Para eventos disjuntos, A_1, A_2, etc, entonces:

$$P\left(\cup_i A_i\right) = \sum_i P(A_i)$$

Eventos mutuamente excluyentes: si sucede uno no puede suceder el otro simultáneamente.

Conceptos de probabilidad

Si dos eventos A y B no son mutuamente excluyentes, entonces:

$$P\left(A \cup B\right) = P(A) + P(B) - P\left(A \cap B\right)$$

Proba de que ocurra el evento A **o** el evento B

Proba de que ocurra el evento A **y** el evento B

$$P\left(A \cap B\right) \equiv P\left(A, B\right)$$

Conceptos de probabilidad

Probabilidad condicional:

Probabilidad (condicional) de que ocurra A dado que sabemos que ocurrió B

$$P\left(A, B\right) = P\left(A \mid B\right) P(B)$$

Si los eventos son independientes, es decir, si la probabilidad de que ocurra A no depende de si ocurrió o no B, entonces:

$$P\left(A \mid B\right) = P(A) \Longrightarrow \boxed{P\left(A, B\right) = P(A) P(B)}$$

Condición de eventos independientes

Ejemplo!

Si lanzamos un dado...

$$P(3) = \frac{1}{6}$$

Proba de obtener un 3

$$P(2 \cup 3 \cup 5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$$

Proba de obtener o un 2 o un 3 o un 5

1/3 1/6

$$P(1 \mid \text{impar}) = \frac{P(1, \text{impar})}{P(\text{impar})}$$

1/2

Proba de obtener un 1 dado que sabemos que el resultado es impar

Ejemplo!

Si lanzamos 2 dados...

Primer dado   Segundo dado

$$P(4,5) = P_1(4)P_2(5) = \frac{1}{6} \times \frac{1}{6}$$

Proba de obtener un 4 en el primer dado y 5 en el segundo

1/6   1/6   1/36

$$P(2 \cup 3) = P_1(2) + P_2(3) - P_1(2)P_2(3)$$

Proba de obtener o un 2 en el primer dado o un 3 en el segundo dado

Conceptos de probabilidad

Como la probabilidad conjunta puede escribirse de varias maneras, encontramos una forma de relacionar las probabilidades condicionales:

$$P\left(A,B\right)=P\left(A\mid B\right)P(B)=P\left(B\mid A\right)P(A)$$

$$P\left(A\mid B\right)=\frac{P\left(B\mid A\right)P(A)}{P(B)}$$

Teorema de Bayes

Conceptos de probabilidad

Teorema de probabilidad total (marginalización): sea A_1, A_2, A_3, etc, una partición del espacio muestral, entonces:

$$P(B) = \sum_i P\left(B \mid A_i\right) P(A_i)$$

Denominador en Bayes

Es decir, la probabilidad de que ocurra B es contemplar todos los posibles eventos que podrían haber ocurrido antes que B.

Ejemplo de Bayes!

Al etiquetar mis mails en 3 categorías descubro que solo el 10% de ellos son muy prioritarios (M), el 20% poco prioritarios (P) y el 70% son spam (S).

Me llega un mail con la palabra “promo” en el asunto que no sé cómo etiquetarlo (en general nunca gano nada, pero bueh, démosle una chance). Revisando mis mails en spam descubro que el 90% de ellos contienen la palabra “promo” en el asunto, mientras que solo el 1% de los muy prioritarios y de los pocos prioritarios la tienen.

¿Qué chance hay que el mail que me llegó sea spam?

Ejemplo de Bayes!

Al etiquetar mis mails en 3 categorías descubro que solo el **10% de ellos son muy prioritarios (M), el 20% poco prioritarios (P) y el 70% son spam (S)**. Me llega un mail con la palabra “promo” en el asunto que no sé cómo etiquetarlo (en general nunca gano nada, pero bueh, démosle una chance). **Revisando mis mails en spam descubro que el 90% de ellos contienen la palabra “promo” en el asunto, mientras que solo el 1% de los muy prioritarios y de los pocos prioritarios la tienen.**

Proba de que el mail sea spam dado que tiene la palabra promo

$$P(S \mid \text{promo}) = \frac{\overbrace{P(\text{promo} \mid S)}^{0.9}\,\overbrace{P(S)}^{0.7}}{P(\text{promo})}$$

$$\underbrace{P(\text{promo} \mid M)}_{0.01}\underbrace{P(M)}_{0.10} + \underbrace{P(\text{promo} \mid P)}_{0.01}\underbrace{P(P)}_{0.20} + \underbrace{P(\text{promo} \mid S)}_{0.9}\underbrace{P(S)}_{0.7}$$

Ejemplo de Bayes!

$$P(S \mid \text{promo}) = \frac{\overbrace{P(\text{promo} \mid S)}^{0.9}\,\overbrace{P(S)}^{0.7}}{\underbrace{P(\text{promo})}_{0.633}}$$

$$\Rightarrow \quad P(S \mid \text{promo}) \sim 0.995$$

Definitivamente, no tuve suerte...

Variable aleatoria

Una **variable aleatoria** es un observable que le asignamos a cada evento del espacio muestral. Cada vez que realizamos un experimento, obtenemos un valor de la variable aleatoria.

espacio muestral

$x = 2$

$x = 3$

Variable aleatoria

Dentro de un mismo espacio muestral, podemos definir diferentes variables aleatorias.

Ej: el espacio muestral son todas las posibles combinaciones de cara y cecas de una moneda al lanzarla N veces.
La variable aleatoria puede ser el número de caras, cantidad de veces que aparecen dos caras seguidas, etc.

Variable aleatoria

Una variable aleatoria puede ser:

- **Discreta:** toma un conjunto finito de valores (ej: número de caras que salen al lanzar 10 veces una moneda), o un número contable de valores (ej: cantidad de autos que pasan por hora en un semáforo).

- **Continua:** toma infinitos valores. Ej: altura de una persona, tiempo que se tarda en atender a un paciente en una clínica, etc.

Variable discreta. Distribuciones de probabilidad.

Cuando tratamos con variables discretas basta con especificar cuál es la probabilidad de que ocurra alguno cada uno de los eventos.

$$P(X = x) = f(x)$$

probabilidad de que la variable aleatoria adopte el valor de x

función de masa de probabilidad

$$\boxed{\sum_i P(x_i) = 1}$$

Normalización

Se define también la función de probabilidad acumulada, aunque va a ser más relevante a la hora de describir variables continuas:

$$P(X < x) = F(x)$$

función de distribución (acumulada) de probabilidad

Variable discreta. Valores medios

Valor medio: $$\langle x \rangle = \sum_i x_i P(x_i)$$

También llamado valor esperado o esperanza matemática

Varianza: $$\sigma^2 = \sum_i (x_i - \langle x \rangle)^2 P(x_i)$$

Desviación: $$\sigma = \sqrt{\sigma^2}$$

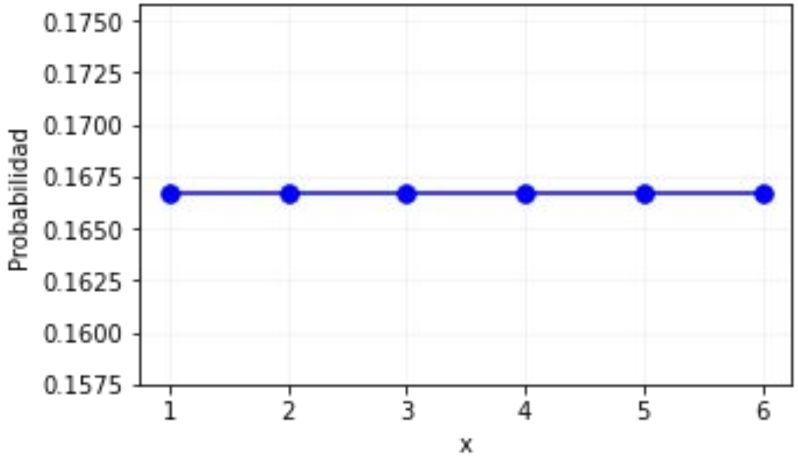
Promedio del cuadrado de las distancias a la media

Algunas distribuciones importantes. Variable discreta

Distribución uniforme discreta: todos los eventos del espacio muestral son equiprobables. Ej: un dado no cargado.

$$P(x) = \frac{1}{n}$$

Cantidad de eventos posibles



Pensar en todos los casos cuál es el espacio muestral y cuál es la variable aleatoria.

Algunas distribuciones importantes. Variable discreta

Distribución binomial: si un evento ocurre con probabilidad *p*, la probabilidad que ocurran *x* eventos en *n* intentos viene dada por la binomial. Ej: cantidad de veces que sale cara al tirar una moneda.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Combinatorio: de cuántas maneras puedo obtener x caras en n tiradas de moneda?

Probabilidad *p* de tener éxito en un único evento (por ejemplo, sacar cara).

Algunas distribuciones importantes. Variable discreta

Distribución binomial:

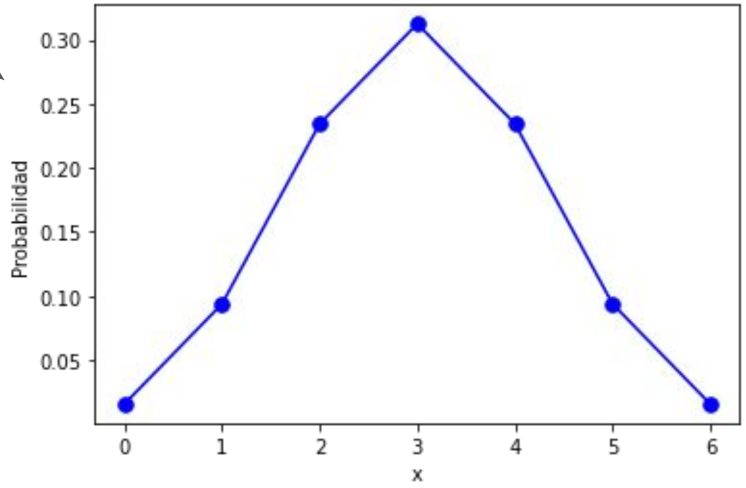
$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

x = 2

x = 3

n = 6

p = 0.5



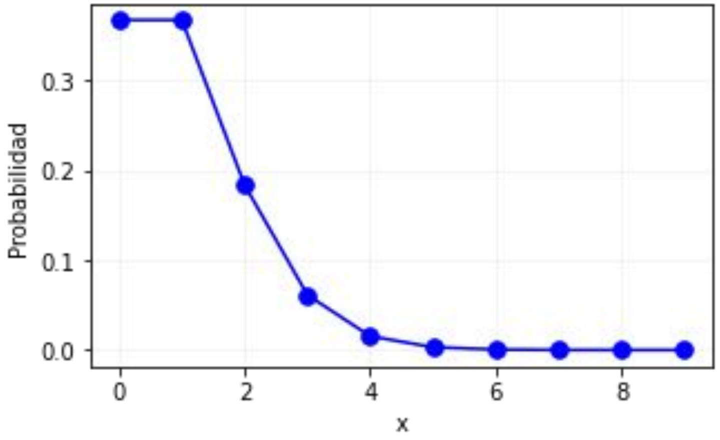
Valor medio: **n * p = 3**

Algunas distribuciones importantes. Variable discreta

Distribución de Poisson: se obtiene como un límite de la binomial cuando la probabilidad del suceso es baja (p yendo a 0) y el número de intentos alta (n a infinito), manteniendo el valor medio constante ($np = \lambda$). Ej: cantidad de autos por hora que pasan por un semáforo.

$$P(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Parámetro y a su vez valor medio de la distribución



Variable continua. Distribuciones de probabilidad.

Cuando tratamos con variables continuas, se define a la función densidad de probabilidad, cuya integral en un dado intervalo nos indica la probabilidad de que la variable aleatoria caiga en dicho intervalo.

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

función densidad de probabilidad

$$\boxed{\int_{-\infty}^{\infty} f(x)dx = 1}$$

Normalización

Como para una variable continua, la probabilidad de que un valor puntual ocurra es exactamente 0 (pensar!), se suele dar también la función acumulada de probabilidad o función de distribución a secas:

$$P(X < x) = F(x)$$

función de distribución (acumulada) de probabilidad

Valores medios. Variable continua

Valor medio: $$\langle x \rangle = \int x f(x) dx$$

También llamado valor esperado o esperanza matemática

Varianza: $$\sigma^2 = \int (x - \langle x \rangle)^2 f(x) dx$$

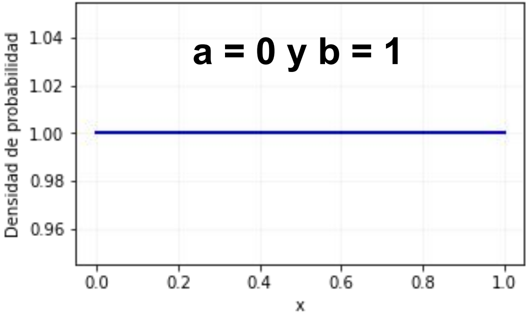
Promedio del cuadrado de las distancias a la media

Desviación: $$\sigma = \sqrt{\sigma^2}$$

Algunas distribuciones importantes. Variable continua

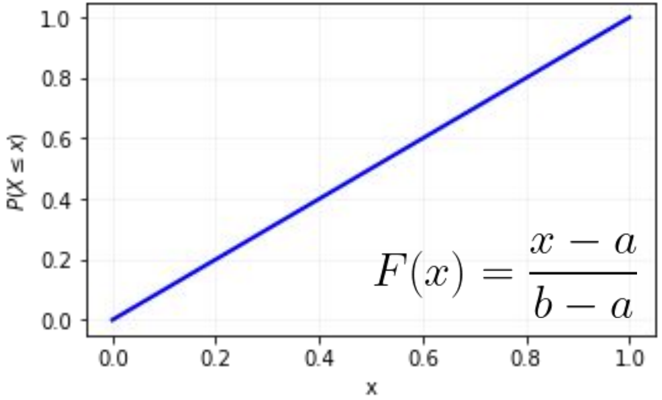
Distribución uniforme: todos los valores en el intervalo [a, b] son equiprobables.

$$f(x) = \frac{1}{b-a}$$



Función densidad de probabilidad

Función distribución (acumulada) de probabilidad



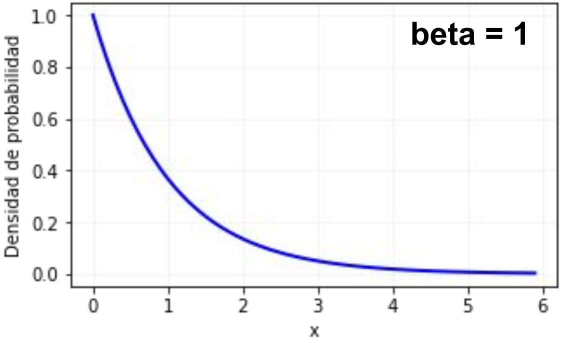
Si x está uniformemente distribuida en [0, 1] entonces (b-a)*x + a lo está en el intervalo [a, b].

Algunas distribuciones importantes. Variable continua

Distribución exponencial: se utiliza para modelar por ejemplo tiempos de espera entre eventos raros.

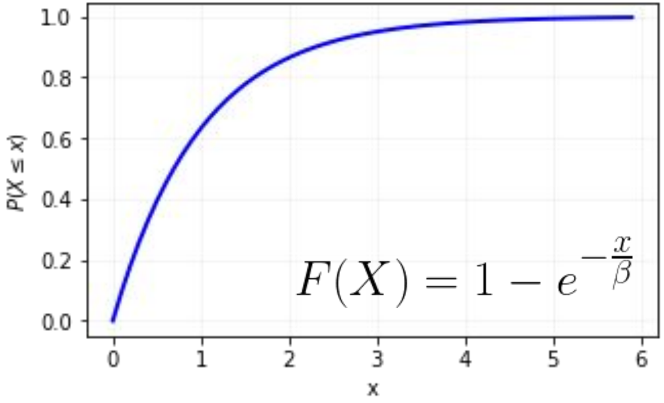
$$f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$$

Valor medio de la distribución



Función densidad de probabilidad

Función distribución (acumulada) de probabilidad



Algunas distribuciones importantes. Variable continua

Distribución normal o gaussiana: esta distribución aparece en múltiples contextos, como resultado de la sumatoria de varios eventos aleatorios (teorema central de límite).

Valor medio

$$\mu \equiv \langle x \rangle$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

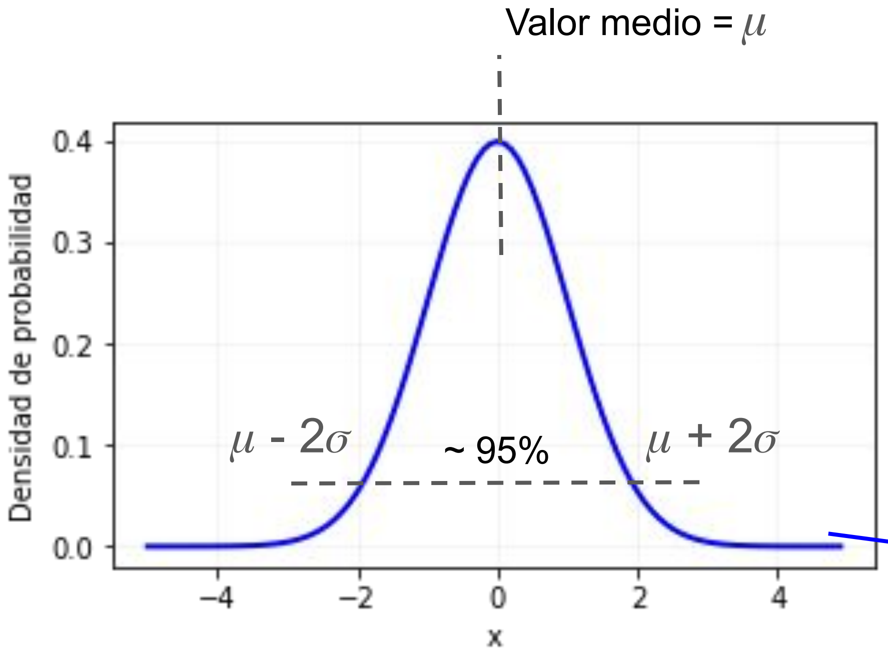
Varianza

Función densidad de probabilidad

En esta distribución, la media y la varianza aparecen explícitamente como parámetros de la distribución.

Propiedades y ¿por qué es tan importante la normal?

Función densidad:



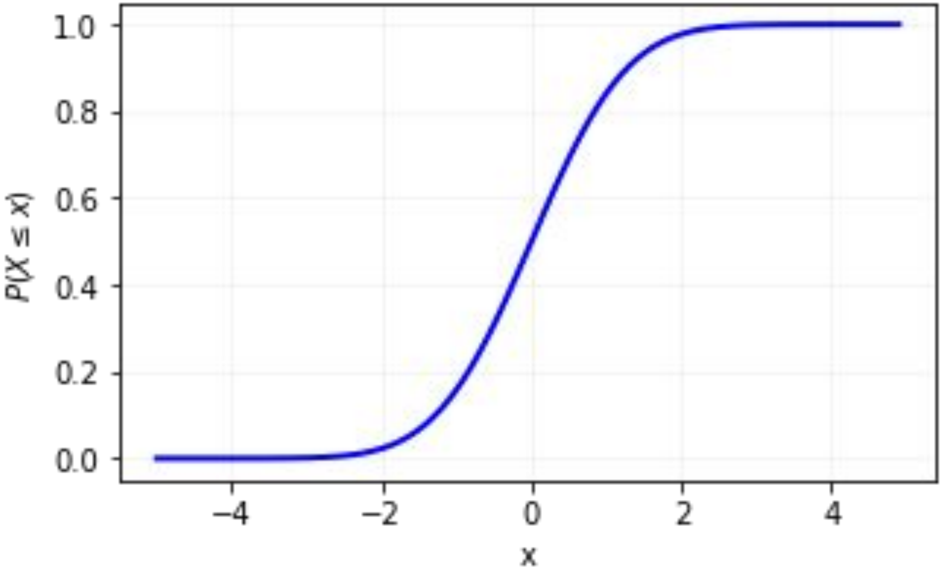
La distribución normal es simétrica y alrededor del 95% de las veces, la variable aleatoria cae a menos de 2 desviaciones de la media.

Para una variable normal estandarizada un valor mayor a 2 (o menor a -2) es raro.

$$Z \sim N(0, 1)$$

Propiedades y ¿por qué es tan importante la normal?

Función acumulada:



$$F(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$$

Función error
https://es.wikipedia.org/wiki/Funci%C3%B3n_error

Función distribución (acumulada) para una variable estandarizada.

Propiedades y ¿por qué es tan importante la normal?

- Si $X \sim N(\mu, \sigma^2)$ y definimos $Z = \dfrac{X - \mu}{\sigma}$ entonces:

$$Z \sim N(0, 1)$$

Estandarización!

- Si $Z \sim N(0, 1)$ y definimos $X = \sigma Z + \mu$ entonces:

$$X \sim N(\mu, \sigma^2)$$

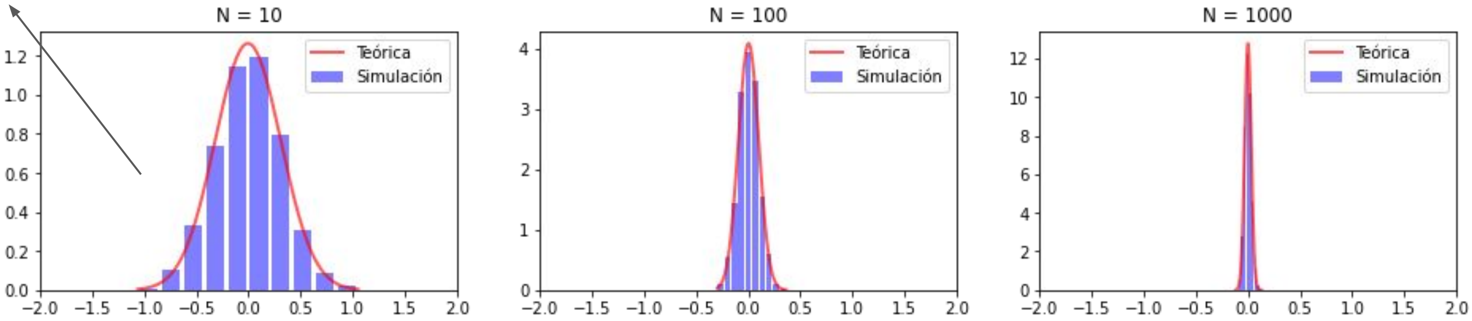
Es decir, una variable normalmente distribuida es fácil de trasladar y escalar, sin perder sus propiedades

Propiedades y ¿por qué es tan importante la normal?

Teorema central del límite: sean X_1, X_2, X_3, … X_N, v.a. independientes e idénticamente distribuidas con media μ y varianza σ^2, entonces para N grande, el promedio está normalmente distribuido.

$$\bar{X} = \frac{1}{N}\sum_i X_i \longrightarrow \bar{X} \sim Norm\left(\mu, \frac{\sigma}{\sqrt{N}}\right)$$

Distribución del promedio



Ver colab

Otras distribuciones para tener presente

Distribución log-normal: Si multiplicamos muchas variables aleatorias y calculamos su logaritmo, la multiplicación se transforma en… una sumatoria de variables aleatorias. ¿La sumatoria cómo se distribuye? A una distribución normal por supuesto. ¿Y la multiplicación? Pues a una log-normal!

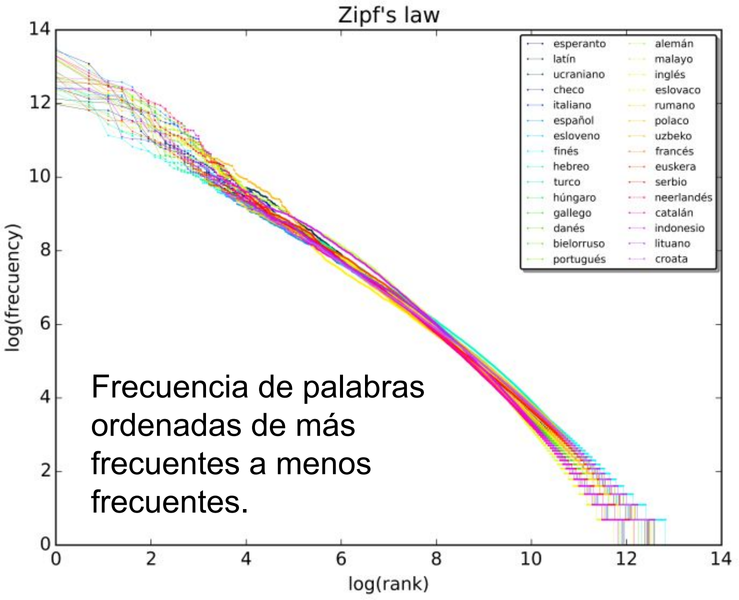
$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}}$$

Jugamos un poco con esta distribución en el colab

Si una variable está distribuida con una log-normal, entonces el logaritmo de dicha variable sigue una distribución normal.

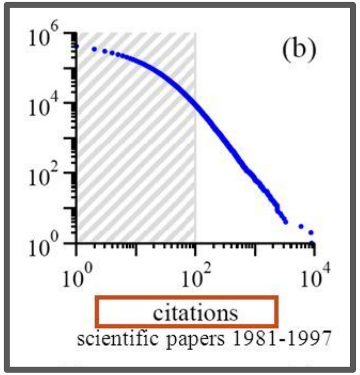
Otras distribuciones para tener presente

Distribución power-law o ley de potencia: $f(x) = ax^{-\nu}$



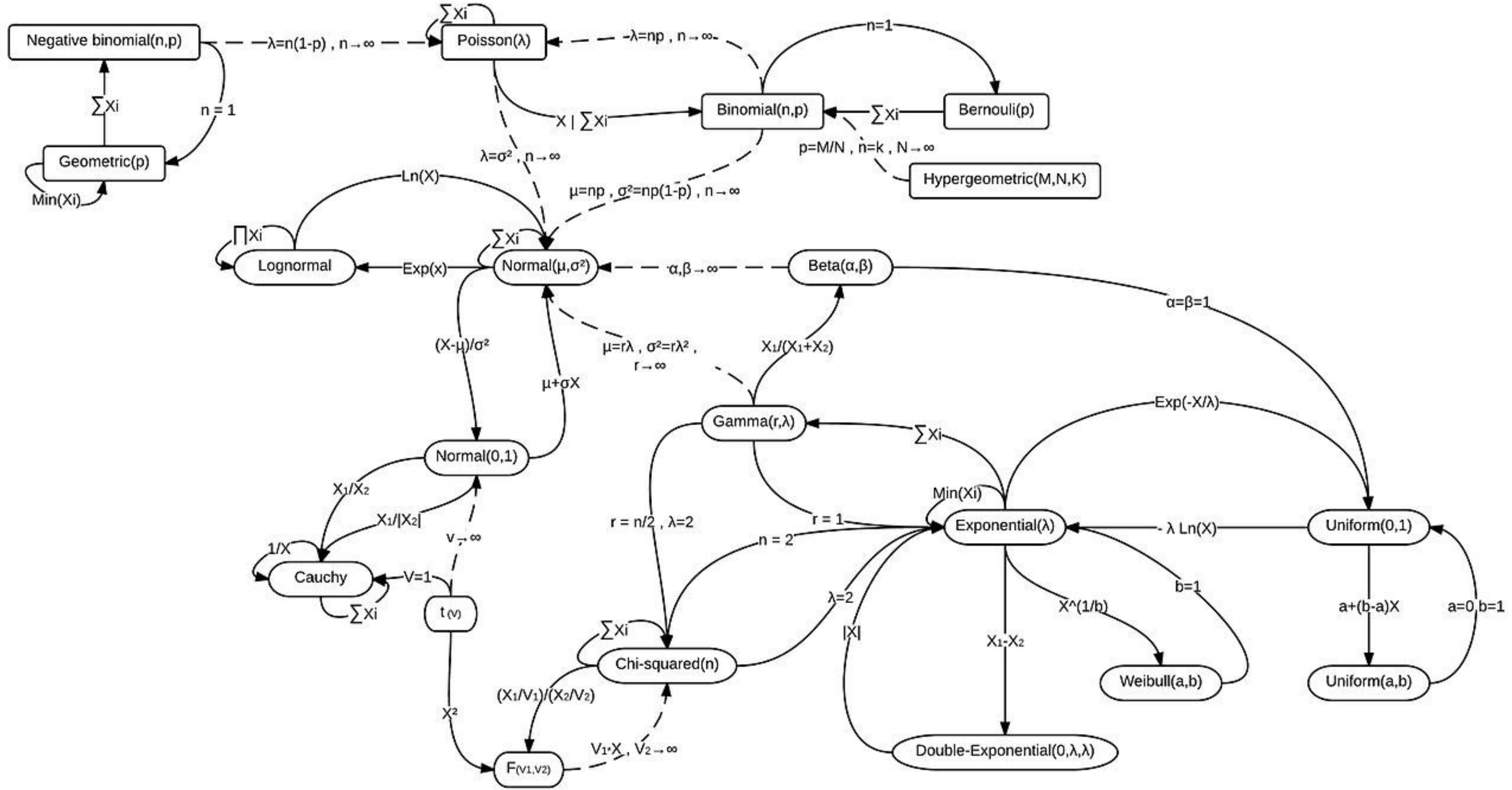
Frecuencia de palabras ordenadas de más frecuentes a menos frecuentes.

De SergioJimenez - Trabajo propio, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=45517440



Citas en papers científicos
Muchos papers reciben pocas citas, unos pocos reciben muchas.

Un rasgo característico de las leyes de potencia es que en escala log-log se ven como una relación lineal (ver colab).



https://en.wikipedia.org/wiki/Relationships_among_probability_distributions

Valores medios y varianza

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial(n, p)	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a, b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ, σ^2)	μ	σ^2
Exponential(β)	β	β^2
Gamma(α, β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α, β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
t_ν	0 (if $\nu > 1$)	$\nu/(\nu-2)$ (if $\nu > 2$)
χ^2_p	p	$2p$
Multinomial(n, p)	np	see below
Multivariate Normal(μ, Σ)	μ	Σ

Desigualdad de Chebyshev

4.2 Theorem (Chebyshev's inequality). *Let* $\mu = \mathbb{E}(X)$ *and* $\sigma^2 = \mathbb{V}(X)$. *Then,*

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad \mathbb{P}(|Z| \geq k) \leq \frac{1}{k^2} \tag{4.2}$$

where $Z = (X-\mu)/\sigma$. *In particular,* $\mathbb{P}(|Z| > 2) \leq 1/4$ *and* $\mathbb{P}(|Z| > 3) \leq 1/9$.

Lo más importante de esto es que la **media define un punto de referencia** y la varianza o la **desviación dá una idea de la escala**, es decir, lejos o cerca de la media siempre es relativo a cuánto mide la desviación.

Momentos de una distribución

Se definen los momentos de una distribución como el valor medio de la variable elevada a una dada potencia:

$$\langle x^n \rangle = \sum_i x_i^n P(x_i)$$

Discreta

$$\langle x^n \rangle = \int x^n f(x) dx$$

Continua

Es usual ver distintas cantidades en términos de momentos:

$$\langle x \rangle \equiv \langle x^1 \rangle$$

La media es el primer momento

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$$

Varianza

Conocer todos los momentos de una distribución es igual a conocer la distribución

Covarianza y correlación

Para un par de variables se puede expresar la relación a través de la covarianza (comparar con la fórmula de la varianza) que se define como:

$$Cov(X,Y) = \int f(x,y)(x - \langle x \rangle)(y - \langle y \rangle)dxdy$$

Función densidad conjunta

Dividiendo por la desviación de cada variable, obtenemos la correlación:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$-1 \leq \rho \leq 1$$

En forma análoga se define para variables discretas

Covarianza y correlación

Si las variables son independientes, entonces:

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

$$\Downarrow$$

$$Cov(X,Y) = 0 \implies \rho(X,Y) = 0$$

Dos variables independientes implica siempre que la correlación es 0.
Al revés no siempre es válido: si la correlación entre dos variables da cero eso no necesariamente indica que las variables sean independientes!

Conceptos de probabilidad y **<u>estadística descriptiva</u>**

Laboratorio de Datos 1°C 2021

Estadística descriptiva

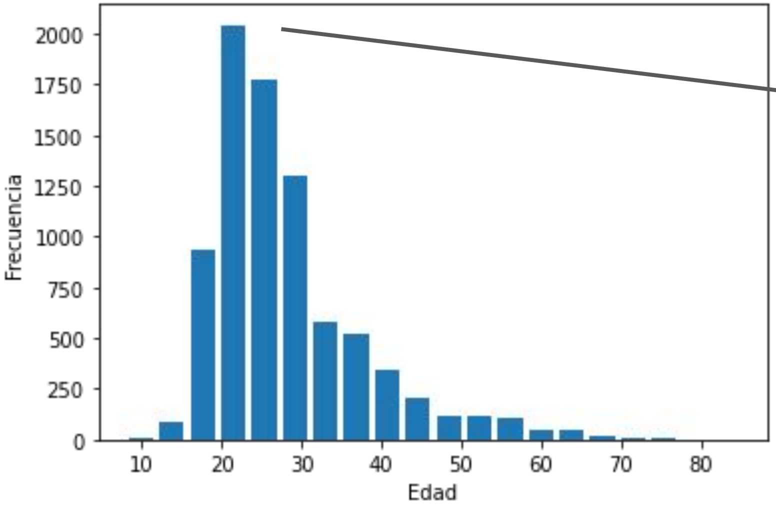
Así como podemos calcular valores medios y momentos de distribuciones de probabilidad, definimos observables en nuestro conjunto de datos.

Con esto podemos:

- **sacar conclusiones** sobre la distribución de probabilidad de la que vienen nuestros datos (estadística inferencial).
- medir estos observables a fin de **resumir y caracterizar** nuestro conjunto de datos (estadística descriptiva).

Histogramas

¿Cómo construimos un análogo de la función densidad o acumulada de nuestros datos? Dividimos el espacio muestral de nuestra variable en segmentos y contamos cuántos datos caen en cada uno.

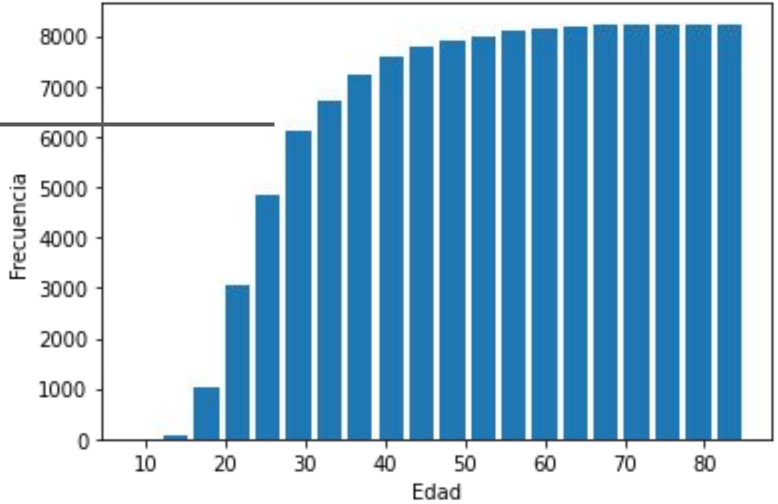


Alrededor de 2000 de individuos tienen entre 20 y 24 años (má’ o meno’)

Histogramas

¿Cómo construimos un análogo de la función densidad o acumulada de nuestros datos? Dividimos el espacio muestral de nuestra variable en segmentos y contamos cuántos datos caen en cada uno.

Más de 6000 participantes (de un total de alrededor 8000) tienen menos de 30 años.

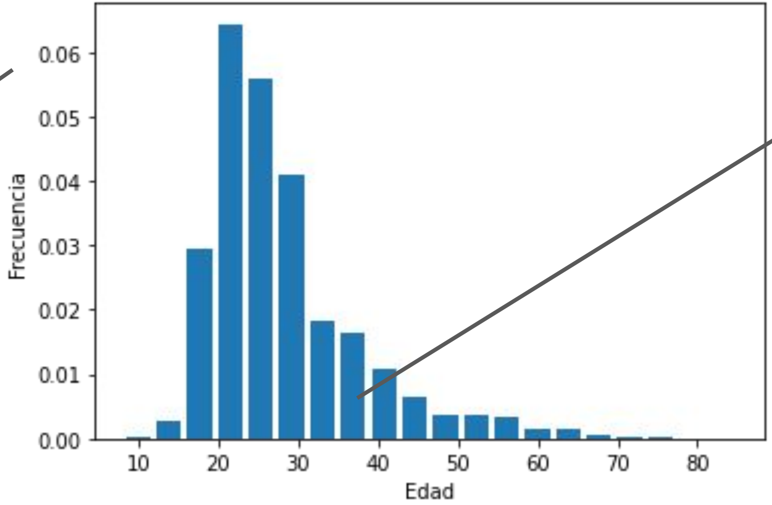


Frecuencia acumulada

Histogramas normalizados

La normalización de los histogramas los hacen más comparables con las distribuciones de probabilidad e independiente de la cantidad de datos que tengamos.

Ojo! estos números pueden ser difícil de interpretar.

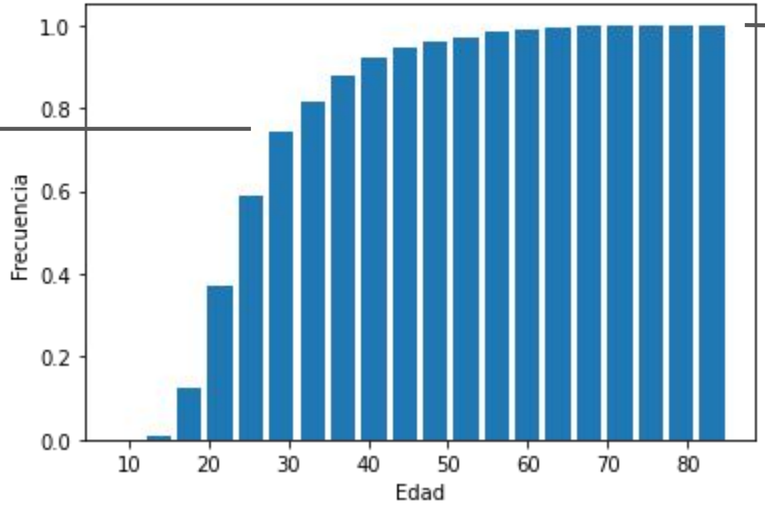


El área cubierta por los bins suma 1.

Histogramas normalizados

La normalización de los histogramas los hacen más comparables con las distribuciones de probabilidad e independiente de la cantidad de datos que tengamos.

Alrededor del 75% de los participantes tienen menos de 30 años.



Este límite siempre es 1.

Observables estadísticos

Análogo a las definiciones de momentos para las distribuciones de probabilidad, para nuestro conjunto de datos podemos definir:

- **Media muestral** (promedio de nuestros datos): $\bar{X} = \frac{1}{N} \sum_i X_i$

- **Varianza muestral:** $S^2 = \frac{1}{N-1} \sum_i (X_i - \bar{X})^2$

- **Desviación muestral:** $S = \sqrt{S^2}$

Observación: el (N-1) surge de pedirle que el estimador la varianza sea estadísticamente no sesgado. En un contexto de muchos datos es prácticamente lo mismo que N.

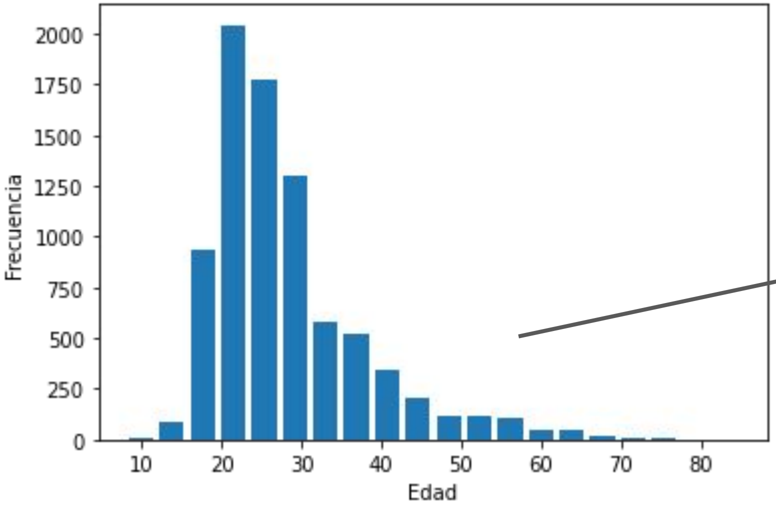
Observables estadísticos

Análogo a las definiciones de momentos para las distribuciones de probabilidad, para nuestro conjunto de datos podemos definir:

- **Moda:** para una variables discreta es el valor más frecuente.

Observables estadísticos

Análogo a las definiciones de momentos para las distribuciones de probabilidad, para nuestro conjunto de datos podemos definir:



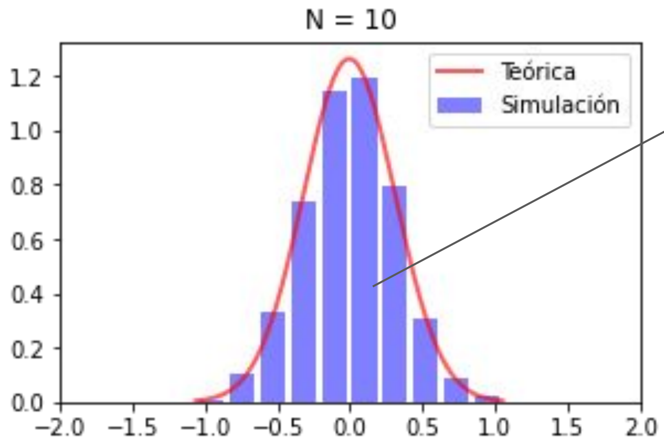
Media: 28.1

Varianza: 90.6

Desviación: 9.5

Error estándar

Cuando calculamos la media muestral de un conjunto de datos muchas veces la interpretamos como una estimación del valor medio de la población del cual vienen nuestros datos. Si es una estimación, en cuánto le estamos errando?



Distribución del promedio

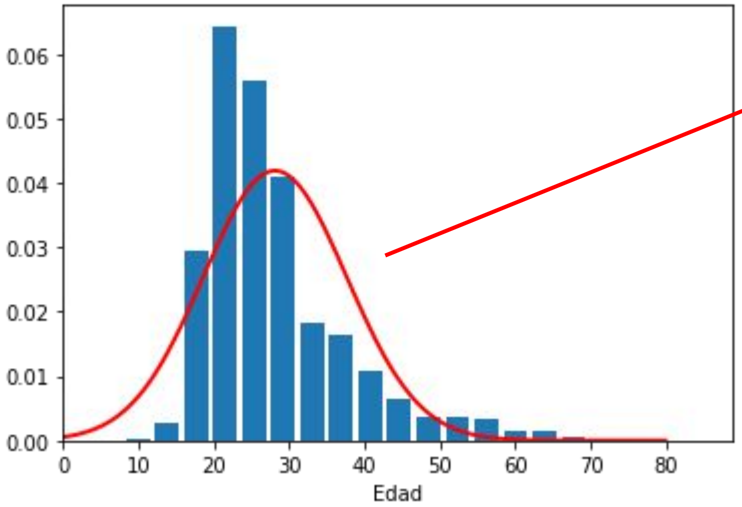
$$\mu \sim \bar{X} \pm \frac{S}{\sqrt{N}}$$

Recordar diapo del teorema central del límite

Para el ejemplo de las edades (N = 8236), la media poblacional quería estimada en: 28.1 ± 0.1

¿Alcanza con la media y la desviación?

Reportar solo la media y la varianza puede dar una idea de que los datos están normalmente distribuidos, lo cual puede ser una muy mala aproximación.

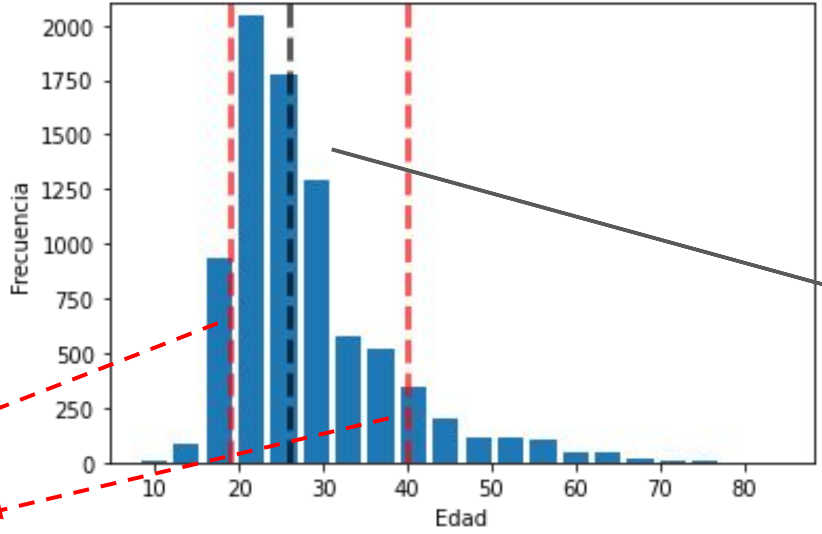


Normal con la media y desviación de los datos

Ojo!: aunque el ejemplo es medio burdo, a veces la normal es una aproximación útil alrededor del valor más probable.

Mediana y cuantiles

Intervalos que contienen una dado porcentaje de los datos. Permite dar una idea más precisa de cómo es la forma de la distribución.



Mediana = 26
C. 80% central = [19 - 40]

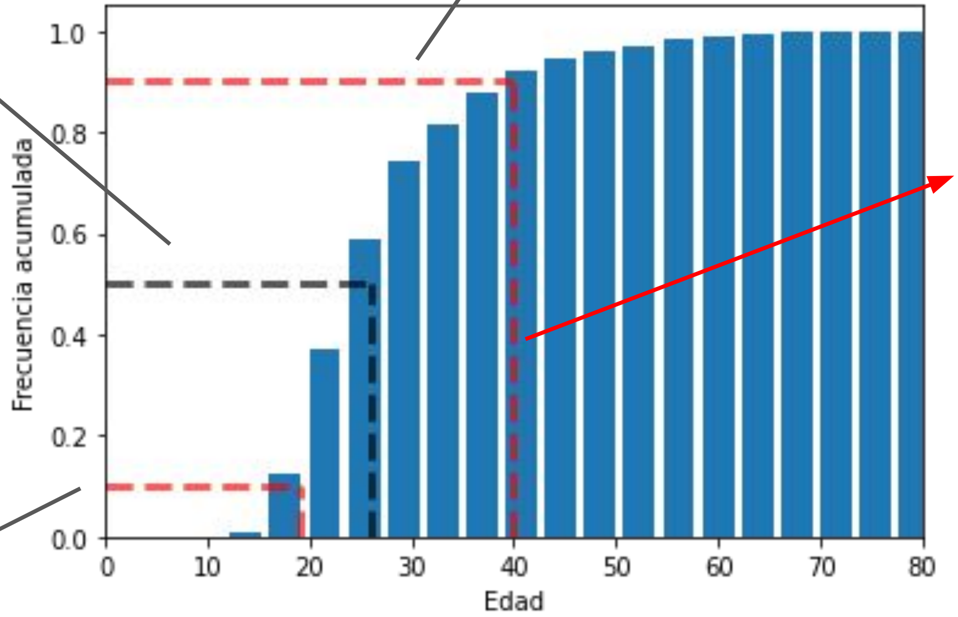
Intervalo central que contiene el 80% de los datos.

Mediana: cuantil del 50%, separa los datos en dos mitades.

Mediana y cuantiles

El 90% de los participantes tienen 40 años o menos.

Mediana: la mitad de los participantes tienen 26 años o menos (o la mitad tienen 26 o más)



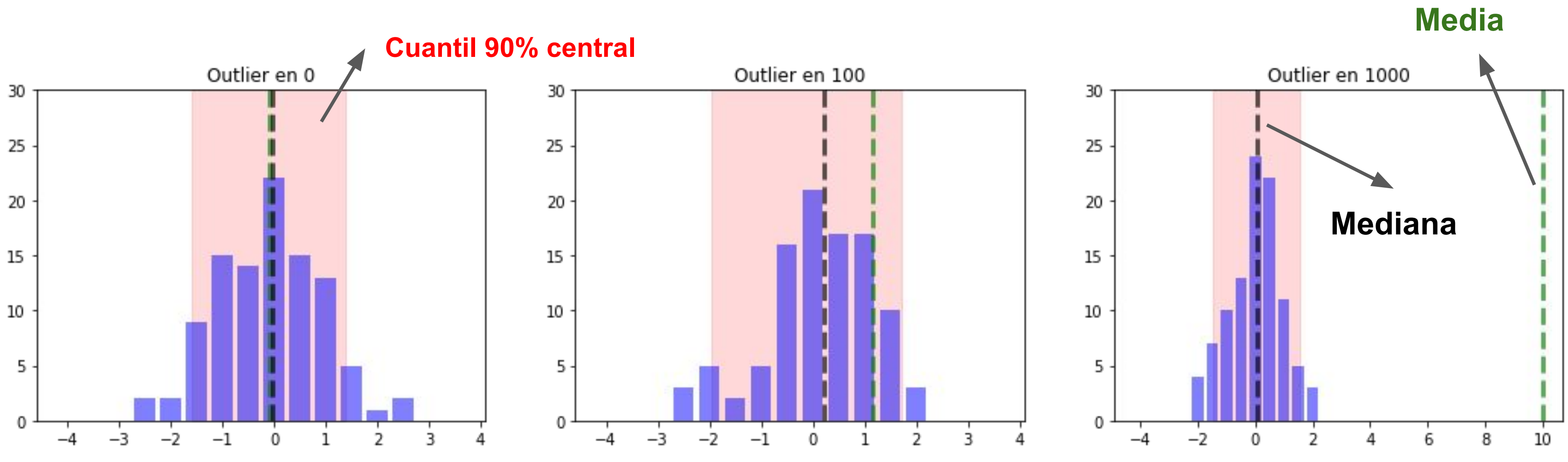
El 80% de los participantes tienen más de 19 años y menos de 40.

Menos del 10% de los participantes tienen 19 años o menos.

La frecuencia acumulada (y normalizada) permite visualizar mejor los cuantiles.

Robustez de estadísticos

La mediana y los cuantiles son estadísticos robustos, mientras que la media y la varianza se ven muy afectadas por la presencia de outliers.



Experimento: a variables normalmente distribuidas, le agregamos un punto que consideramos como outlier (ver colab).

Correlación de dos variables. Coeficiente de Pearson.

La correlación de Pearson mide la relación lineal entre dos variables.

X	Y
1.3	4.3
2.4	8.9
1.5	10.1
1.2	3.9
5.1	5.3

$$\bar{X} = \frac{1}{N}\sum_i X_i$$

$$r = \frac{1}{N}\sum_i \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{S_X S_Y}$$

$$S_X^2 = \frac{1}{N-1}\sum_i (X_i - \bar{X})^2$$

Coeficiente de Pearson.

La correlación de Pearson mide la relación lineal entre dos variables.

X	Y
1.3	4.3
2.4	8.9
1.5	10.1
1.2	3.9
5.1	5.3

$$r = \frac{1}{N}\sum_i \left(\frac{X_i - \bar{X}}{S_X}\right)\left(\frac{Y_i - \bar{Y}}{S_Y}\right)$$

La correlación de Pearson es el promedio de la multiplicación entre las variables estandarizadas.

Coeficiente de Spearman.

En vez de calcular la correlación entre los datos, lo hace entre sus rankings (ordenando los datos de menor a mayor, por ejemplo).

X	R_X	Y	R_Y
1.3	2	4.3	2
2.4	4	8.9	4
1.5	3	10.1	5
1.2	1	3.9	1
5.1	5	5.3	3

Máximo de y

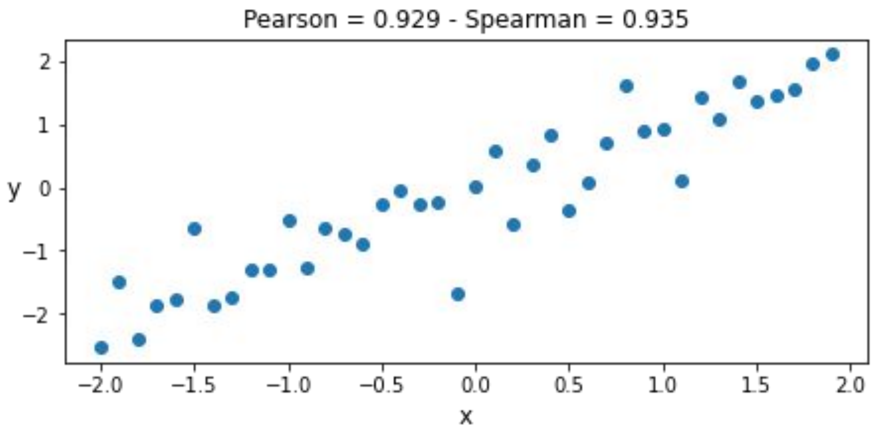
Mínimo de y

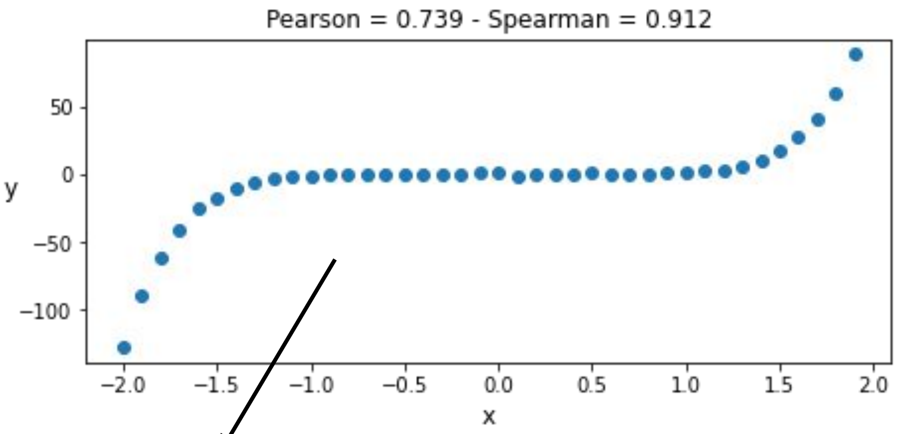
Capta relaciones monótonas entre las variables.

$$r_S = \frac{1}{N}\sum_i \frac{(R_{X_i} - \bar{R}_X)(R_{Y_i} - \bar{R}_Y)}{S_{R_X} S_{R_Y}}$$

Comparación de coeficientes de correlación

Mientras Pearson capta bien la relación lineal entre dos variables, Spearman es más general y encuentra correlaciones entre variables monótonamente relacionadas.

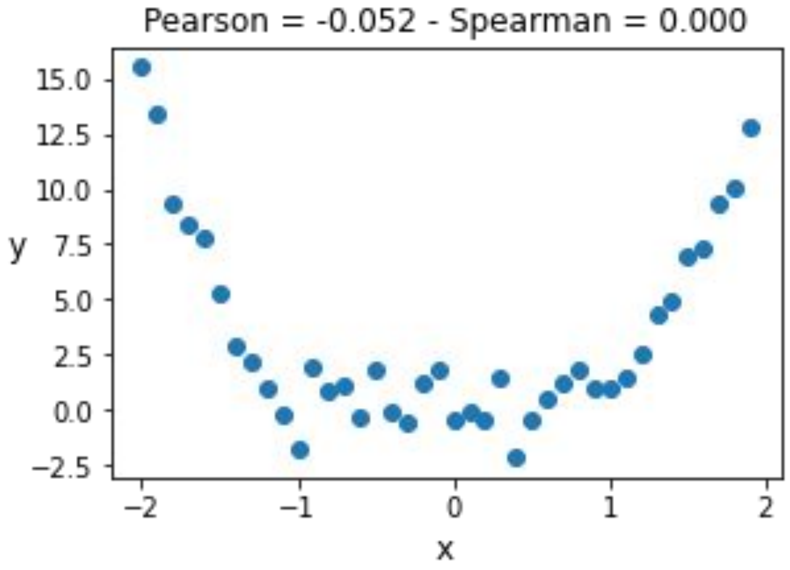




Relación no-lineal pero monótona (cuando una crece o decrece, la otra también o crece o decrece)

Comparación de coeficientes de correlación

Si la relación es no monótona tanto Pearson como Spearman no son buenas medidas de qué tan ligadas están las variables.



Recordar: una correlación baja no necesariamente implica que las variables no están relacionadas.

OJO! Dos variables correlacionadas no necesariamente implica una relación causal, podemos encontrar correlaciones espurias! https://www.tylervigen.com/spurious-correlations

Referencias

- Gráficos interactivos sobre estadística y probabilidad para afianzar conceptos básicos https://seeing-theory.brown.edu/
- Estadística, machine learning y buenos temas? Sí, StatQuest! https://www.youtube.com/c/joshstarmer/playlists
- Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media. Mucho más técnico que lo que exige la materia, pero muy conciso y preciso.