Curso Machine Learning

Tema 2: Probabilidad y Estadística, conceptos básicos.

Magdalena Lucini, Sebastián Filipigh, Luis Duarte

FaCENA - UNNE - 2023

Introducción

Estadística

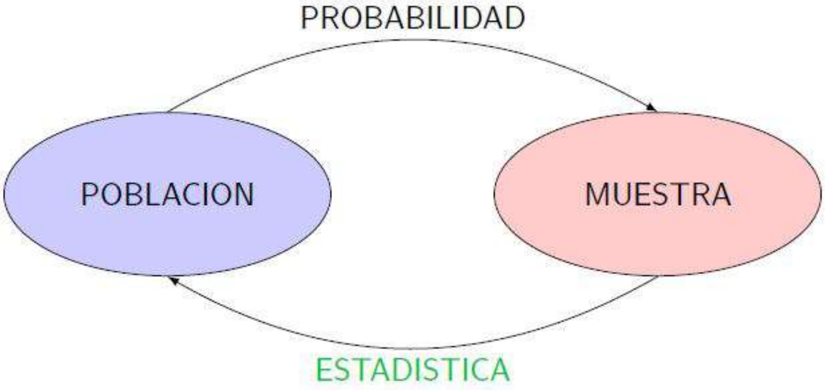
Algunas definiciones:

- Disciplina científica que se ocupa de la obtención, orden y análisis de un conjunto de datos con el fin de obtener explicaciones y predicciones sobre fenómenos observados.
- Disciplina que estudia la variabilidad, recolección, organización, análisis, interpretación y presentación de los datos, así como el proceso aleatorio que los genera siguiendo las leyes de la probabilidad
- Ciencia que se ocupa de la recolección, resumen, análisis, e interpretación de hechos o datos numéricos.

Algunos objetivos

- **Extraer conocimiento** a partir de un conjunto de datos, con el fin de **tomar decisiones** en base a la mejor información posible (siempre existe incertidumbre).
- Obtener información de una **población**, sin necesidad de estudiar todos los elementos que la componen.
- Sacar conclusiones sobre alguna característica de una **población** en base a una **muestra** respresentativa de la misma

Introducción



Introducción

Estadística Descriptiva:

- Herramientas para resumir un conjunto de datos (modelados como realizaciones de una variable aleatoria), a través de ciertas medidas numéricas. Representa la información de una manera distinta para facilitar su interpretación, pero no permite realizar predicciones o inferencias
- Muchos datos (modelados como realizaciones de alguna variable/vector aleatorio): recolectar/resumir/presentar/graficar/analizar.
- Métodos no supervisados (reducción de dimensionailidad)
- Facilita interpretación, NO permite realizar predicciones o inferencias.

Estadística Inferencial

- Usar datos disponibles para aprender, predecir, hacer generalizaciones , etc.
- Estimar, predecir, inferir, pronosticar, ajustar, etc
- Pruebas de hipótesis, regresión, clasificación, etc.

Algunos conceptos de Estadística Descriptiva

Organización y Presentación de resultados

- Representaciones tabulares (tablas): 1^{er} paso en la organización de datos, que se ordenan en filas y columnas para documentar y comunicar la información.
- Representaciones gráficas (histogramas, gráficos de barras, circulares, etc): brindan un resumen visual de los datos.
- Medidas descriptivas numéricas (variables cuantitativas): medidas de tendencia central, posición, dispersión, forma.
- Dependiendo del tipo de datos e información a comunicar se elegirá qué tipo de representación utilizar.

Ejemplo: Base de datos diabetes gestacional

Base de datos con información de un estudio observacional de factores de riesgo en la presencia de Diabetes Gestacional (DG), llevado a cabo en un grupo de 48 mujeres en el año 2018 en la ciudad de Córdoba, Argentina.

caso_control	Respuesta: Presencia o ausencia de Diab. Gestacional (DG)	0= no tiene DG 1= tiene DG
Edad	Edad en años	
estacivil	Estado civil:	1 casada 2=concubina 3= separada
Pp	Peso pregestacional (Kg)	
Pa	Peso actual (Kg)	
Talla	Estatura (m)	
enpregest	Estado nutricional pregestacional	1= bajo peso 2= peso normal 3= sobrepeso 4= obesidad
Enactual	Estado nutricional actual	1= bajo peso 2= peso normal 3= sobrepeso 4= obesidad
Obrasoc	Obra social	0= no tien 1= si tiene
Estudios	Máximo nivel de estudios alcanzados	1= primario 2= secundario 3= terciario o universitario
Estrsoc	Estrato social	1= bajo 2= medio 3=alto
Actfpre	Actividad física pregestacional	0=NO 1=Si
Antdb	Antecedentes familiares de diabetes	0=NO 1= Si
Fumaud	Estatus de Fumadora previo al embarazo	0=NO 1= Si
Eg	Edad estacional (semanas de gestación)	
comidas_dia	Cantidad de comidas realizadas al día	
Vet	Valor energético total de la dieta (calorías/día)	
Fe	Hierro dietario (mg/día)	
Calico	Calcio dietario (mg/día)	
Cho	Carbohidratos dietarios (g/día)	
beb_azucar	Consumo de bebidas azucaradas (cc/día)	

Base de datos diabetes gestacional

	A	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	i	caso_control	edad	eg	estacivil	pp	pa	talla	IMC_pg	IMC_actual	enpregest	enactual	obrasoc	estudios	estrsoc	actfpre	antdb	fumaud	comidas_d	vet	fe	calcio	cho	beb_azucar
2		Tiene DG	30	35	casada	70,00	78,00	1,60	27,34	30,47	3	2	1	2	Medio	0	1	0	5	2243,34	16,29	621,92	450,00	987
3		No tiene DG	31	28	casada	58,50	63,00	1,65	21,49	23,14	2	2	1	3	Alto	1	1	1	4	2449,85	17,83	621,79	223,04	114,29
4		No tiene DG	33	36	casada	55,00	66,00	1,53	23,50	28,19	2	2	1	3	Alto	1	0	0	3	2321,03	16,96	426,21	310,97	228,57
5		Tiene DG	30	28	Concubina	83,00	90,00	1,60	32,42	35,16	4	3	1	3	Bajo	0	1	1	4	5034,36	35,96	1134,84	483,74	914,29
6		No tiene DG	32	35	casada	79,00	86,00	1,60	30,86	33,59	4	3	1	2	Bajo	0	0	1	2	2546,82	19,28	355,20	197,48	500
7		No tiene DG	31	39	Concubina	61,00	75,00	1,60	23,83	29,30	2	2	1	1	Alto	1	0	0	5	2314,75	13,21	772,41	265,15	342,86
8		Tiene DG	21	38	casada	80,00	88,00	1,65	29,38	32,32	3	3	1	1	Bajo	0	1	0	2	3662,79	26,34	721,94	391,02	320
9		No tiene DG	23	34	casada	49,00	54,00	1,56	20,13	22,19	2	1	1	2	Bajo	0	1	0	4	2251,56	16,92	594,60	266,94	120
10		No tiene DG	18	37	casada	49,00	64,00	1,73	16,37	21,38	1	1	1	2	Bajo	1	1	0	3	2106,17	19,77	635,10	235,01	85,71
11		Tiene DG	25	33	casada	60,00	53,00	1,58	24,03	21,23	2	1	1	1	Bajo	0	1	1	4	2540,21	14,23	854,70	304,35	914
12		No tiene DG	23	37	casada	56,00	65,00	1,55	23,31	27,06	2	2	1	2	Medio	0	1	0	3	3522,14	17,67	1180,48	367,18	400
13		No tiene DG	26	37	casada	57,00	72,00	1,62	21,72	27,43	2	2	1	2	Medio	0	1	0	4	2082,93	11,17	791,40	242,73	571,43
14		Tiene DG	27	37	casada	91,00	100,00	1,73	30,41	33,41	4	3	1	3	Alto	0	1	0	4	2496,39	24,55	876,59	311,40	875
15		No tiene DG	29	28	casada	48,00	55,50	1,60	18,75	21,68	2	1	1	2	Medio	1	0	0	5	3575,05	19,56	790,66	368,00	428,57
16		No tiene DG	29	39	casada	55,00	63,00	1,59	21,76	24,92	2	2	1	2	Medio	0	1	0	4	3698,53	14,92	952,02	307,74	668
17		Tiene DG	24	30	casada	62,00	70,50	1,50	27,56	31,33	3	3	0	2	Bajo	1	1	1	3	2916,38	13,06	978,96	241,45	457,14
18		No tiene DG	26	37	casada	58,00	74,00	1,67	20,80	26,53	2	2	0	3	Alto	0	1	1	4	2994,17	22,68	266,88	301,74	0
19		No tiene DG	21	38	casada	68,00	80,00	1,60	26,56	31,25	3	2	0	2	Medio	1	1	0	3	3550,62	29,40	937,04	369,79	235
20		Tiene DG	32	39	casada	68,00	78,00	1,73	22,72	26,06	2	2	1	2	Medio	0	1	1	4	3212,56	24,87	399,52	383,47	800

Generalidades

En general, una base de datos es del tipo:

$$\mathbf{x} = \begin{bmatrix} x_1(t_1) & \dots & x_p(t_1) \\ \vdots & \ddots & \vdots \\ x_1(t_n) & \dots & x_p(t_n) \end{bmatrix}$$

- x_i= variables, $i = 1 \dots p$ (Cualquier característica susceptibles de tomar distintos estados entre unidades elementales, o que varían dentro de una misma unidad elemental a través del tiempo)
- t_j = tiempo, observaciones, realizaciones, etc. $j = 1, .., n$

Es usual agupar los elementos de **x** por

- Filas (clustering)
- Columnas (correlación)

Estadística descriptiva:Tipos de variables

Cualitativas

Expresan una cualidad o propiedad que el objeto en estudio tiene o no, o bien lo tiene en distinto grado. Pueden ser **DICOTÓMICAS**:dos categorías o clases (ej: género al nacer) ó **POLICOTÓMICAS**:más de dos categorías (ej: nivel de estudios alcanzado)

Cuantitativas

Asumen valores numéricos. Expresan una cantidad.

- **Discretas**: Surgen de contar. Son aquellas que sólo toman valores discretos dentro de su campo de variación (ej: cantidad de materias aprobadas)
- **Continuas**: Surgen de medir. Toman cualquier valor dentro de su rango de variación (ej: altura de un alumno).

Algunos gráficos variables cualitativas:

Gráfico de sectores

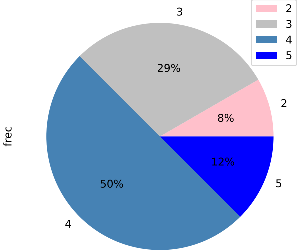


Figura: Número de comidas al día

Gráfico de barras

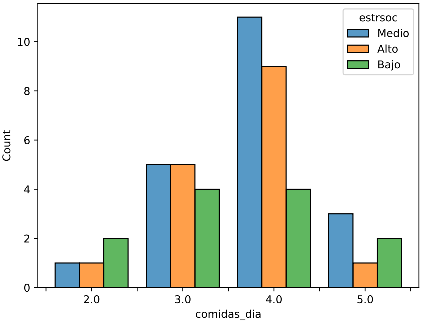
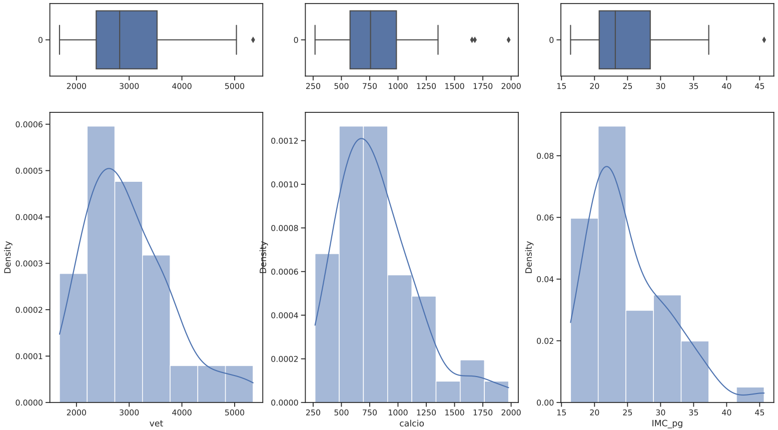
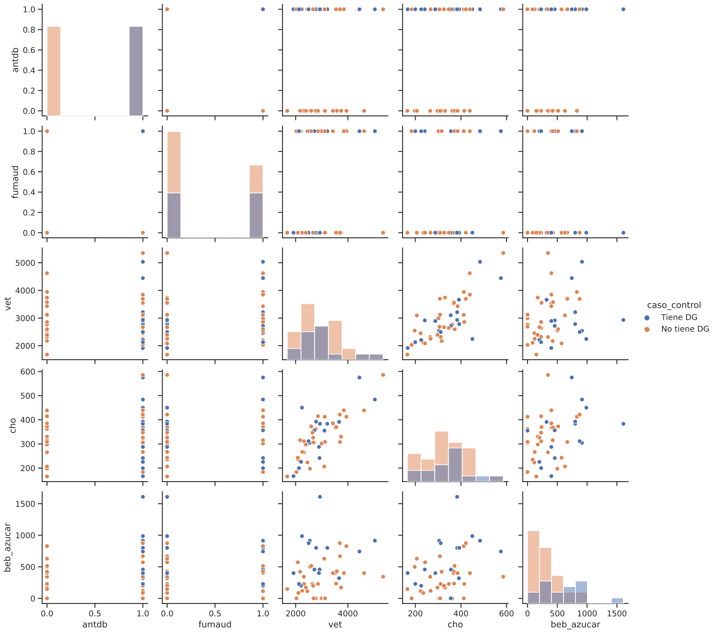


Figura: Número de comidas al día, discriminados por estrato social

Algunos gráficos variables cuantitativas: Histograma y boxplot



Gráficos


antdb
fumaud
vet
cho
beb_azucar
caso_control
Tiene DG
No tiene DG
0.0
0.5
1.0
2000
4000
200
400
600
0
500
1000
1500
0.2
0.4
0.6
0.8
3000
5000
300
500

Medidas descriptivas numéricas - Caso univariado

- Medidas de Tendencia Central Valores numéricos que se obtienen de variables cuantitativas y cuyos resultados se localizan por el centro de la distribución. Ej: Media (promedio aritmético), Mediana, Moda.
- Medidas de Posición: Valores numéricos que permiten dividir la distribución de datos en partes iguales. Ej: Cuartiles, Deciles, Percentiles.
- Medidas de Dispersión: Valores numéricos que proporcionan una idea sobre cuan esparcidos o concentrados están los datos correspondientes a una variable. Ej: Rango, Rango intercuartílico, varianza, desviación estandar, coeficiente de variación.
- Medidas de Forma: Dan una idea de la forma de la distribución de la variable. Ej: Coeficiente de asimetría, de kurtosis (apuntamiento).

Media y Varianza: Caso univariado, datos sin agrupar

X variable en estudio, n tamaño de la muestra, $x_1; \ldots; x_n$ valores que asume la variable,

- Media: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- Varianza: $Var(x) = s_x^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
- Desviación estándar: $\sqrt{s_x^2}$

Caso multivariado: Matriz de Covarianza

$$\mathbf{S} = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{x}-\bar{\mathbf{x}})^T(\mathbf{x}-\bar{\mathbf{x}})$$

donde :

Matriz de medias

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 & \dots & \bar{x}_p \\ \vdots & \ddots & \vdots \\ \bar{x}_1 & \dots & \bar{x}_p \end{bmatrix}$$

Matriz de varianzas-covarianzas

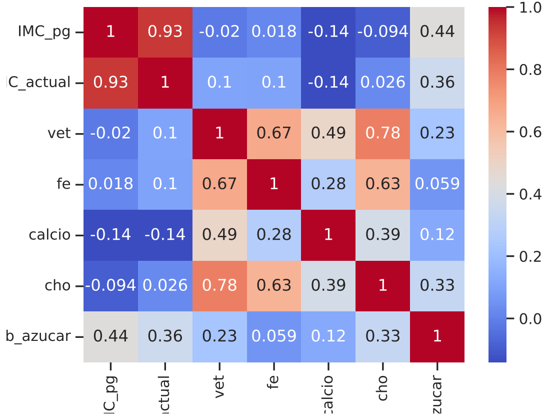
$$\mathbf{S} = \begin{bmatrix} Var(x_1) & \dots & Cov(x_1, x_p) \\ \vdots & \ddots & \vdots \\ Cov(x_p, x_1) & \dots & Var(x_p) \end{bmatrix}$$

$$Cov(x_i, x_j) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_i(t_k) - \bar{x}_i)(x_j(t_k) - \bar{x}_j)$$

De forma análoga, y teniendo en cuenta que

$Corr(x_i, x_j) = \dfrac{Cov(x_i, x_j)}{s_{x_i} s_{x_j}}$ se construye la matriz de correlación.

Matriz de correlación - heatmap



Conceptos básicos de Probabilidad

Variable aleatoria

Una Variable Aleatoria (v.a.) unidimensional es una regla o función X, que asigna a cada elemento del espacio muestral (espacio de estados) **S** un número (real),

$$X : \mathbf{S} \to \mathbb{R}_{\mathbb{X}}$$

R_X es el rango de la variable aleatoria.

- Si R_X finito o infinito numerable $\Rightarrow$ X v.a. discreta
- Si existe (a, b) intervalo de números reales $(a, b) \subset R_X$ $\Rightarrow$ X v.a. continua

Variables aleatorias

Función de distribución acumulada (cdf)

Sea $(\mathbf{S}, \mathcal{A}, P)$ un espacio de probabilidad y sea X una v.a. definida sobre dicho espacio. La función $F : \mathbb{R} \to [0, 1]$ definida por

$$F(x) = P(X \leq x)$$

es llamada Función de distribución (de probabilidades) de X

Función de probabilidad de masa (f.p.m)

Si X v.a. discreta que puede asumir los distintos valores $x_1, x_2, \ldots x_n$ (o $x_1, x_2, \ldots$ si X puede asumir una cantidad infinita numerable de valores distintos) .

La función de probabilidad de masade X es una función $p : I \subseteq \mathbb{R} \to [0, 1]$ definida como $p(x_i) = P(X = x_i) = P(\{\omega \ \in \mathbf{S} : X(\omega) = x_i\})$

Función de densidad de probabilidades (pdf)

Si X v.a continua , $F_X(x) = P(X \leq x)$ su función de distribución acumulada, y si existe una función $f(x) \geq 0$ tal que $F_X(x) = \int_{-\infty}^{x} f(s)ds$, $\forall x \in \mathbb{R}$, entonces f recibe el nombre de función de densidad (de probabilidades) de la v.a. X.

Algunas distribuciones discretas

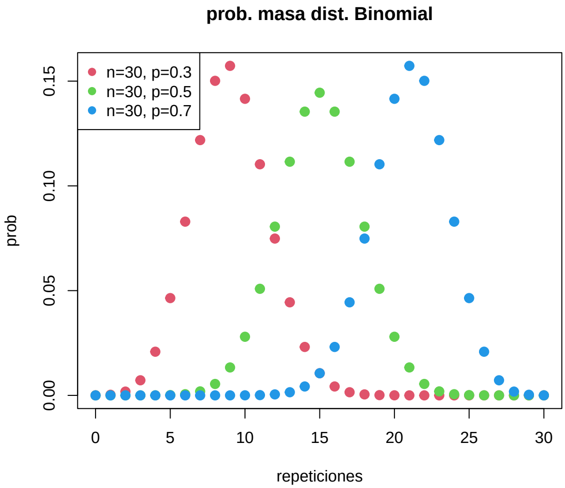
Distribución Binomial

$X \sim \mathcal{B}(n, p)$

X = número de éxitos observados en las n repeticiones de un experimento binomial, p= probabilidad de éxito en cada repetición

$$p(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}$$



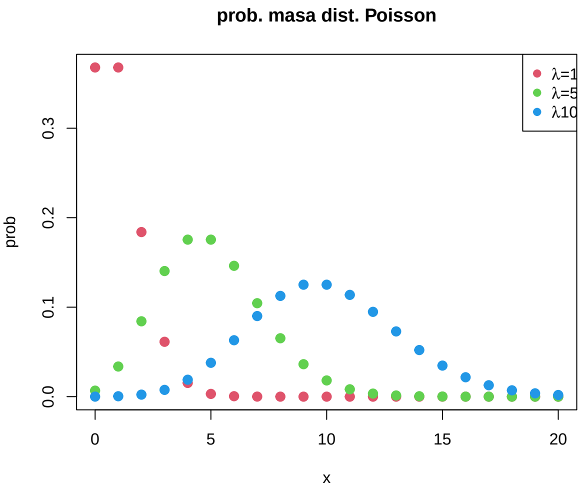
Algunas distribuciones discretas

Distribución de Poisson

X v.a que cuenta el número de eventos que ocurre en un intervalo de tiempo (espacio, volumen) $\Rightarrow X$ tiene distribución de Poisson de parámetro λ , λ es la tasa media de ocurrencia del evento en un intervalo de tiempo, (espacio, volumen, etc.)

$$P(X = r) = \frac{\lambda^r}{r!} e^{-\lambda}, r = 0, 1, 2, \ldots$$

$$E(X) = \lambda, \quad Var(X) = \lambda$$



Distribuciones continuas

Distribución Gaussiana (Normal)

También conocida como distribución gaussiana, es una de las distribuciones más comunmente utilizadas por físicos, químicos e ingenieros ya que

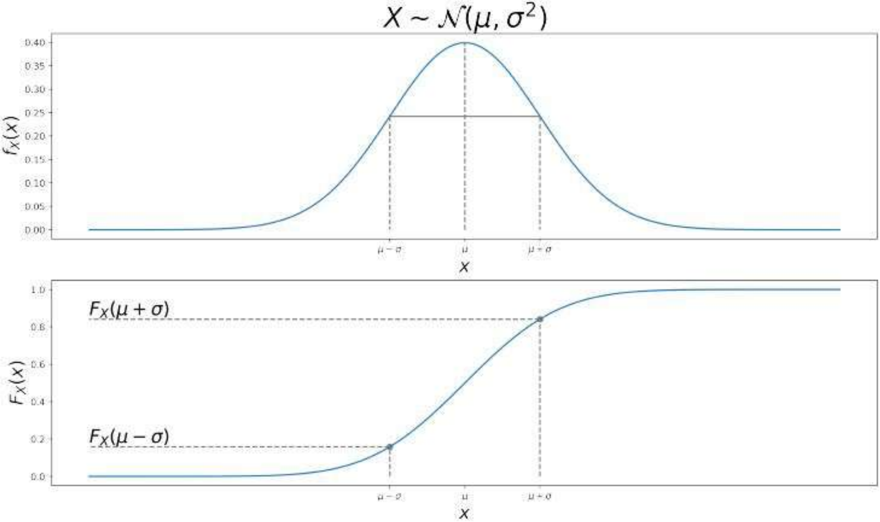
- Si repite un experimento una cierta cantidad de veces entonces la variable que representa el promedio de los resultados tiene aproximadamente una distribución normal.
- Aparece en el estudio de numerosos fenómenos físicos (por ejemplo: velocidad de moléculas (Maxwell))
- Se la denomina “normal” porque representa un gran número de fenómenos en la naturaleza

Distribución Normal

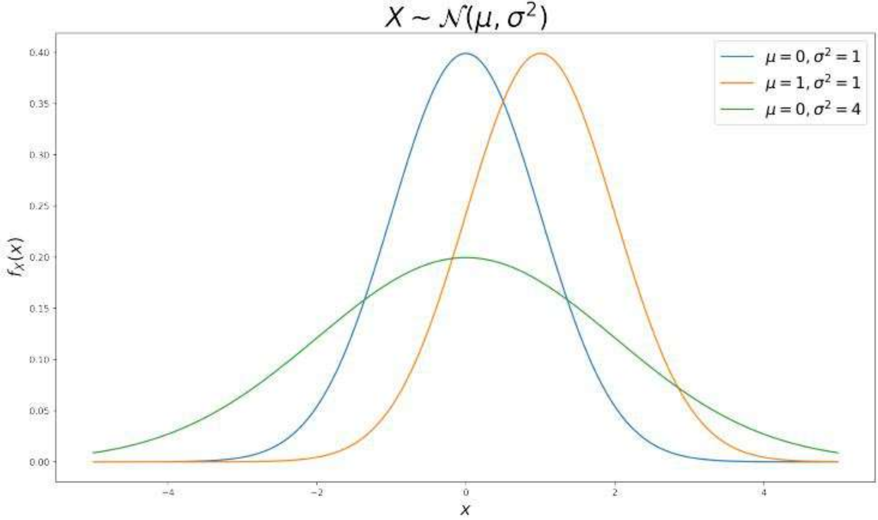
Si X es una v.a. con Distribución Normal con media μ y desviación estándar σ, entonces $X \sim N(\mu, \sigma^2)$

- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty$
- $F(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\mathrm{x}-\mu}{\sigma\sqrt{2}}\right)\right)$, con $\mathrm{erf(x)} = \frac{2}{\pi}\int_0^{\mathrm{z}} \mathrm{e}^{-\mathrm{t}^2}\mathrm{dt}$
- $E(X) = \mu$
- $V(X) = \sigma^2$

Distribución Normal



Distribución Normal



Estadística Inferencial: Estimación de parámetros
Máxima Verosimilitud (estimación puntual)

<u>Definiciones:</u>
Sean $x_1, \cdots, x_n$ realizaciones de $X_1, \cdots, X_n$, muestra aleatoria de una v.a. con función de densidad (o probabilidad de masa) $f(x)$ que involucra a algún parámetro θ (la denotaremos $f(x;\theta)$).

1. Se define la función de verosimilitud $\mathcal{L}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$
2. El estimador de máxima verosimilitud del parámetro θ es el valor de θ que maximiza la función $\mathcal{L}(x_1, \ldots, x_n; \theta)$

Distribución Gaussiana univariada: estimación de μ y σ^2 por MV

Sean $x_1, \cdots, x_n$ observaciones de $X_1, \cdots, X_n$ muestra aleatoria de una v.a. con distribución normal $N(\mu, \sigma^2)$.
Recordar que $f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$
Entonces:

$$\begin{aligned}
\mathcal{L}(x_1, \ldots, x_n; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left(-(1/2\sigma^2) \sum_{i=1}^{n} (x_i-\mu)^2\right)
\end{aligned}$$

Estimación de μ y σ^2 por MV - continuación

- Los EMV se denotan por $\hat{\mu}$ y $\hat{\sigma^2}$
- $(\hat{\mu}, \hat{\sigma^2}) = \arg\max_{(\mu,\sigma^2)} \mathcal{L}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}; \mu, \sigma^2)$
- Resolver $\frac{d \ln \mathcal{L}(x_1,\ldots,x_n;\mu,\sigma^2)}{d\mu} = 0$ y $\frac{d \ln \mathcal{L}(x_1,\ldots,x_n;\mu,\sigma^2)}{d\sigma^2}$
- Resultan
  - $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
  - $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Vectores Aleatorios

Vectores Aleatorios

Sea $(\mathbf{S}, \mathcal{A}, P)$ un espacio de probabilidad y sea $n \in \mathbb{N}$. Un Vector Aleatorio n-dimensional $(X_1, X_2, \ldots, X_n)$ es una aplicación del espacio muestral **S** en $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \ldots \mathbb{R}$, es decir

$$\begin{aligned} \mathbf{S} &\to \mathbb{R} \times \mathbb{R} \ldots \times \mathbb{R} \\ \omega &\to (X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) \end{aligned}$$

Cada componente del vector aleatorio es una variable aleatoria unidimensional.

Vectores aleatorios bidimensionales

- Pueden modelarse dos v.a X e Y conjuntamente
- $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$ es la función de distribución acumulada conjunta
- $f_{XY}(x, y)$ ($f_X(x)$, $f_Y(y)$) es la densidad conjunta (marginales) de X e Y

Distribuciones Marginales

Si bien (X, Y) es un “vector aleatorio”, tanto X como Y son variables aleatorias unidimensionales, y como tales tienen su propia distribución.

Distribuciones Marginales

Dada $F_{XY}(x, y)$ función de distribución conjunta de (X, Y) se define la función de distribución marginal de X como

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \leq +\infty), \quad x \in \mathbb{R}$$

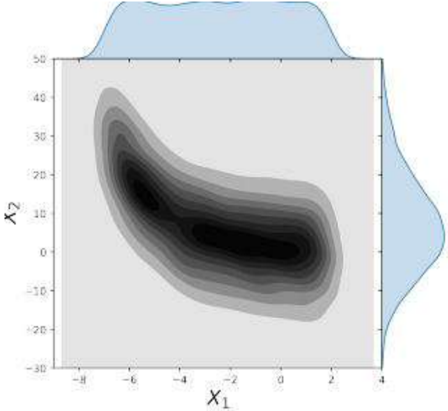
Análogamente, se define la función de distribución marginal de Y como

$$F_Y(x) = P(Y \leq y) = P(X \leq +\infty, Y \leq y), \quad y \in \mathbb{R}$$

Densidad marginal caso continuo

(X, Y) va continuo con función de densidad conjunta $f(x, y)$.
Las funciones de densidad marginales de X e Y están dadas por

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy, \ x \in \mathbb{R}, f_Y(y) = \int_{-\infty}^{\infty} f(x, y)dx \ y \in \mathbb{R}$$



Distribuciones condicionadas (continuas)

Distribuciones Condicionadas Continuas

Se definen la función de densidad de X *condicionada* al valor *y* de Y (*y* fijo) como

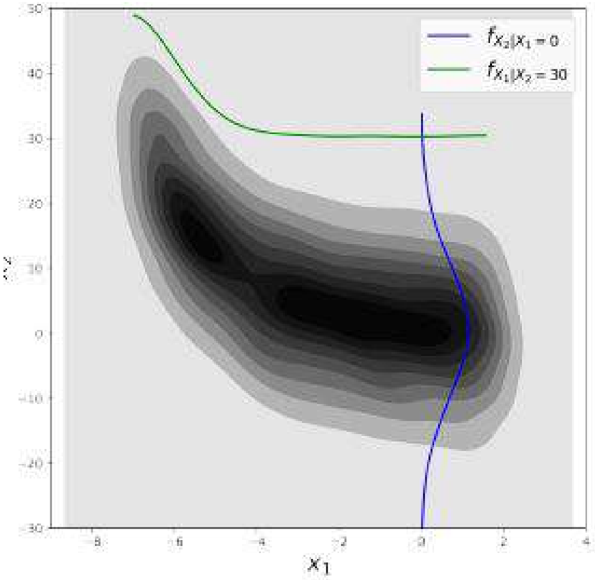
$$f_{X/Y=y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}, \quad x \in \mathbb{R}$$

y la función de densidad de Y *condicionada* al valor *x* de X (*x* fijo) como

$$f_{Y/X=x})(y) = \frac{f_{XY}(x,y)}{f_X(x)}, \quad y \in \mathbb{R}$$

donde $f_X(x)$ y $f_Y(y)$ son las densidades marginales de X e Y, respectivamente.

Distribuciones condicionadas



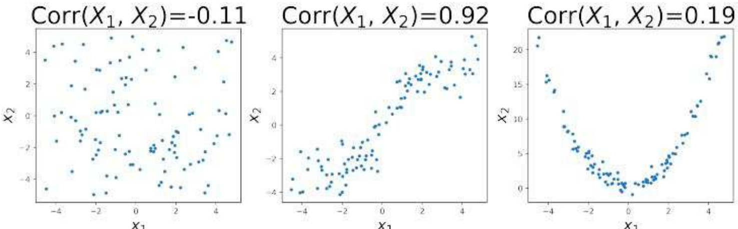
Covarianza y correlación

Covarianza entre X e Y

$$\sigma_{XY} = Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E(XY) - E(X)E(Y)$$

Correlación entre X e Y

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



Dos variables X e Y son independientes si
$f_{XY}(x, y) = f_X(x) f_Y(y)$
Observaciones:

1. Si $\sigma_{XY} = 0$ se dice que X e Y son no correlacionadas o incorreladas.
2. ρ_{XY} indica el grado de relación lineal entre las variables X e Y.
3. Si X e Y son independientes, entonces:
    - son no correlacionadas. (La recíproca no es cierta)
    - $f_{X/Y=y}(x) = f_X(x)$ y $f_{Y/X=x}(y) = f_Y(y)$
    - E(XY)=E(X)E(Y)
    - Var(aX+Y)=aVar(X)+Var(Y)

Distribución Gaussiana Bivariada $\mathcal{N}(\mu, \Sigma)$

$$\mu = E(\mathbf{X}) = [\mu_X, \mu_Y]^T \qquad \Sigma = \begin{bmatrix} \sigma_X^2 & Cov(X,Y) \\ Cov(Y,X) & \sigma_Y^2 \end{bmatrix}$$

Función de densidad de probabilidades

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \right)^T \Sigma^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \right) \right)$$

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2\sigma_X^2\sigma_Y^2(1-\rho^2)} \left(\sigma_Y^2(x-\mu_X)^2 + \sigma_X^2(y-\mu_Y)^2 - \right.\right.$$

$$\left.\left. -2\sigma_X\sigma_Y\rho(x-\mu_X)(y-\mu_Y)\right)\right\}$$

Distribución Gaussiana Bivariada $\mathcal{N}(\mu, \Sigma)$

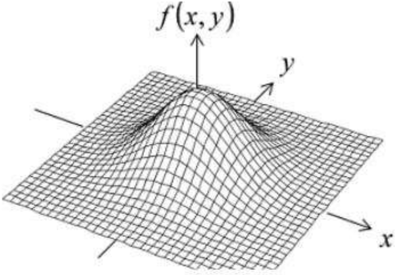


Figura: pdf distribución Gaussiana bivariada

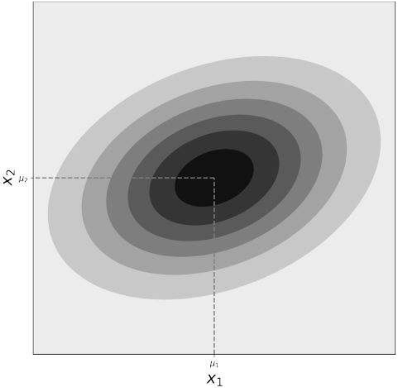
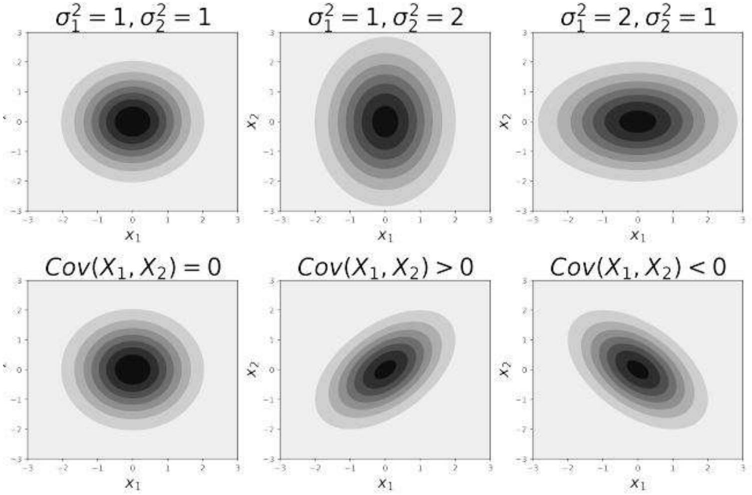


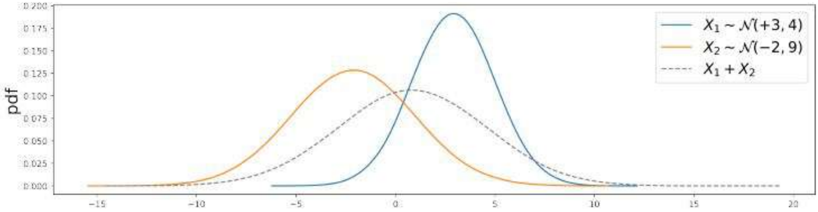
Figura: curvas nivel pdf distribución Gaussiana bivariada

Distribución Gaussiana bivariada: matriz de covarianza Σ



Propiedades de distribuciones Gaussianas

- Sean X_1 y X_2 dos variables aleatorias gaussianas
- $Y = X_1 + X_2$ también es gaussiana! (la suma de gaussianas, es gaussiana)



Propiedades de la suma de variables aleatorias

- $E(X_1 + X_2) = E(X_1) + E(X_2) = \mu_1 + \mu_2$
- $Var(X_1 + X_2) = Var(X_1) + Va(X_2) + 2Cov(X_1, X_2)$

Distribución Gaussiana multivariada $\mathcal{N}(\mu, \Sigma)$

El vector $\mathbf{X} = (X_1, \ldots, X_n)$ tiene distribución Normal (Gaussiana) multivariada si la función de densidad conjunta está dada por:

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

donde:
$\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, $\mu = [\mu_{X_1}, \ldots, \mu_{X_n}]^T$ es el vector de medias y Σ es a matriz de varianza-covarianza de las X_i, y $|\Sigma|$ es el determinante de esa matriz.

Observaciones

- No hay una expresión analítica para las funciones de verosimilitud.
- Para estimar μ y Σ deben usarse otros métodos o técnicas (ej. algoritmo EM)

Comentarios

- Ejercicios y ejemplos de esta unidad en notebooks jupyter disponibles en página curso
- Bibliogafía sugerida: Bishop C., 2006: Pattern Recognition and Machine Learning. Springer