# Variables aleatorias multidimensionales: caso 2D

Antonio Sala

Dept. Ing. Sistemas y Automatica (DISA)
**Universitat Politècnica de València (UPV)**

Video-presentación disponible en:

http://personales.upv.es/asala/YT/V/**va2d**.html

# Presentación

**Motivación:**

El objetivo del análisis multivariante es describir la relación entre múltiples variables aleatorias. El caso más sencillo es 2 variables.

**Objetivos:**

Comprender los conceptos básicos asociados a la distribución de probabilidad conjunta de dos variables: distribución marginal, covarianza.

**Contenidos:**

Función de densidad conjunta. Probabilidad marginal. Covarianza y correlación. Conclusiones

# Caso bidimensional $\zeta \in \mathbb{R}^2$

Consideremos una variable aleatoria bidimensional

$$\zeta := (x, y) \in \Omega := \mathbb{X} \times \mathbb{Y} \subseteq \mathbb{R}^2$$

con una función de densidad $f : \Omega \mapsto \mathbb{R}^+$ tal que $\int_{\mathbb{X}} \int_{\mathbb{Y}} f(x, y)\, dxdy = 1$.

**Ejemplo:** distribución normal bidimensional

$$f(x, y) = A e^{-\frac{1}{2}(a(x-x_0)^2 + 2b(x-x_0)(y-y_0) + c(y-y_0)^2)}$$

$$f(\zeta) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}(\zeta-\zeta_0)^T \Sigma^{-1}(\zeta-\zeta_0)}$$

# Distribución de probabilidad Marginal

Consideradas como variables "separadas", si sólo prestamos atención a una de ellas, $x$ ($y$), esta variable tiene una **densidad marginal** dada por

$$f_x(x) := \int_{\mathbb{Y}} f(x,y)dy, \qquad f_y(y) := \int_{\mathbb{X}} f(x,y)dx.$$

En efecto $\int_{\mathbb{X}} f_x(x) = \int_{\mathbb{X}} \left(\int_{\mathbb{Y}} f(x,y)dy\right) dx = \int_{\mathbb{X}} \int_{\mathbb{Y}} f(x,y)\, dxdy = 1$

- **Medias marginales:**

$$E(x) := \int_{\mathbb{X}} x f_x(x)dx = \int_{\mathbb{X}} \int_{\mathbb{Y}} x \cdot f(x,y)\, dxdy,$$
$$E(y) := \int_{\mathbb{Y}} y f_y(y)dy = \int_{\mathbb{X}} \int_{\mathbb{Y}} y \cdot f(x,y)\, dxdy.$$

- **Varianzas marginales**:

$$\sigma_x^2 := \int_{\mathbb{X}} \int_{\mathbb{Y}} (x - E(x))^2 f(x,y)dxdy, \qquad \text{similar con } \sigma_y^2.$$

# Ejemplo:

Una empresa fabrica balones con la siguiente distribución de "color" y "tamaño":

| Tam. \ Col. | Rojo | Verde |
|---|---|---|
| 10 cm | 10% | 45% |
| 15 cm | 30% | 15% |

El par (tamaño, color) de un balón al azar es una vble. aleatoria bidimensional discreta.

La distribución marginal de la variable aleatoria 1-D "*tamaño*" es:

$$p(10cm) = 0.1 + 0.45 = 0.55, \quad p(15cm) = 0.3 + 0.15 = 0.45$$

La distribución marginal de la variable aleatoria 1-D "*color*" es:

$$p(Rojo) = 0.1 + 0.3 = 0.4, \quad p(Verde) = 0.45 + 0.15 = 0.6$$

# **Novedad:** Medidas de **interacción** entre las variables

- **Covarianza:** $\sigma_{xy} := \int_{\mathbb{X}} \int_{\mathbb{Y}} (x - E(x))(y - E(y)) f(x, y) dxdy.$
  - $\sigma_{xy} > 0$ indica que es más probable que "$x$ e $y$ estén <simultáneamente por encima> o <simultáneamente por debajo> de su media" (producto positivo de incrementos) que "una esté por encima y otra por debajo" (producto negativo de incrementos).
  - $\sigma_{xy} < 0$ indica lo contrario, que es más probable que "que una esté por encima y otra por debajo".
- **Correlación:** $r_{xy} := \dfrac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$.

  *$r_{xy}$ es adimensional, mientras que $\sigma_{xy}$ no.

  Significado similar a covarianza, pero independiente de "unidades".

**Cov. muestral:** $\sigma_{xy,N} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)$, tiende a $\sigma_{xy}$ si $N$ grande.

$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i$, $\mu_y = \frac{1}{N} \sum_{i=1}^{N} y_i$

# Conclusiones

- En variables aleatorias bidimensionales $(x, y)$, hay dos distribuciones **marginales** (mirar sólo $x$ o sólo $y$, sin tener en cuenta la existencia de la otra variable), con sus medias y varianzas "marginales".
- Como novedad, la "**interacción**" entre ambas variables puede medirse mediante la **covarianza** o la **correlación**.
- La idea de "*interacción*" lanza la posibilidad de "predecir una en función de la otra", en aplicaciones donde sólo una de ellas es fácilmente accesible (predecir "*temperatura mediodía mañana*" a partir de "*temperatura mediodía hoy*").
  - Necesidad de estudiar con más detalle si las variables son o no "independientes".
  - Si no son "independientes", estudiar en detalle dicha "predicción" estadística.

Esto será objetivo de otros materiales.