DEPARTAMENTO DE GEOFÍSICA
UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Definiciones Básicas de Probabilidades

Septiembre de 2018

Profesor Francisco Hernán Ortega Culaciati

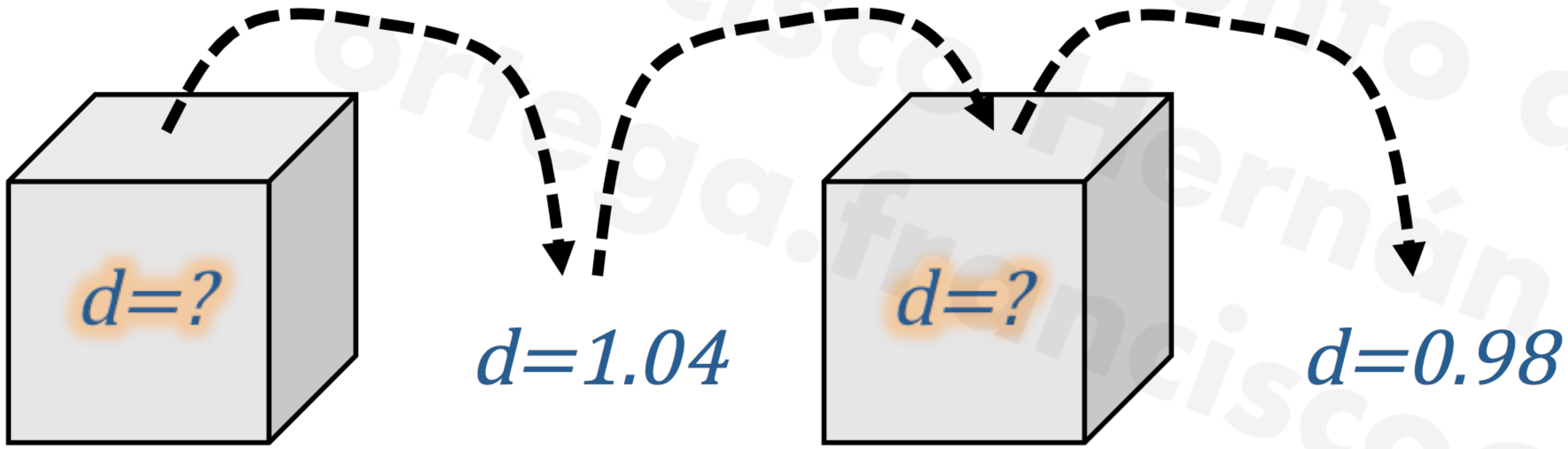
Repaso de Probabilidades

Variables aleatorias y sus FDP

FDP = Funciones de Densidad de Probabilidad
(PDF = Probability Density Functions)

Variables aleatorias y sus FDP

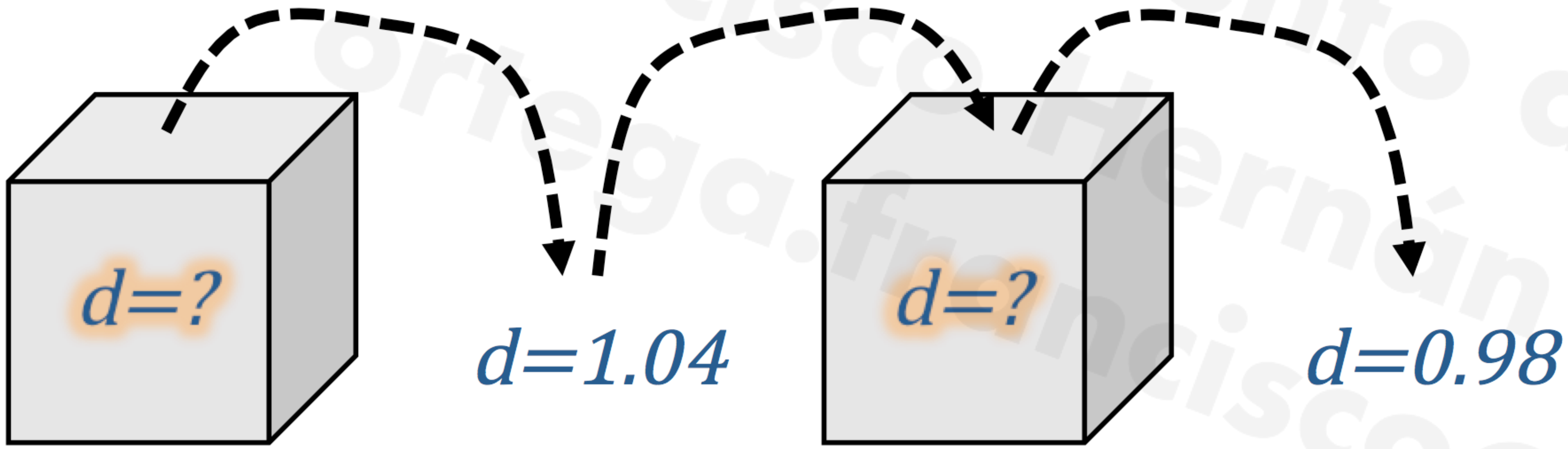
Variable aleatoria (random variable), d



No tiene un valor fijo hasta que se ha hecho una realización, o se ha "realizado" (realized) un valor de esta.

Variables aleatorias y sus FDP

Variable aleatoria (random variable), d



No tiene un valor fijo hasta que se ha hecho una realización, o se ha “realizado” (realized) un valor de esta.

Por lo general hay una tendencia a realizar algunos valores más que otros

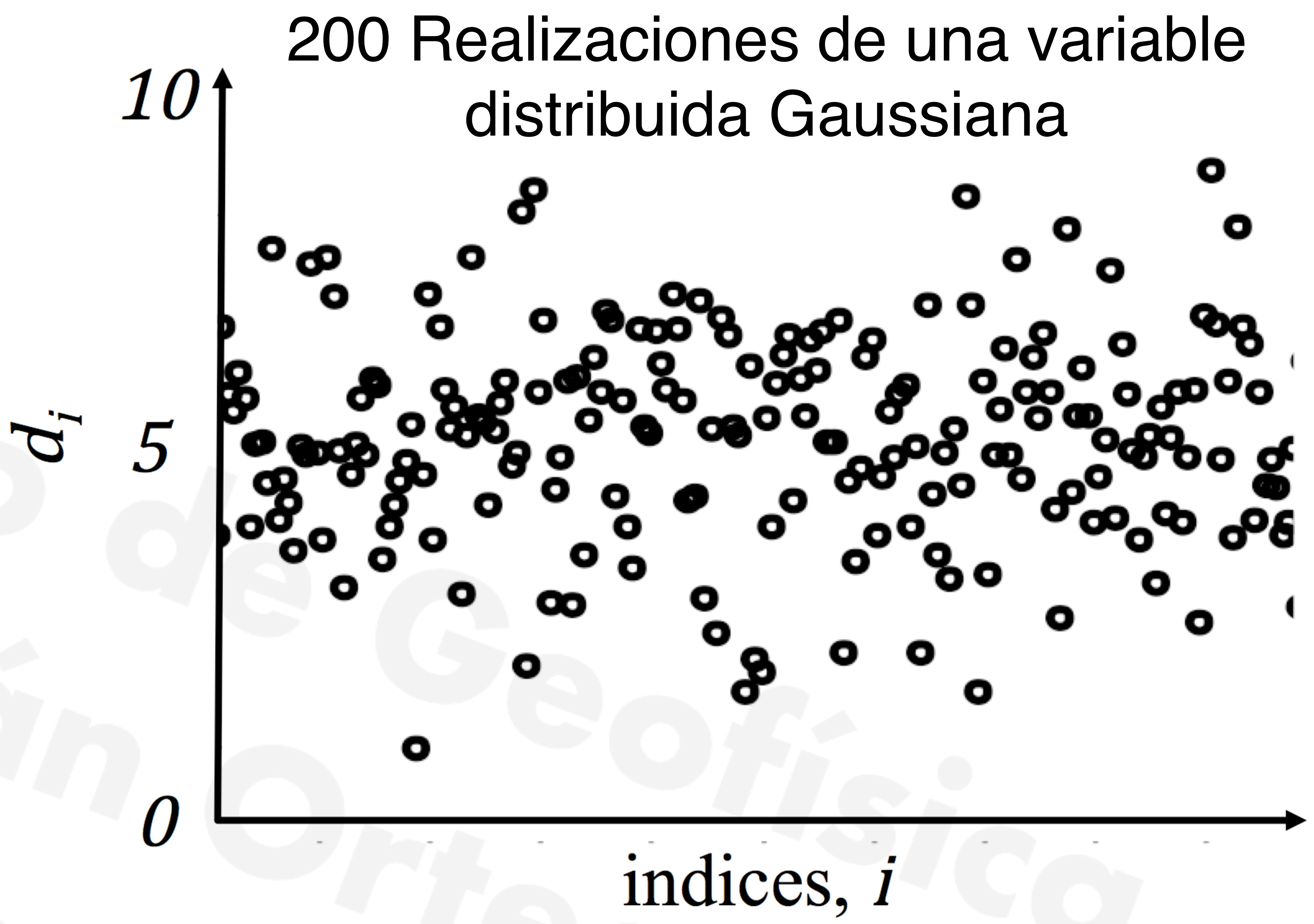
Esto depende de la función de densidad de probabilidad (fdp) asociada a la variable aleatoria

Variables aleatorias y sus FDP

Variable aleatoria (random variable), d

Por lo general hay una tendencia a realizar algunos valores más que otros

Esto depende de la función de densidad de probabilidad (fdp) asociada a la variable aleatoria

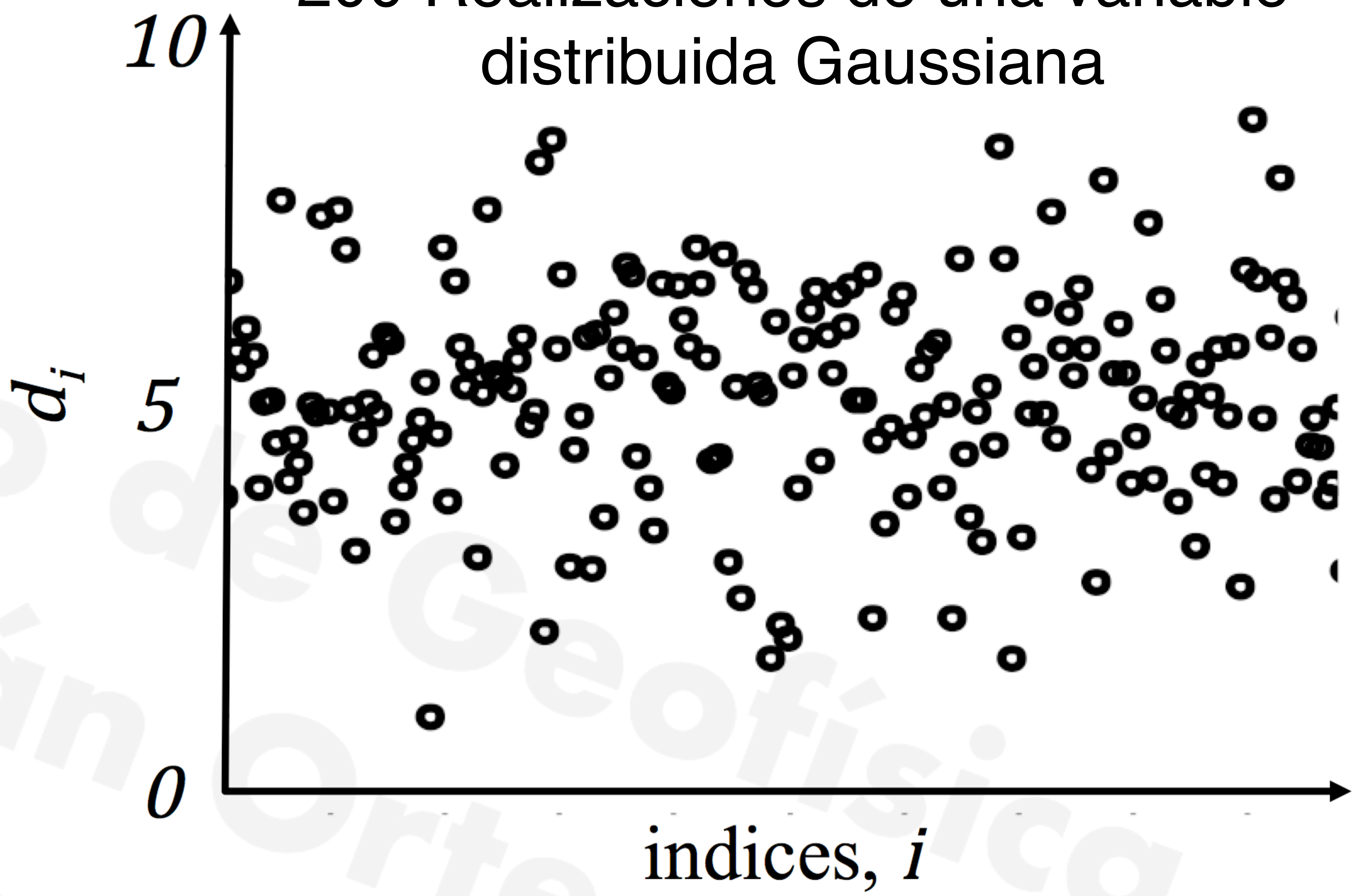


Variables aleatorias y sus FDP

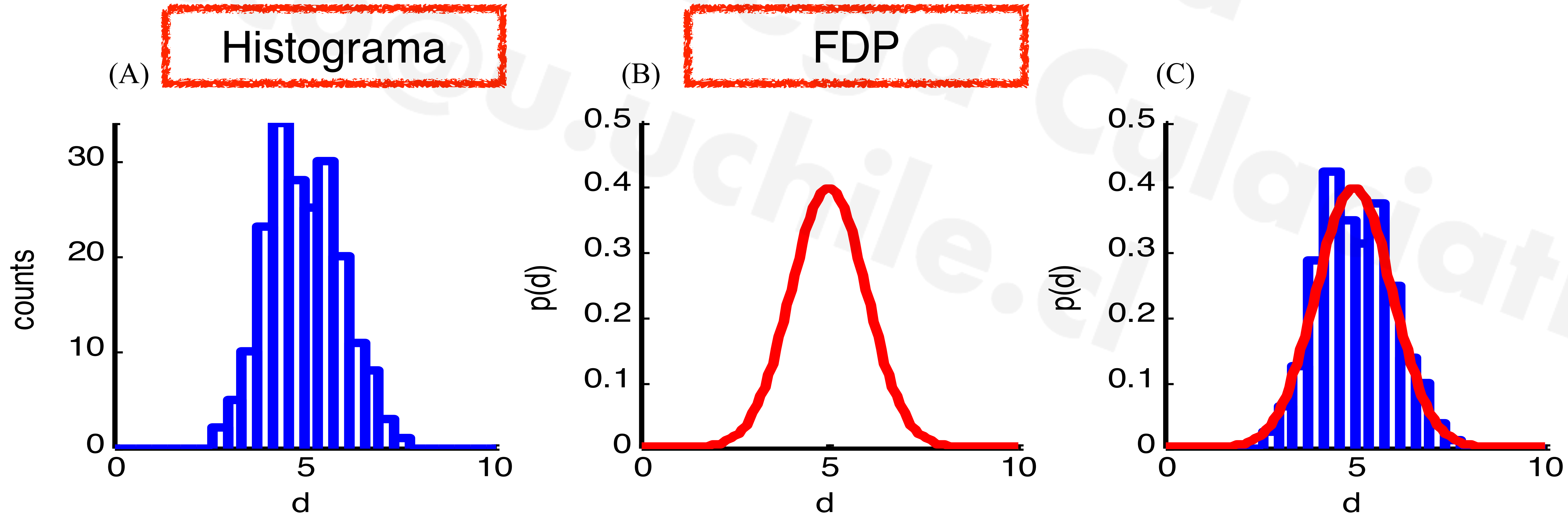
Variable aleatoria (random variable), d

Por lo general hay una tendencia a realizar algunos valores más que otros

200 Realizaciones de una variable distribuida Gaussiana



Esto depende de la función de densidad de probabilidad (fdp) asociada a la variable aleatoria

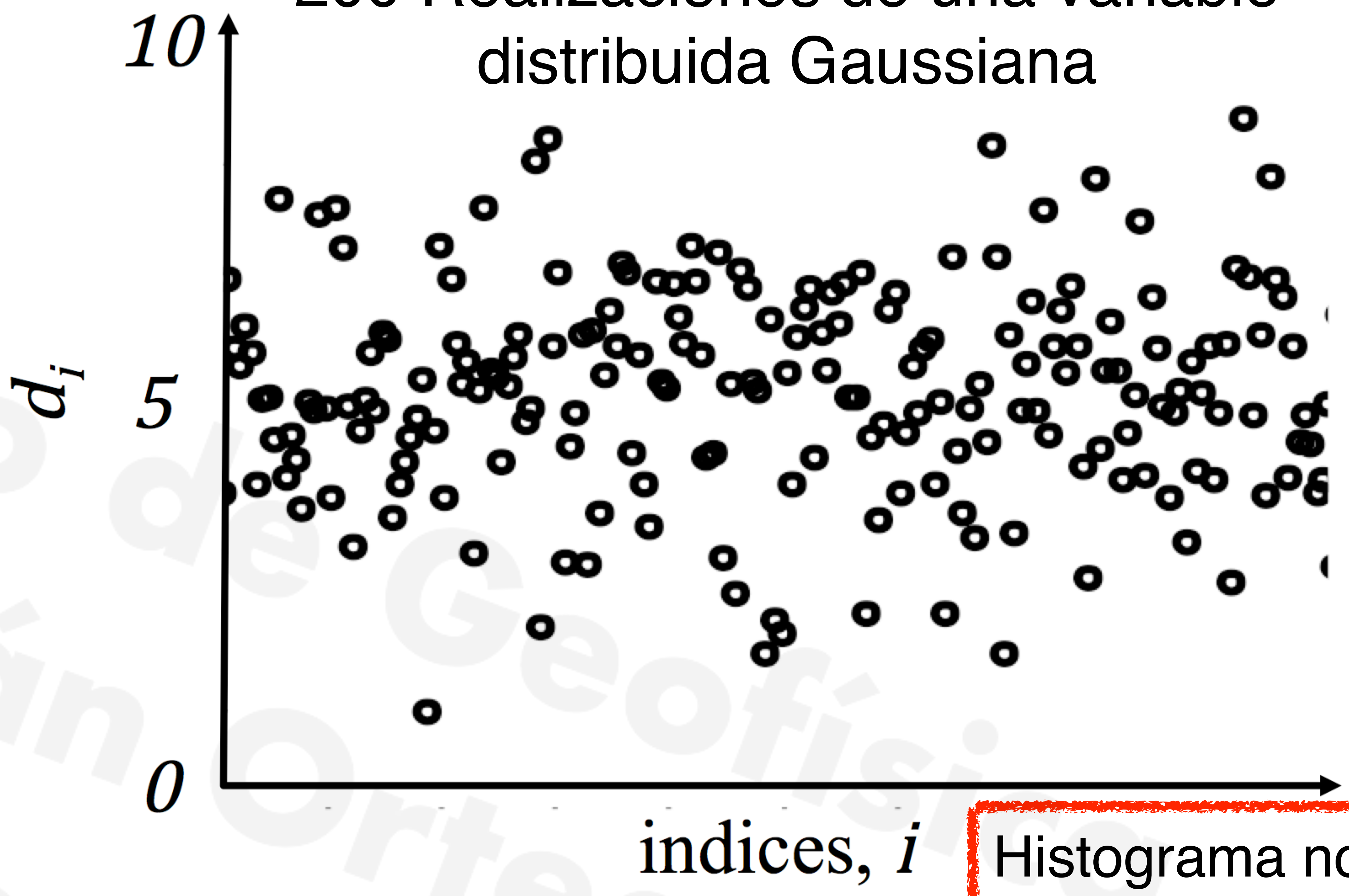


Variables aleatorias y sus FDP

Variable aleatoria (random variable), d

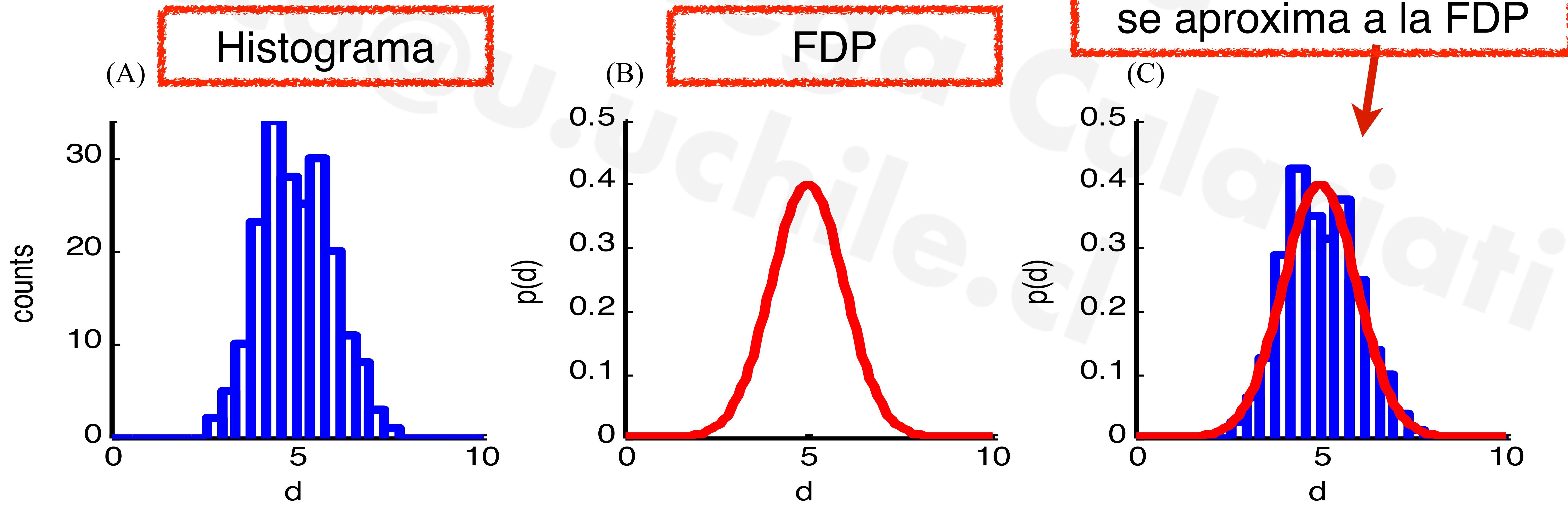
Por lo general hay una tendencia a realizar algunos valores más que otros

200 Realizaciones de una variable distribuida Gaussiana



Esto depende de la función de densidad de probabilidad (fdp) asociada a la variable aleatoria

Histograma normalizado se aproxima a la FDP



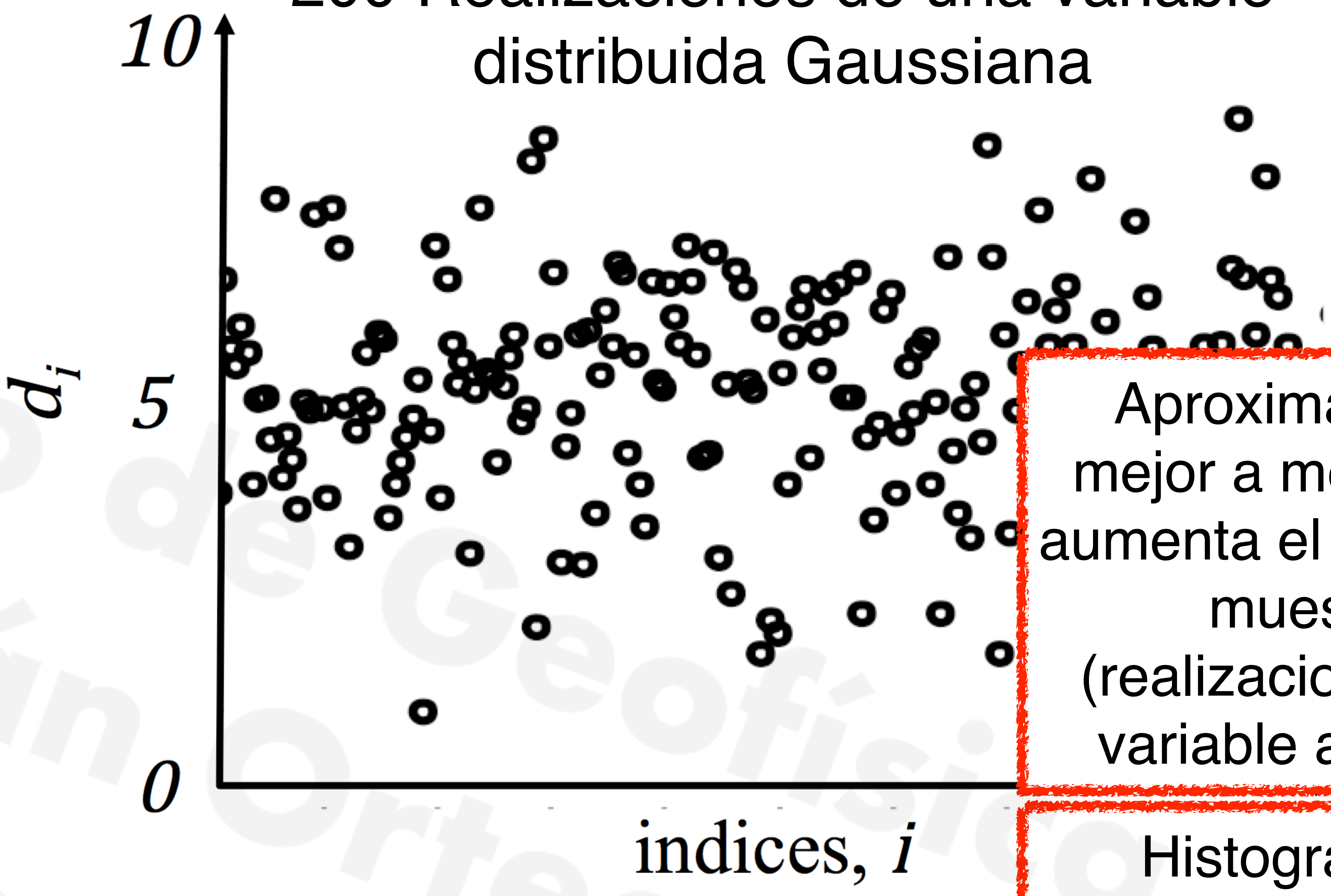
Variables aleatorias y sus FDP

Variable aleatoria (random variable), d

Por lo general hay una tendencia a realizar algunos valores más que otros

Esto depende de la función de densidad de probabilidad (fdp) asociada a la variable aleatoria

200 Realizaciones de una variable distribuida Gaussiana

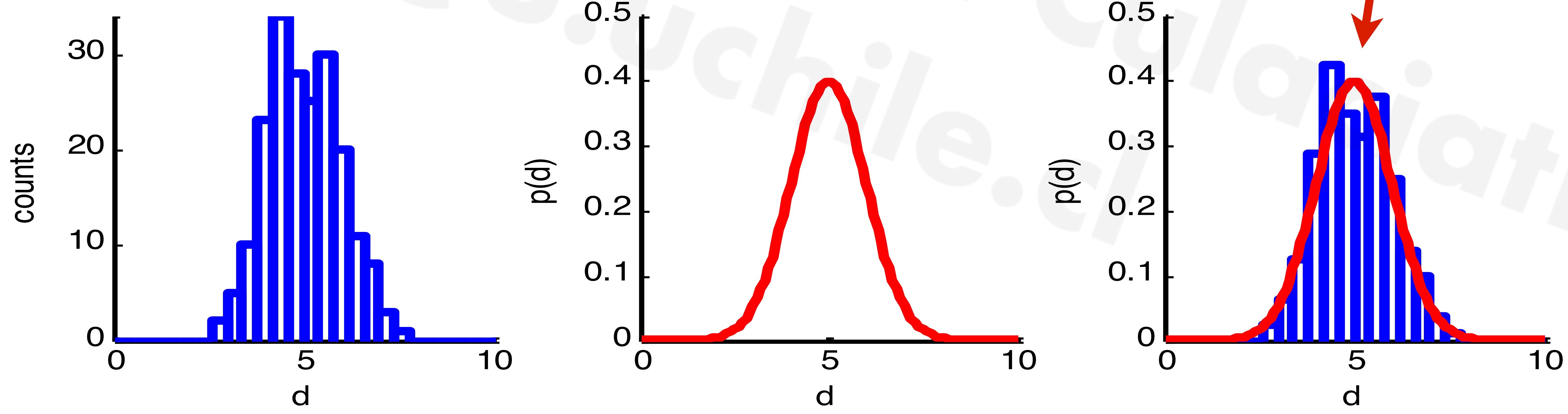


Aproximación es mejor a medida que aumenta el número de muestras (realizaciones de la variable aleatoria)

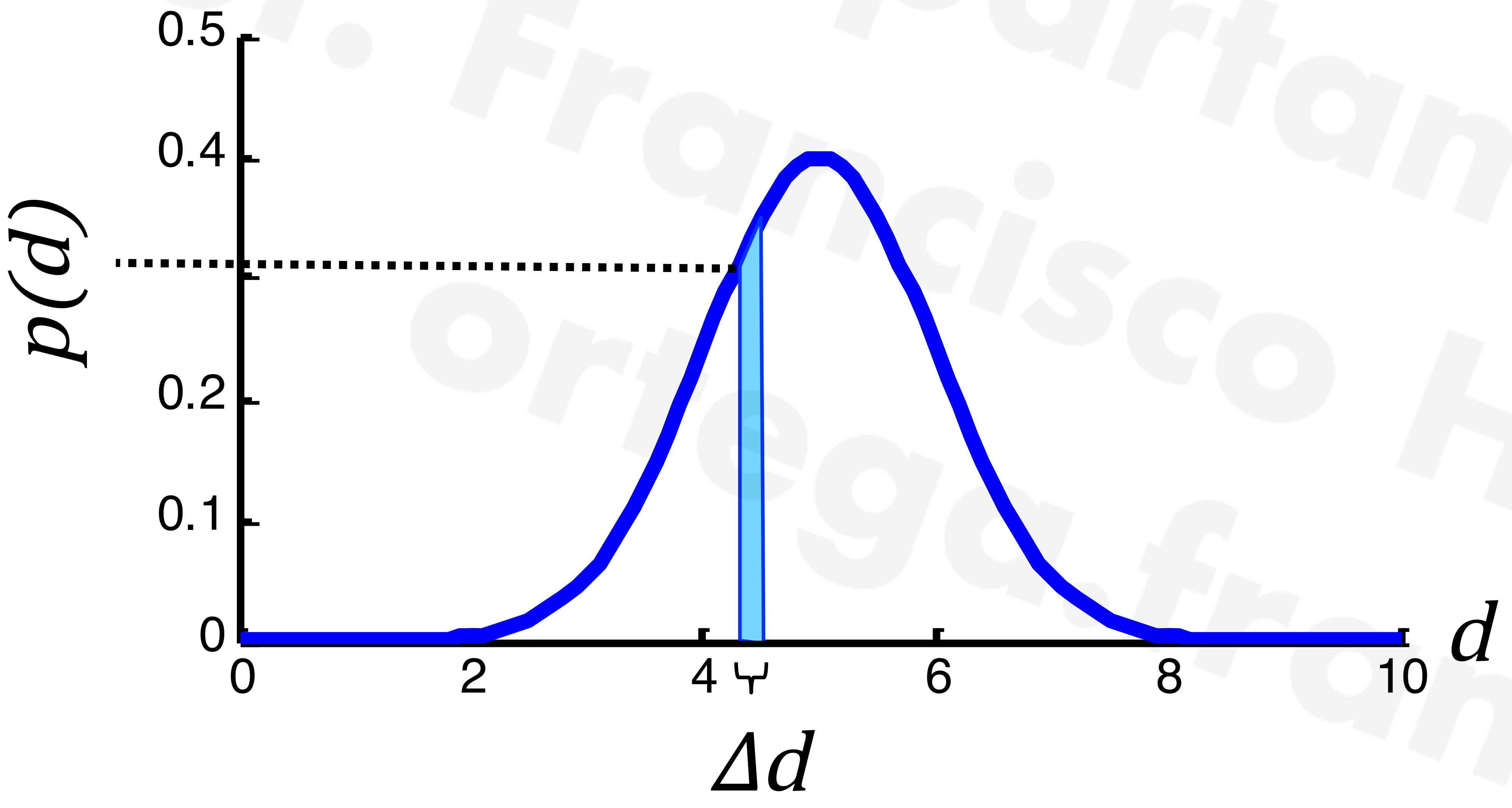
Histograma se aproxima a la FDP

Histograma

FDP



¿Cómo calculamos probabilidades?



La probabilidad de la variable de pertenecer al intervalo

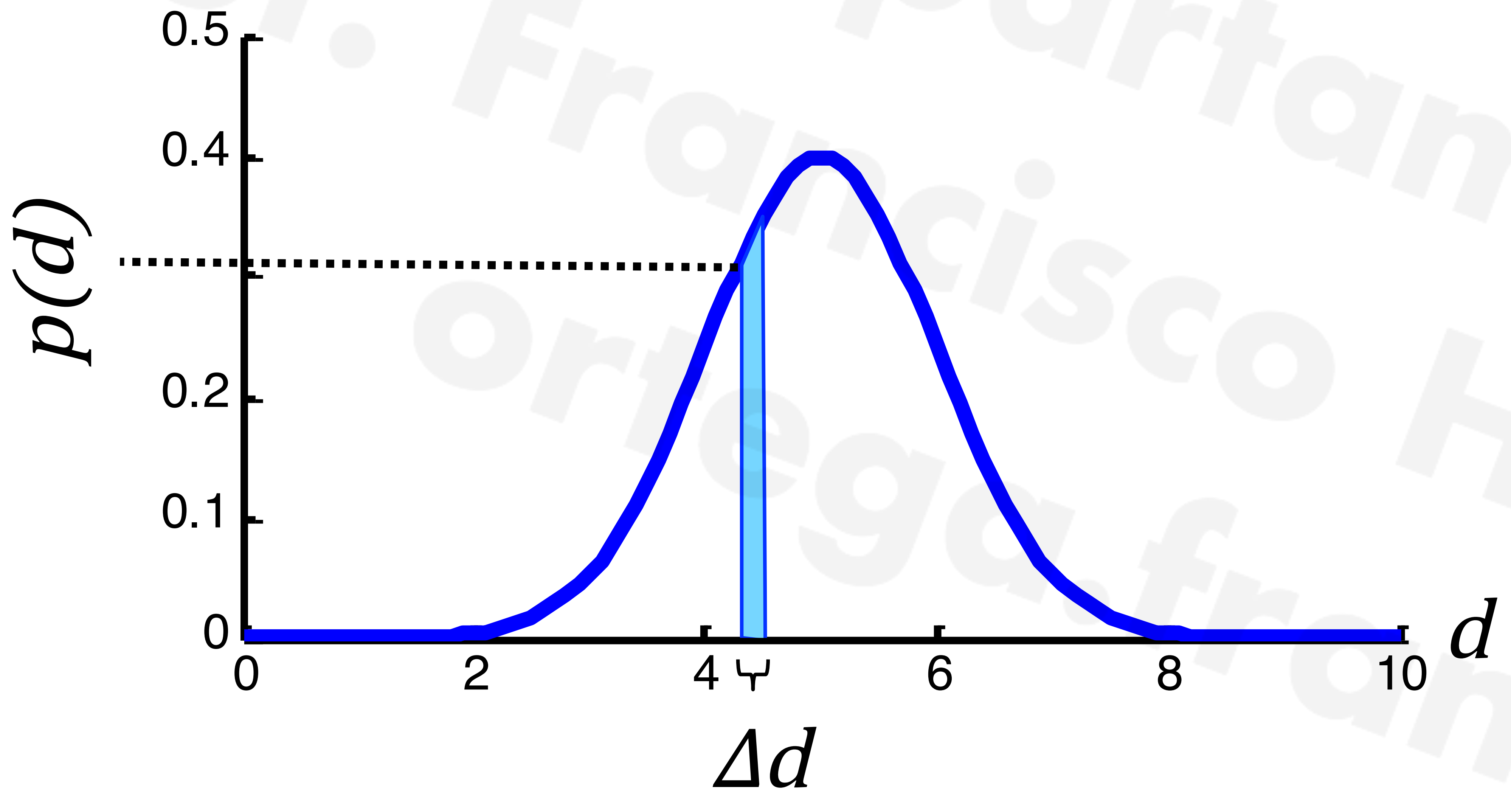
$$[d, d + \Delta d]$$

es

$$p(d)\Delta d$$

Variables aleatorias y sus FDP

¿Cómo calculamos probabilidades?



La probabilidad de la variable de pertenecer al intervalo

$$[d, d + \Delta d]$$

es

$$p(d)\Delta d$$

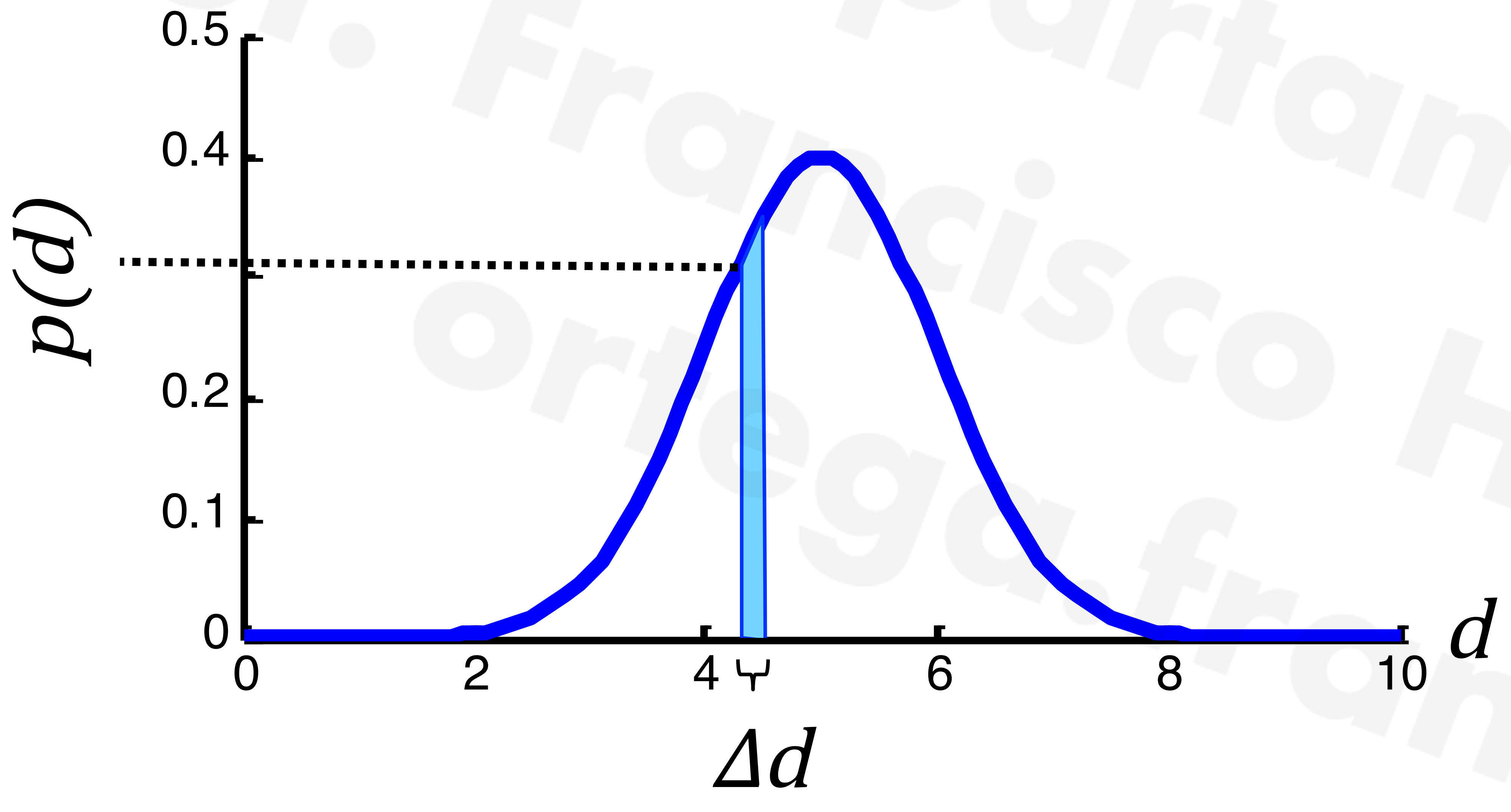
En general la probabilidad de un evento se define por la integral

$$P(d_1, d_2) = \int_{d_1}^{d_2} p(d)\, \mathrm{d}d$$

Que es la probabilidad de que "d" se encuentre en el intervalo [d_1, d_2]

Variables aleatorias y sus FDP

¿Cómo calculamos probabilidades?



La probabilidad de la variable de pertenecer al intervalo

$$[d, d + \Delta d]$$

es

$$p(d)\Delta d$$

En general la probabilidad de un evento se define por la integral

$$P(d_1, d_2) = \int_{d_1}^{d_2} p(d)\, \mathrm{d}d$$

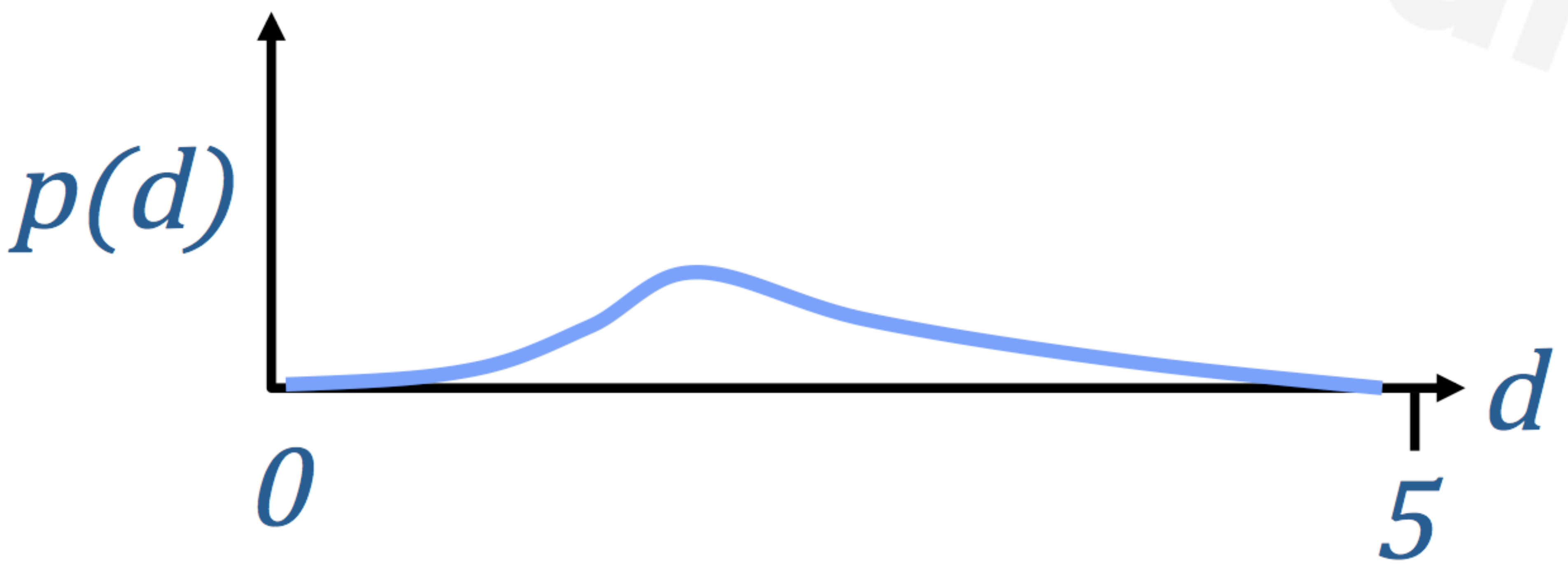
Y se exige una condición de normalización

$$P(d_{min}, d_{max}) = \int_{d_{min}}^{d_{max}} p(d)\, \mathrm{d}d = 1$$

Si no hay información adicional, por lo general se asume $d_{min} = -\infty, d_{max} = +\infty$

Variables aleatorias y sus FDP

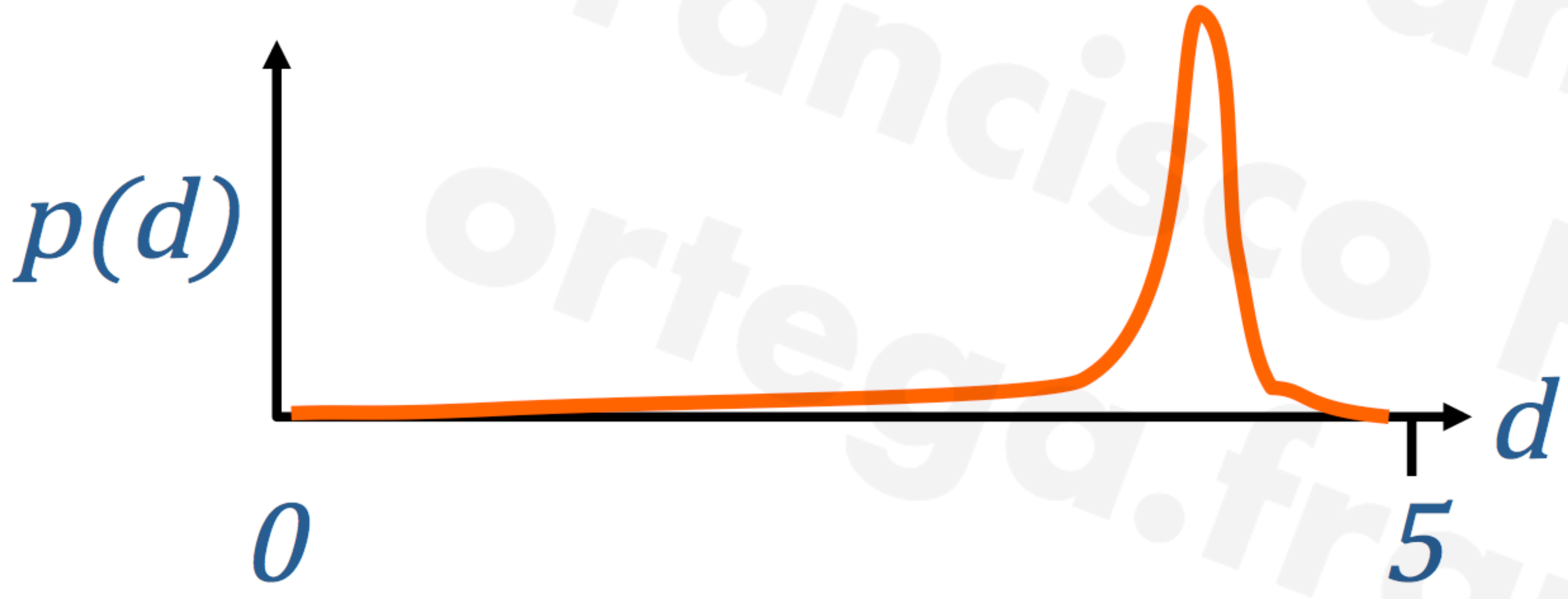
¿Cómo comparamos 2 FDP ?

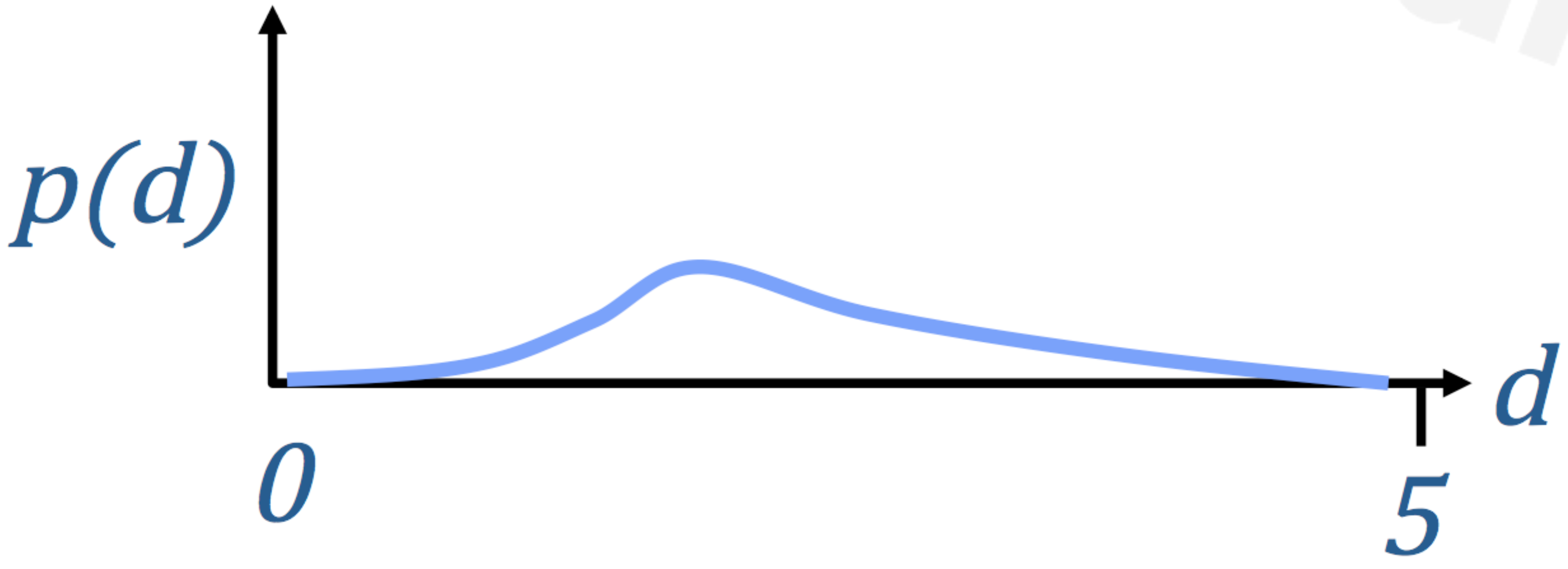


Podemos calcular valores estándar de la FDP (Ej: los momentos de la FDP)

Variables aleatorias y sus FDP

¿Cómo comparamos 2 FDP ?





Podemos calcular valores estándar de la FDP, muchas posibilidades
(Ej: los momentos de la FDP)

Ejemplos de "medidas" para comparar:

- El valor típico :
  "el centro de la FDP"

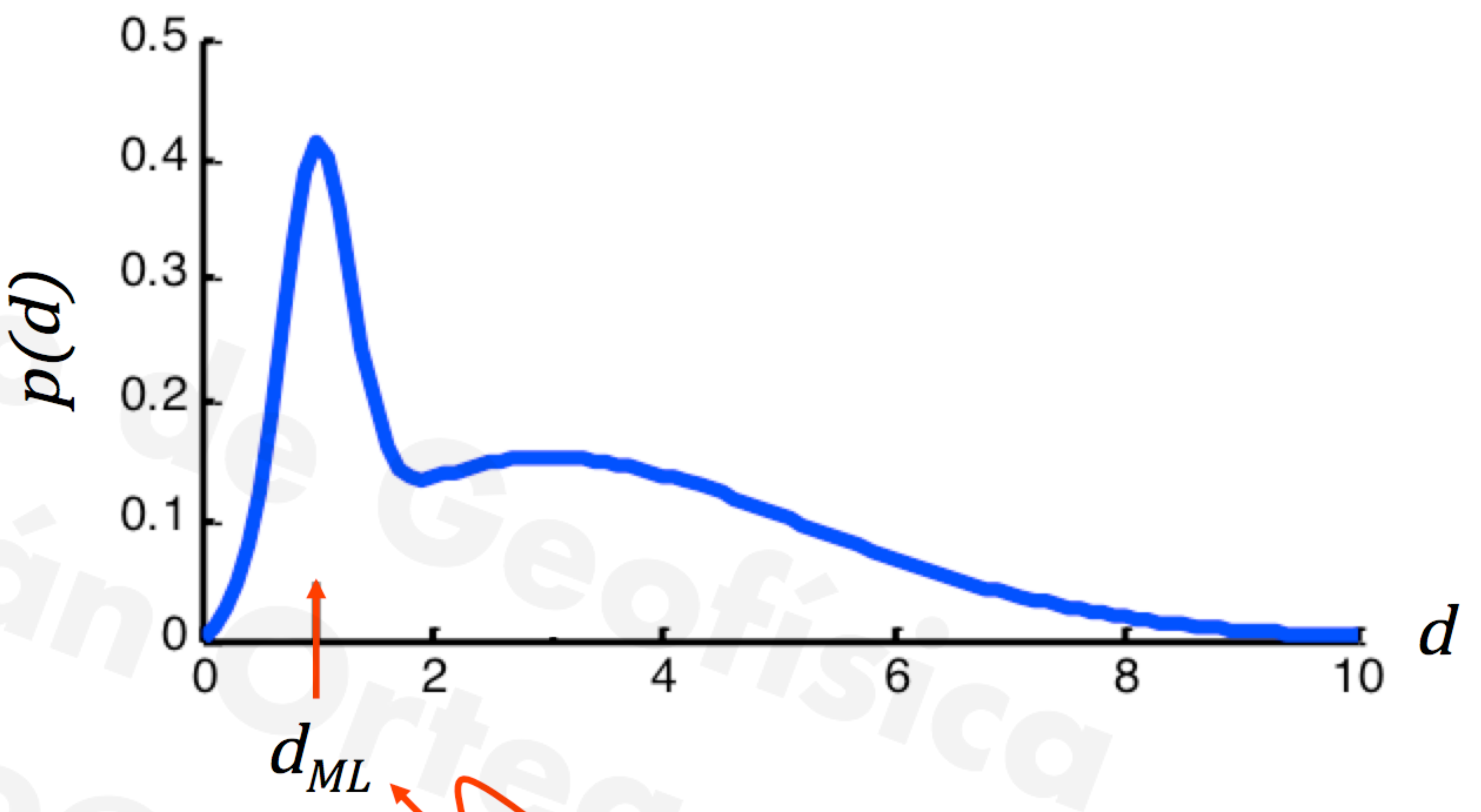
- Una medida de dispersión alrededor de un valor tipico
  "El 'ancho' de la FDP"

Variables aleatorias y sus FDP

¿Cómo comparamos 2 FDP ?

Ejemplos de "medidas" para comparar:

- El valor típico :
  "el centro de la FDP"

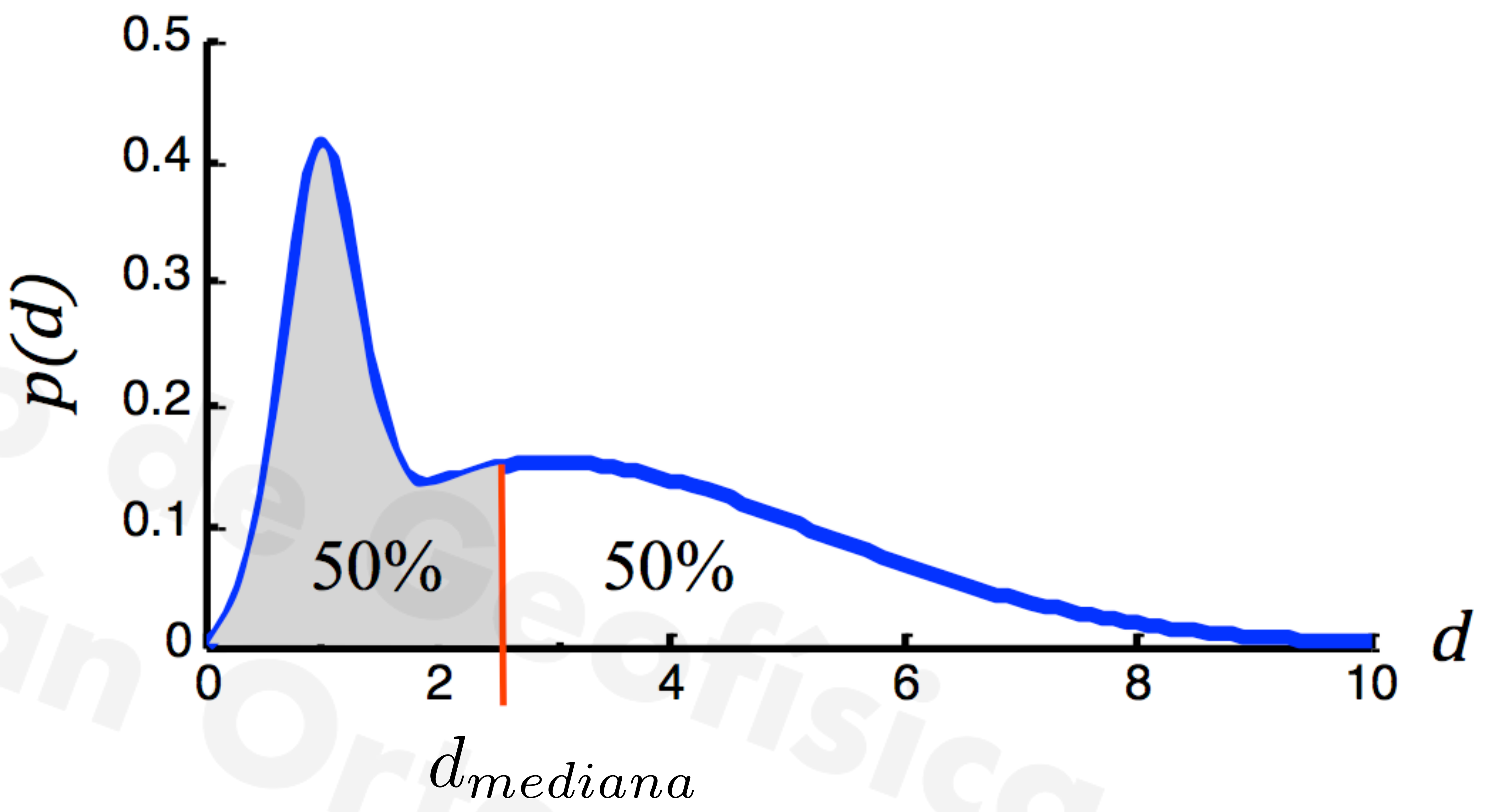


punto bajo el máximo de la FDP
(moda de la variable aleatoria) o
"punto de máxima verosimilitud"
(maximum likelihood point)

¿Cómo comparamos 2 FDP ?

Ejemplos de “medidas” para comparar:

- El valor típico :
  “el centro de la FDP”



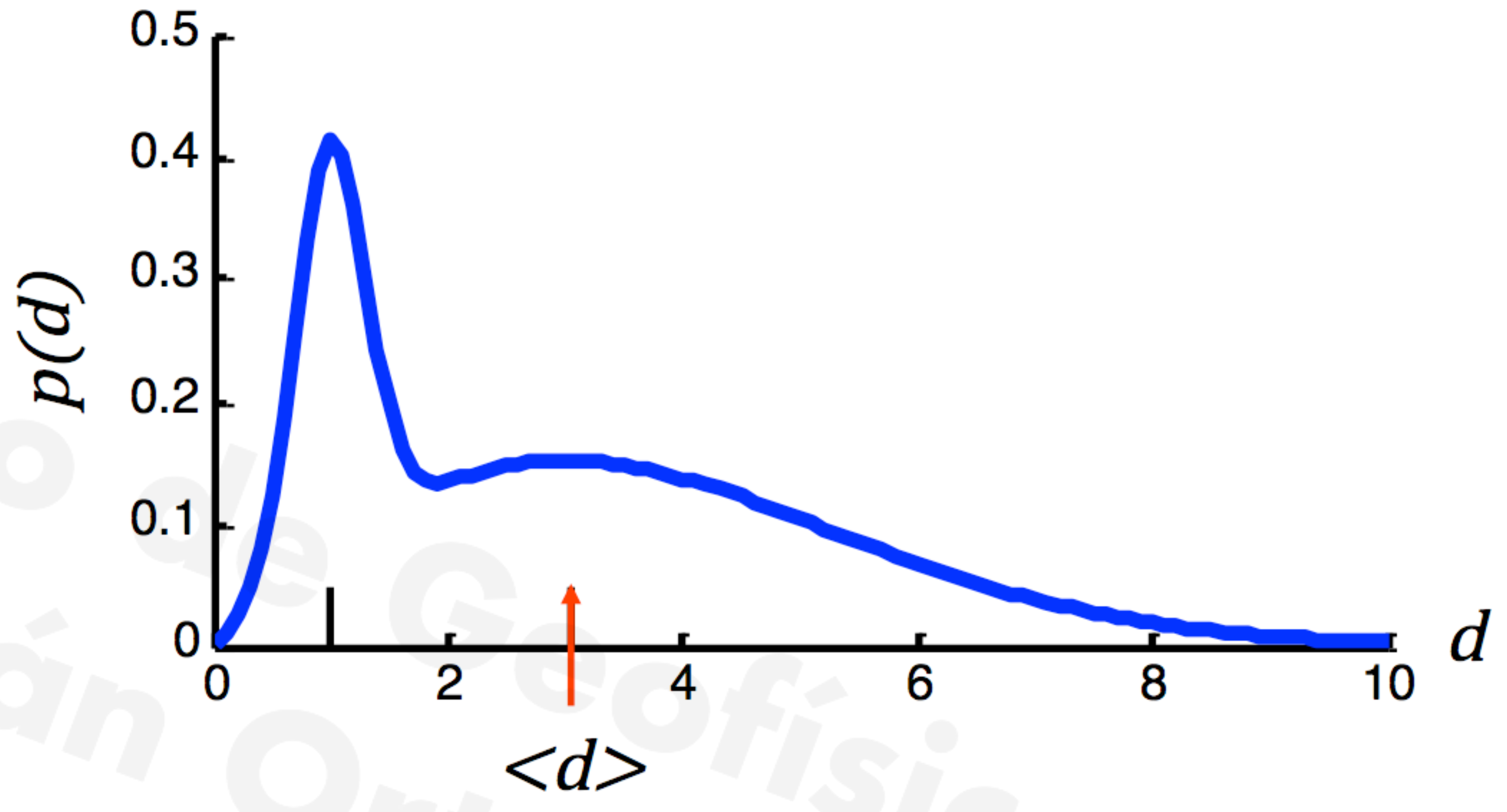
mediana (median): punto dividiendo el área en dos regiones con igual probabilidad (50%, mitad de la probabilidad)

Variables aleatorias y sus FDP

¿Cómo comparamos 2 FDP ?

Ejemplos de “medidas” para comparar:

- El valor típico :
  “el centro de la FDP”



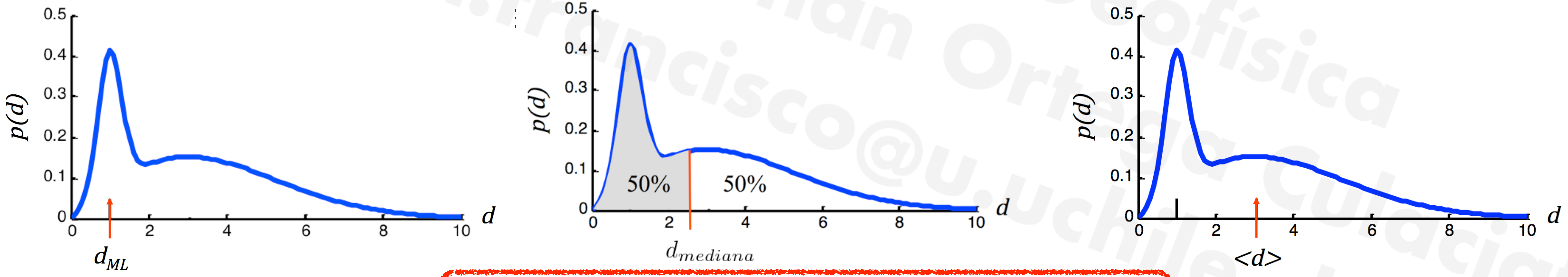
valor esperado (expectation) o valor promedio (mean)

Variables aleatorias y sus FDP

¿Cómo comparamos 2 FDP ?

Ejemplos de “medidas” para comparar:

- El valor típico :
  “el centro de la FDP”



$$d_{ML} \neq d_{mediana} \neq < d >$$

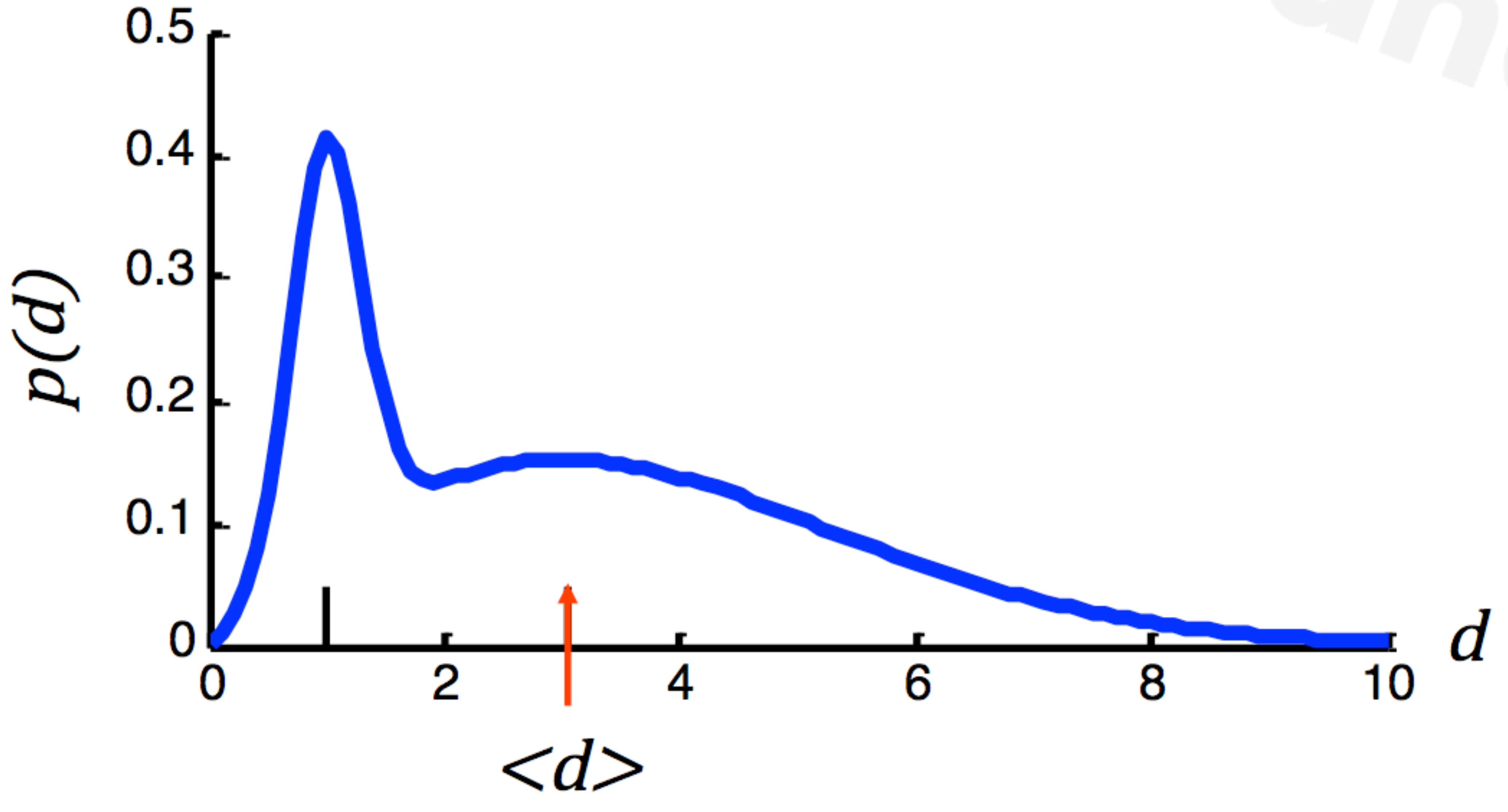
ESTOS 3 VALORES PUEDEN SER DIFERENTES !!!

Variables aleatorias y sus FDP

¿Cómo comparamos 2 FDP ?

Ejemplos de “medidas” para comparar:

- El valor típico :
  “el centro de la FDP”



Para el caso general, el valor esperado o promedio tiene una formula cerrada

En el caso continuo (**d**), el valor esperado o promedio se calcula como la siguiente integral

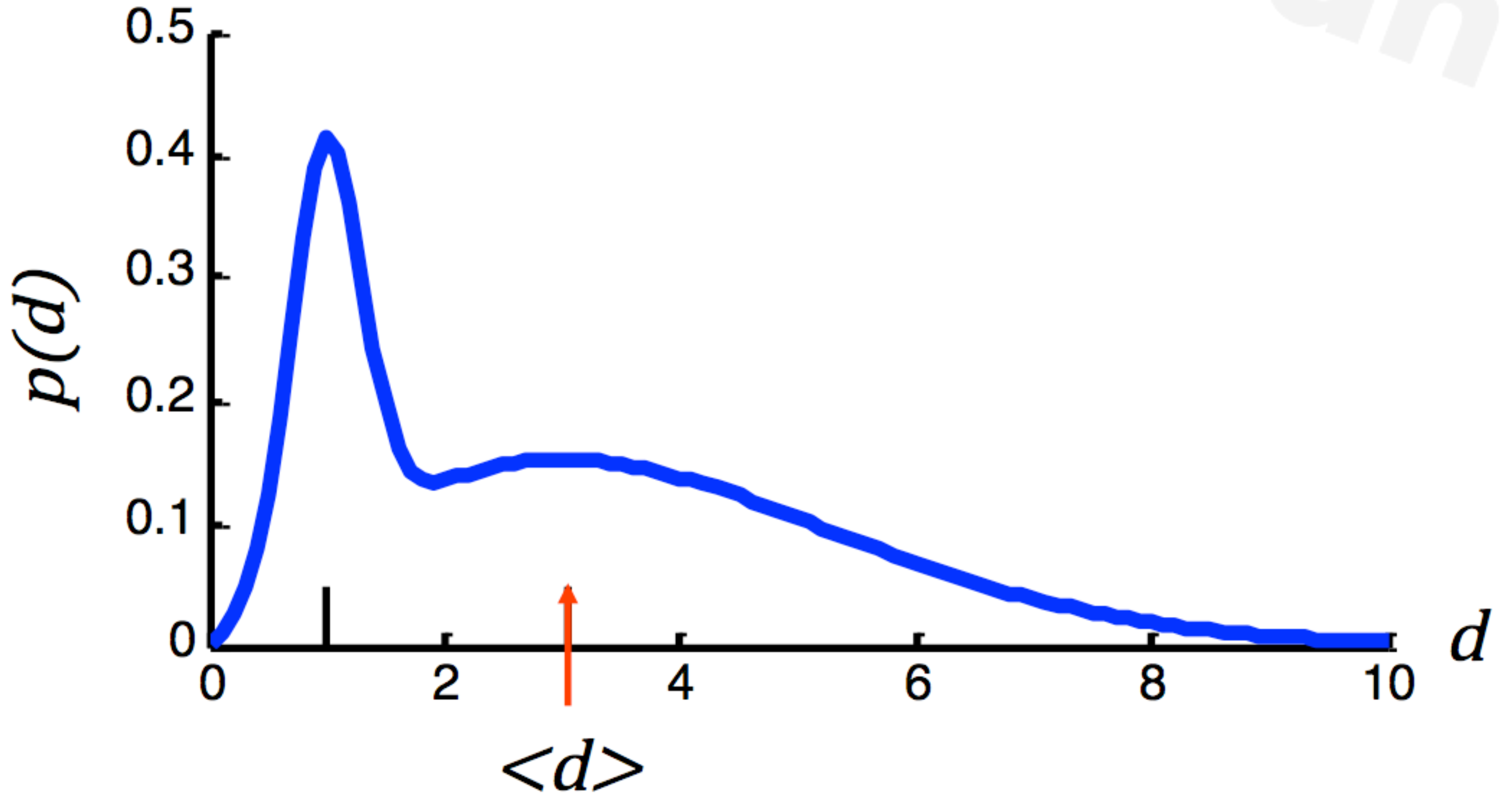
$$<d> = \int_{d_{min}}^{d_{max}} d\, p(d)\, \mathrm{d}d$$

Variables aleatorias y sus FDP

¿Cómo comparamos 2 FDP ?

Ejemplos de "medidas" para comparar:

- El valor típico :
  "el centro de la FDP"



Para el caso general, el valor esperado o promedio tiene una formula cerrada

Para el caso discreto (N realizaciones de **d**) un histograma normalizado aproxima la FDP

Fórmula para el promedio

$$<d> = \frac{1}{N}\sum_{i=0}^{N} d_i$$

Si se calcula el histograma de los valores de **d,** se puede aproximar <d> usando el valor de la muestra en el centro de cada celda del histograma y el número de realizaciones de **d** que se ubican en cada celda (las cuentas del histograma, i.e., cuantos valores realizados de d caen en cada celda).

$$<d> \approx \frac{1}{N}\sum_{s=0}^{M} d^{(s)} N_s$$

Se puede aproximar la FDP contínua de la variable d por una FDP discreta formada normalizando el histograma por el número total de realizaciones

$$<d> \approx \sum_{s=0}^{M} d^{(s)} \frac{N_s}{N}$$

$$\approx \sum_{s=0}^{M} d^{(s)} P(d_s)$$

Variables aleatorias y sus FDP

¿Cómo comparamos 2 FDP ?

Ejemplos de "medidas" para comparar:

- El valor típico :
  "el centro de la FDP"

- Una medida de dispersión alrededor de un valor típico
  "El 'ancho' de la FDP"

¿Cómo cuantificar el "ancho" o dispersión de una FDP?

una forma es usar una medida de la "distancia" entre un valor de la variable aleatoria y un valor típico.

Ej: la norma euclidiana al cuadrado de la diferencia entre la variable aleatoria d y su valor esperado <d>

$$q(d) = (d - \langle d \rangle)^2$$

si *p(d)* es la FDP de la variable aleatoria d, el área bajo la función *q(d)p(d)* es pequeña si una gran porción de la probabilidad esta concentrada cerca de *<d>* (*i.e., p(d)* es angosta); y el área es grande si gran parte de la probabilidad esta lejos de *<d>* (*i.e., p(d)* es ancha)

Variables aleatorias y sus FDP

¿Cómo cuantificar el “ancho” o dispersión de una FDP?

una forma es usar una medida de la “distancia” entre un valor de la variable aleatoria y un valor típico.

Ej: la norma euclidiana al cuadrado de la diferencia entre la variable aleatoria y su valor esperado

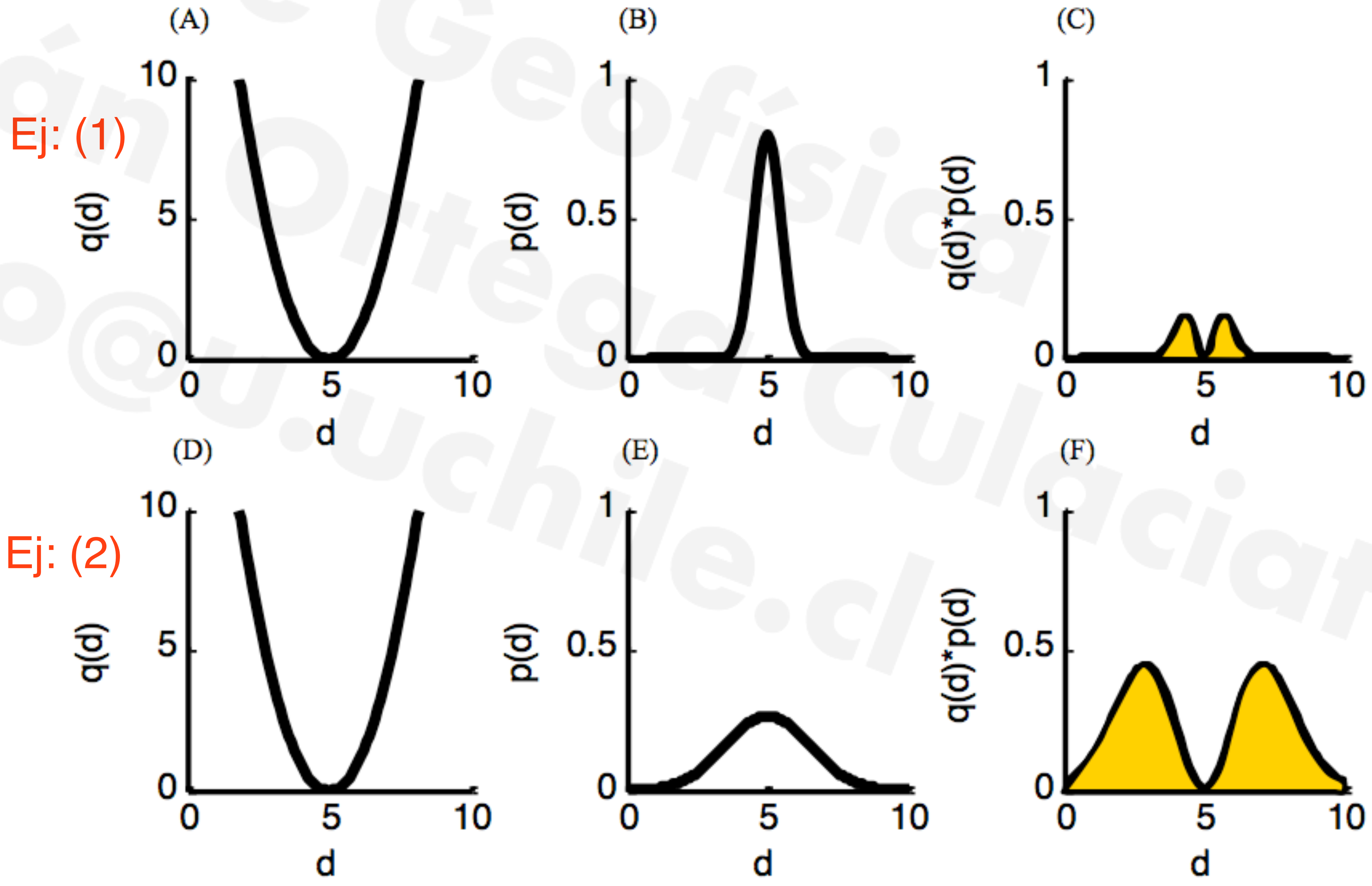
$$q(d) = (d-\langle d\rangle)^2$$

si *p(d)* es la FDP de la variable aleatoria d, el área bajo la función *q(d)p(d)* es pequeña si una gran porción de la probabilidad esta concentrada cerca de *<d>* (*i.e., p(d)* es angosta); y el área es grande si gran parte de la probabilidad esta lejos de *<d>* (*i.e., p(d)* es ancha)

Entonces, una posibilidad es cuantificar el “ancho” o dispersión de la FDP calculando el área bajo la función q(d)p(d) (area amarilla en la figura)

$$\text{dispersión} = \int_{-\infty}^{+\infty} q(d)p(d)\mathrm{d}d$$

Cálculo de la dispersión de *d* para 2 casos de FDP



Variables aleatorias y sus FDP

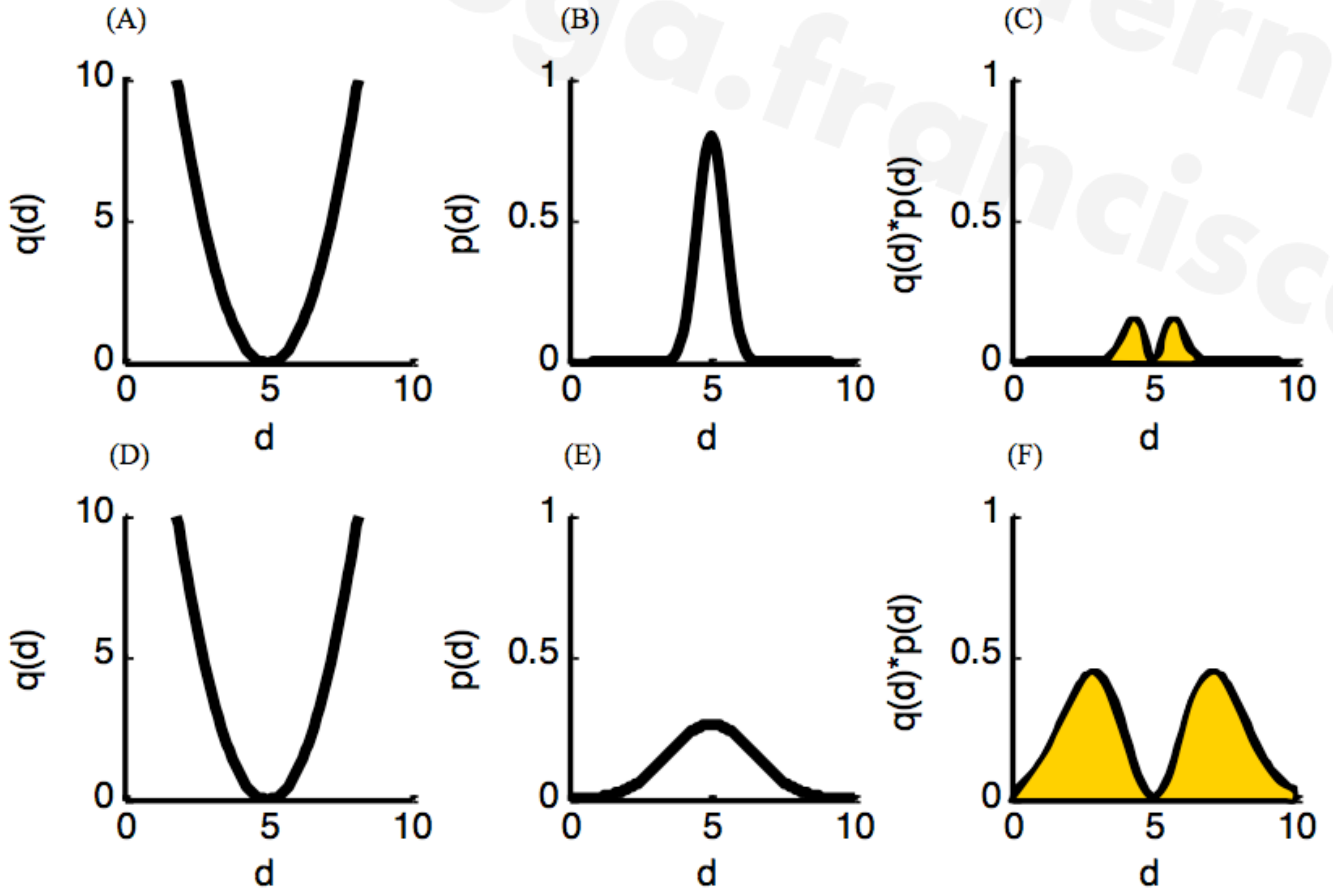
Entonces, una posibilidad es cuantificar el "ancho" o dispersión de la FDP calculando el área bajo la función q(d)p(d) (area amarilla en la figura)

$$\text{dispersión} = \int_{-\infty}^{+\infty} q(d)p(d)\mathrm{d}d$$

Cálculo de la dispersión de *d* para 2 casos de FDP

Ej: (1)

Ej: (2)



si $\quad q(d) = (d - <d>)^2$

la medida de dispersión es la varianza (variance) de la FDP (segundo momento de la FDP)

$$\sigma^2 = \int_{-\infty}^{+\infty} (d - <d>)^2\, p(d)\mathrm{d}d$$

y la medida de "ancho" de la FDP sería la raiz cuadrada de la varianza: la desviación estandar σ

Variables aleatorias y sus FDP

Fórmulas para estimar el promedio y la varianza de un conjunto de datos

$$\langle d \rangle^{est} = \frac{1}{N}\sum_{i=1}^{N} d_i$$ “promedio de la muestra”

$$(\sigma^2)^{est} = \frac{1}{N-1}\sum_{i=1}^{N} (d_i - \langle d \rangle^{est})^2$$ “varianza de la muestra”

Variables aleatorias y sus FDP

Fórmulas para estimar el promedio y la varianza de un conjunto de datos

$$\langle d \rangle^{est} = \frac{1}{N}\sum_{i=1}^{N} d_i$$

"promedio de la muestra"

$$(\sigma^2)^{est} = \frac{1}{N-1}\sum_{i=1}^{N}(d_i - \langle d \rangle^{est})^2$$

"varianza de la muestra"

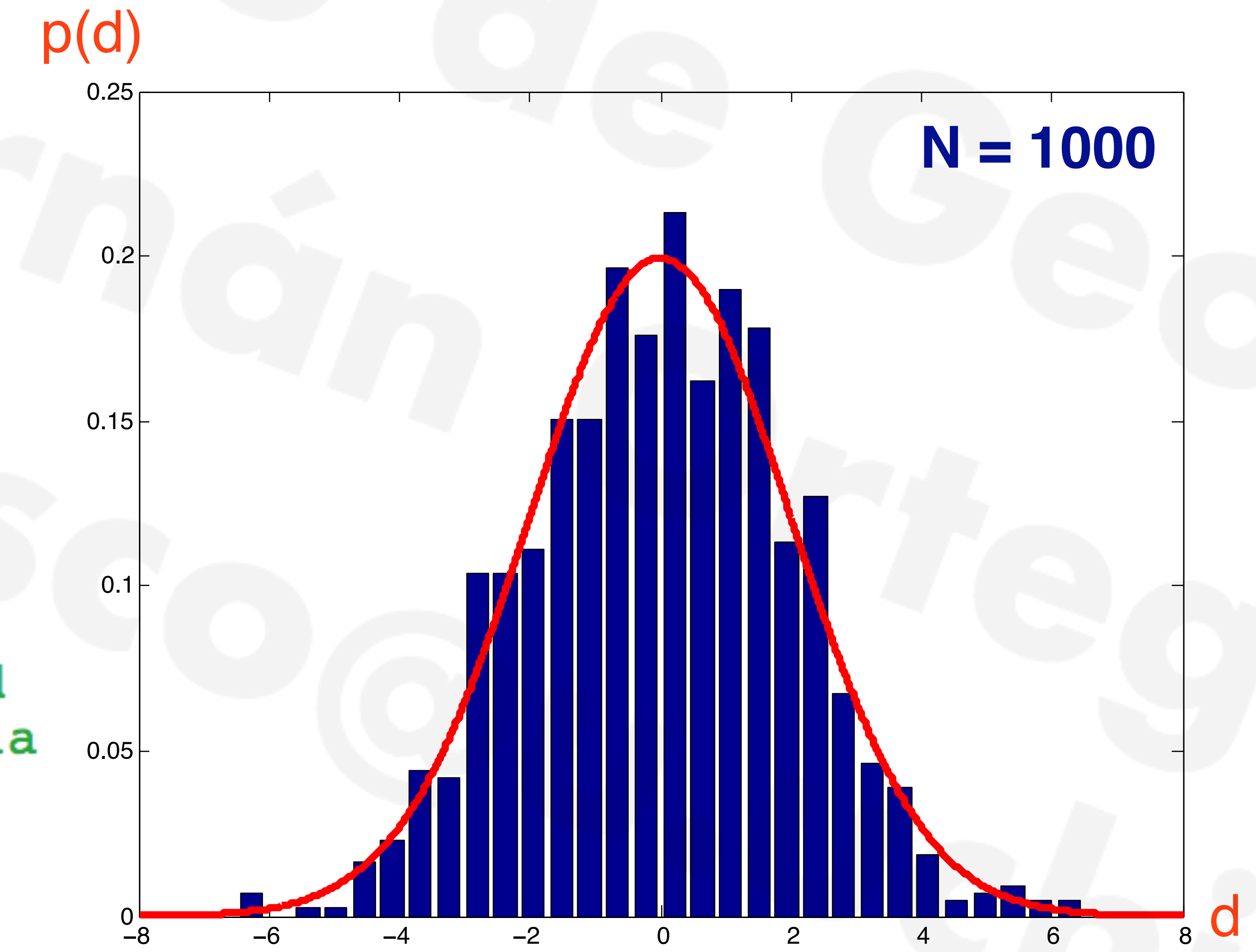
```
% asignar una discretización de los valores posibles de la variable d
% asumiremos d_min y d_max iguales a 4 sigma para minimizar el % de la
% probabilidad que queda fuera del intervalo.
mhuD = 0;
sigmaD = 2;
dmin = mhuD -4 * sigmaD;
dmax = mhuD + 4 * sigmaD;
Nceldas = 30;
% primero generar N muestras
N = 1000 ;
d_muestras = mhuD + randn(N,1)*sigmaD;
% obtener valores de la probabilidad en cada casillero.
[Ns, dS] = hist(d_muestras, Nceldas);
Dd = dS(2) - dS(1); % paso de la discretización

p = Ns ./ (Dd * sum(Ns)); %vuelvo a normalizar por la discretización
```

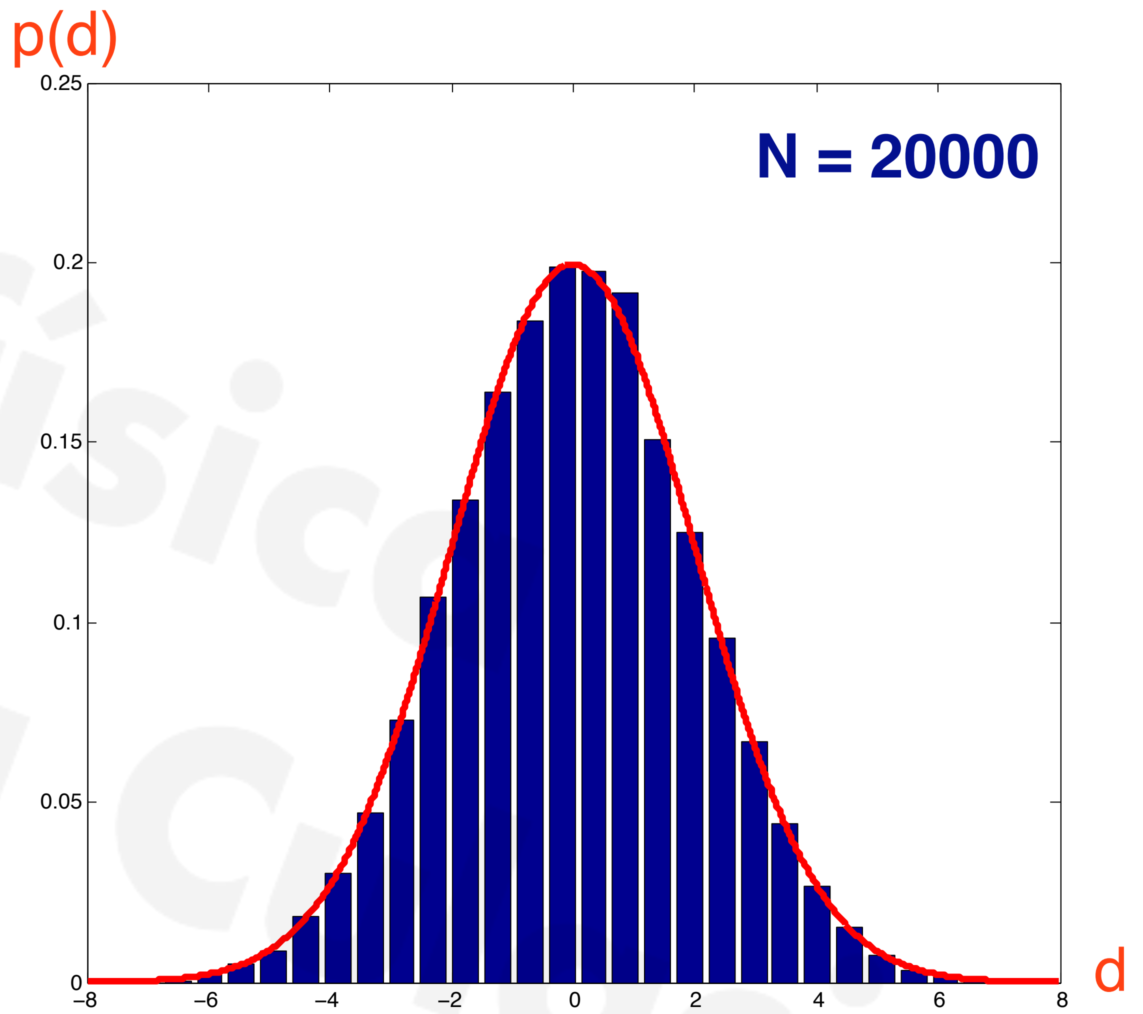
Se generan muestras de una normal con media "0" y desviación estandar "2"

```
%%%%%% Obtener promedio y varianza usando la tabulación de la PDF
disp(['Promedio y varianza usando la tabulación de la PDF']);
d_promedio = Dd * sum (dS.*p);
q = (dS - d_promedio).^2 ;
Varianza = Dd * sum(q.*p) ;
DesvEstandar = sqrt(Varianza) ;

%%%% Obtener promedio y varianza usando N muestras de la distribución
dM_promedio = mean(d_muestras) ;
DevEstandar_M = std(d_muestras) ;
Varianza_M = DevEstandar_M.^2 ;
```

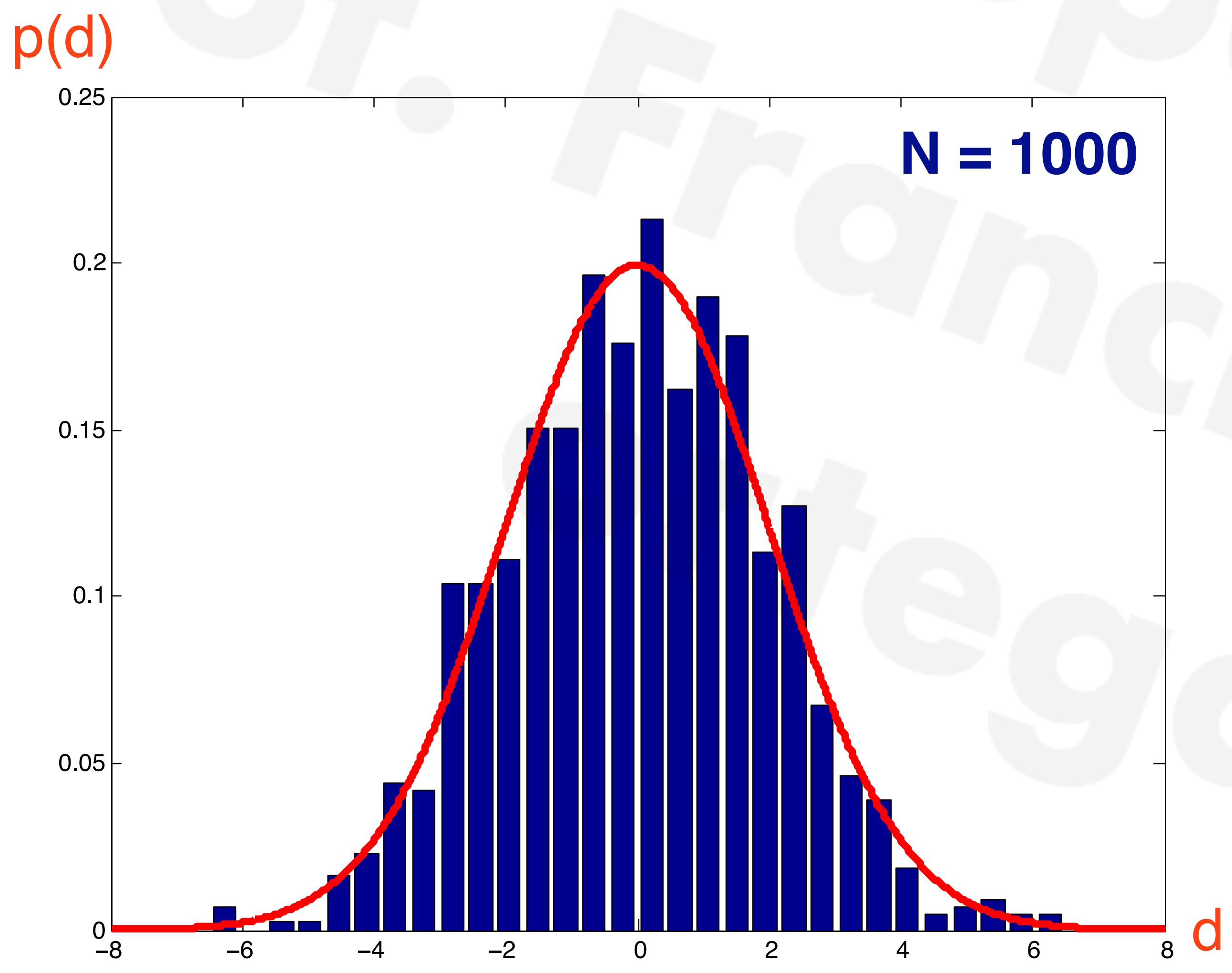



```
% Usando N = 1000 realizaciones
% de la variable aleatoria "d"
% Promedio y Varianza a partir de
% la PDF tabulada usando N muestras
d_promedio =     -0.0205
Varianza =        3.8074
DesvEstandar =    1.9513
% de las N muestras directamente
dM_promedio =    -0.0191
DevEstandar_M =   1.9501
Varianza_M =      3.8030
```

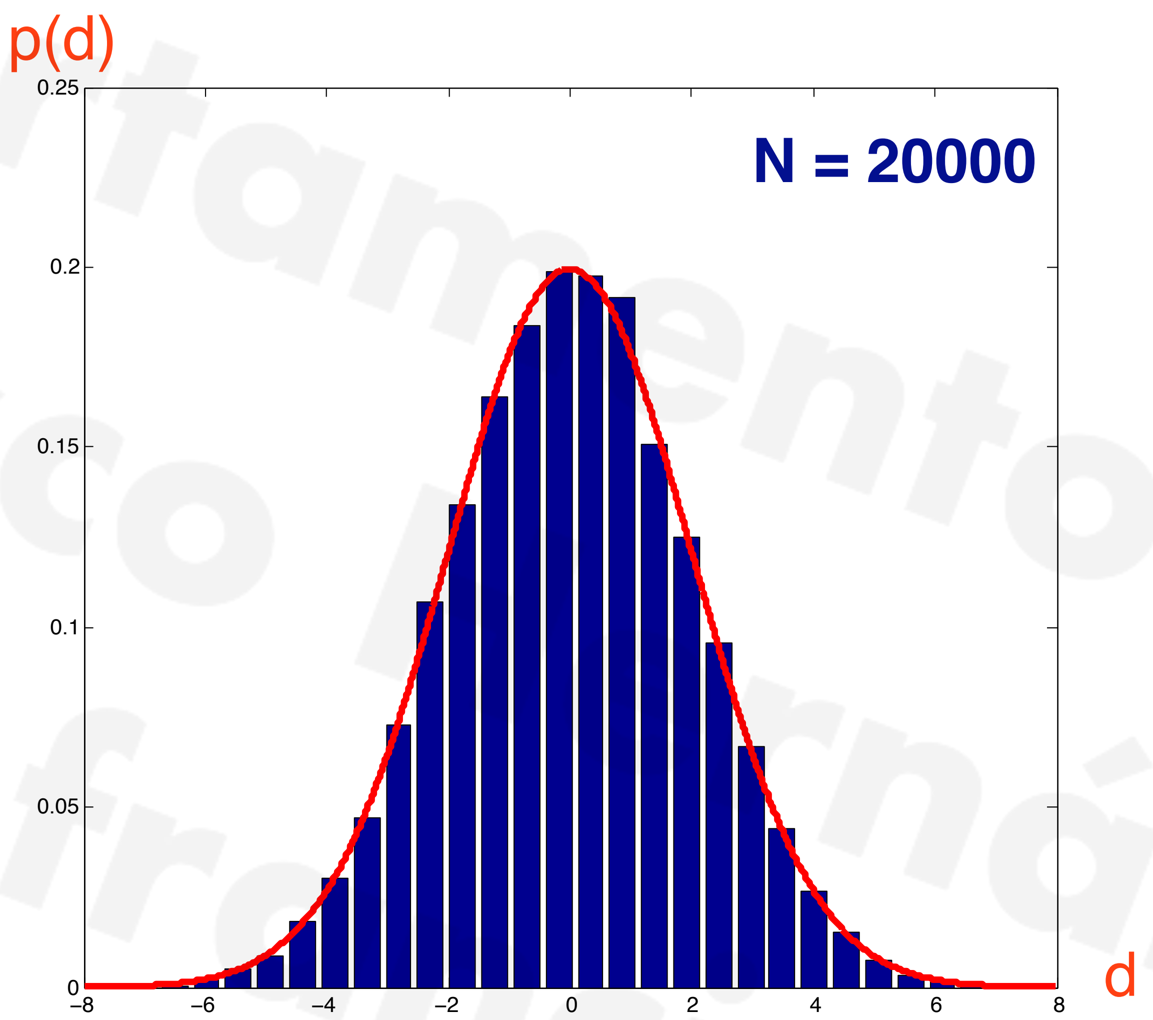



```
% Usando N = 20000 realizaciones
% de la variable aleatoria "d"
% Promedio y Varianza a partir de
% la PDF tabulada usando N muestras
d_promedio =      0.0066
Varianza =        3.9554
DesvEstandar =    1.9888
% de las N muestras directamente
dM_promedio =     0.0081
DevEstandar_M =   1.9838
Varianza_M =      3.9355
```

Variables aleatorias y sus FDP




```
% Usando N = 1000 realizaciones
% de la variable aleatoria "d"
% Promedio y Varianza a partir de
% la PDF tabulada usando N muestras
d_promedio =     -0.0205
Varianza =        3.8074
DesvEstandar =    1.9513
% de las N muestras directamente
dM_promedio =    -0.0191
DevEstandar_M =   1.9501
Varianza_M =      3.8030
```




```
% Usando N = 20000 realizaciones
% de la variable aleatoria "d"
% Promedio y Varianza a partir de
% la PDF tabulada usando N muestras
d_promedio =       0.0066
Varianza =         3.9554
DesvEstandar =     1.9888
% de las N muestras directamente
dM_promedio =      0.0081
DevEstandar_M =    1.9838
Varianza_M =       3.9355
```

Es importante notar que las estimaciones del valor esperado y varianza son diferentes si se obtienen estimando y utilizando la FDP de la variable ó si se obtienen utilizando directamente las realizaciones o “muestras” de la variable aleatoria.

En ambos casos la calidad de la estimación va a depender de la cantidad de muestras de la variable aleatoria (N).

En el caso particular de estimar y utilizar la FDP discreta de la variable aleatoria se introduce (más) incertidumbre en los resultados ya que al estimar la FDP hay un proceso de discretización de la variable aleatoria, donde hay que escoger el tamaño de las celdas del histograma que posteriormente se normaliza para aproximar la FDP.

Pueden jugar con el script “PromedioVarianza.py” (Python 3) para darse cuenta de la dependencia de los resultados de el tamaño de la muestra N y del tamaño de la discretización del histograma Nceldas.

Variables aleatorias y sus FDP

Otras formas en que se pueden comparar las FDP de variables aleatorias son métodos que vienen derivados de conceptos de Teoría de la Información

- Divergencia de Kullback-Liebler
- Entropía de la variable aleatoria
- Información mútua entre dos variables aleatorias

Estos términos requieren de una discusión más acabada y dedicada y no serán cubiertos en este tutorial.

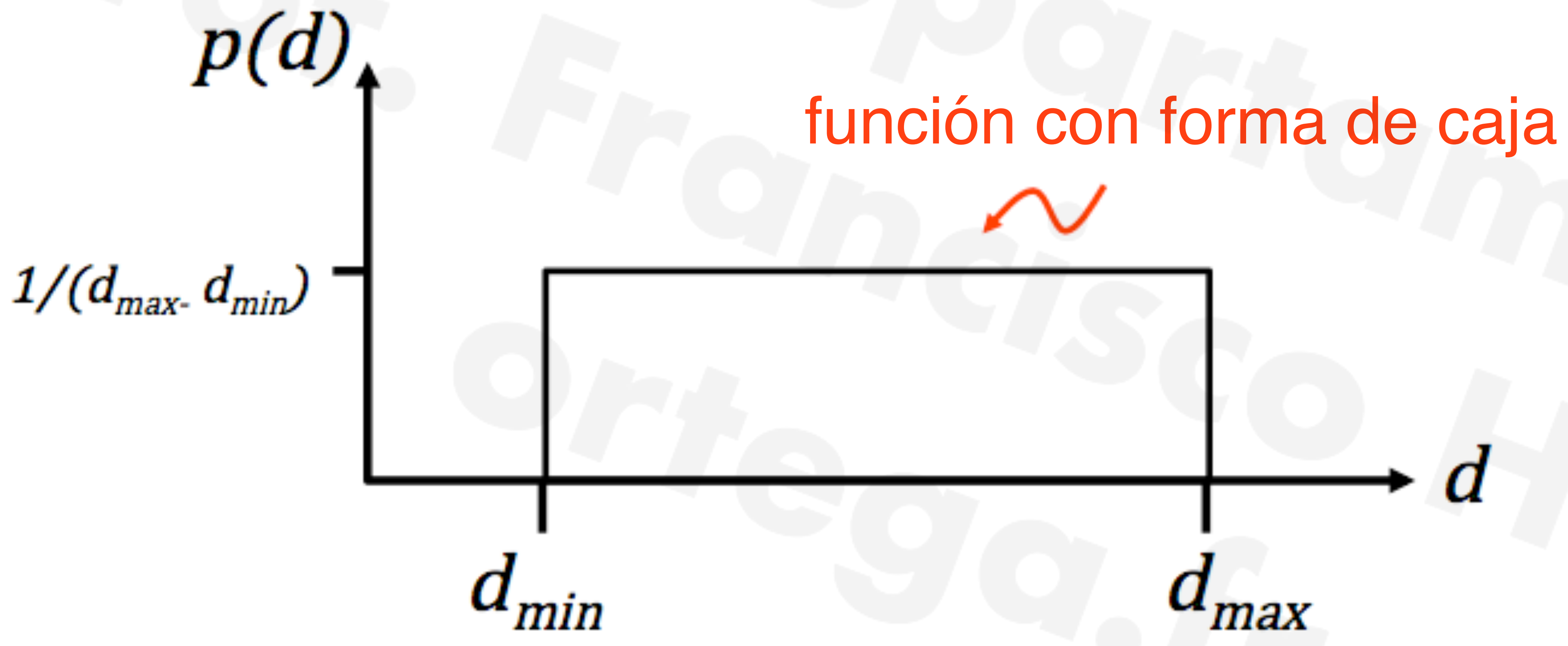
Dos Funciones de Densidad de Probabilidad Importantes

FDP Uniforme

FDP Gaussiana (o Normal)

Dos FDP importantes

FDP Uniforme $d \sim \mathcal{U}\left(d_{min}, d_{max}\right)$

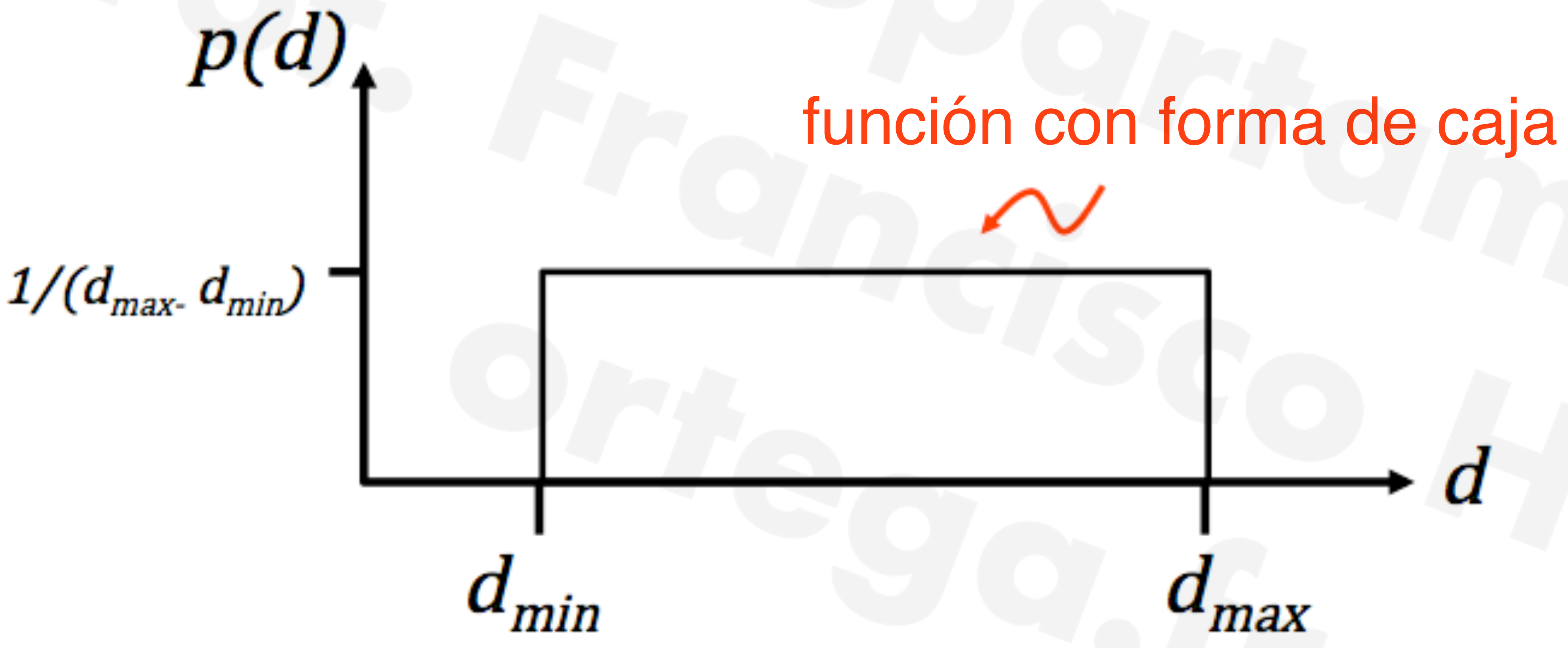


Una realización de *d* puede ser cualquier valor entre d_{min} y d_{max} con igual probabilidad.

Esta FDP es útil para caracterizar un estado de información “homogéneo” de un parámetro cartesiano, lo que algunos autores llaman “estado de no información”. (para los que deseen adentrarse más en este concepto, se recomienda leer la sección 1.2.4 de Tarantola, 2005).

Dos FDP importantes

FDP Uniforme $d \sim \mathcal{U}(d_{min}, d_{max})$



Una realización de *d* puede ser cualquier valor entre d_{min} y d_{max} con igual probabilidad.

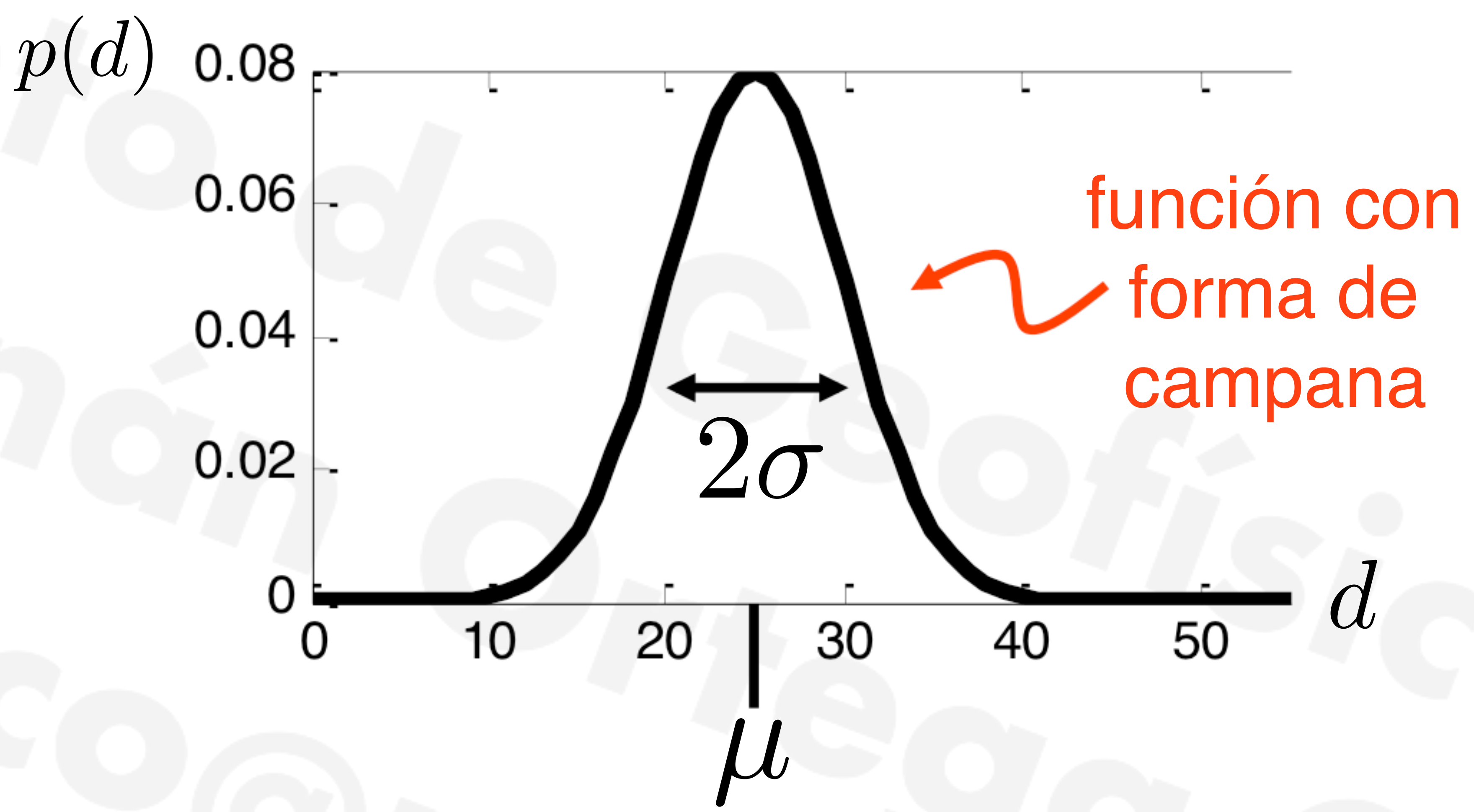
Esta FDP es útil para caracterizar un estado de información "homogéneo" de un parámetro cartesiano, lo que algunos autores llaman "estado de no información". (para los que deseen adentrarse más en este concepto, se recomienda leer la sección 1.2.4 de Tarantola, 2005).

FDP Normal o Gaussiana $d \sim \mathcal{N}(\mu, \sigma)$

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

μ =valor esperado, promedio

σ =desviación estandar



Una realización de d puede tener un valor cercano a μ con mucho más probabilidad que uno alejado de éste. La probabilidad de que una muestra tome un valor alejado de μ decrece exponencialmente con dicha distancia.

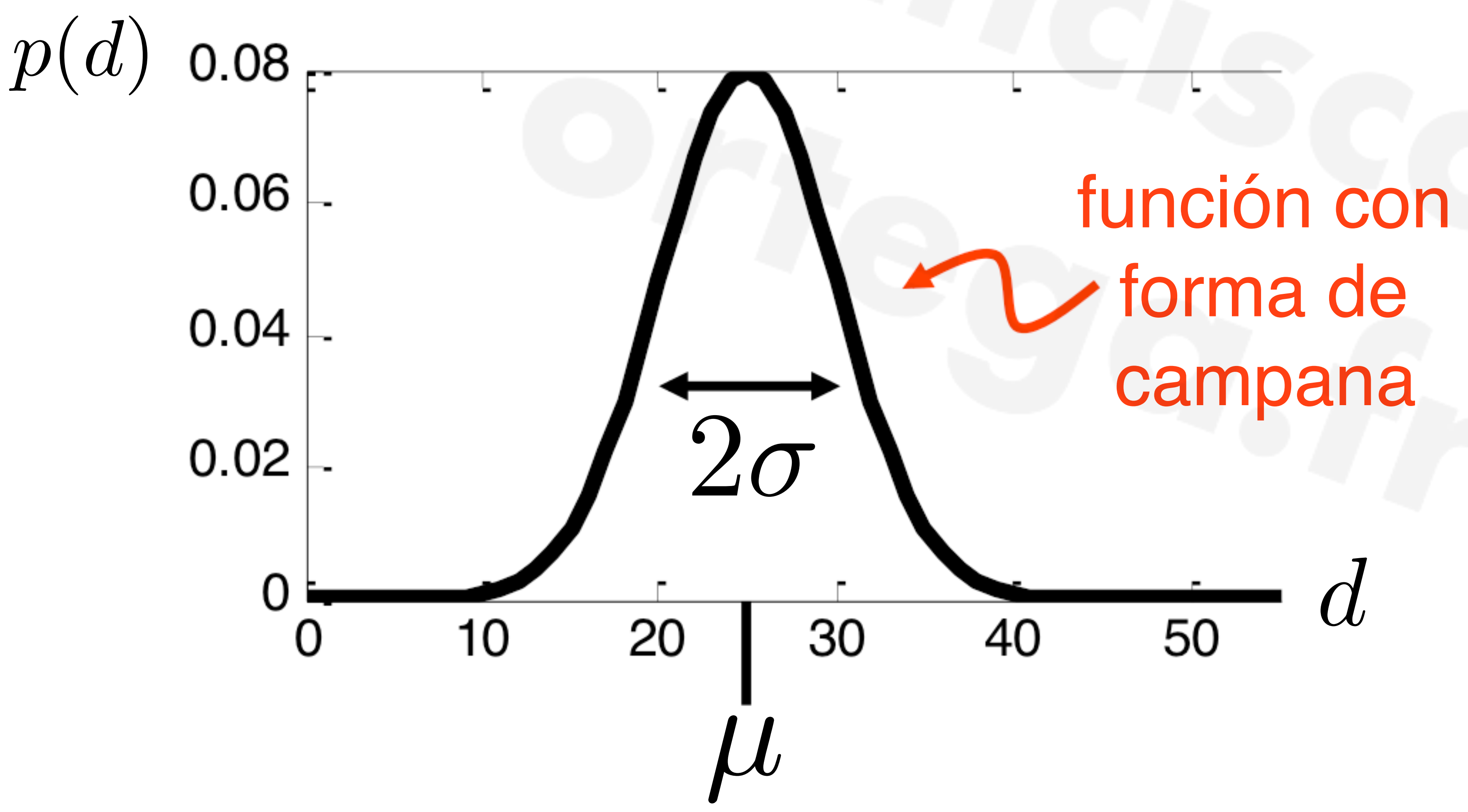
Dos FDP importantes

FDP Normal o Gaussiana $d \sim \mathcal{N}(\mu, \sigma)$

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

μ =valor esperado, promedio

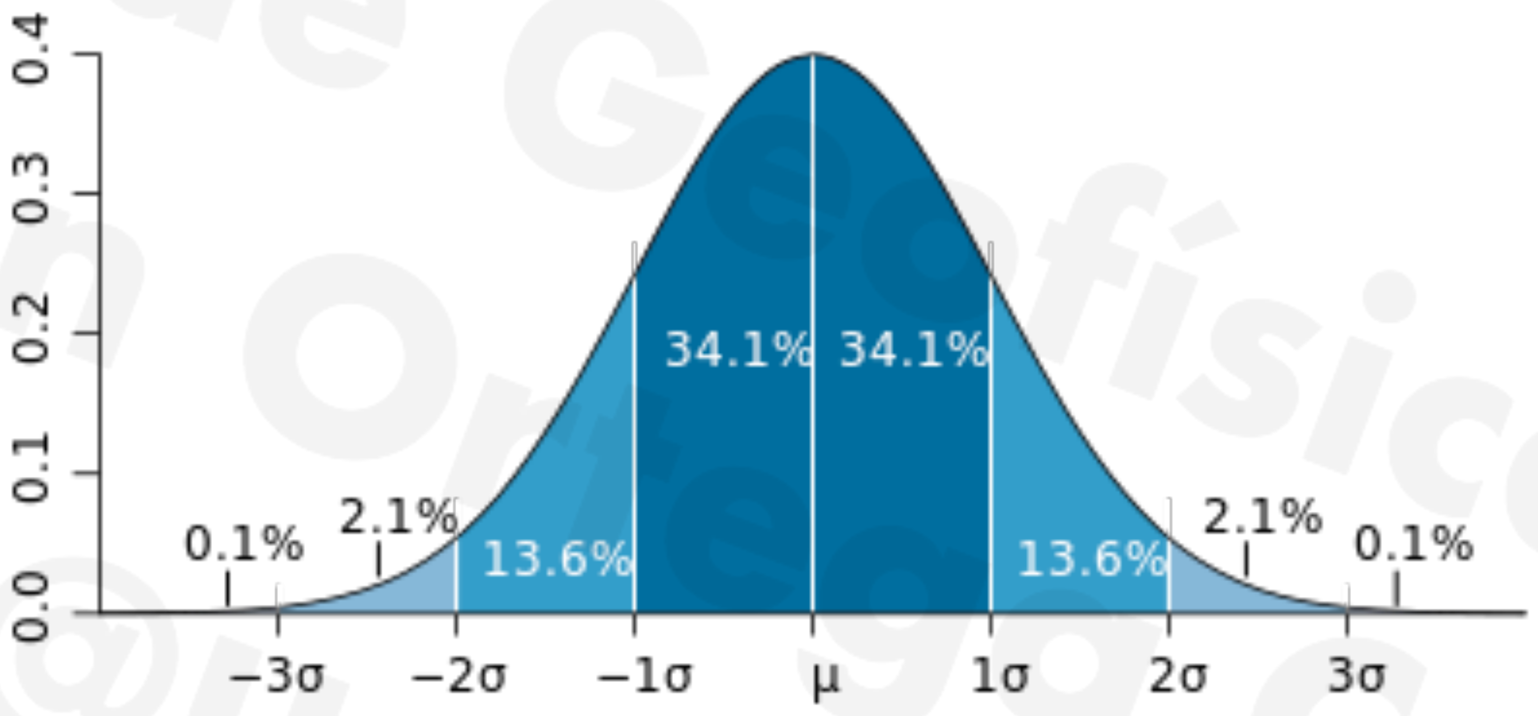
σ =desviación estandar



Una realización de d puede tener un valor cercano a μ con mucho más probabilidad que uno alejado de éste. La probabilidad de que una muestra tome un valor alejado de μ decrece exponencialmente con dicha distancia.

Qué tan rápido decae la probabilidad depende de la varianza de la distribución.

La tabla siguiente muestra que porcentaje de la probabilidad queda encerrada en intervalos simétricos con respecto a μ . $[\mu - n\sigma, \mu + n\sigma]$



n	P, %
1	68.27
2	95.45
3	99.73

Por ejemplo, hay un 99,73% de probabilidad de que una realización de *d* tenga un valor en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$

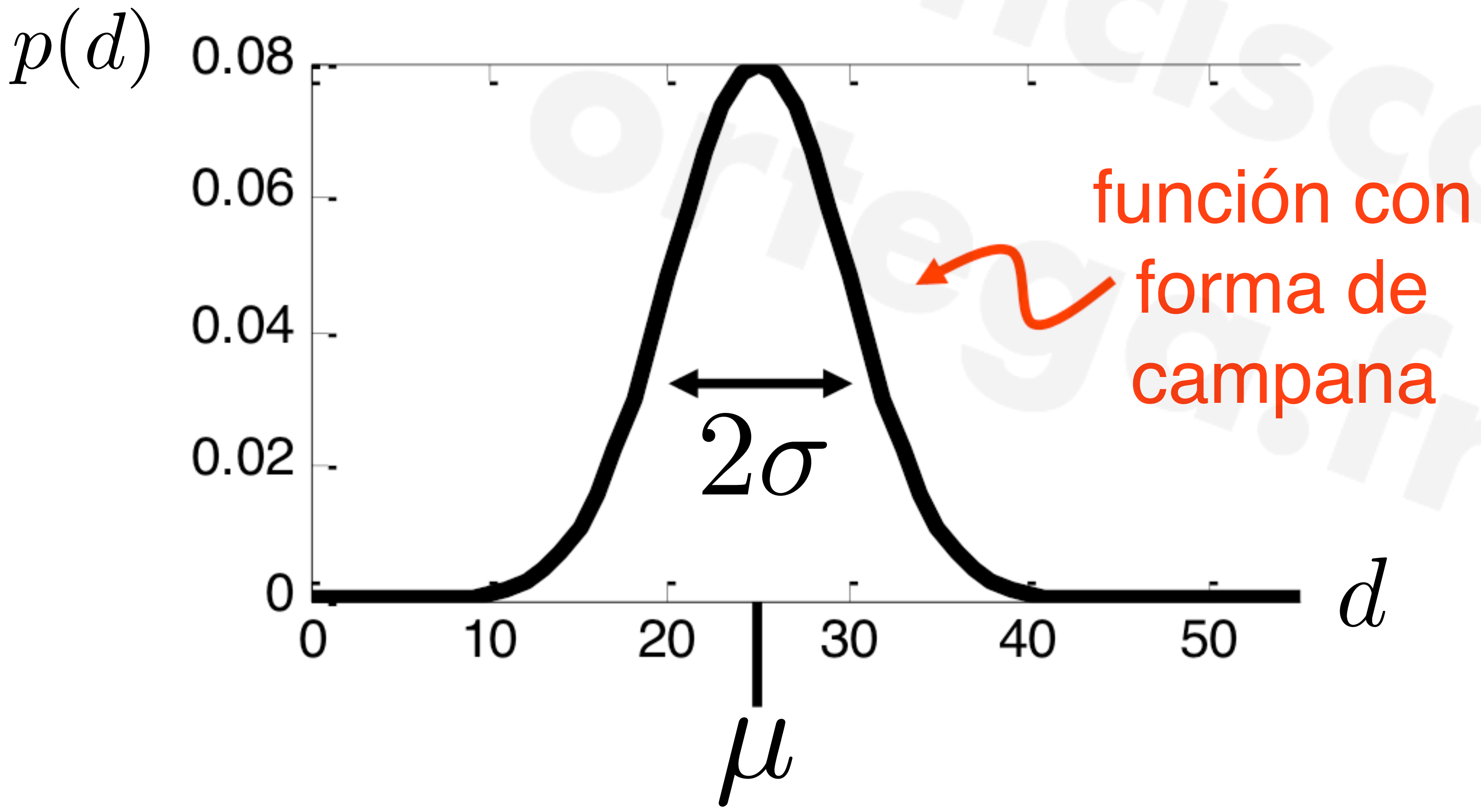
Dos FDP importantes

FDP Normal o Gaussiana $d \sim \mathcal{N}(\mu, \sigma)$

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

$\mu =$valor esperado, promedio

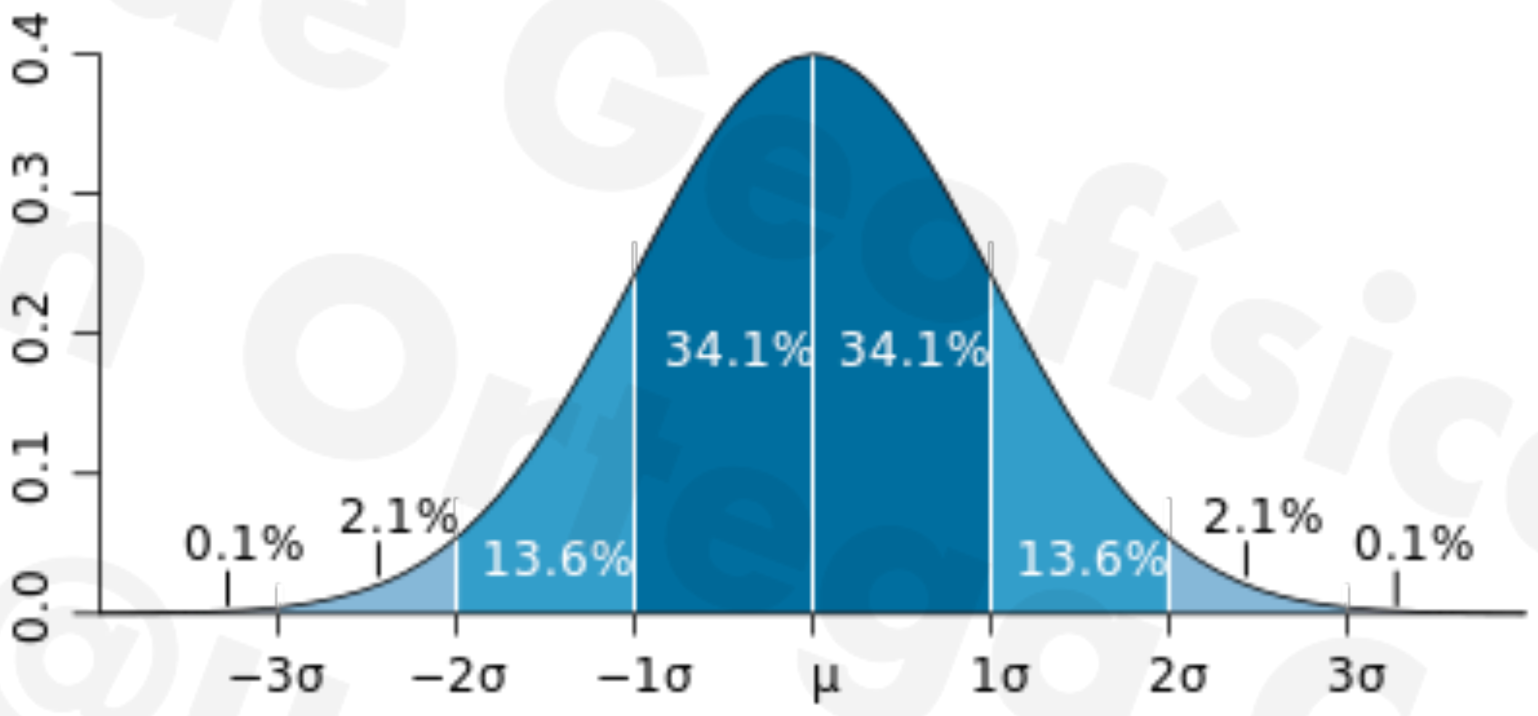
$\sigma =$desviación estandar



Una realización de d puede tener un valor cercano a μ con mucho más probabilidad que uno alejado de éste. La probabilidad de que una muestra tome

Qué tan rápido decae la probabilidad depende de la varianza de la distribución.

La tabla siguiente muestra que porcentaje de la probabilidad queda encerrada en intervalos simétricos con respecto a μ . $[\mu - n\sigma, \mu + n\sigma]$



n	P, %
1	68.27
2	95.45
3	99.73

Por ejemplo, hay un 99,73% de probabilidad de que una realización de d tenga un valor en el

Conocer la FDP de una variable aleatoria es poder CUANTIFICAR INCERTEZA, lo que es útil para poder tomar decisiones.

Todo tiene incerteza asociada, desde la talla de unos zapatos que quieren comprar, hasta el cálculo de la magnitud e hipocentro de un terremoto, la cantidad de sal que va a salir del salero cuando lo sacudan, etc.

Intervalos de Confianza: Una forma de cuantificar incerteza

Intervalos de Confianza

Supongamos que encontramos el siguiente resultado en la literatura

$$m_1 = 50 \pm 2 \ (95\%) \quad ; m_2 = 30 \pm 1 \ (95\%)$$

¿Que significa?

Intervalos de Confianza

Supongamos que encontramos el siguiente resultado en la literatura

$$m_1 = 50 \pm 2 \ (95\%) \quad ; m_2 = 30 \pm 1 \, (95\%)$$

¿Que significa?

Aquí tenemos lo siguiente:
1) m1 y m2 son mediciones independientes.
2) m1 va a tomar valores entre 50-2 = 48, y 50+2 = 52 con una probabilidad de un 95%.
3) m2 va a tomar valores entre 30-1 = 29, y 30+1 = 31 con una probabilidad de un 95%.

Intervalos de Confianza

Supongamos que encontramos el siguiente resultado en la literatura

$$m_1 = 50 \pm 2 \ (95\%) \quad ; m_2 = 30 \pm 1 \ (95\%)$$

¿Que significa?

si la FDP de m1 es p(m1)

si la FDP de m2 es p(m2)

calcular el promedio $< m_1 >$

calcular el promedio $< m_2 >$

calcular la varianza σ_1

calcular la varianza σ_2

Intervalos de Confianza

Supongamos que encontramos el siguiente resultado en la literatura

$$m_1 = 50 \pm 2 \ (95\%) \quad ; m_2 = 30 \pm 1 \ (95\%)$$

¿Que significa? $\longrightarrow$ Si m1 y m2 siguen una distribucion Normal/Gaussiana

si la FDP de m1 es p(m1)

calcular el promedio $< m_1 >$

calcular la varianza σ_1

Hay una probabilidad de un 95% de que m1 tome un valor entre 48 y 52

si la FDP de m2 es p(m2)

calcular el promedio $< m_2 >$

calcular la varianza σ_2

Hay una probabilidad de un 95% de que m2 tome un valor entre 29 y 31

$$m_1 = \underbrace{50}_{< m_1 >} \pm \underbrace{2}_{2\sigma_1} \ (95\%) \quad ; m_2 = \underbrace{30}_{< m_2 >} \pm \underbrace{1}_{2\sigma_2} \ (95\%)$$

Relaciones entre dos o más variables aleatorias

Funciones de densidad de probabilidad conjunta

FDP conjunta (joint PDF)

$$p(\mathbf{d}) = p(d_1, d_2, \ldots, d_N)$$ probabilidad de que las observaciones estan “cerca” de **d**

$$p(\mathbf{m}) = p(m_1, m_2, \ldots, m_M)$$ probabilidad de que los parámetros están “cerca” de **m**

Relaciones entre dos o más variables aleatorias

Se habla de correlación o dependencia entre 2 o más variables aleatorias cuando existe una relación entre estas (o según algunos autores entre "el ruido" de estas).

Ejemplo: Diseñamos un experimento en que se escoge a un grupo de personas para que mida el ancho y el largo de un rectángulo. A cada uno se le pasa una regla graduada diferente, las cuales tienen errores en su graduación debido al método que se utiliza para fabricarlas.

Después del experimento vamos a tener un conjunto de pares de medidas de ancho y largo del rectángulo. Si tomamos una medida hecha con una regla que estaba más larga de lo que debería, tanto el largo como el ancho medidos serán más largos de lo que deberían. De la misma manera si la regla es más corta, tanto el largo y el ancho medidos son más cortos.

Luego, se espera que las mediciones de ancho y largo estén con un error en la misma dirección (más largo o más corto, ambos), esto es que las variables ancho y largo están correlacionadas o dependientes una de la otra. En general, la dependencia o correlación es un aspecto no deseado en mediciones experimentales.

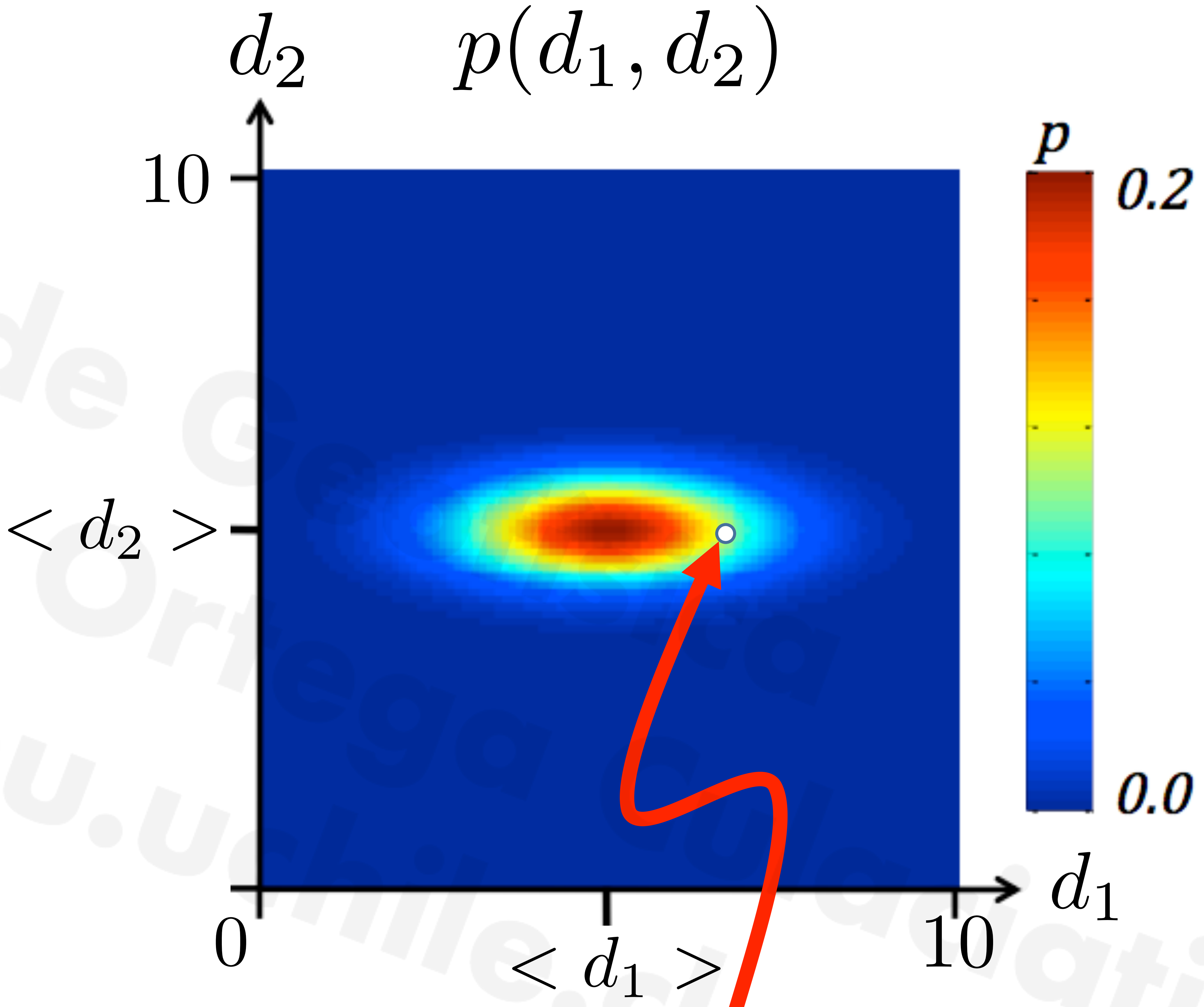
Variables aleatorias no correlacionadas (independientes)

Cuando dos variables aleatorias no están correlacionadas, no se observa un patrón de comportamiento entre los valores de una de las variables y otra

Variables aleatorias no correlacionadas (independientes)

Cuando dos variables aleatorias no están correlacionadas, no se observa un patrón de comportamiento entre los valores de una de las variables y otra

En la figura de la derecha, cuando d1 toma valores más altos que su valor promedio, d2 puede tomar valores más altos o bajos que su valor promedio con igual probabilidad.



Cuando d_1 toma un valor alto, no hay una tendencia que indique que d_2 sea alto o bajo

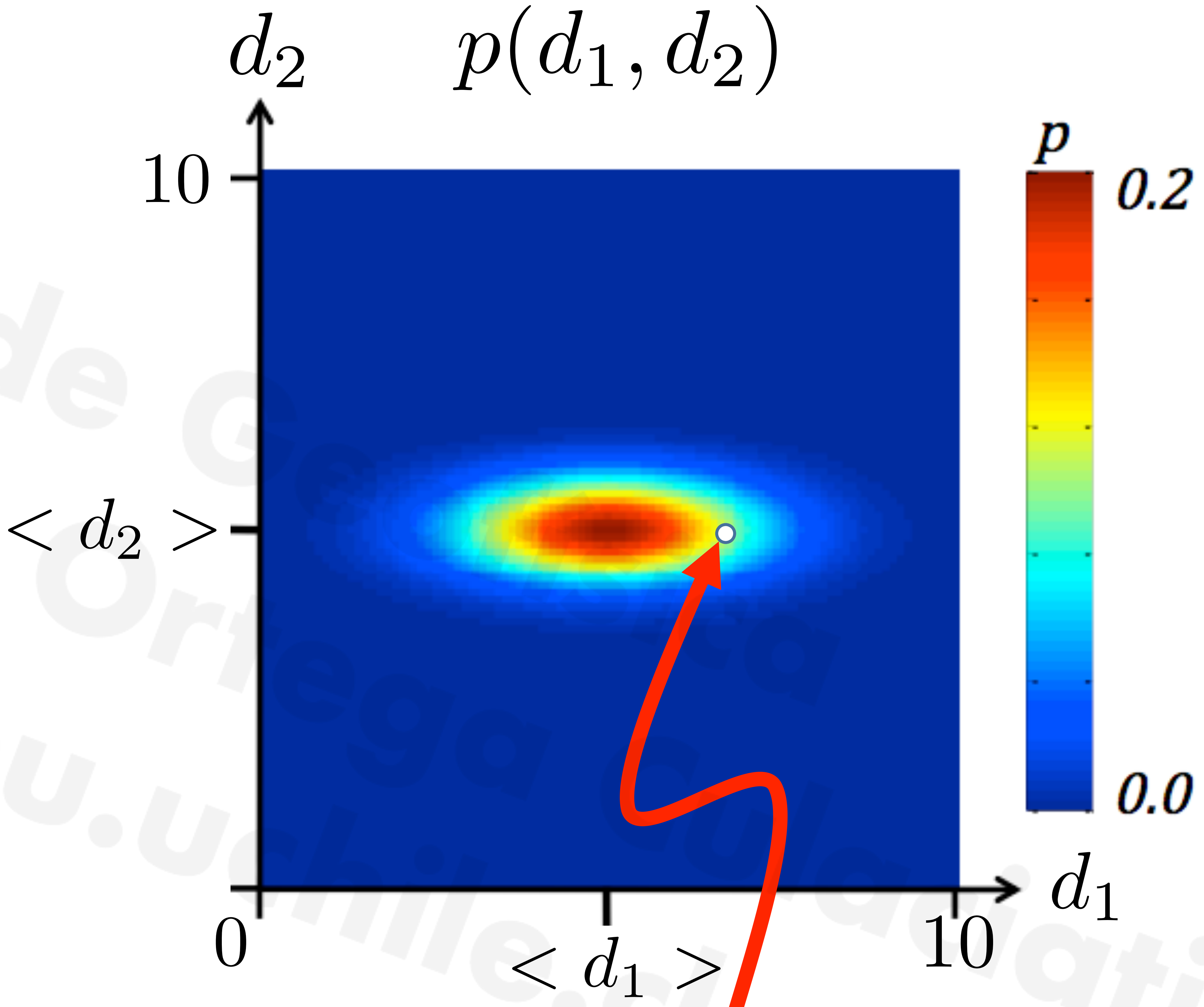
Variables aleatorias no correlacionadas (independientes)

Cuando dos variables aleatorias no están correlacionadas, no se observa un patrón de comportamiento entre los valores de una de las variables y otra

En la figura de la derecha, cuando d1 toma valores más altos que su valor promedio, d2 puede tomar valores más altos o bajos que su valor promedio con igual probabilidad.

En el caso de la figura, como las variables son independientes, la FDP conjunta es el producto de las FDP de las dos variables

$$p(d_1, d_2) = p(d_1)p(d_2)$$



Cuando d_1 toma un valor alto, no hay una tendencia que indique que d_2 sea alto o bajo

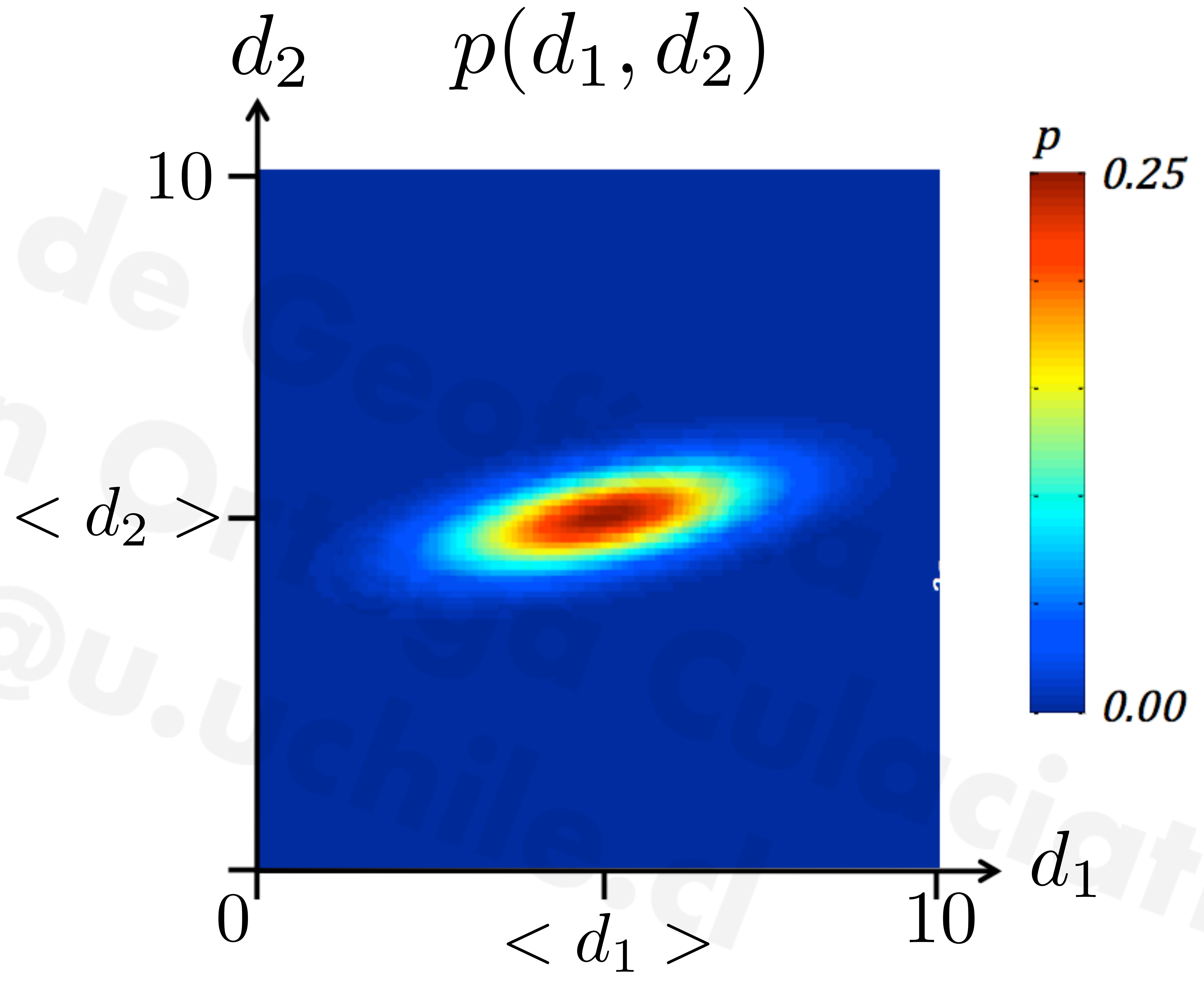
Variables aleatorias no correlacionadas (independientes)

En el caso general de variables aleatorias independientes (no correlacionadas), la FDP conjunta se escribe como el producto de las FDP individuales

$$p(\mathbf{d}) = p(d_1)p(d_2)p(d_3)\dots p(d_N)$$

Variables aleatorias correlacionadas (dependientes)

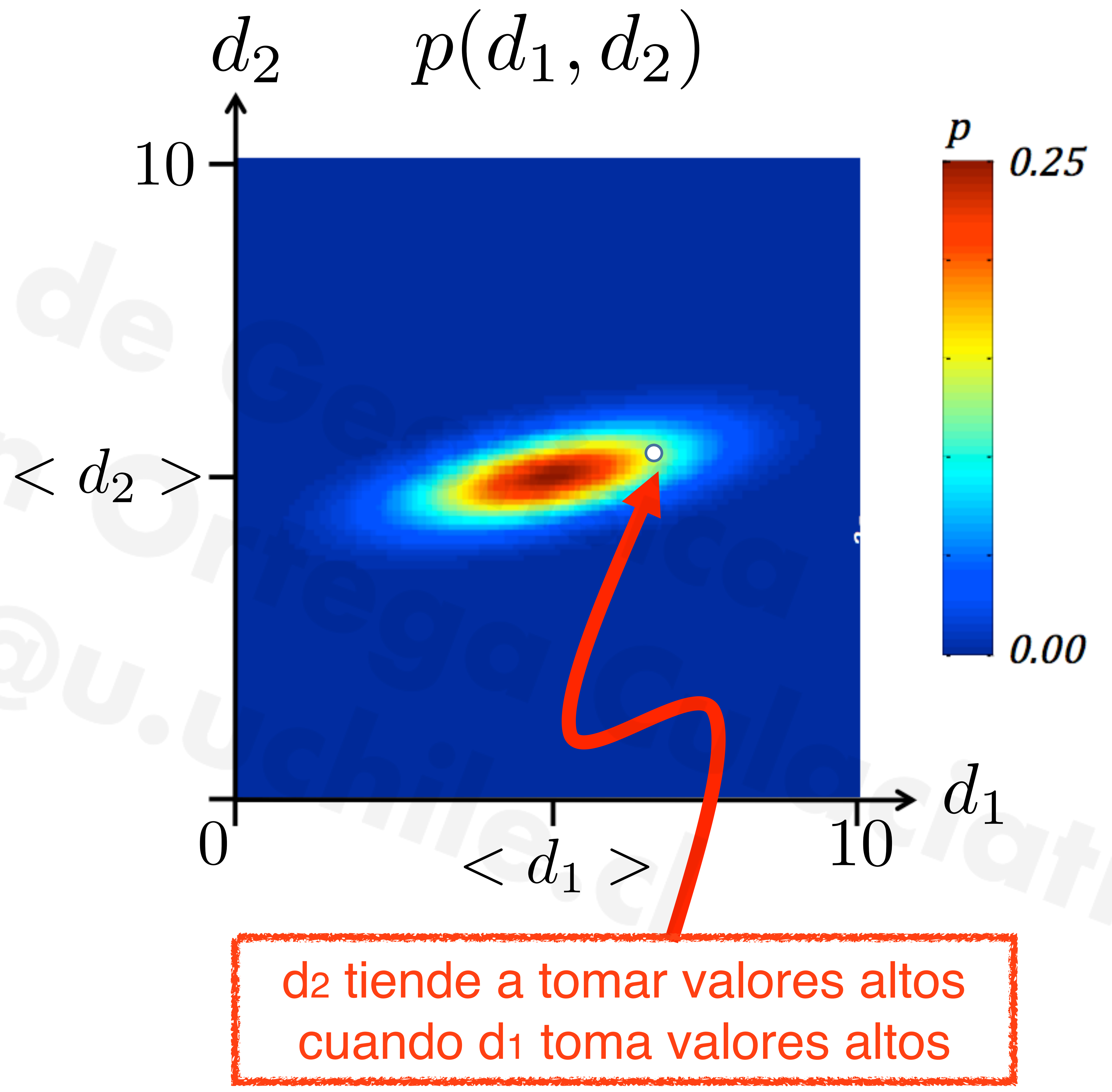
NOTA: La FDP conjunta, de las variables aleatorias d1 y d2 en el caso de la figura, dada la realización (d1*, d2*), cuantifica la probabilidad de que d1 se ecuentre en una vecindad de d1* y de que d2 se ecuentre en una vecindad de d2*



Variables aleatorias correlacionadas (dependientes)

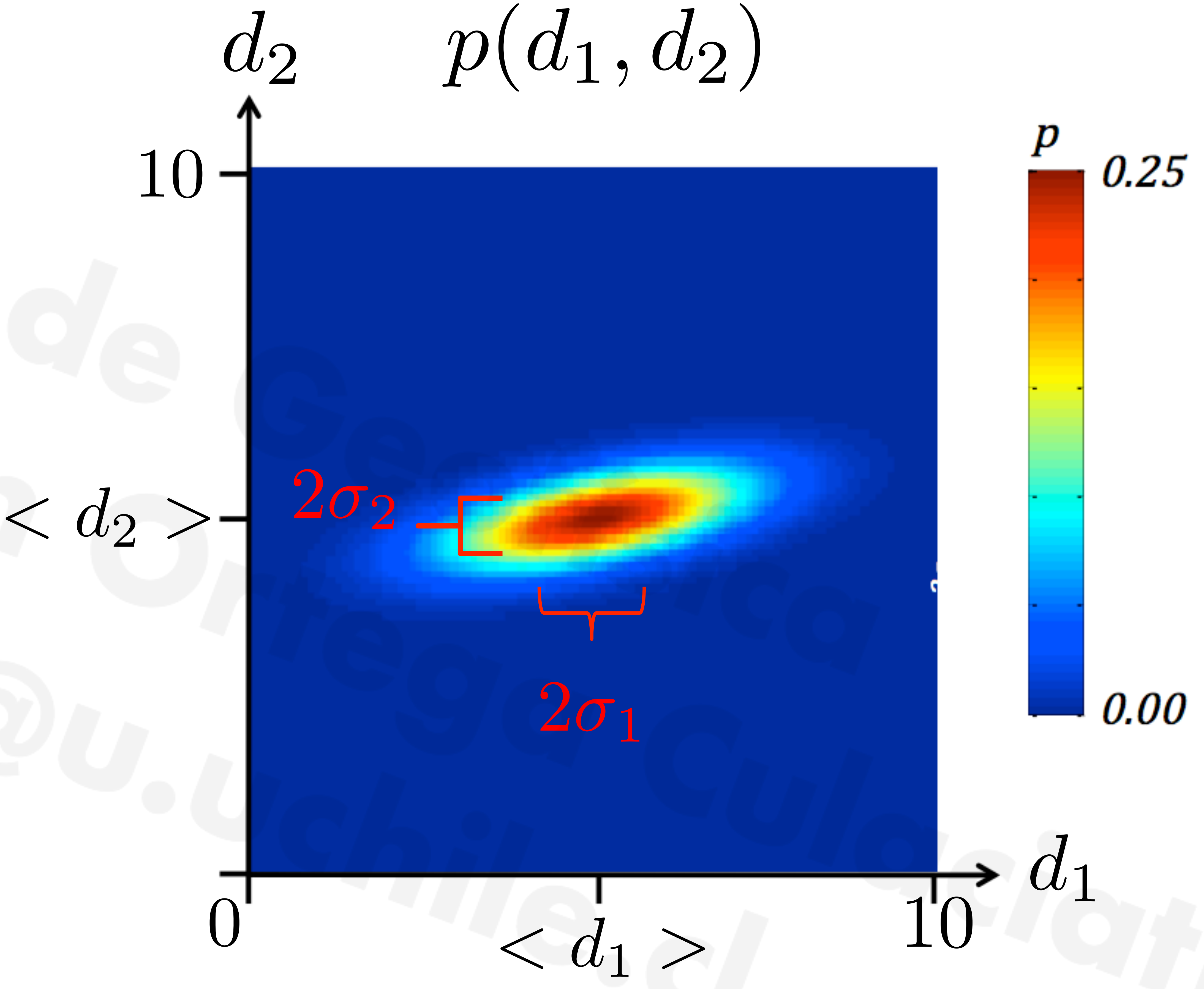
En la figura, la FDP denota una correlación *positiva* entre las variables d_1 y d_2

es decir, dado un valor de d_1 mayor que su promedio, se espera un valor de d_2 mayor que su promedio



Variables aleatorias correlacionadas (dependientes)

Las varianzas miden el ancho de la distribución en los dos ejes coordenados.

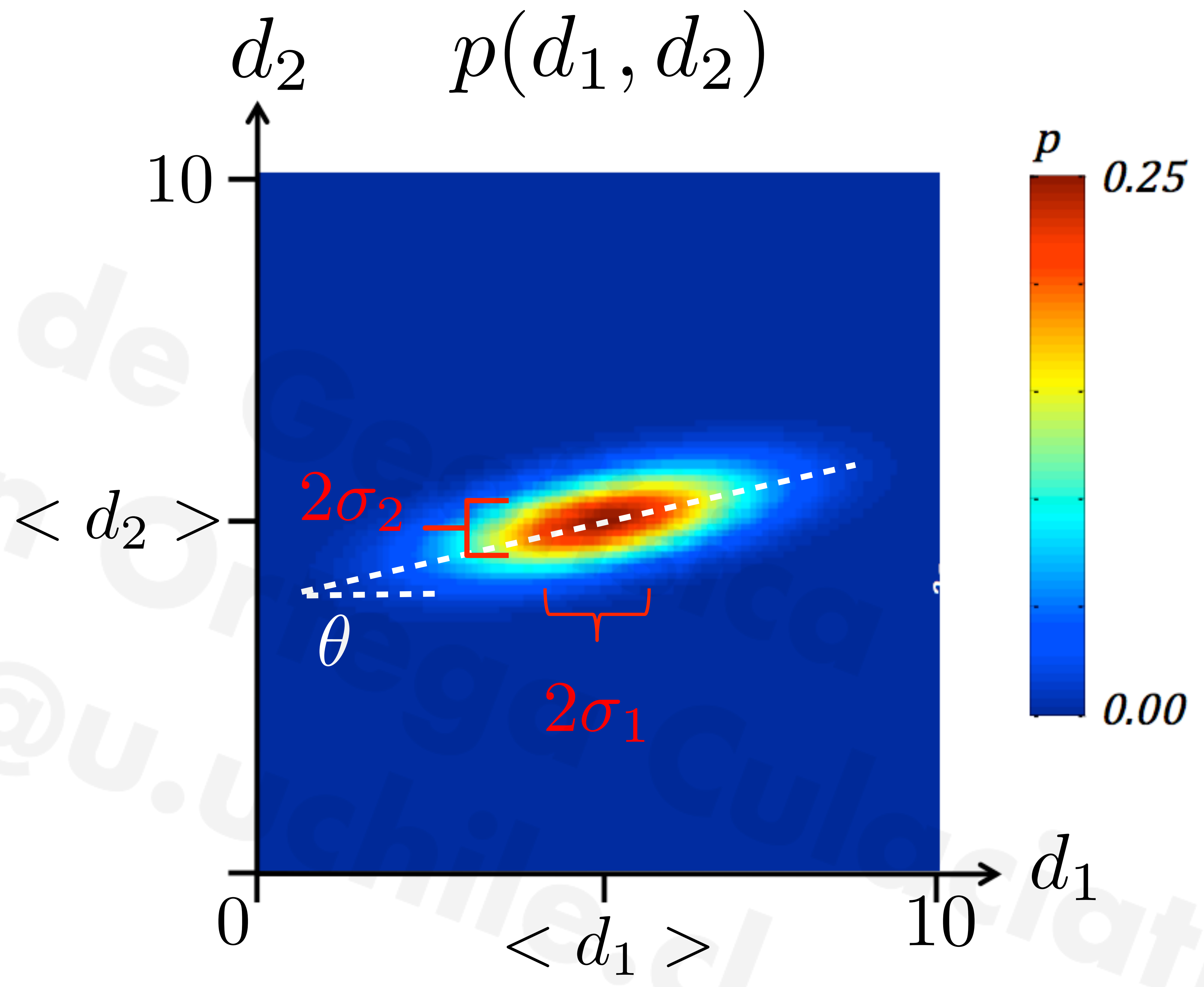


Variables aleatorias correlacionadas (dependientes)

Las varianzas miden el ancho de la distribución en los dos ejes coordenados.

Para medir la correlación necesitamos medir el “ángulo” en que esta inclinada la FDP.

Veremos más adelante que en el caso de FDPs que son muy similares a una Normal multivariada, la covarianza, o el coeficiente de correlación nos da una “medida” de que tan dependientes son las variables aleatorias.

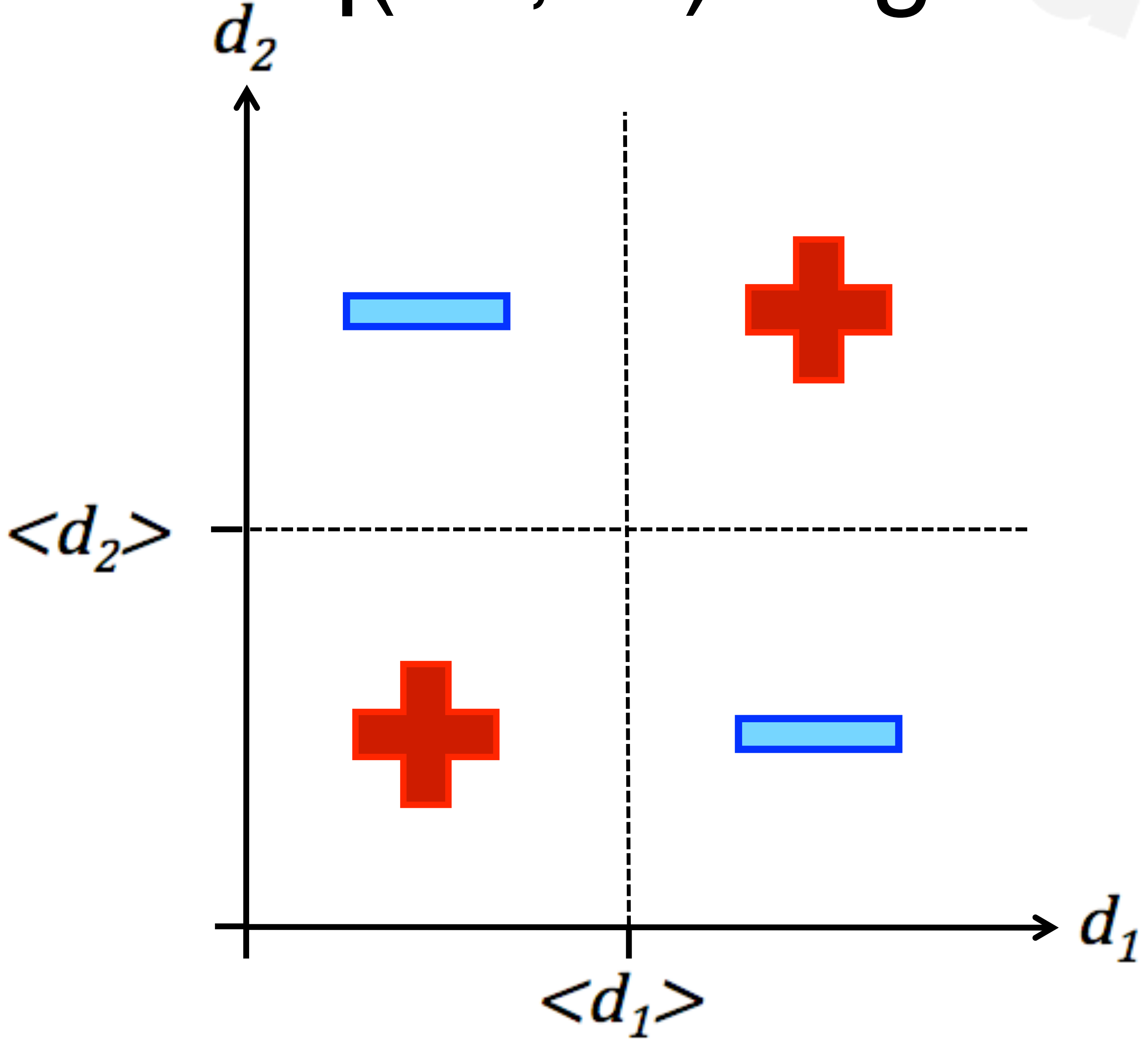


Variables aleatorias correlacionadas (dependientes)

para estimar la dispersión de una PDF de dos variables podemos usar la medida de distancia

$$q(d_1, d_2) = (d_1 - < d_1 >)(d_2 - < d_2 >)$$

Los signos de q(d1,d2) según cuadrante



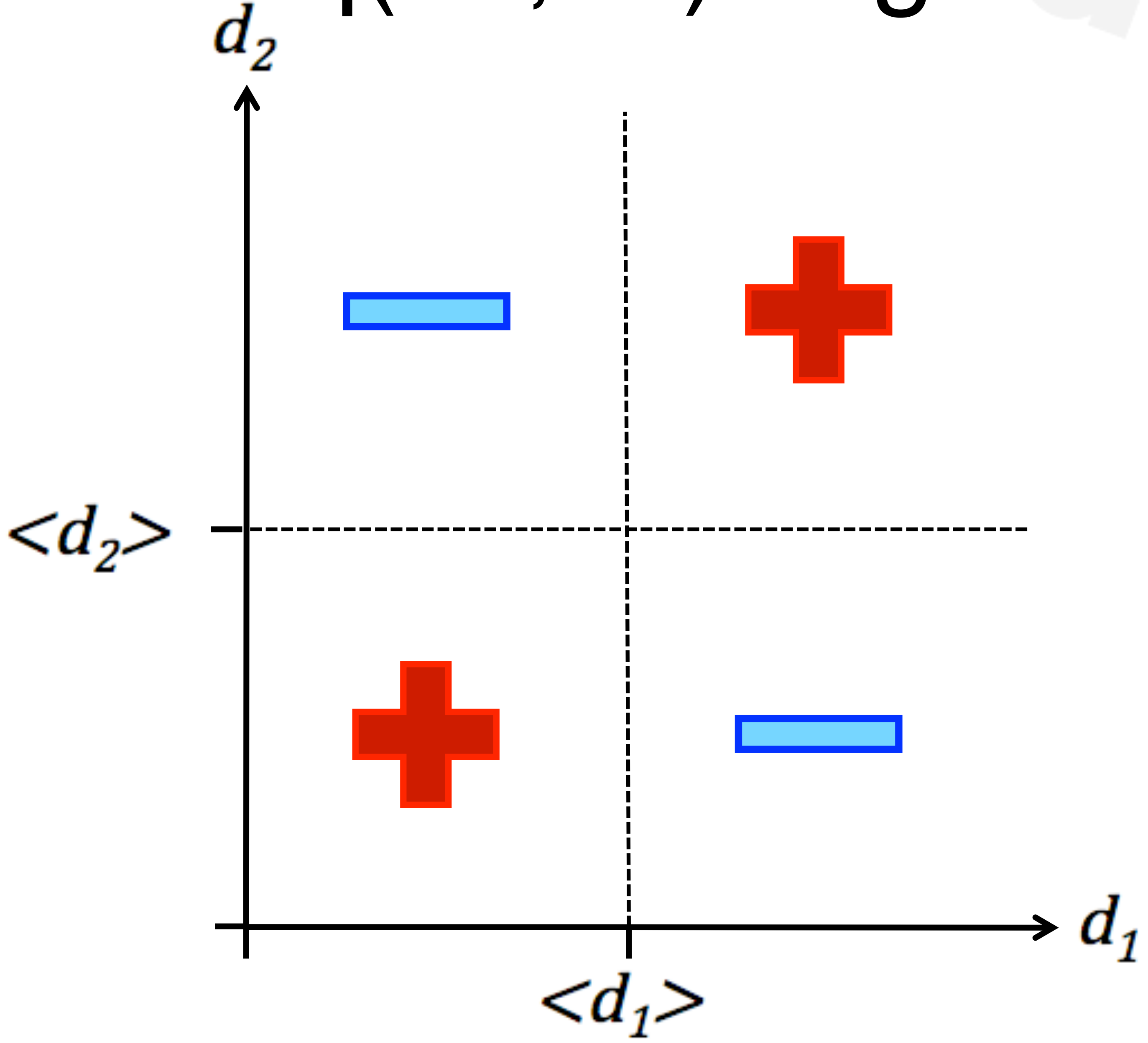
Relaciones entre dos o más variables aleatorias

Variables aleatorias correlacionadas (dependientes)

para estimar la dispersión de una FDP de dos variables podemos usar la medida de distancia

$$q(d_1, d_2) = (d_1 - < d_1 >)(d_2 - < d_2 >)$$

Los signos de q(d1,d2) según cuadrante

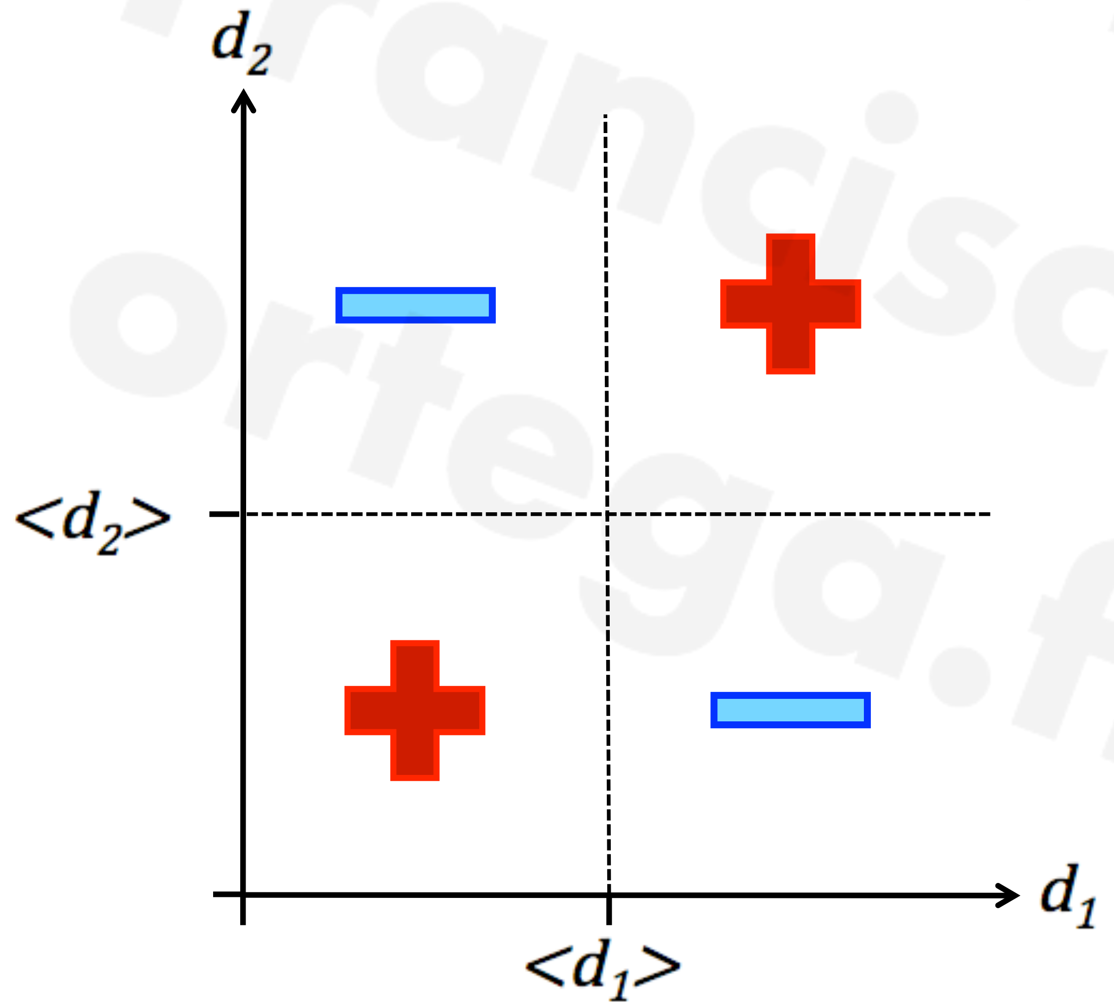


Siguiendo la misma idea que para el cálculo de la varianza, se multiplica la FDP por esta función, para obtener una estimación de la correlación de las variables aleatorias.

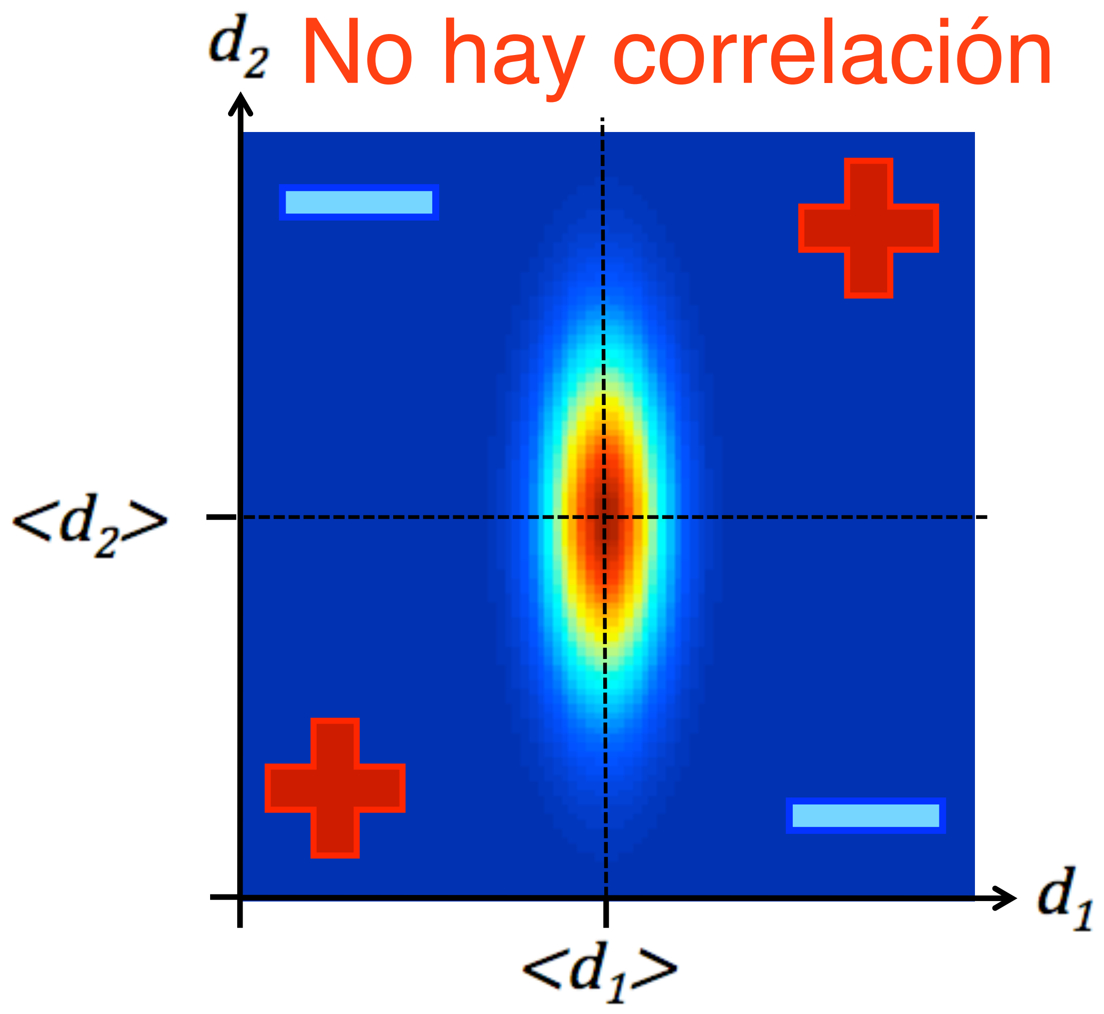
Relaciones entre dos o más variables aleatorias

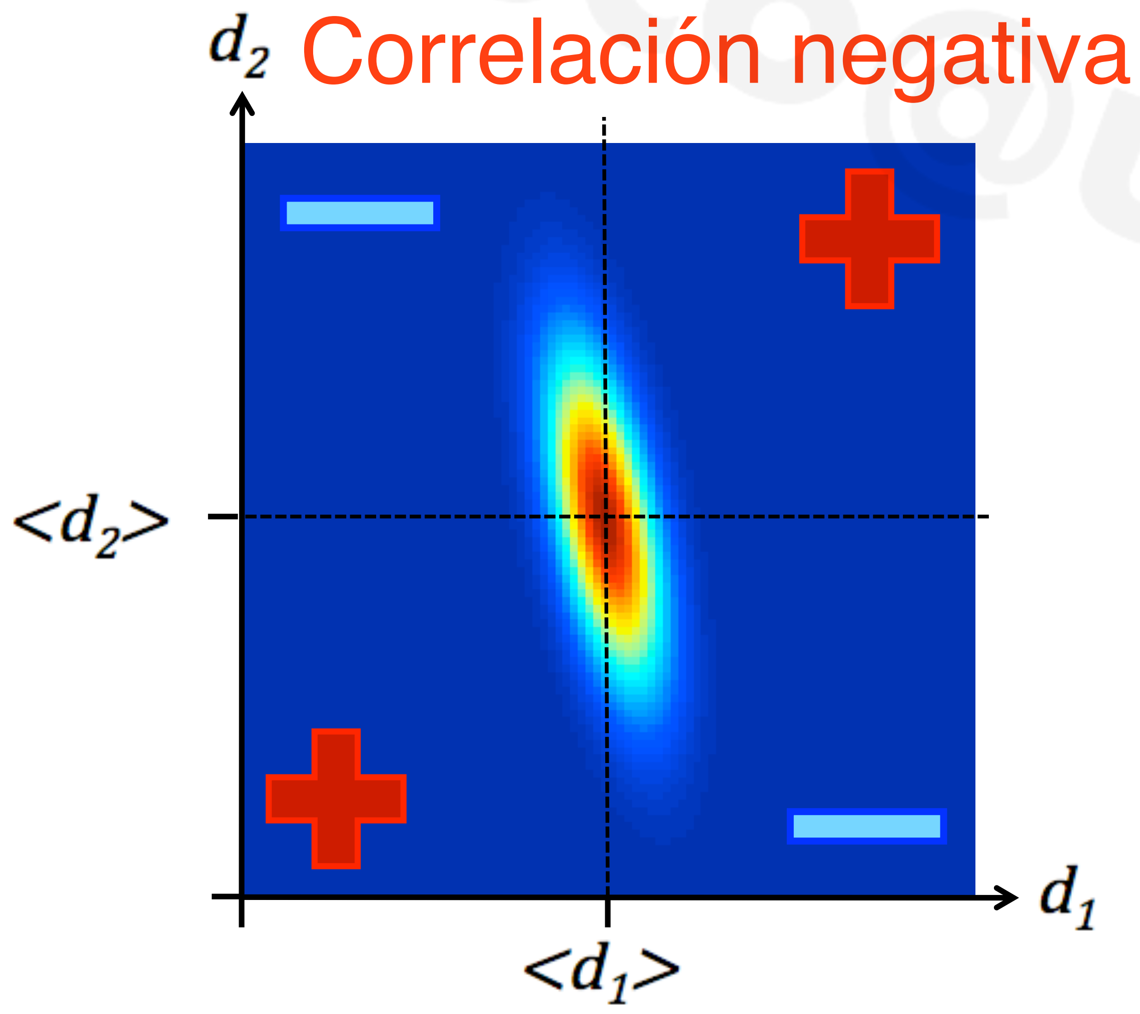
Variables aleatorias correlacionadas (dependientes)

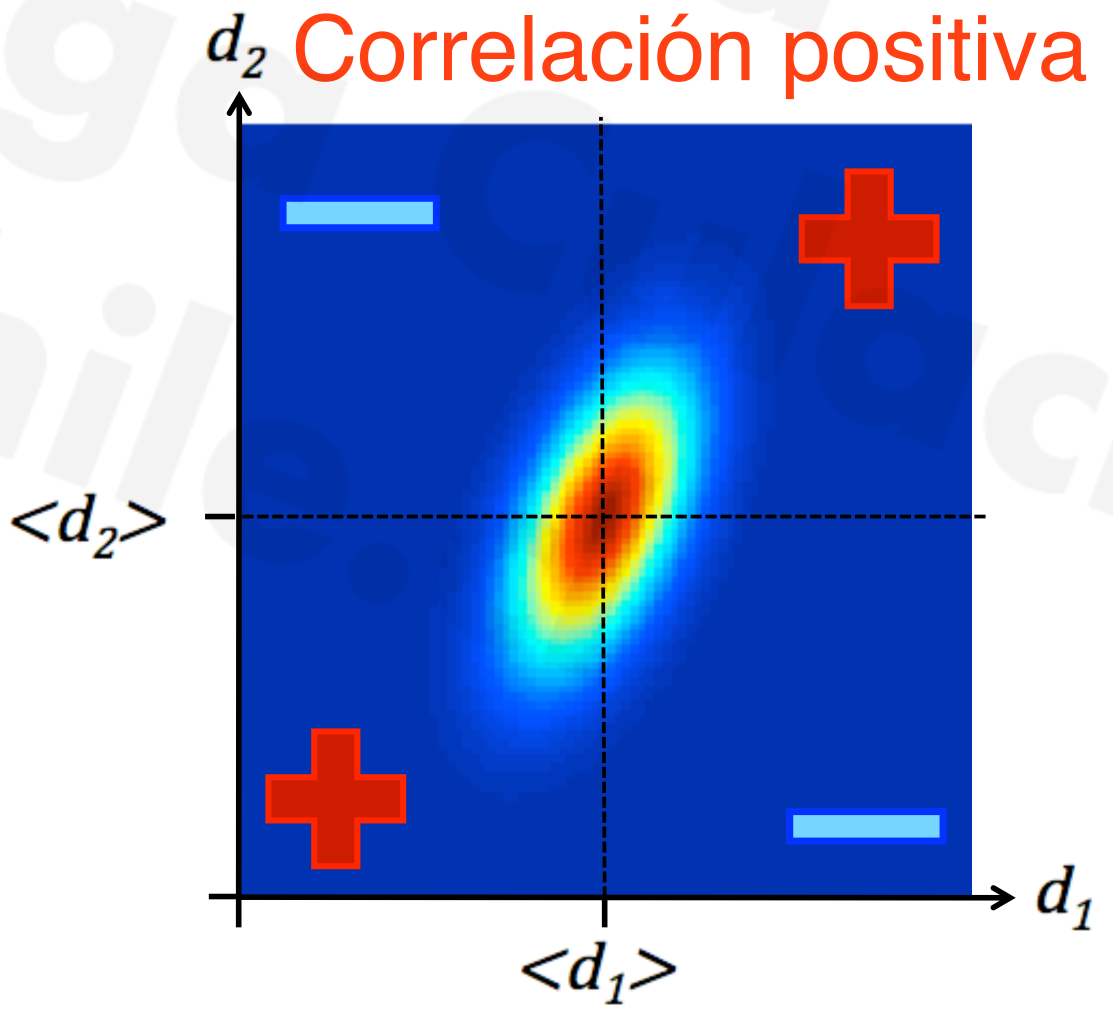
$$q(d_1, d_2) = (d_1 - < d_1 >)(d_2 - < d_2 >)$$



Siguiendo la misma idea que para el cálculo de la varianza, se multiplica la FDP por esta función, para obtener una estimación de la correlación de las variables aleatorias.







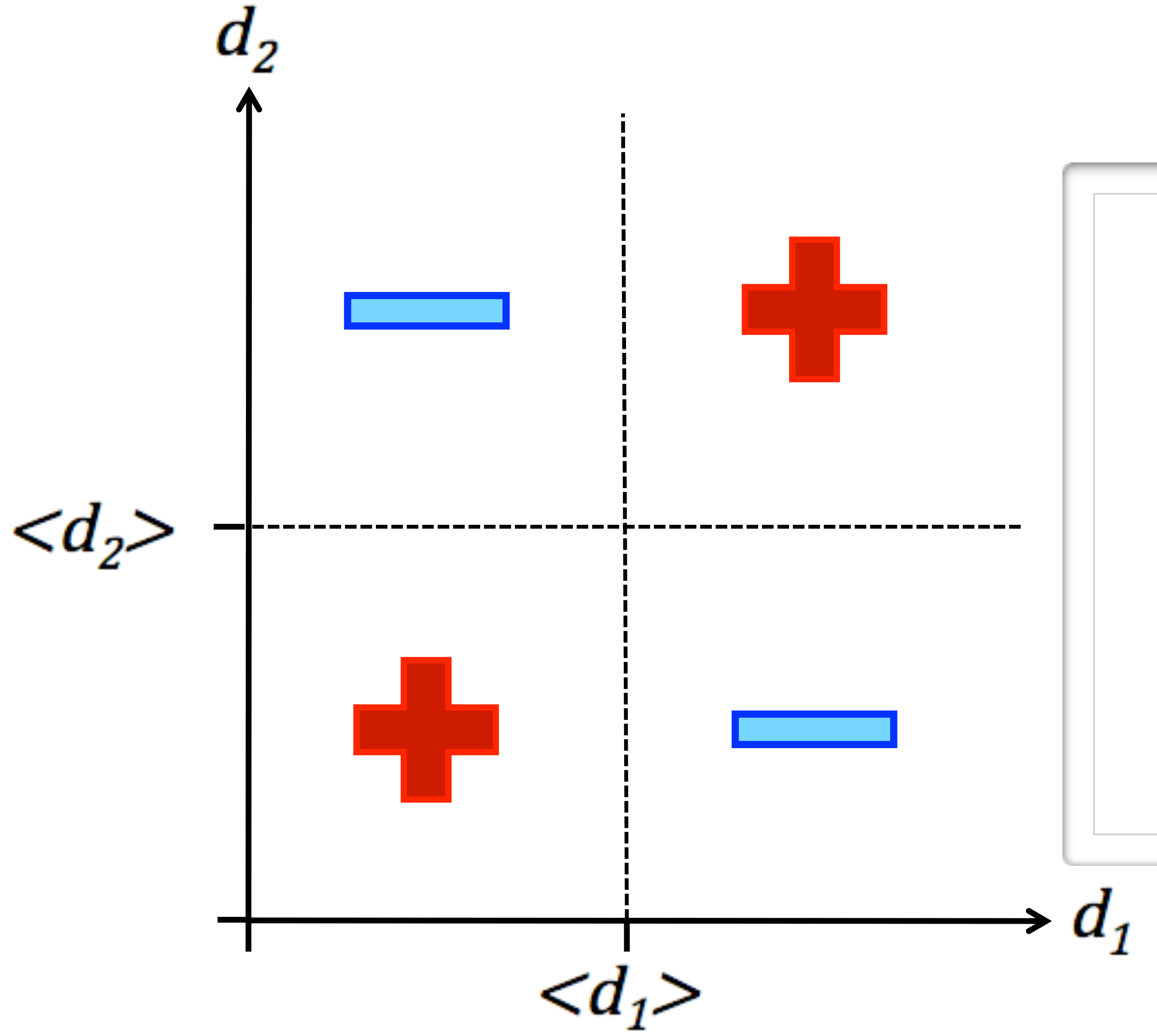
Relaciones entre dos o más variables aleatorias

Variables aleatorias correlacionadas (dependientes)
Fórmula para calcular la covarianza

(en el caso de una FDP de dos variables aleatorias, $p(d_1, d_2)$)

$$\mathrm{cov}(d_1, d_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [d_1 - \langle d_1 \rangle]\,[d_2 - \langle d_2 \rangle]\, p(\mathbf{d})\, \mathrm{d}d_1\, \mathrm{d}d_2$$

$$q(d_1, d_2) = (d_1 - < d_1 >)(d_2 - < d_2 >)$$



Covarianza positiva => correlación positiva - un valor de d_1 mayor que $<d_1>$ esta asociado a un valor de d_2 mayor que $<d_2>$

Covarianza negativa => correlación negativa - un valor de d_1 mayor que $<d_1>$ esta asociado a un valor de d_2 menor que $<d_2>$

Promedio y Covarianza para una FDP conjunta de N variables

El promedio o valor esperado se puede escribir como un vector de N elementos

$$\langle d \rangle_i = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} d_i \, p(\mathbf{d}) \, \mathrm{d}d_1 \cdots \mathrm{d}d_N$$

Covarianza se puede escribir como una matriz simétrica de NxN elementos

$$[\operatorname{cov} \mathbf{d}]_{ij} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} [d_i - \langle d_i \rangle] \, [d_j - \langle d_j \rangle] \, p(\mathbf{d}) \, \mathrm{d}d_1 \cdots \mathrm{d}d_N$$

Los elementos de la diagonal son las varianzas de cada variable aleatoria

Los elementos fuera de la diagonal son las covarianzas entre las diferentes variables aleatorias

Covarianza obtenida de un conjunto de K realizaciones de N variables aleatorias

$$[\text{cov}\,\mathbf{d}]_{ij}^{est} = \frac{1}{K}\sum_{k=1}^{K}\left(D_{ki} - \langle D_i\rangle^{est}\right)\left(D_{kj} - \langle D_j\rangle^{est}\right)$$

$$D_{ki} = \text{realización } k \text{ del tipo de dato } i$$

en Matlab, Covarianza = cov(D)
en Python 3, Covarianza = scipy.cov(D.T)

Ejemplo: En la FDP Gaussiana o Normal multivariada el vector de valores esperados o promedio y la matriz de covarianza aparecen directamente como parámetros de la distribución

$$p(\mathbf{d}) = \frac{1}{(2\pi)^{\frac{N}{2}} \ |[\mathrm{cov}\ \mathbf{d}]|^{½}} \ \exp\left(-½\,[\mathbf{d} - \langle\mathbf{d}\rangle]^{\mathrm{T}}\,[\mathrm{cov}\ \mathbf{d}]^{-1}\,[\mathbf{d} - \langle\mathbf{d}\rangle]\right)$$

Funciones de variables aleatorias

Funciones de variables aleatorias

Tenemos una variable aleatoria **d** con FDP conjunta p(**d**)

Un modelo con el cual procesamos la variable aleatoria definido por la relación **m** = **f**(**d**)

¿Cuál es la FDP conjunta p(**m**) de la variable aleatoria **m**?

Observaciones con errores asociados (errores observacionales, instrumentales, etc)

↓

proceso de análisis y modelamiento de datos

↓

Inferencias de parámetros con una incerteza asociada

¿ p(**d**) $\longrightarrow$ **m** = **f**(**d**) $\longrightarrow$ p(**m**) ?

Funciones Lineales de variables aleatorias

¿ $p(\mathbf{d}) \longrightarrow \mathbf{m} = \mathbf{f}(\mathbf{d}) \longrightarrow p(\mathbf{m})$?

Ejemplo de una FDP Gaussiana (Normal) multivariada

Observaciones con errores asociados (errores observacionales, instrumentales, etc)

$$\mathbf{m} = \mathbf{M}\mathbf{d} + \mathbf{v}$$

relación lineal

parámetros de modelos **m**

Funciones de variables aleatorias

¿ p(**d**) → **m** = **f**(**d**) → p(**m**) ?

Ejemplo de una FDP Gaussiana (Normal) multivariada

Observaciones con errores asociados (errores observacionales, instrumentales, etc)

$$\mathbf{m} = \mathbf{M}\mathbf{d} + \mathbf{v}$$

relación lineal

parámetros de modelos **m**

dada la FDP de la Normal multivariada

$$p(\mathbf{d}) = \frac{1}{(2\pi)^{\frac{N}{2}} \;|[\mathrm{cov}\;\mathbf{d}]|^{½}} \;\exp\left(-½\,[\mathbf{d} - \langle\mathbf{d}\rangle]^{T}\,[\mathrm{cov}\;\mathbf{d}]^{-1}\,[\mathbf{d} - \langle\mathbf{d}\rangle]\right)$$

y la transformación lineal

$$\mathbf{m} = \mathbf{M}\mathbf{d} + \mathbf{v}$$

¿ Cuál es la FDP p(**m**) ?

Funciones de variables aleatorias

¿ p(**d**) → **m** = **f**(**d**) → p(**m**) ?

Ejemplo de una FDP Gaussiana (Normal) multivariada

Observaciones con errores asociados (errores observacionales, instrumentales, etc)

↓

$$\mathbf{m} = \mathbf{M}\mathbf{d} + \mathbf{v}$$

relación lineal

↓

parámetros de modelos **m**

dada la FDP de la Normal multivariada

$$p(\mathbf{d}) = \frac{1}{(2\pi)^{\frac{N}{2}} \; |[\mathrm{cov}\,\mathbf{d}]|^{½}} \; \exp\left(-½\,[\mathbf{d} - \langle\mathbf{d}\rangle]^{T}\,[\mathrm{cov}\,\mathbf{d}]^{-1}\,[\mathbf{d} - \langle\mathbf{d}\rangle]\right)$$

y la transformación lineal

$$\mathbf{m} = \mathbf{M}\mathbf{d} + \mathbf{v}$$

¿ Cuál es la FDP p(**m**) ?

$$p(\mathbf{m}) = \frac{1}{(2\pi)^{\frac{M}{2}} \; |[\mathrm{cov}\,\mathbf{m}]|^{½}} \exp\left(-½[\mathbf{m} - \langle\mathbf{m}\rangle]^{T}[\mathrm{cov}\,\mathbf{m}]^{-1}[\mathbf{m} - \langle\mathbf{m}\rangle]\right)$$

donde

$$\langle\mathbf{m}\rangle = \mathbf{M}\,\langle\mathbf{d}\rangle + \mathbf{v}$$

$$[\mathrm{cov}\,\mathbf{m}] = \mathbf{M}\,[\mathrm{cov}\,\mathbf{d}]\,\mathbf{M}^{T}$$

Funciones de variables aleatorias

¿ p(**d**) → **m** = **f**(**d**) → p(**m**) ?

Ejemplo de una FDP Gaussiana (Normal) multivariada

Observaciones con errores asociados (errores observacionales, instrumentales, etc)

$$\mathbf{m} = \mathbf{M}\mathbf{d} + \mathbf{v}$$

relación lineal

parámetros de modelos **m**

Para una relación lineal entre observaciones **d** y parámetros del modelo **m** tenemos que si la FDP de las observaciones es Gaussiana, la de los parámetros también lo es.

Además obtenemos también una fórmula para la propagación de errores (de las observaciones a los parámetros del modelo)

$$[\mathrm{cov}\ \mathbf{m}] = \mathbf{M}\,[\mathrm{cov}\ \mathbf{d}]\,\mathbf{M}^{\mathrm{T}}$$

Funciones de variables aleatorias

Ejemplo: cálculo del promedio de N variables aleatorias independientes y siguiendo una distribución Gaussiana con igual varianza

Las N variables aleatorias (d_i) tienen varianza igual a σ_d^2

Fórmula para el promedio muestral

$$m_1 = \frac{1}{N}\sum_{i=1}^{N} d_i$$
$$= \frac{1}{N}[1, 1, 1, \ldots, 1]\mathbf{d}$$
$$= \mathbf{M}\mathbf{d}$$

$$\mathbf{M} = \frac{1}{N}[1, 1, 1, \ldots, 1]$$

las N variables aleatorias d_i tienen igual varianza y son indepencientes

$$[\mathrm{cov}\,\mathbf{d}\,] = \sigma_d{}^2\mathbf{I}$$

Luego usando la fórmula para la propagación del error

$$[\mathrm{cov}\,\mathbf{m}] = \mathbf{M}\,[\mathrm{cov}\,\mathbf{d}]\,\mathbf{M}^T$$

$$[\mathrm{cov}\,\mathbf{m}\,] = \sigma_d{}^2\mathrm{MM}^\mathrm{T} = \sigma_d{}^2\mathrm{N}/\mathrm{N}^2 = (\sigma_d{}^2/\mathrm{N})\mathrm{I} = \sigma_m{}^2\mathrm{I}$$

o equivalentemente

$$\sigma_m{}^2 = (\sigma_d{}^2/\mathrm{N})$$

Funciones de variables aleatorias

Ejemplo: cálculo del promedio de N variables aleatorias independientes y siguiendo una distribución Gaussiana con igual varianza

$$m_1 = \mathbf{M}\mathbf{d}$$

$$\mathbf{M} = \frac{1}{N}[1, 1, 1, \ldots, 1]$$

$$[\operatorname{cov} \mathbf{d}] = \sigma_d^2 \mathbf{I}$$

$$[\operatorname{cov} \mathbf{m}] = \mathbf{M}\,[\operatorname{cov} \mathbf{d}]\,\mathbf{M}^{\mathrm{T}}$$

$$[\operatorname{cov} \mathbf{m}] = (\sigma_d^2/N)\mathbf{I} = \sigma_m^2 \mathbf{I}$$

$$\sigma_m^2 = (\sigma_d^2/N)$$

luego el error del promedio muestral se reduce a medida de que se aumenta el número de datos

$$\sigma_m = \frac{\sigma_d}{\sqrt{N}}$$

Volviendo a Funciones de Densidad de Probabilidad

FDP conjunta (joint PDF)

$p(\mathbf{d}) = p(d_1, d_2, \ldots, d_N)$ probabilidad de que las observaciones están "cerca" de **d**

$p(\mathbf{m}) = p(m_1, m_2, \ldots, m_M)$ probabilidad de que los parámetros del modelo están "cerca" de **m**

Notar que **m** y **d** son vectores de dimensión N y M respectivamente.

$\mathbf{d} = [d_1, d_2, \ldots, d_N]^T \qquad \mathbf{m} = [m_1, m_2, \ldots, m_M]^T$

Función de Densidad de Probabilidad Marginal

La FDP marginal es una FDP de una variable formada a partir de una FDP conjunta de varias variables aleatorias. La idea es obtener el comportamiento de una de las variables de la FDP conjunta, independientemente del comportamiento de todas las otras variables aleatorias

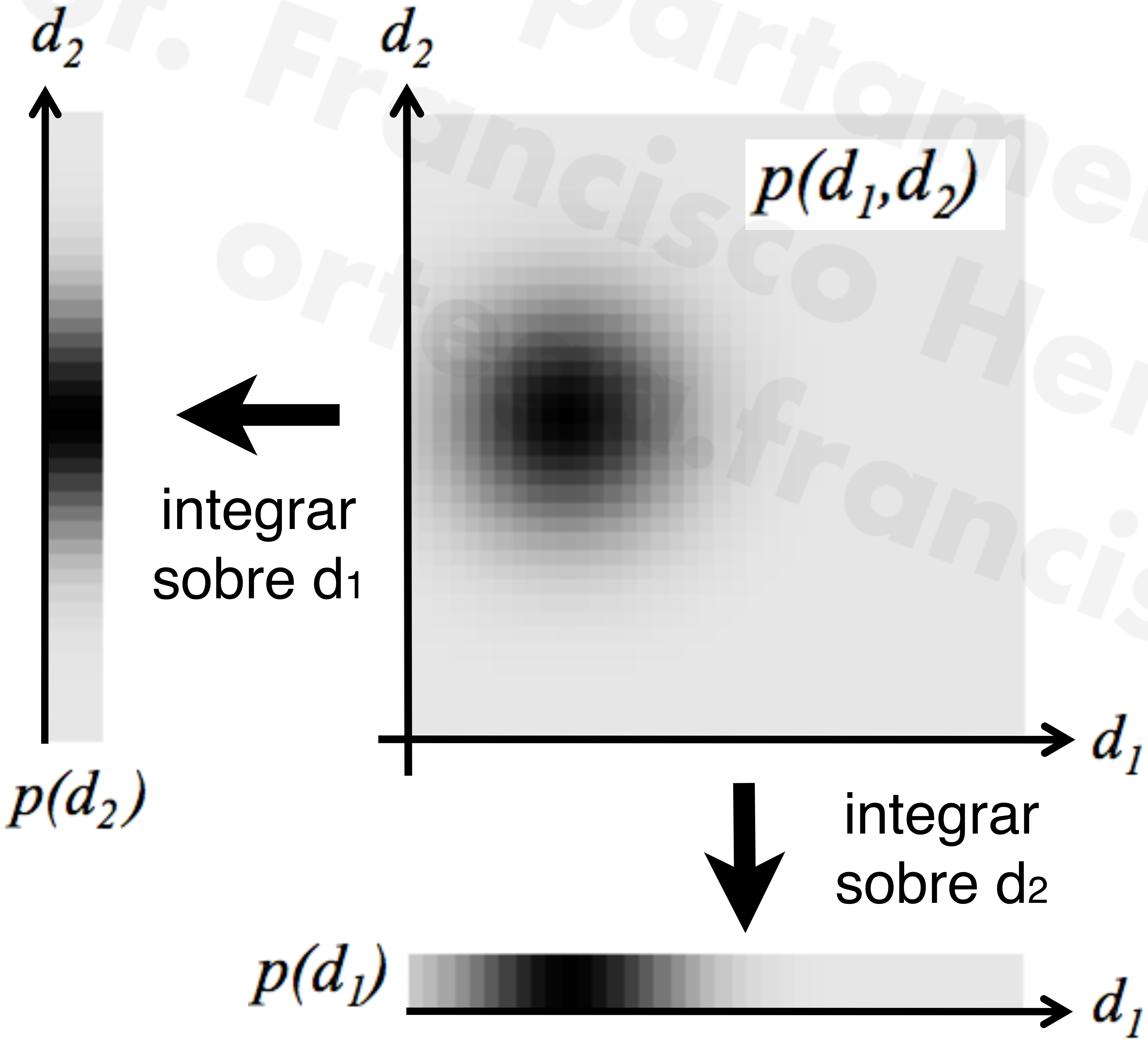
$$p(\mathbf{d}) \rightarrow p(d_i)$$

$$p(d_i) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p(\mathbf{d})\, \mathrm{d}d_j\, \mathrm{d}d_k \cdots \mathrm{d}d_l$$

(N-1 integrales)

Se integra con respecto a todas las variables aleatorias menos la variable di

Función de Densidad de Probabilidad Marginal



Ejemplo gráfico de marginalización de variables aleatorias.

La marginalización de la FDP conjunta con respecto a una variable considera el comportamiento posible de todas las otras variables aleatorias.

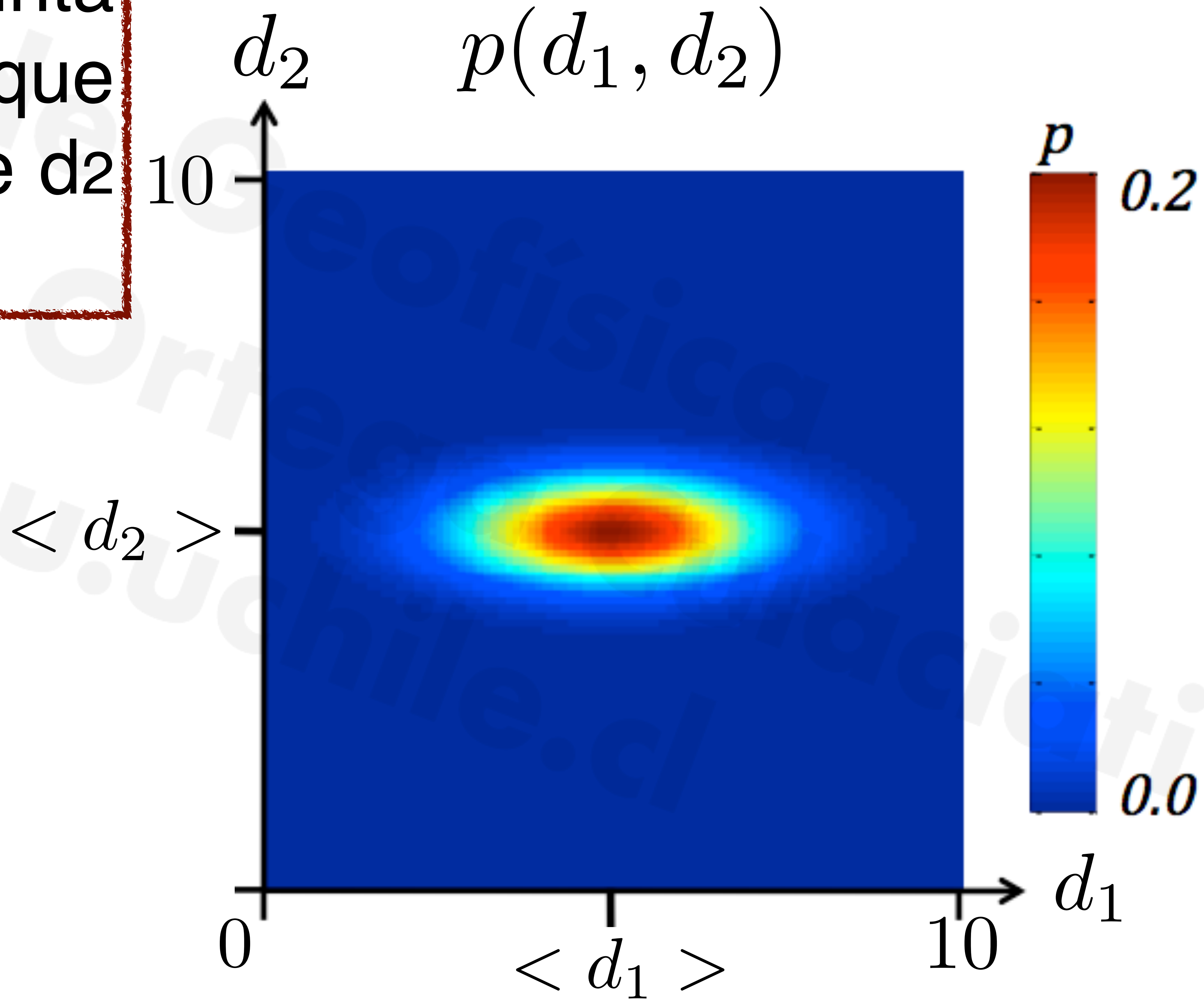
FDP Condicional y el Teorema de Bayes

FDP Condicional y el Teorema de Bayes

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1^*, d_2^*)

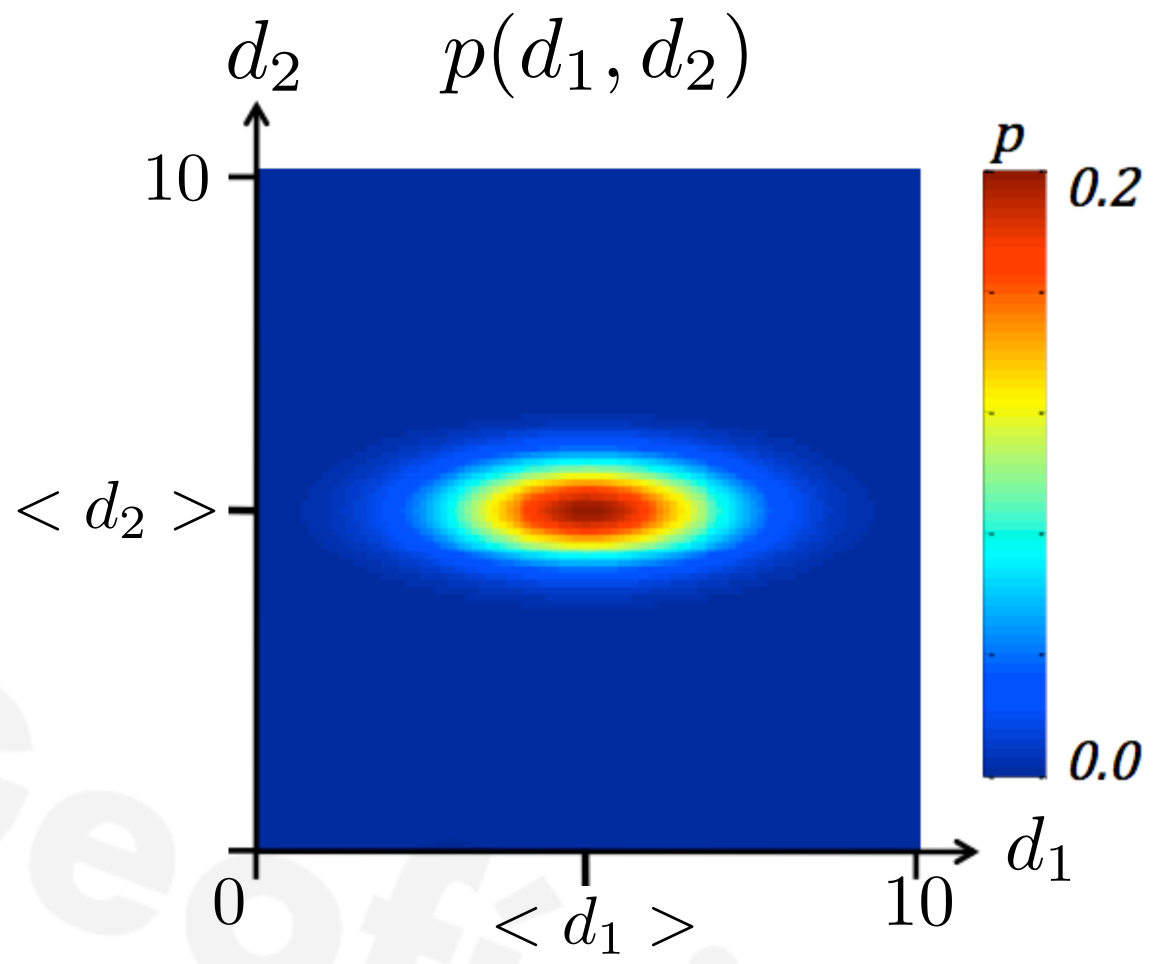
FDP conjunta $p(d_1,d_2)$

La Función de Densidad de Probabilidad Conjunta $p(d_1=d_1^*,d_2=d_2^*)$ cuantifica la probabilidad de que d_1 se ecuentre en una vecindad de d_1^* **y** de que d_2 se ecuentre en una vecindad de d_2^*.



FDP Condicional y el Teorema de Bayes

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1^*, d_2^*)



FDP conjunta $p(d_1,d_2)$

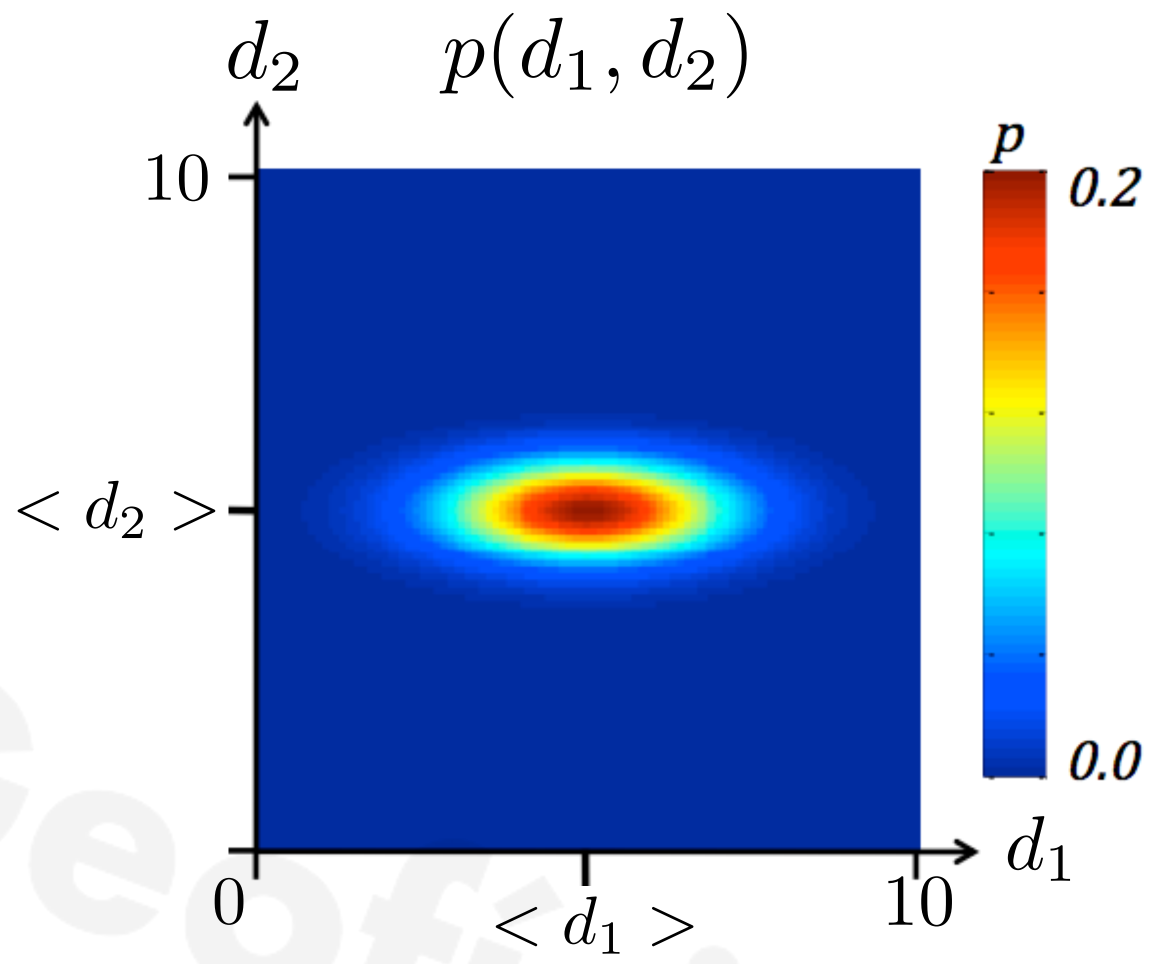
La Función de Densidad de Probabilidad Conjunta $p(d_1=d_1^*,d_2=d_2^*)$ cuantifica la probabilidad de que d_1 se ecuentre en una vecindad de d_1^* **y** de que d_2 se ecuentre en una vecindad de d_2^*.

FDP condicional $p(d_1 \mid d_2)$

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1^*, d_2^*), la Función de Densidad de Probabilidad Condicional cuantifica la probabilidad de que d_1 se ecuentre en una vecindad de d_1^* **dado que se sabe que** d_2 se secuentra en una vecindad de d_2^*.

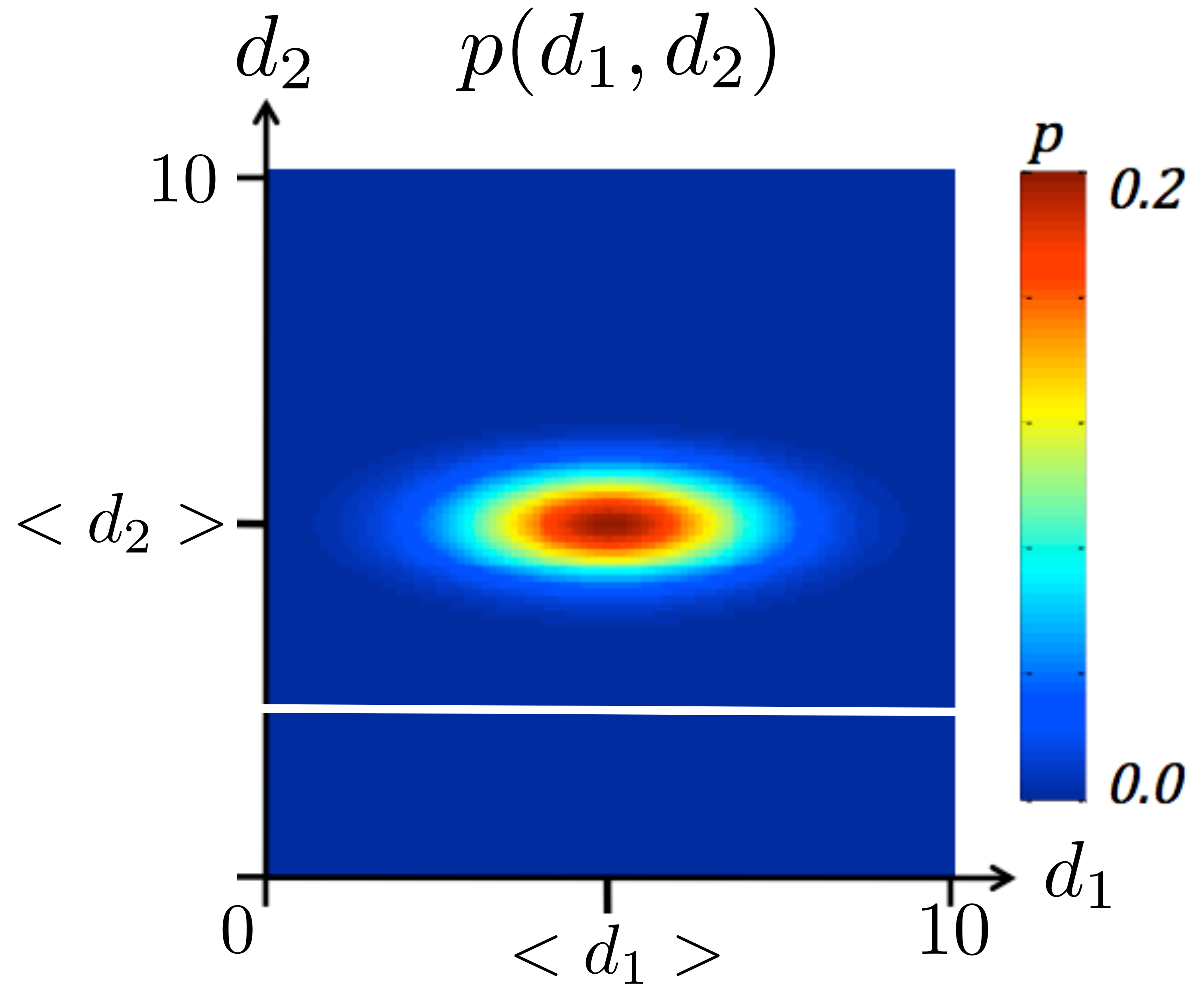
FDP Condicional y el Teorema de Bayes

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1^*, d_2^*)



FDP conjunta $p(d_1,d_2)$

La Función de Densidad de Probabilidad Conjunta $p(d_1=d_1^*,d_2=d_2^*)$ cuantifica la probabilidad de que d_1 se encuentre en una vecindad de d_1^* **y** de que d_2 se ecuentre en una vecindad de d_2^*.

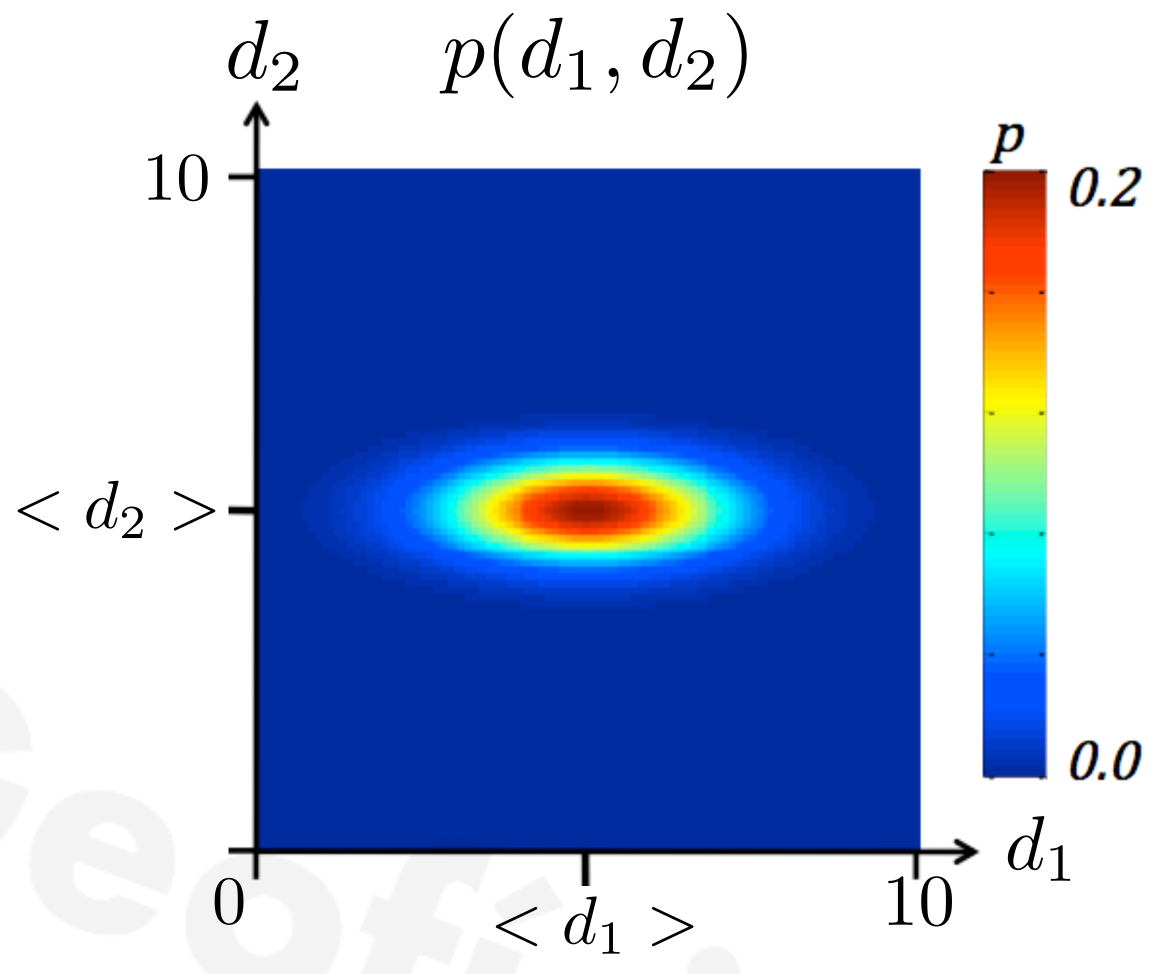


FDP condicional $p(d_1 \mid d_2)$

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1^*, d_2^*), la Función de Densidad de Probabilidad Condicional cuantifica la probabilidad de que d_1 se encuentre en una vecindad de d_1^* **dado que se sabe que** d_2 se secuentra en una vecindad de d_2^*.

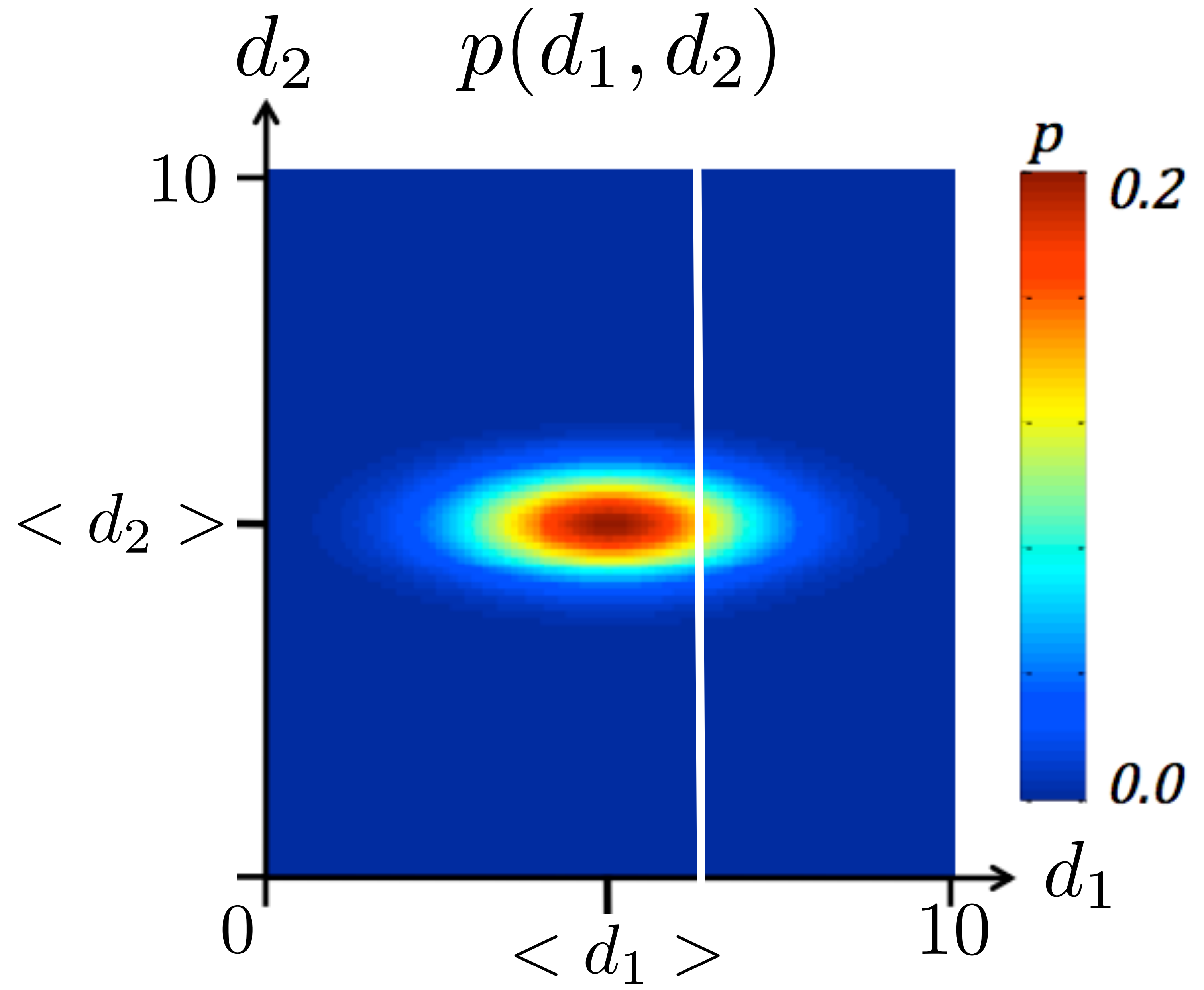
FDP Condicional y el Teorema de Bayes

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1^*, d_2^*)



FDP conjunta $p(d_1,d_2)$

La Función de Densidad de Probabilidad Conjunta $p(d_1=d_1^*,d_2=d_2^*)$ cuantifica la probabilidad de que d_1 se encuentre en una vecindad de d_1^* **y** de que d_2 se ecuentre en una vecindad de d_2^*.

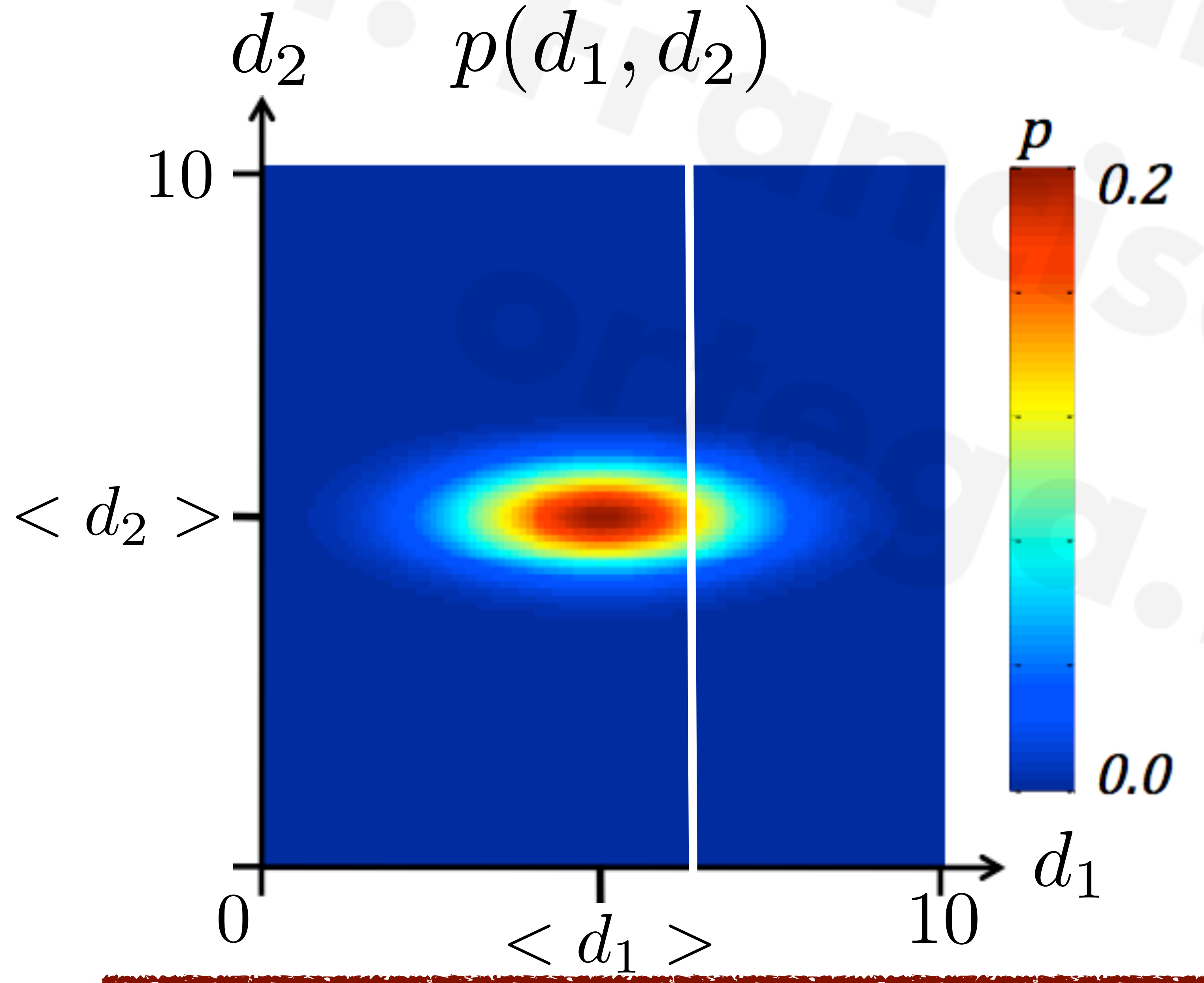


FDP condicional $p(d_1 \mid d_2)$

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1^*, d_2^*), la Función de Densidad de Probabilidad Condicional cuantifica la probabilidad de que d_1 se encuentre en una vecindad de d_1^* **dado que se sabe que** d_2 se secuentra en una vecindad de d_2^*.

FDP Condicional y el Teorema de Bayes

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1*, d_2*)



FDP condicional p(d_1 | d_2)

Definamos las variables aleatorias d_1 y d_2 y un valor dado de estas variables (d_1*, d_2*), la Función de Densidad de Probabilidad Condicional cuantifica la probabilidad de que d_1 se encuentre en una vecindad de d_1* **dado que se sabe que** d_2 se secuentra en una vecindad de d_2*.

Entonces, para convertir una FDP conjunta p(d_1,d_2) a una FDP condicional *p(d_1|d_2),* basta evaluar la FDP conjunta en d_2 y normalizar el resultado, para obtener una probabilidad condicional total igual a 1.

$$p(d_1|d_2) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\, dd_1}$$

FDP Condicional y el Teorema de Bayes

Entonces, para convertir una FDP conjunta p(d_1,d_2) a una FDP condicional $p(d_1|d_2)$, basta evaluar la FDP conjunta en d_2 y normalizar el resultado, para obtener una probabilidad condicional total igual a 1.

$$p(d_1|d_2) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\, \mathrm{d}d_1}$$

área bajo la FDP para un valor fijo de d_2

FDP Condicional y el Teorema de Bayes

Entonces, para convertir una FDP conjunta p(d_1,d_2) a una FDP condicional $p(d_1|d_2)$, basta evaluar la FDP conjunta en d_2 y normalizar el resultado, para obtener una probabilidad condicional total igual a 1.

$$p(d_1|d_2) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\, \mathrm{d}d_1} = \frac{p(d_1, d_2)}{p(d_2)}$$

área bajo la FDP para un valor fijo de d_2

probabilidad marginal de d_2

FDP Condicional y el Teorema de Bayes

Entonces, para convertir una FDP conjunta p(d_1,d_2) a una FDP condicional $p(d_1|d_2)$, basta evaluar la FDP conjunta en d_2 y normalizar el resultado, para obtener una probabilidad condicional total igual a 1.

$$p(d_1|d_2) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\, \mathrm{d}d_1} = \frac{p(d_1, d_2)}{p(d_2)}$$

de manera similar podemos derivar la FDP condicional $p(d_1|d_2)$

$$p(d_2|d_1) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\, \mathrm{d}d_2} = \frac{p(d_1, d_2)}{p(d_1)}$$

FDP Condicional y el Teorema de Bayes

$$p(d_1|d_2) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\, \mathrm{d}d_1} = \frac{p(d_1, d_2)}{p(d_2)} \qquad p(d_2|d_1) = \frac{p(d_1, d_2)}{\int p(d_1, d_2)\, \mathrm{d}d_2} = \frac{p(d_1, d_2)}{p(d_1)}$$

y obtenemos el resultado conocido como el Teorema de Bayes

$$p(d_1|d_2) = \frac{p(d_2|d_1)\, p(d_1)}{p(d_2)}$$

$$p(d_2|d_1) = \frac{p(d_1|d_2)\, p(d_2)}{p(d_1)}$$

FDP Condicional y el Teorema de Bayes

IMPORTANTE

$$p(d_1|d_2) \neq p(d_2|d_1)$$

Ejemplo:

la probabilidad de que una persona muera dado que tiene cáncer pancreático es 90%

(la tasa de fatalidad del cáncer pancreático es bastante alta)

pero

la probabilidad de que un difunto haya muerto de cáncer pancreático es 1.3%

(la mayoría de la gente muere de otras causas)

Referencias y Agradecimientos

Gran parte del contenido de este apunte esta basado en el libro de William Menke “Geophysical Data Analysis: Discrete Inverse Theory” (Elsevier) y muchas de las figuras han sido modificadas del material para el instructor proveído por el autor del libro.