Mondragon Unibertsitatea
Goi Eskola Politeknikoa

Fundamentos estadísticos I

Fundamentos del Aprendizaje Automático

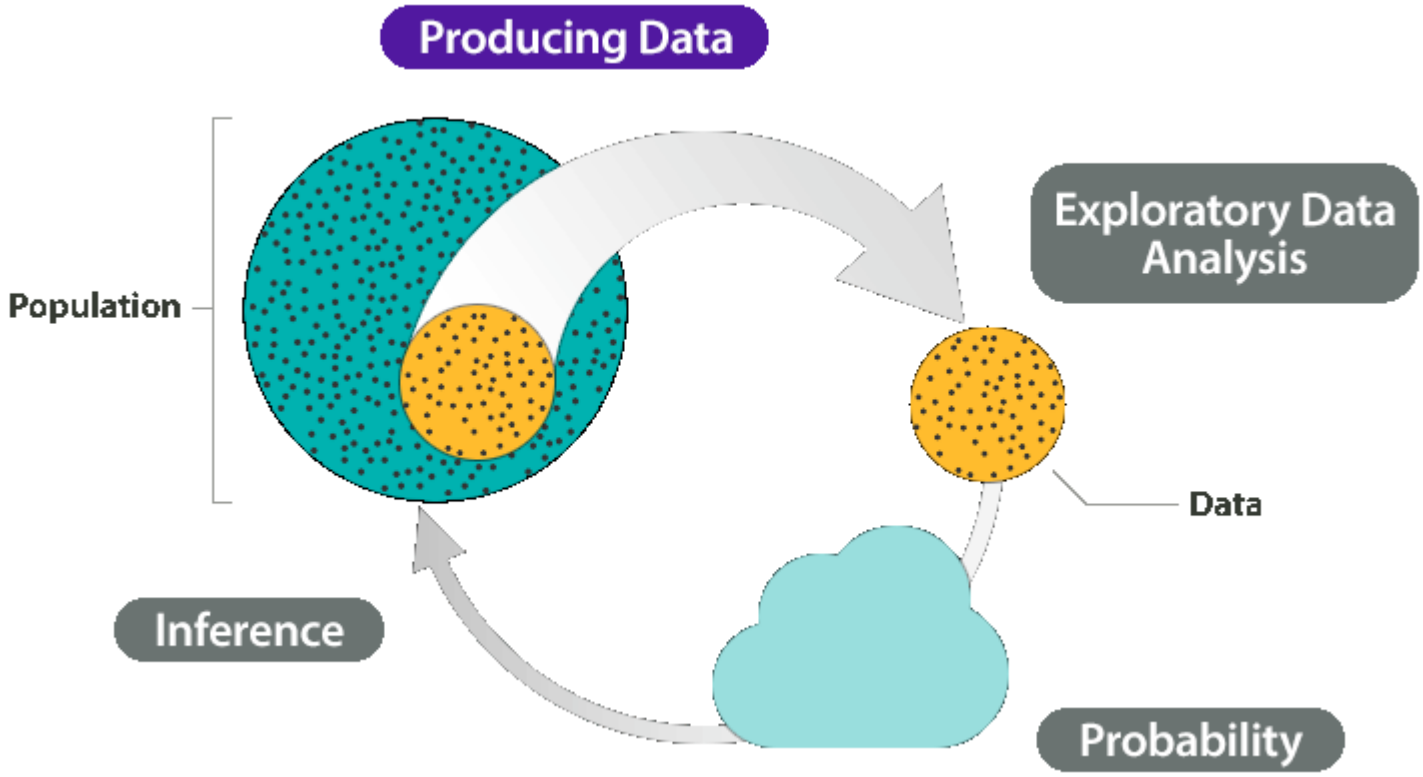
Índice

Mondragon Unibertsitatea

Goi Eskola Politeknikoa

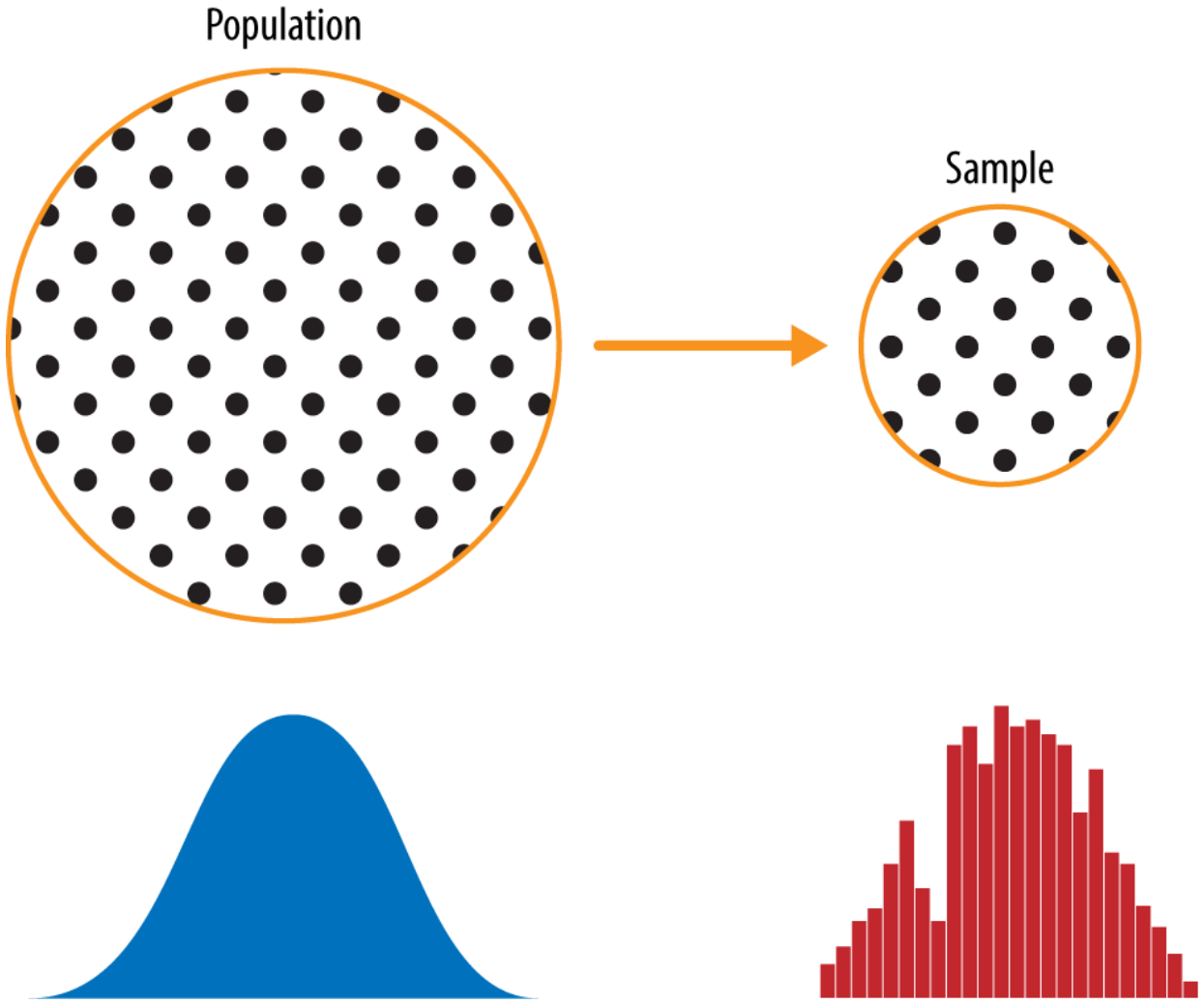
Muestreo de datos

Muestreo de datos



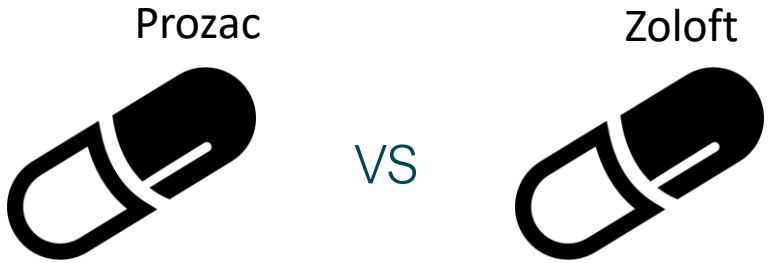
src: standford.edu

Muestreo de datos



Muestreo de datos - DoE

- Tratamiento de la depresión



  - Se compara:
    - Nivel de depresión personas que toman Prozac VS Zoloft

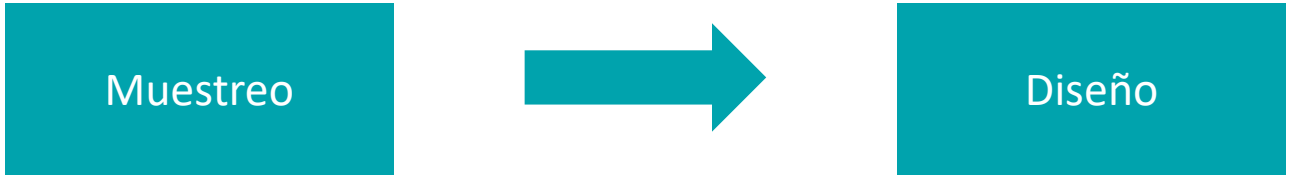
➡ Dato: las personas que toman Prozac están menos deprimidas

  - Conclucisón: Prozac disminuye más la depresión que Zoloft

¿Es esta deducción correcta?

Muestreo de datos

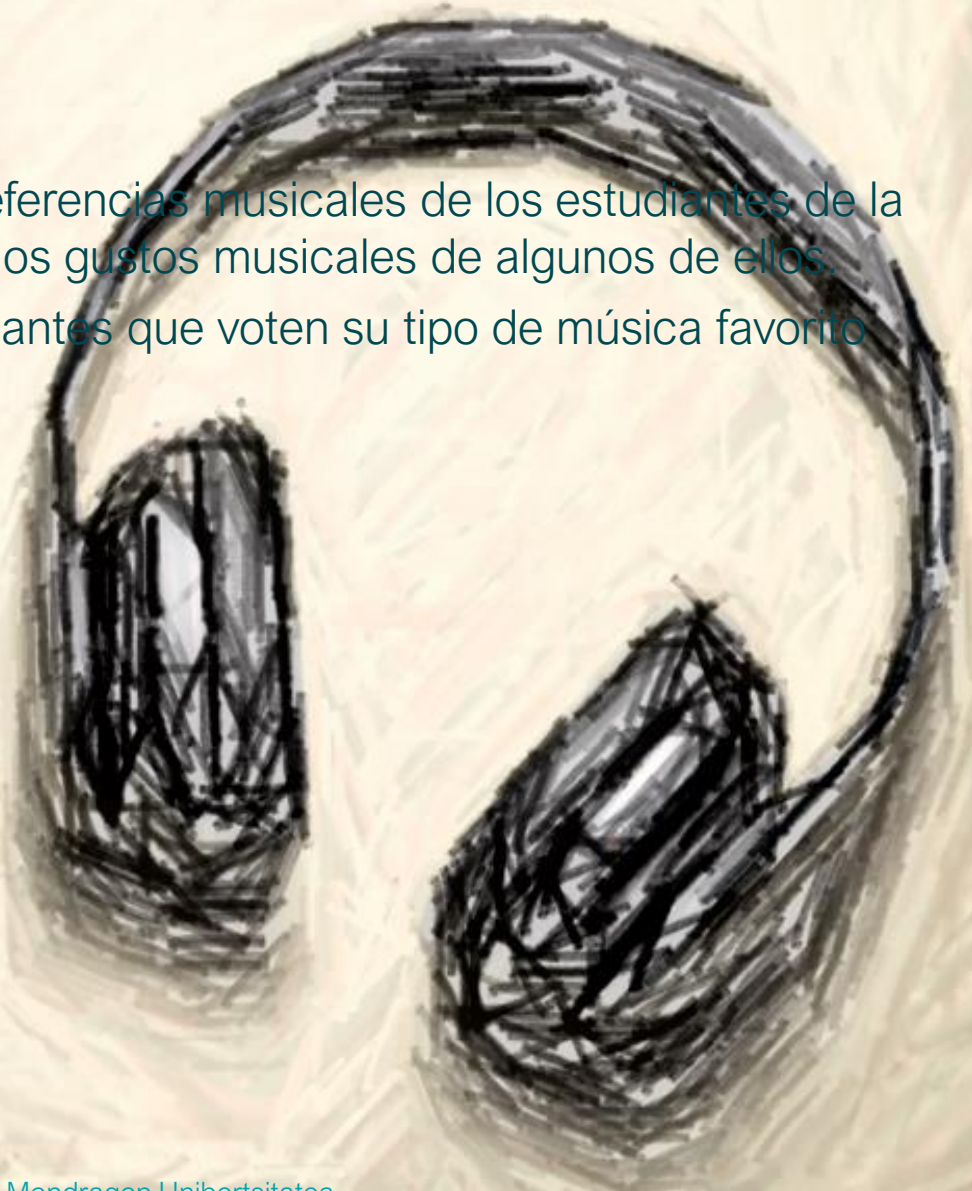
- El diseño para producir datos tiene que ser considerado cuidadosamente
  - Cuidado con las relaciones causales
    - La asociación no siempre implica causalidad!!!



  - Una muestra que crea datos que no son representativos se denomina como **muestra sesgada (biased sample)**

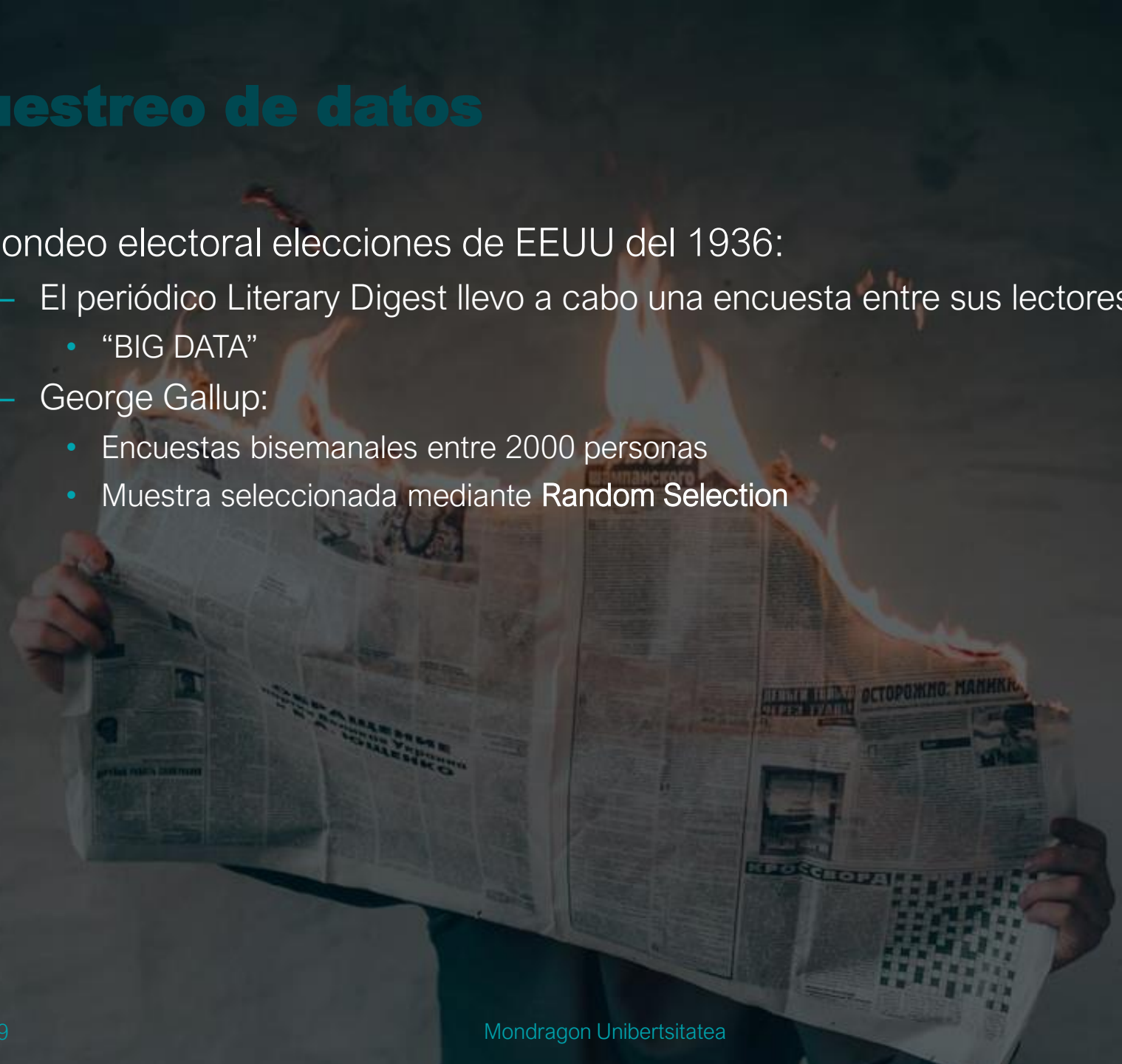
Muestreo de datos

- Ejemplo :
  - Queremos determinar las preferencias musicales de los estudiantes de la universidad, basándonos en los gustos musicales de algunos de ellos.
  - Para ello, se pide a los estudiantes que voten su tipo de música favorito.
  - ¿Sería una buena muestra?



Muestreo de datos

- Sondeo electoral elecciones de EEUU del 1936:
  - El periódico Literary Digest llevo a cabo una encuesta entre sus lectores
    - “BIG DATA”
  - George Gallup:
    - Encuestas bisemanales entre 2000 personas
    - Muestra seleccionada mediante **Random Selection**



Random Selection

Mondragon Unibertsitatea
Goi Eskola Politeknikoa

- Random Selection
  - Simple Random Sampling:
    - Se escoge cualquier “individuo” de forma aleatoria, tienen la misma probabilidad de ser escogidos
  - Stratified Sampling:
    - Se utiliza cuando la población está dividida naturalmente en subpoblaciones (*strata*)
      - Ejemplo elecciones:
        - » Dividir por clase social, género, cultura o raza…
        - » Coger una muestra aleatoria de cada subconjunto

Tamaño de muestra

- Calidad vs cantidad
  - Big Data:
    - Algunas veces, conviene tener menos datos => random sampling
      - Ejemplos:
        - » Capacidad de cómputo
        - » Datos faltantes y outliers: puede resultar muy costoso (o imposible) buscar valores faltantes cuando tenemos millones de registros
    - Otras veces, necesitamos inmensas cantidades de datos:
      - Búsquedas de google
        - » 150 000 palabras en inglés, más de un trillón de consultas al año

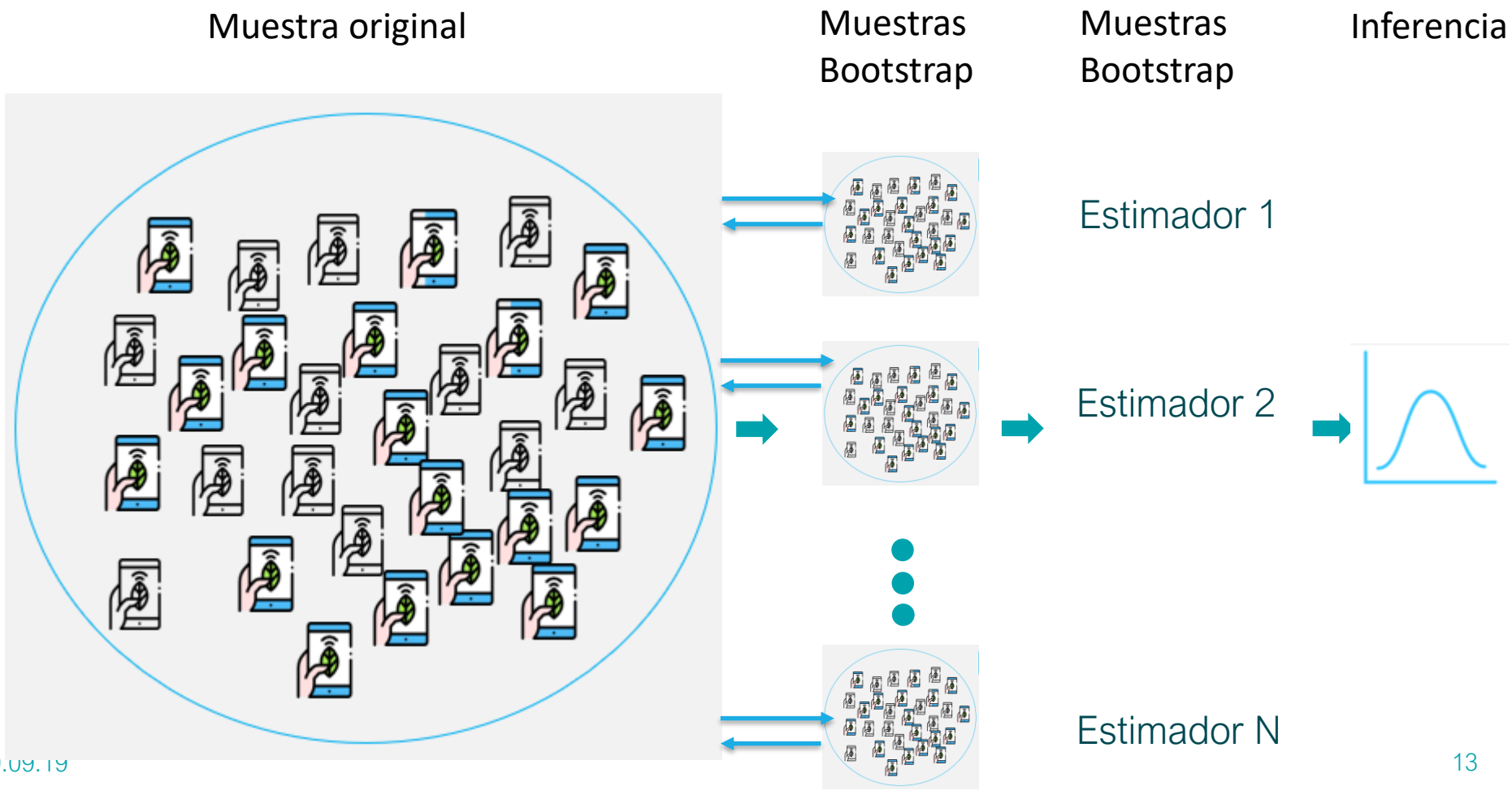
A	Image	Word	Lorry	Square	Ride
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0

Sesgos

- “If you torture the data long enough, sooner or later it will confess”
  - Si se especifica una hipótesis y se valida
    - Se puede obtener una conclusión de cierta confianza
  - Si se trabaja con los datos disponibles y buscamos patrones…seguro que encontramos patrones
    - ¿Será concluyente? ¿O serán fruto del azar?
    - Deberíamos tener uno o varios conjuntos de datos reservados para validar el funcionamiento

Bootstrapping

- Estadística: medida tomada en una **muestra (no población)**
- La idea básica consiste en hacer una inferencia **sobre un estimador o estadística** (por ejemplo, media de la muestra) de una muestra



Bootstrapping

- 1. Cogemos x registros, los guardamos, los reemplazamos
- 2. Repetimos el paso 1. N veces
- 3. Calculamos la media de las muestras
- 4. Repetimos el proceso varias veces
- 5. Usamos los resultados para:
  - Calcular la desviación estándar del estimador (en este caso la media)
  - Visualizamos el resultado + mostramos un **intervalo de confianza**

Bagging

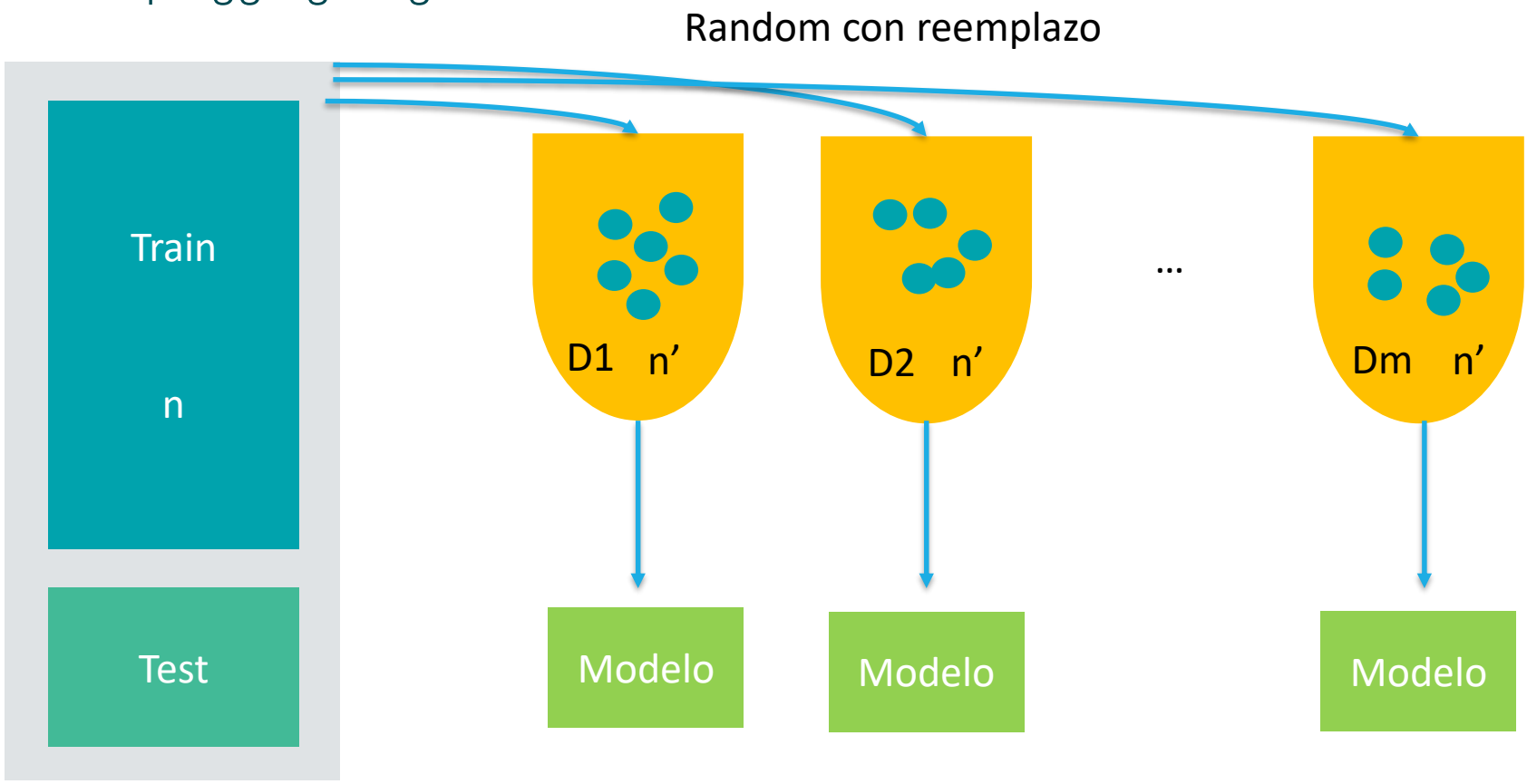
n = número de instancias
n' = número de instancias en cada bag
m = número de bolsas

- Bootstrap aggregating:

Bagging

n = número de instancias
n' = número de instancias en cada bag
m = número de bolsas

- Bootstrap aggregating:



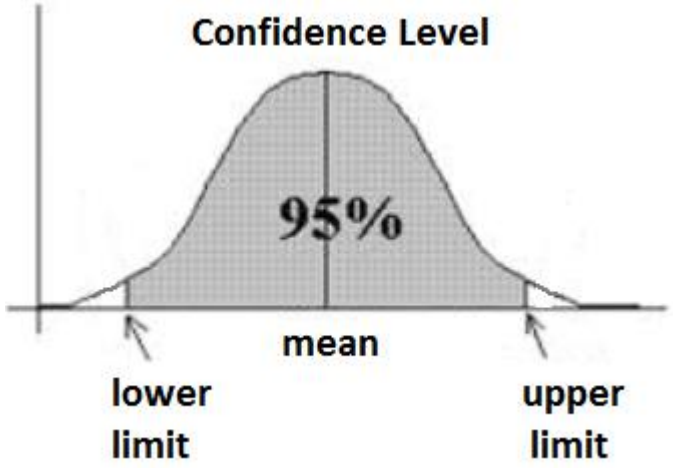
Mondragon Unibertsitatea

Goi Eskola Politeknikoa

Significación estadística

Intervalos de confianza

- Un par (o varios) de valores entre cuales se estima que estará un cierto valor desconocido con una determinada probabilidad de acierto



- Línea central: sería la hipotética media real de la población
- Podemos afirmar con un 95% de confianza que la estadística (por ejemplo, la media) estará entre el valor superior y el inferior

Test de hipótesis

- Se denomina al proceso de hacer inferencias sobre la población basándose en un test estadístico sobre la muestra
- La **hipótesis nula** y la **hipótesis alternativa** son maneras de validar si una asunción es significativa o no estadísticamente
  - Test A/B:
    - Hipótesis: el precio B genera más beneficios
      - ¿Por qué necesitamos una hipótesis?¿Por qué no elegimos la opción que nos de más beneficio?

Test de hipótesis

Mondragon Unibertsitatea
Goi Eskola Politeknikoa

Lanzamientos reales…

Cruz	Cara	Cruz	Cara	Cruz	Cruz	Cruz	Cruz	Cara	Cruz
Cara	Cara	Cruz	Cara	Cara	Cara	Cara	Cara	…	…
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…

Test de hipótesis

- Se diseñaría un experimento A/B de tal manera que la diferencia entre A y B sea por:
  - Asignación aleatoria de sujetos
  - Diferencia real entre A y B
- Una hipótesis estadística evalúa si la aleatoriedad es una explicación válida para la diferencia entre los grupos A y B

Hipótesis nula

- Objetivo de una hipótesis:
  - Demostrar que la diferencia entre grupos es más extrema que la que la aleatoriedad podría producir
  - La asunción de base: los dos “tratamientos” son equivalentes
    - Hipótesis nula
      - Tenemos que demostrar que esta hipótesis es incorrecta

Significación estadística y el p-valor

- Significación estadística:
  - Modo en el que se mide si un experimento produce un resultado más extremo del que la aleatoriedad podría producir
  - P-valor: La frecuencia con la que el modelo de aleatoriedad produce un resultado más extremo que el resultado observado

Ejemplo 1

- Un neurólogo quiere testar el efecto de una droga en el tiempo de respuesta inyectando 100 ratas con una unidad de la dosis.
  - La media del tiempo de respuesta de las ratas que no están inyectadas es 1.2 segundos
  - La media de las ratas inyectadas es 1.05 segundos (std = 0.5s)
- ¿Tiene la droga algún efecto en el tiempo de repuesta?

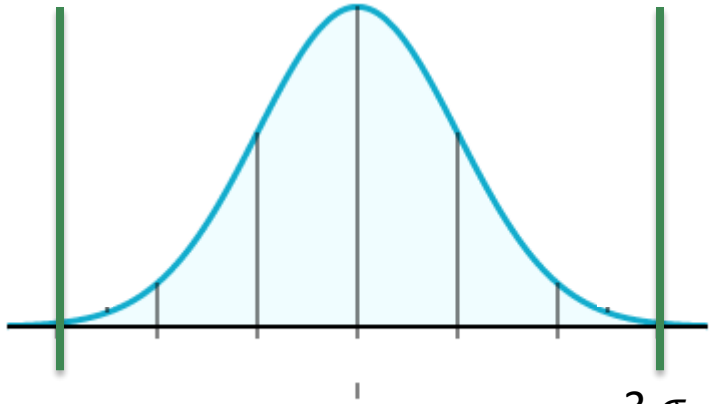
- $H_0 = La\ droga\ no\ tiene\ efecto \Rightarrow\ \mu = 1.2s$
- $H_1 = La\ droga\ tiene\ efecto \Rightarrow\ \mu\ != 1.2s$ al darle la droga (hip. Alternativa)

Ejemplo 1

- Asumimos que la hipótesis nula es cierta
  - Siendo esta hipótesis cierta, ¿cuál es la probabilidad de que obtengamos estos resultados en la muestra?
  - Si la probabilidad es lo suficientemente pequeña => seguramente la hipótesis nula no será cierta.

- Comprobémoslo…

Ejemplo 1



$$\mu_{\bar{x}} = \mu = 1.2\ s$$

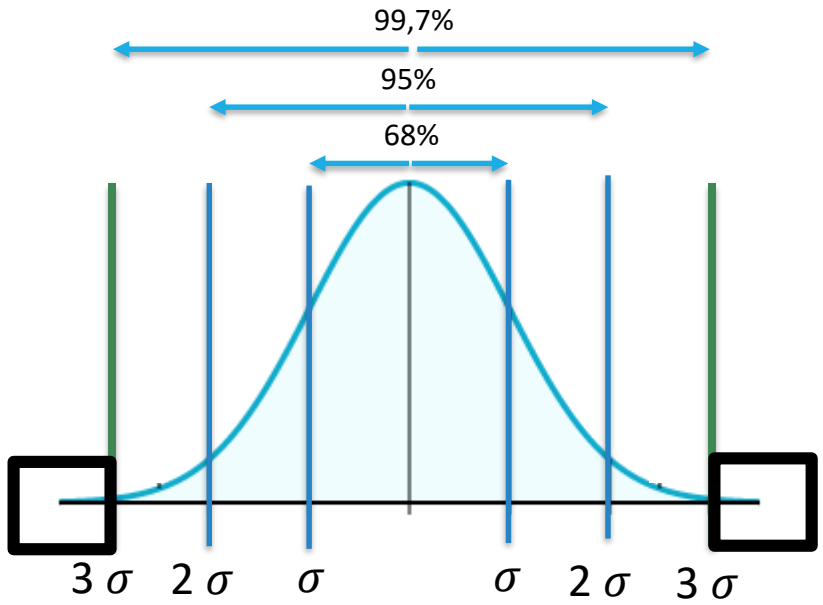
$$\sigma_{\bar{x}} = \frac{0}{\sqrt{100}} \approx \frac{s}{\sqrt{100}} = \frac{0.5}{10} = 0.05$$

Z-score: el número de desviaciones estándar del que se encuentra un punto desde la media

$$\Rightarrow \quad z = \frac{1.2 - 1.05}{0.05} = 3$$

¿Cuál sería la probabilidad de que obtengamos un valor tan extremo de forma aleatoria?

Ejemplo 1



¿Cuál sería la probabilidad de que obtengamos un valor tan extremo de forma aleatoria?

➡ Si la hipótesis nula es cierta => hay un 0.3% de probabilidad de que sea cierta

- Podemos rechazar la hipótesis nula

P-valor = 0.003

0.003 < **0.05** (5%)

Ejemplo 2:

- Un fabricante de frenos de disco afirma que los discos fabricados en su fábrica pesan al menos 1000g. Tenemos la sensación de que esto puede no ser cierto.
  - Recogemos una muestra de 30 discos, siendo la media de los discos 990g y la desviación estándar 12.5 g. Tomando como el nivel de significancia 0.05 ¿podemos rechazar la hipótesis del fabricante?

- *Hipótesis nula* => $\mu 0 \geq 1000$

Ejemplo 2

$$\mu_{\bar{x}} = \mu = 990\ s$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{30}} \approx \frac{s}{\sqrt{30}} = \frac{12.5}{\sqrt{30}} = 2{,}28$$

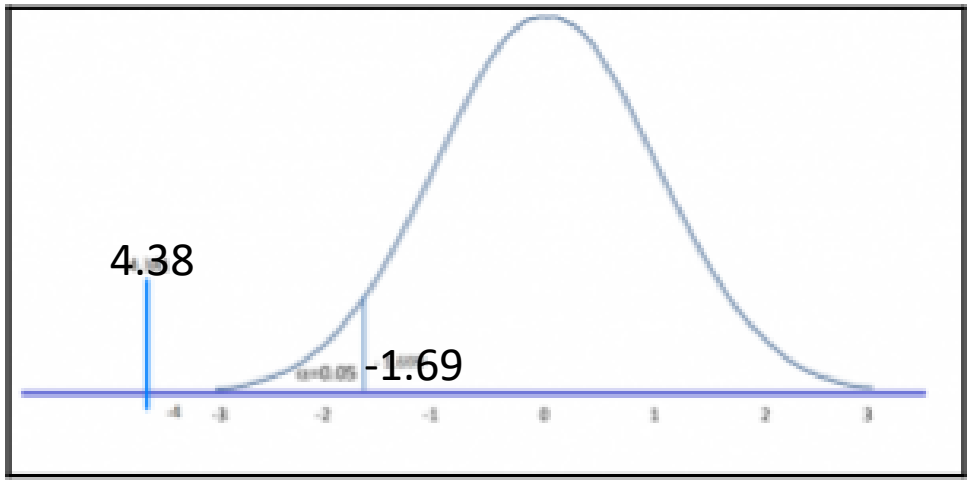
$$z = \frac{990\ - 1000}{2{,}28} = -4.38$$

Si P- valor = 0.05 => z = ±1.699 **(empírico)**

-4.38 < -1.699

P- valor = 7.03 e^{-05}

Podemos rechazar la hipótesis nula



Chi-square test

- Test de independencia
  - Test de hipótesis en datos ordinales
  - Partiendo de dos variables aleatorias X e Y: el test determina si existe dependencia entre ellos

  - $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$ $\quad o_i = observed \;\; e_i = \text{expected}$

Ejemplo

- Tenemos un profesor que pone exámenes tipo test, donde las opciones siempre son A, B, C o D. Este profesor nos indica que todas las respuestas correctas están distribuidas de forma equitativa (25%), pero tenemos la sensación de que no es así…

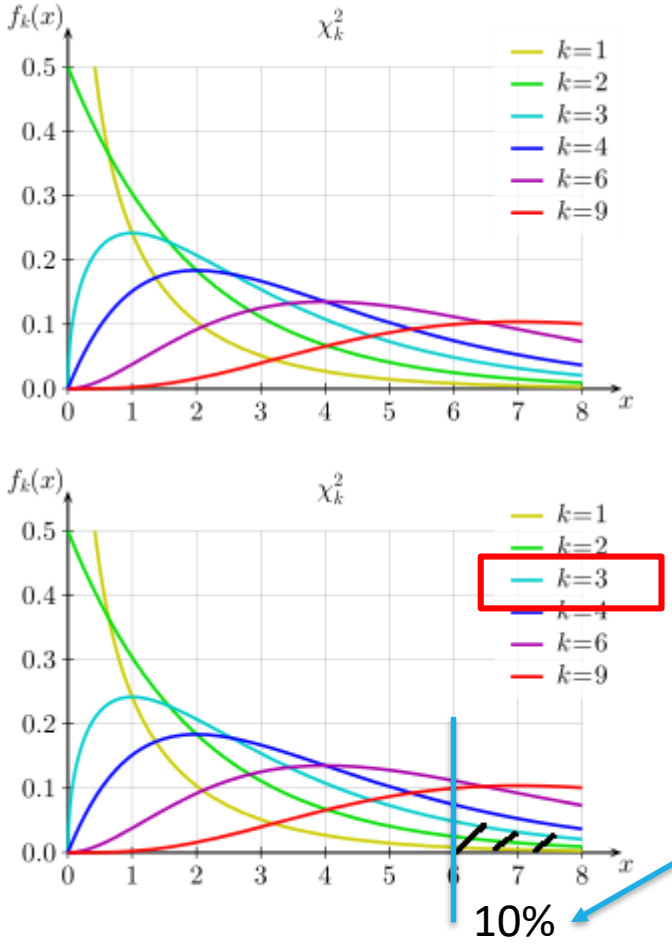
- *H0 = Las respuestas están distribuidas de forma equitativa*
- *Ha = Las respuestas no están distribuidas de forma equitativa*

Res. correcta	Res. Esperada	Res. observada
A	25	20
B	25	20
C	25	25
D	25	35

$$\chi^2 = \frac{(25-20)^2}{25} + \frac{(25-20)^2}{25} + \frac{(25-25)^2}{25} + \frac{(25-35)^2}{25} = 6$$

¿Cuál es la probabilidad de que obtengamos un valor tan extremo como el χ^2?

Ejemplo



P-valor = $P(\chi^2) \geq 6 > 0.1$

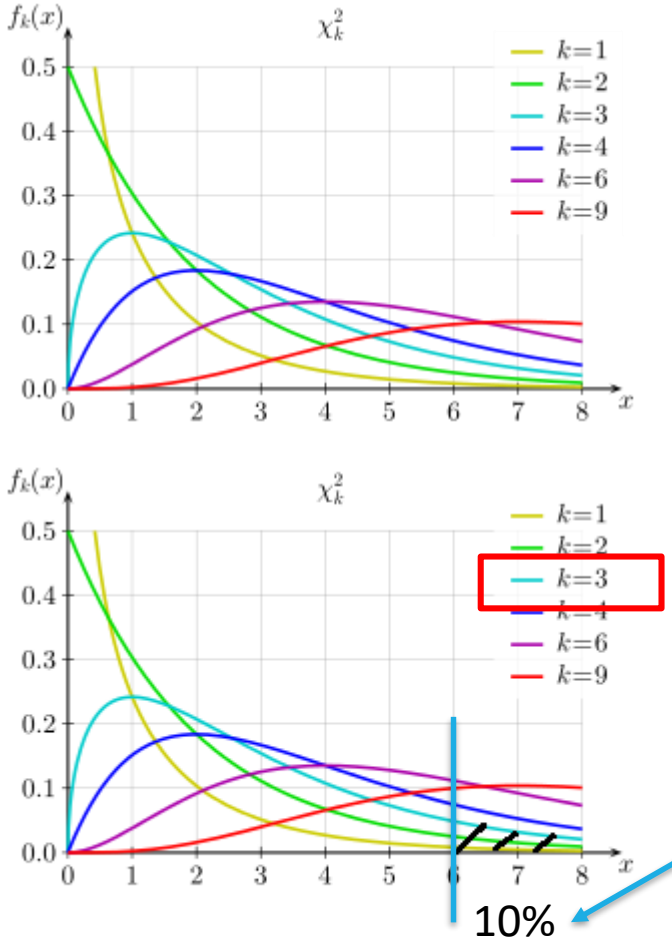
k = grados de libertad

k = 3 => tenemos 4 variables, con k = 3 podemos deducir el cuarto valor

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Ejemplo



k = grados de libertad

k = 3 => tenemos 4 variables, con k = 3 podemos deducir el cuarto valor

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

P-valor = $P(\chi^2) \geq 6 > 0.1$

No podemos rechazar la H0

Otros test

- ANOVA: Analysis of Variance,
- T-test
- Mann-Whitney U-test
- …

Uso de los tests en ML

- En el machine learning nos servirán de poco, se utilizan más en estadística
  - Determinar el p-valor adecuado para las publicaciones
  - Feature selection:
    - Identificar si la prevalencia de una clase es inusualmente alta => no aleatoria

P-hacking...

- Atención investigadores…:
  - Los p-valores pueden indicar cuánto de compatible es el dato con un modelo estadístico concreto
  - ¡No miden la probabilidad de que una hipótesis sea cierta!
  - No nos podemos basar solo en el p-valor para tomar decisiones
    - Si se utiliza, hay que combinarla con otras evidencias

P-hacking

- P-valor 0.05 => ¿por qué?...
- Analizamos nuestros datos
  - Vemos que rozamos la significación estadística (0,051)…
  - ¡Cogemos más datos!
    - P-valor = 0.049
- ¿aseguramos con esto que los resultados son significantes?

Mondragon Unibertsitatea
Goi Eskola Politeknikoa

Distancias estadísticas

Distancias estadísticas

- Uso:
  - Entender patrones en los datos de entrada
  - Reconocer similitudes entre los datos
- Elegir una buena métrica de distancia puede mejorar el funcionamiento de los algoritmos de **clasificación y clusterización**

- **Importancia:**
  - Saber distinguir cuál es el más apropiado

Distancias estadísticas: numéricos

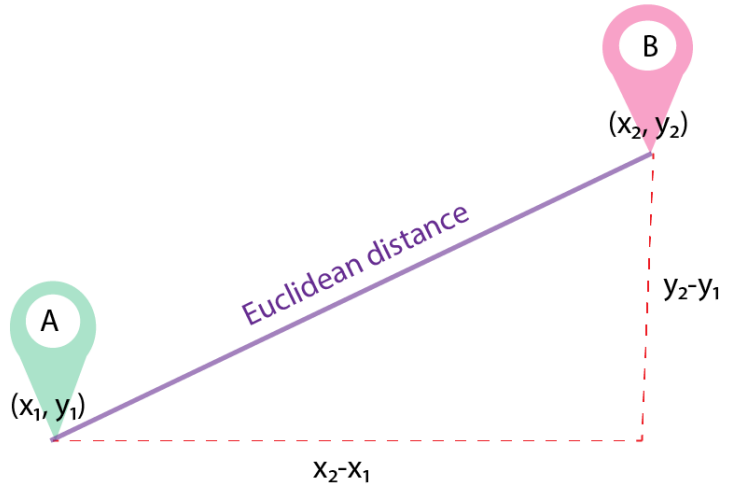
- Distancia euclidea: distancia mínima entre dos puntos

  Distancia entre dos puntos 2D

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

  Distancia entre dos puntos nD

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



- Distancia más usada en ML
  - ¿Es siempre la más adecuada?
  - Depende de los datos

Distancias estadísticas: numéricos

- *"The biggest problem in machine learning is the curse of dimensionality"*
  - *"Bellman"*
- Cada problema tiene su propia noción semántica de la similaridad, que normalmente se captura erroneamente con métricas estándares
- En altas dimensiones, lo que es intuitivo en 2 o 3 dimensiones, pierde sentido

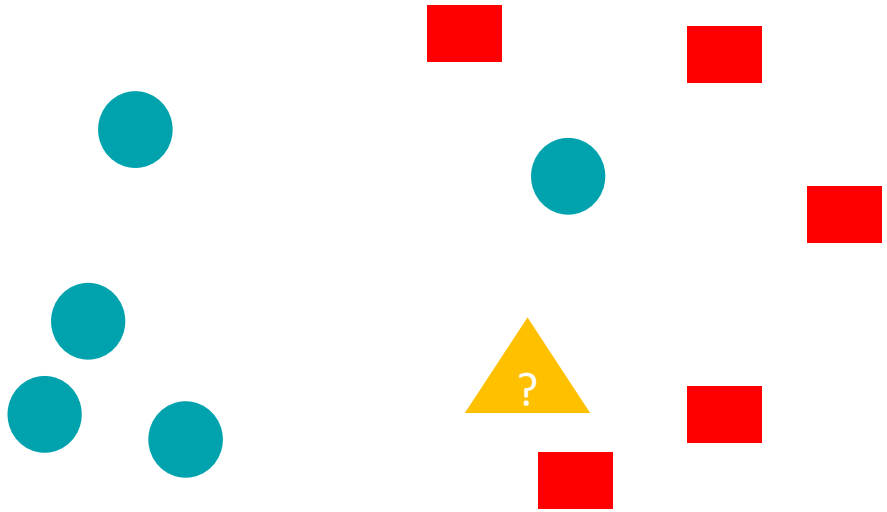
Distancias estadísticas: numéricos

Query

Imágenes similares

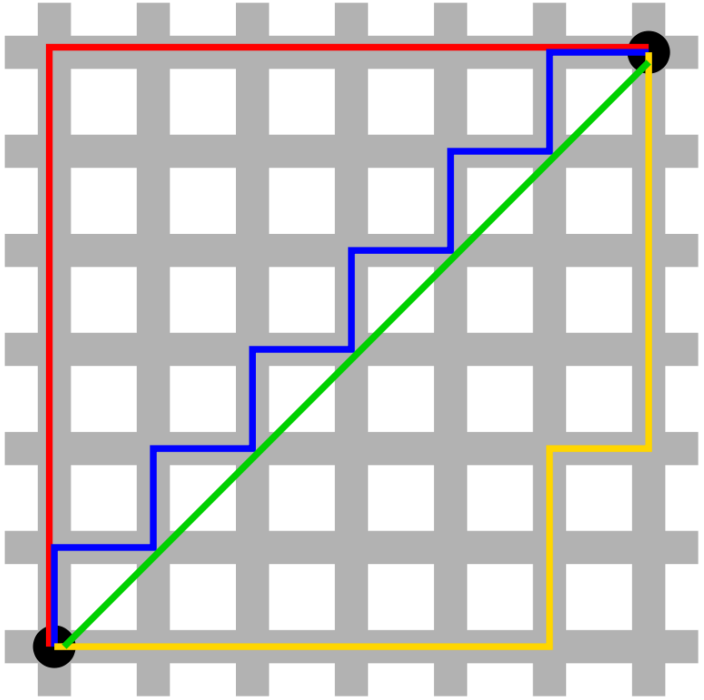
Distancias estadísticas: numéricos

- Distancia euclidea, desventajas:
  - Trata todas las dimensiones de la misma forma => espacio simétrico, esférico
  - Sensible cuando hay valores extremos en un atributo



Distancias estadísticas: numéricos

- Distancia de Manhattan:



$$\sum_{i=1}^{n} |x_i - y_i|$$

- Recomendable en altas dimensiones
- Ayuda a reducir el impacto de los outliers
  - Más robusto
- Dataset con atributos binarios
- Ref: https://bib.dbvis.de/uploadedFiles/155.pdf

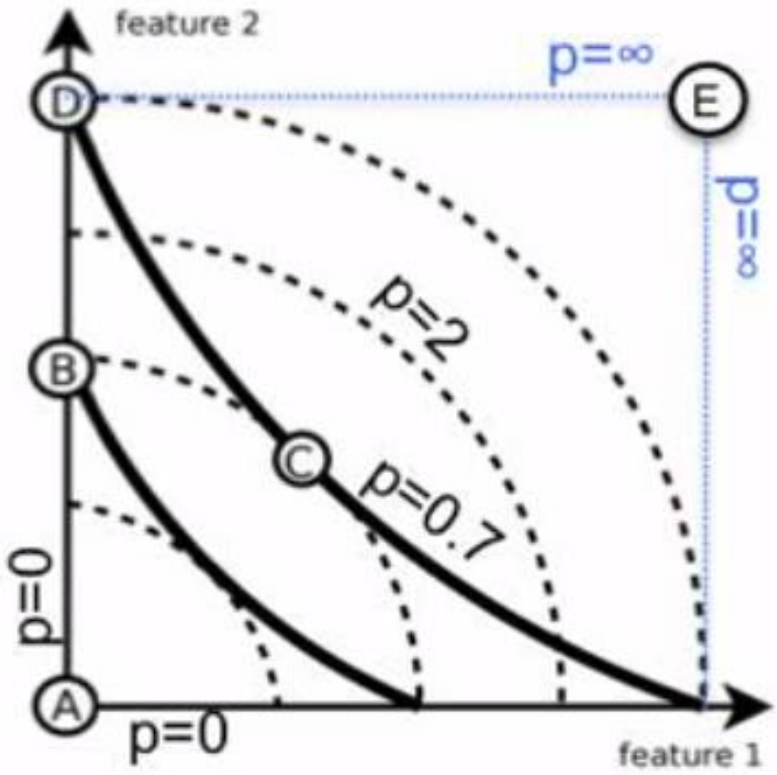
Distancias estadísticas: numéricos

- Distancia de Minkowski
  - Generalización de las distancias Euclidea y Manhattan

$$\sqrt[\lambda]{\sum_{i=1}^{n} |x_i - y_i|^{\lambda}}$$

  - $\lambda = 1$ => distancia de Manhattan
  - $\lambda = 2$ => distancia Euclidea
  - $\lambda = \infty$ => distancia Chebyshev

Distancias estadísticas: numéricos



---------- P = 2

——— P = 0.7

$$\sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

Distancias desde A:

- Cuando P = 2:
  - B y C están a la misma distancia que D
  - B difiere de A solo en el feature 1 …
- Cuando P = 0.7:
  - D y C están a la misma distancia

Distancias estadísticas: numéricos

- Distancia de Mahalanobis:
  - Se utiliza para calcular la distancia entre dos puntos en un espacio multivariable
  - Para la distancia Euclidea:
    - Tenemos un sistema de coordenadas cartesiano, lanzamos líneas perpendiculares
      - Calculamos distancias en base a ese sistema de coordenadas
    - ¿Qué pasa si nuestras variables están correlacionadas? Las distancias no se podrián medir con una línea recta

Distancias estadísticas: numéricos

Salario	m^2
75	19.6
52.8	20.8
64.8	17.2
43.2	20.4
84.0	17.6
49.2	17.6

Salario	m^2
75,000	19.6
52,800	20.8
64,800	17.2
43,200	20.4
84,000	17.6
49,200	17.6

Salario	m^2
75	19,600
52.8	20,800
64.8	17,200
43.2	20,400
84.0	17,600
49.2	17,600

- Problema 1: Tendrán las 3 tablas la misma distancia Euclidea?
  - Están afectadas por al escala
    - Estandarización

Distancias estadísticas: numéricos

Salario	m^2	Ahorro
75	20	27.9
52.8	21	23.3
64.8	17	28.6
43.2	20	19.3
84.0	18	34.8
49.2	18	23.0

- **Problema 2:** puede existir correlación entre las variables…
  - Si calculamos la distancia Euclidea, podríamos estar "sumando" el mismo impacto varias veces

- Mahalanobis nos ayuda a solucionar estos dos problemas

Distancias estadísticas: numéricos

- La distancia de Mahalanobis sigue el siguiente proceso:
  - Transforma las variables en variables no correlacionadas
  - Hace que su varianza sea = 1 (estandariza los datos)
  - Calcula la distancia Euclidea

$$D^{2=} \quad \sqrt{(x_A - x_B)^T * C^{-1} * (x_A - x_B)^T}$$

$D^{2=}$ Distancia de Mahalanobis
x = vector de datos
C = Matriz de covarianza

Distancias estadísticas: categóricos

- Distancia de Hamming
  - Medimos cuántos atributos hay que cambiar para que un valor sea igual que el valor al que se calcula la distancia
  - Ejemplo:

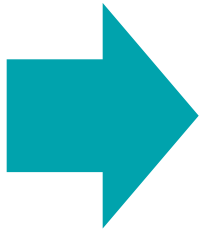
0	1	1	0	1	0	1	0
1	1	0	1	1	0	1	1

  - d(01101010,11011011) = 4
  - Aplicación más comun: similaridad entre textos

Distancias estadísticas: categóricos

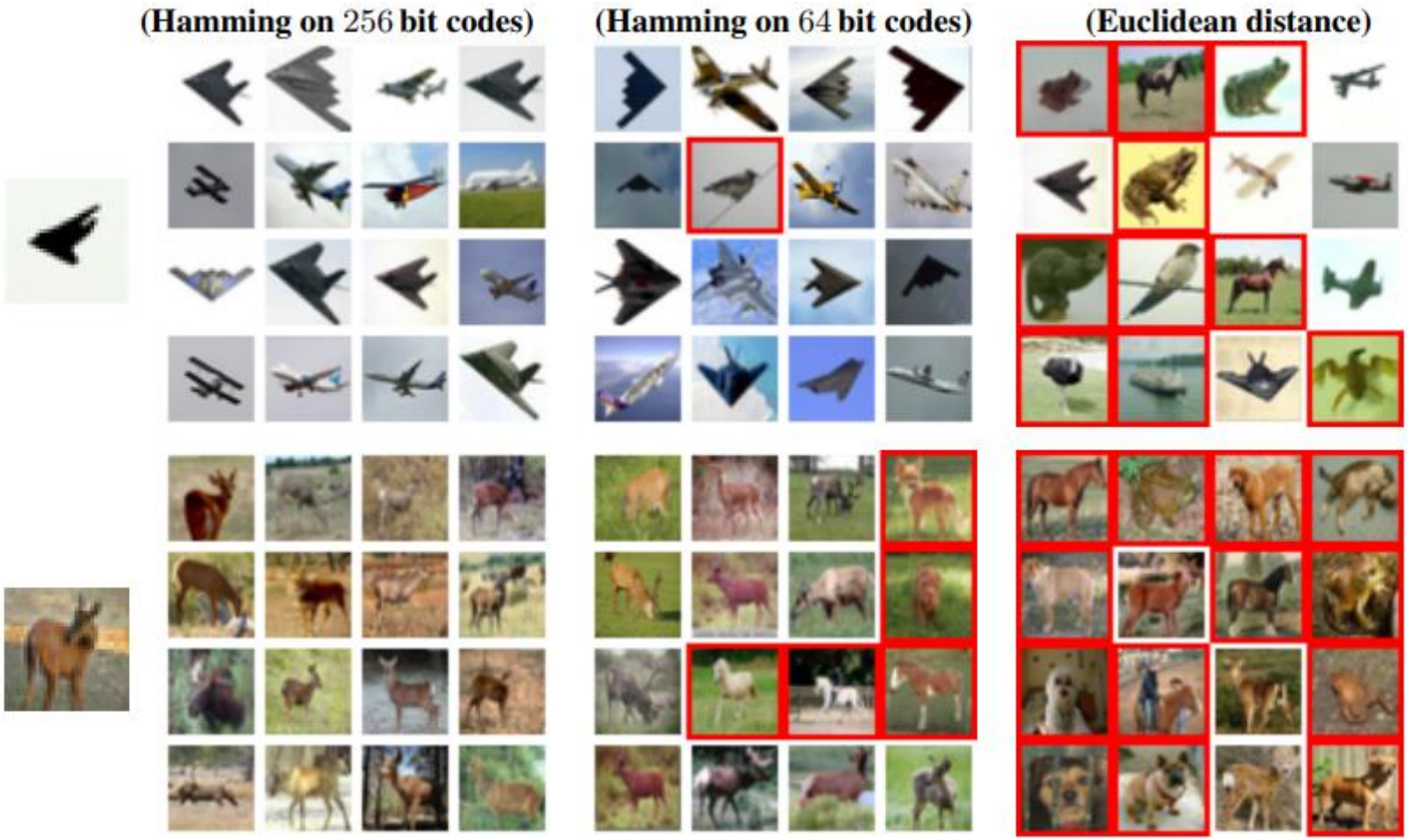
- Aunque tengamos valores numéricos, podemos transformar el dataset para utilizer la distancia de Hamming:

Máquina	Temperatura	Presión	Resultado
M1	100	80	Ok
M1	120	90	Nok
M1	123	80	Ok
M2	140	90	Ok
M2	132	91	Ok
M2	110	87	Nok
M3	110	89	Nok
M3	110	88	Ok
M3	100	79	Ok



M1	M2	M3	Temperatura	Presión	Resultado
1	0	0	100	80	Ok
1	0	0	120	90	Nok
1	0	0	123	80	Ok
0	1	0	140	90	Ok
0	1	0	132	91	Ok
0	1	0	110	87	Nok
0	0	1	110	89	Nok
0	0	1	110	88	Ok
0	0	1	100	79	Ok

Distancias estadísticas: categóricos



*Los rectangulos rojos indican errores

https://papers.nips.cc/paper/4808-hamming-distance-metric-learning.pdf

Distancias estadísticas: mixtos

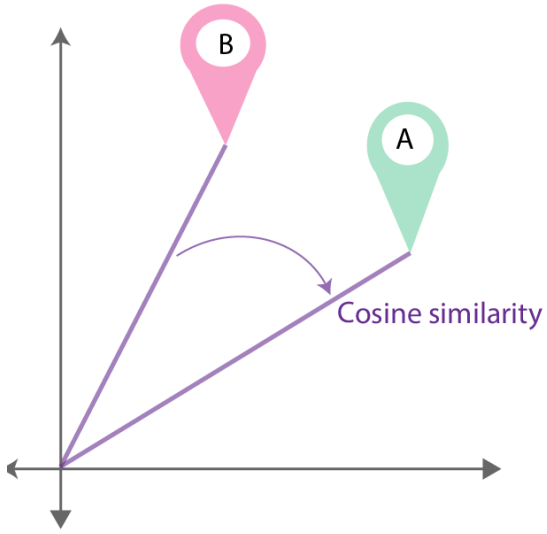
- Similaridad del coseno:
  - Normalmente, esta métrica se utiliza para buscar similitudes entre distintos documentos
  - Medimos el ángulo entre distintos documentos / vectores
  - Cuanto menor sea el ángulo, más se parecerán los dos vectores
    - Valor máximo de similitud => 1
    - Vectores ortogonales => 0
    - Vectores sin similitud (direcciones opuestas) => -1

Cos 0′ = 1 Cos 90′ = 0

Cos 180′ = -1

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$



png src: towardsdatascience.com

Distancias estadísticas: mixtos

- Ejemplo: ¿cuánto se parecen estos textos?
  - Julia me quiere más de lo que me quiere Julio
  - Jacinto me gusta más de lo que me quiere Julia
  1. Creamos una lista de palabras que se utilizan
     1. Julia me quiere más de los que Julio Jacinto gusta
     2. Creamos dos vectores:

Julia	1	1
Me	2	2
Quiere	2	1
Más	1	1
De	1	1
Lo	1	1
Que	1	1
Julio	1	0
Jacinto	0	1
gusta	0	1

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

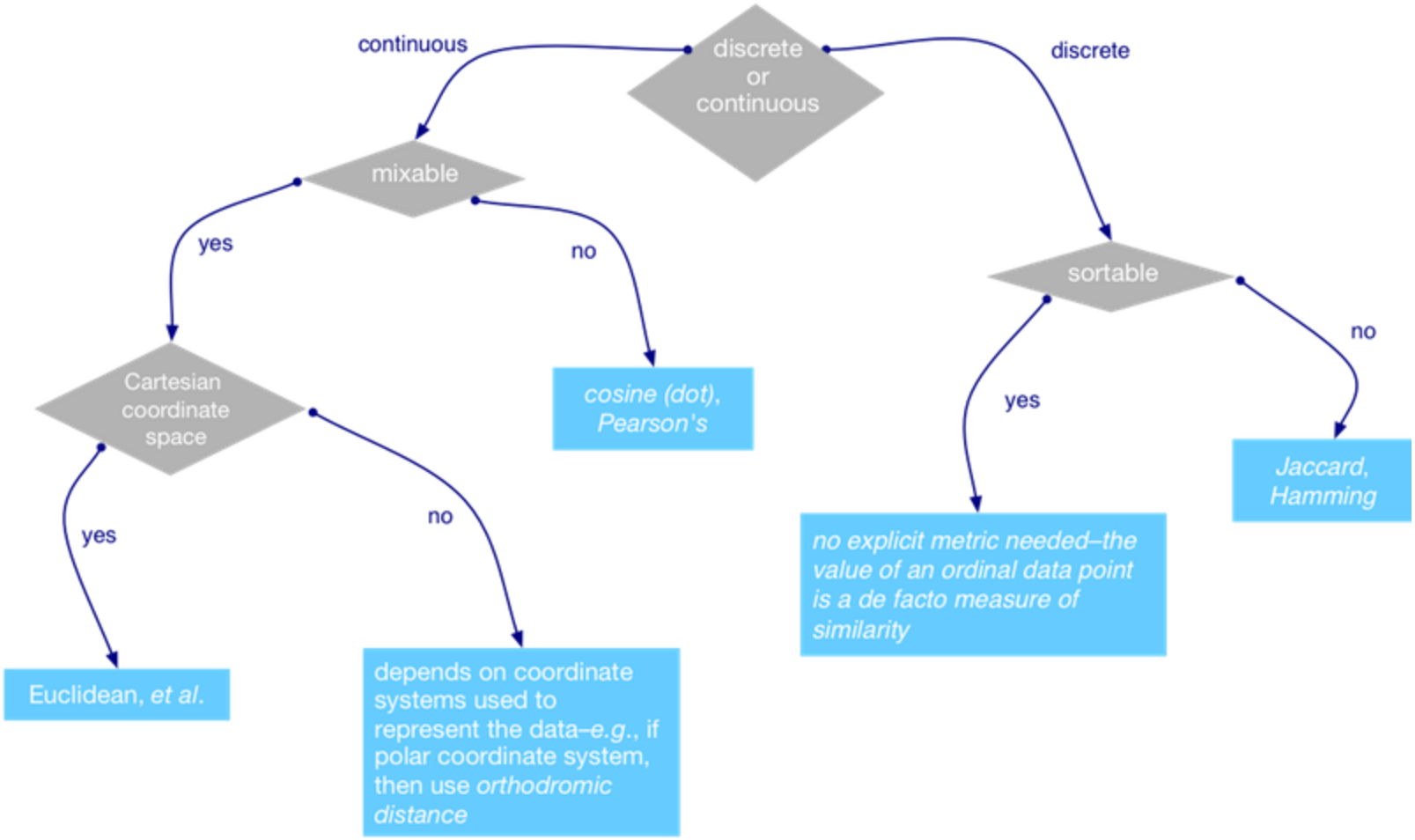
0.848

```
import numpy as np
from scipy import spatial

a = np.array([1,2,2,1,1,1,1,1,0,0])
b = np.array([1,2,1,1,1,1,1,0,1,1])

result = 1 - spatial.distance.cosine(a,b)
```

Distancias estadísticas



src.: https://github.com/alexland/similarity-metrics

Mondragon Unibertsitatea
Goi Eskola Politeknikoa

Aitor Agirre
aaguirre@mondragon.edu