Conceptos Fundamentales

Curso de Estadística

TAE, 2005

J.J. Gómez-Cadenas

Análisis de datos en física de partículas

Experimento en física de partículas: Observación de n sucesos de un cierto tipo (colisiones e^+ e^-, interacciones de neutrinos, colisiones pp, etc.)

Teoría: Modelos (en general efectivos) que describen los datos experimentales en términos de una serie de parámetros (secciones eficaces, constantes de acoplo, masas)

Análisis de datos. Contraste entre teoría y experimento: Extracción de los parámetros de la teoría u observación de fenómenos no predichos por la teoría:

Aplicaciones de la estadística al análisis de datos:

Estimar parámetros

Cuantificar los errores en dichas estimaciones

Cuantificar el grado de acuerdo entre datos y teoría

Definición de Probablilidad (Kolmogorov 1933)

Sea S un conjunto con cierto número de elementos (espacio de muestras)

Sean A,B,…. Subconjutos de S:

Para todo subconjuto A de S puede definirse un número real P(A) al que llamaremos probabilidad a partir de los siguientes tres axiomas:

$$\forall A \subset S \rightarrow P(A) \geq 0$$

$$Si\ A \cap B = 0 \rightarrow P(A \cup B) = P(A) + P(B)$$

$$P(S) = 1$$

Propiedades de las funciones de probabilidad

Pueden derivarse a partir de los tres axiomas anteriores

$$P(\bar{A}) = 1 - P(A)$$

$$P(A \cup \bar{A}) = 1$$

$$P(\varnothing) = 0$$

$$Si\ A \subset B \rightarrow P(A) \leq P(B)$$

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

Probabilidad condicional

La probabilidad de A dado B (donde P(B)≠0) es:

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

Decimos que los subconjutos A y B son independientes cuando

$$P(A \cap B) = P(A)P(B)$$

En cuyo caso:

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

No confundir conjuntos independientes con conjuntos disjuntos

$$A \cap B = \varnothing$$

Interpretación de la Probabilidad

I. **Frecuencia relativa**

A,B,…son los resultados de un experimento reproducible

$$P(A) = \lim_{n\to\infty} \frac{ocurrencias\ de\ A}{n}$$

II. **Probabilidad subjetiva (Bayesiana)**

A,B,…son hipótesis (sentencias que son verdaderas o falsas)

P(A) expresa el grado de credibilidad de que A sea cierta

Ambas interpretaciones son consistentes con los axiomas de Kolmogorov

Ambas interpretaciones son aplicables en física de partículas.

A menudo no se especifica (ni se tiene claro) que interpretación se está usando

Teorema de Bayes

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

$$P(A \cap B) = P(B \cap A)$$

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

Ley de la probabilidad total

Supongamos que S puede descomponerse en términos de conjuntos disjuntos:

$$S = \cup_i A_i \rightarrow A_i \cap A_j = \varnothing, i \neq j$$

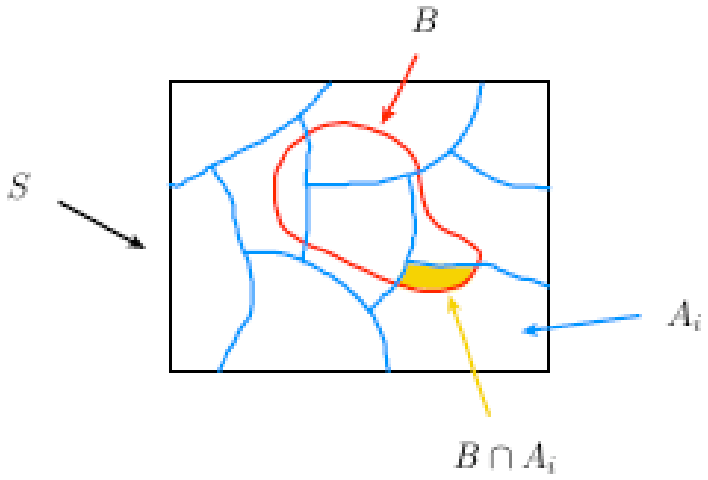
Además: $P(A_i) \neq 0, \forall i$

$$B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$$

Considerar un subconjunto B de S

$$P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

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



$$P(A \mid B) = \frac{P(B \mid A) P(A)}{\sum_i P(B \mid A_i) P(A_i)}$$

Función de densidad de probabilidad (PDF)

Considerar un experimento cuyo resultado es x (variable continua)

El espacio de muestra corresponde al conjunto de valores que x puede tomar

La función de densidad de probabilidad, pdf, da la probabilidad de observar un valor de x en el intervalo infinitesimal $[x, x+dx]$

$$P(x:[x,x+dx]) = f(x)d(x)$$

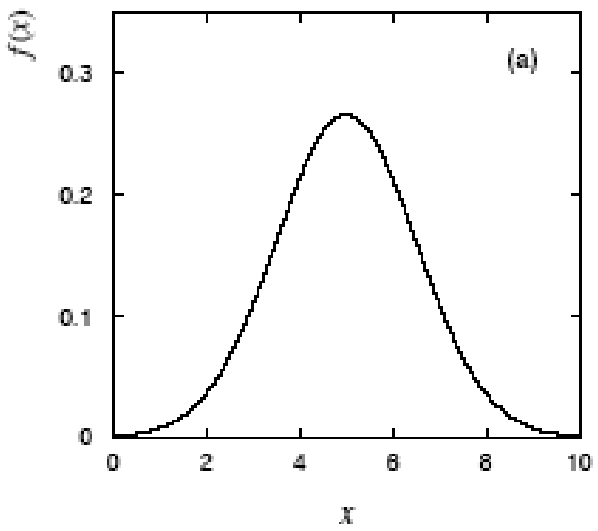
Por definición f(x) está normalizada a la unidad:

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

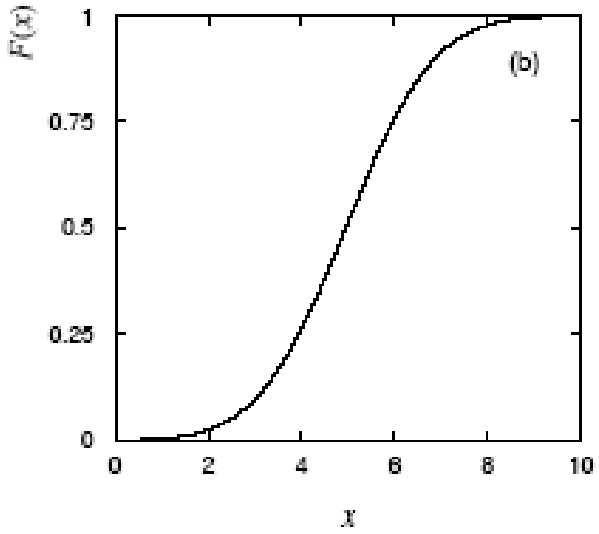
Definimos la función de acumulación F(x) como:

$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

f(x) (Gauss)



F(x) (Gauss)



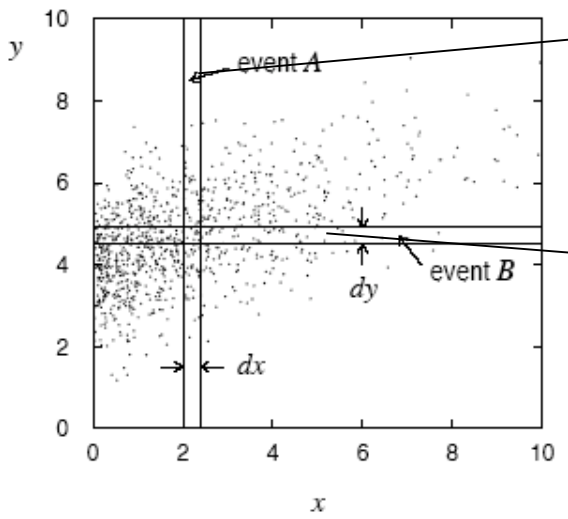
La pdf de una variable *x* puede interpretarse como la función que especifica el modelo que sigue *x*.

El concepto (clásico) es: “Si medimos x muchas veces y representamos las frecuencias a las que aparece un determinado resultado (normalizado a la integral total) obtenemos la pdf”.

El concepto Bayesiano es: “f(x) nos da la probabilidad de que ocurra un determinado valor de x”

Variables multidimensionales: PDF conjunta

Resultado de la medida: Vector multidimensional de variables aleatorias



A se observa con (x,y):

x entre [x,x+dx]

y arbitrario

B se observa con (x,y):

x arbitrario

y entre [y,y+dy]

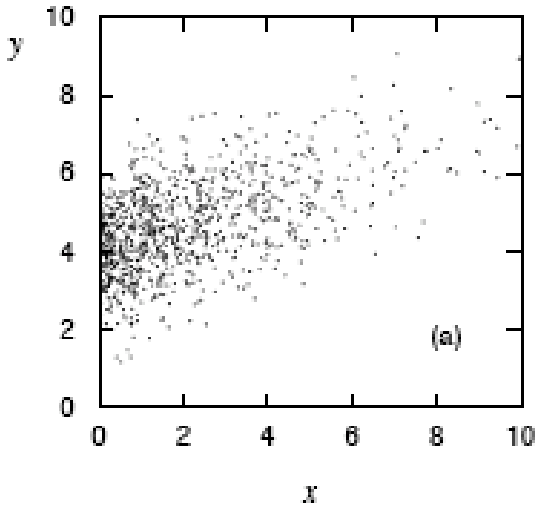
$$\iint_S f(x,y)dxdy = 1$$

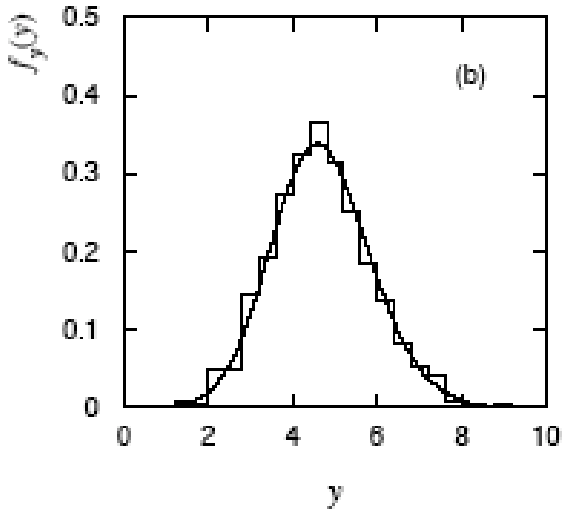
La probabilidad de que un determinado punto (x,y) se observe en el rectángulo que define la intersección (A∩B) es la pdf conjunta f(x,y) multiplicada por el elemento de área.

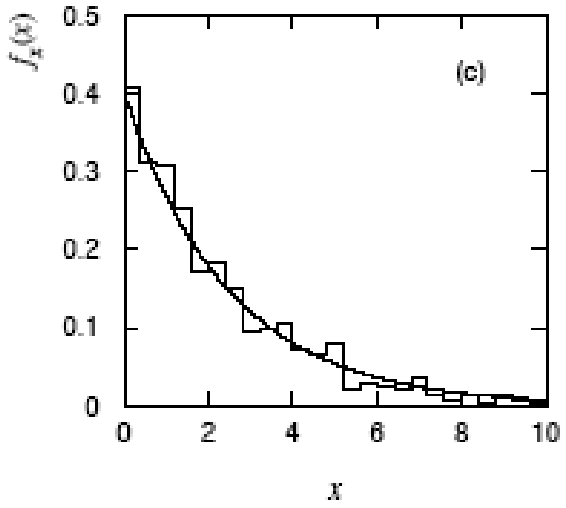
Corresponde a la densidad de puntos en un scatter plot (x,y) en el límite de infinitos puntos

P(A∩B) =probabilidad de que x⊂[x,x+dx], y⊂[y,y+dy]=f(x,y)dxdy

Distribuciones marginales





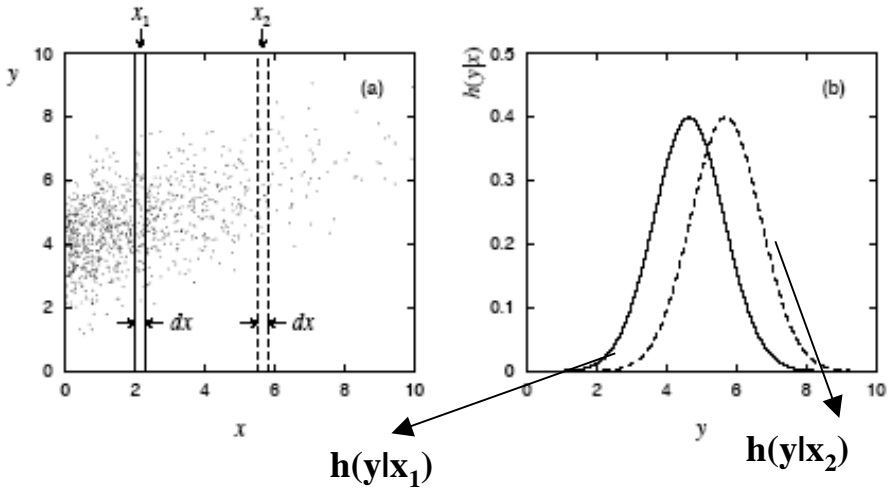


Proyección de la pdf conjunta en los ejes x,y

$$f_x = \int f(x,y)dx$$

$$f_y = \int f(x,y)dy$$

PDF condicional



Probabilidad de que y esté en $[y,y+dy]$ para cualquier x, dado que x está en $[x,x+dx]$ para cualquier y

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{f(x,y)dxdy}{f_x(x)dx}$$

La pdf condicional para y, dado x se define como:

$$h(y \mid x) = \frac{f(x,y)}{f_x(x)} = \frac{f(x,y)}{\int f(x,y')dy'}$$

NB: $h(y|x)$ es una pdf de y, en la que x se trata como un parámetro constante. Se obtiene a partir de f(x,y), manteniendo x constante y renormalizando la función de tal manera que se obtenga área unidad cuando se integra sólo sobre y.

Corresponde al histograma normalizado de y obtenido a partir de la proyección en el eje y de un elemento diferencial de x en un scatter plot

Relación entre PDF condicionales para (x,y)

$$h(y \mid x) = \frac{f(x,y)}{f_x(x)} = \frac{f(x,y)}{\int f(x,y')dy'} \qquad g(x \mid y) = \frac{f(x,y)}{f_y(y)} = \frac{f(x,y)}{\int f(x',y)dx'}$$

Combinando ambas pdfs Obtenemos el teorema de Bayes para variables continuas :

La pdf marginal puede expresarse en términos de la condicional:

Corresponde a la ley de la probabilidad total para variables aleatorias continuas

Decimos que (x,y) son independientes si:

$$g(x \mid y) = \frac{h(y \mid x) f_x(x)}{f_y(y)}$$

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

$$f_x = \int f(x,y)dx = \int g(x \mid y) f_y(y)dy$$

$$f_y = \int f(x,y)dy = \int h(y \mid x) f_x(x)dx$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{\sum_i P(B \mid A_i)P(A_i)}$$

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

$$P(A \cap B) = P(A)P(B)$$

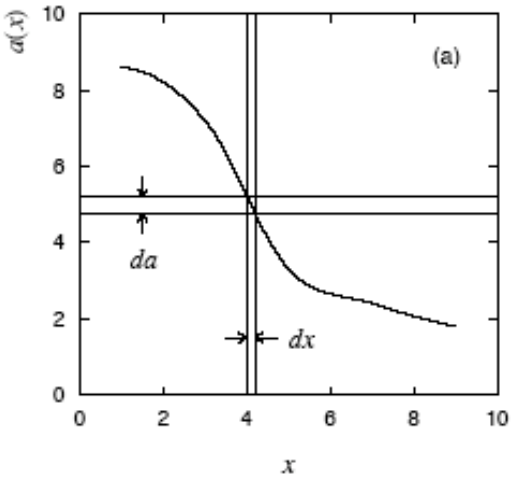
Funciones de variables aleatorias

Una función de una variable aletoria es una variable aleatoria

x es una variable continua y aleatoria, distribuida de acuerdo a la pdf *f(x)*

a(x) es una función continua de **x**

¿Cuál es la pdf *g(a)* que describe la distribución de *a*?



Imponemos que la probabilidad de que *x* ocurra entre *[x,x+dx]* sea igual a la probabilidad de que *a* ocurra entre *[a,a+da]*

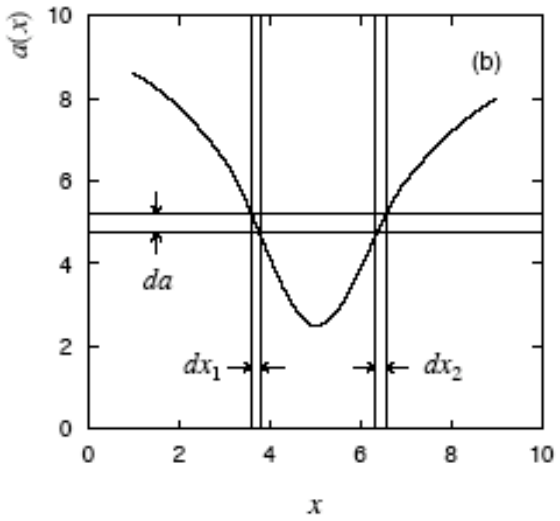
$$g(a)da = \int_S f(x)dx$$

dS = región del espacio en x donde a está en [a,a+da]

Si **a(x)** puede invertirse (para obtener **x(a)**) entonces:

$$g(a)da = f(x(a))\frac{dx}{da}$$

Si **a(x)** no tiene una única inversa debemos incluir todos los dx en dS que correspondan a da:



Example: $a = x^2, \quad x = \pm\sqrt{a}, \quad dx = \pm\frac{da}{2\sqrt{a}}$

$g(a)\,da = \int_{dS} f(x)\,dx$

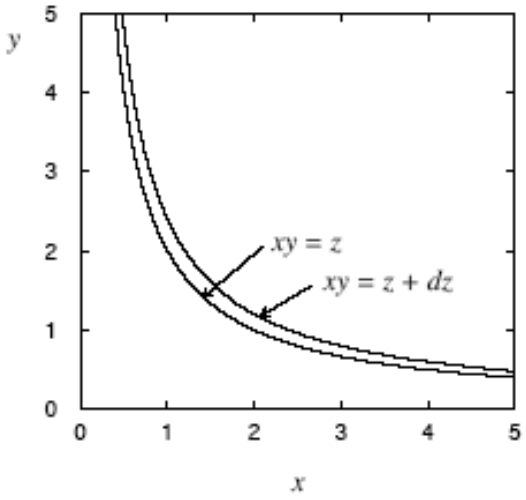
$dS = \left[\sqrt{a}, \sqrt{a} + \frac{da}{2\sqrt{a}}\right] \cup \left[-\sqrt{a} - \frac{da}{2\sqrt{a}}, -\sqrt{a}\right]$

$g(a) = \frac{f(\sqrt{a})}{2\sqrt{a}} + \frac{f(-\sqrt{a})}{2\sqrt{a}}$

Funciones de más de una variable aleatoria

Considerar $\vec{x} = (x_1, x_2, ..., x_n)$; $pdf : f(x_1, x_2, ..., x_n)$: $a = a(\vec{x})$

$$g(a)da = \int ... \int_{dS} f(x_1, x_2, ..., x_n) dx_1, dx_2, ..., dx_n$$



Ejemplo: x, e independientes, con pdf **g(x)** y **h(x).** Por lo tanto la pdf conjunta **f(x,y)=g(x)h(y)**

¿Cuál es la pdf de su producto **z(x,y) = xy**?

$$f(z)dz = \int ... \int_{dS} f(x,y)dxdy = \int ... \int ... \int_{dS} g(x)h(y)dxdy = \int_{-\infty}^{\infty} g(x)dx \int_{\frac{z}{|x|}}^{\frac{(z+dz)}{|x|}} h(y)dy$$

$$f(z) = \int_{-\infty}^{\infty} g(x)h(z/x)\frac{dx}{|x|} = \int_{-\infty}^{\infty} g(z/y)h(y)\frac{dy}{|y|}$$

Jacobiano de una transformación de variables

Considerar $\vec{x} = (x_1, x_2, ..., x_n)$ con pdf conjunta $f(\vec{x})$

Formar n funciones linealmente independientes $\vec{y}(\vec{x}) = (\vec{y}_1(\vec{x}), ..., \vec{y}_n(\vec{x}))$,

tales que las funciones inversas $x_1(\vec{y}), ..., x_n(\vec{y})$ existen.

La pdf conjunta de $\vec{y}$ es entonces:

$$g(\vec{y}) = |J| f(\vec{x})$$

donde:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

Valores esperados: Media

El valor esperado **E[x]** de una variable aleatoria x distribuida de acuerdo con la pdf **f(x)** es (corresponde a la media):

$$E[x] = \int_{-\infty}^{+\infty} xf(x)dx = \mu$$

Para variables discretas: $E[x] = \sum_i x_i P(x_i)$

NB: **E[x]** no es una función de x sino un parámetro (depende de la forma) de **f(x).**

Para una función **a(x)** con pdf **g(a)**

$$E[a] = \int_{-\infty}^{+\infty} ag(a)da \quad \text{Pero} \quad g(a)da = \int_S f(x)dx$$

$$ag(a)da = \int_S a(x)f(x)dx \rightarrow E[a] = \int_{-\infty}^{+\infty} a(x)f(x)dx$$

Valores esperados: Momentos y Varianza

Momento algebraico de orden n: $E[x^n] = \int_{-\infty}^{+\infty} x^n f(x)dx = \mu_n'$

Momento central de orden n: $E[(x - E[x])^n] = \int_{-\infty}^{+\infty} (x-\mu)^n f(x)dx = \mu_n$

En particular el momento central de orden 2 es la varianza:

$$E[(x - E[x])^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \sigma^2 = V[x]$$

La varianza mide la dispersión de x entorno a su valor medio. La raíz cuadrada de la varianza, σ, se denomina desviación estándar de x.

Valores esperados: Más de una variable

Para el caso de una función a que depende de más de una variable aleatoria la media es:

$$E[a(\vec{x})] = \int_{-\infty}^{+\infty} a g(a) da = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} a(\vec{x}) f(\vec{x}) dx_1 \cdots dx_n = \mu_a$$

Mientras que la varianza es:

$$V[\alpha] = E[(a-\mu_a)^2] = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (a(\vec{x})-\mu_a)^2 f(\vec{x}) dx_1 \cdots dx_n = \sigma_a^2$$

La covarianza de dos variables aleatorias se define como:

$$V_{xy} = E[(x-\mu_x)(y-\mu_y)] = E[xy] - \mu_x\mu_y = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy f(x,y) dxdy - \mu_x\mu_y$$

Matriz de covarianza

Dadas dos funciones $a(\boldsymbol{x})$ y $b(\boldsymbol{x})$ de n variables aleatorias $(x_1,\ldots x_n)$ la matriz de covarianza (o matriz de error) V_{xy} $(cov[x,y])$ es:

$$V_{xy} = \mathrm{cov}[a,b] = E[(a-\mu_a)(b-\mu_b)] = E[ab] - \mu_a\mu_b = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} abg(a,b)dadb - \mu_a\mu_a$$

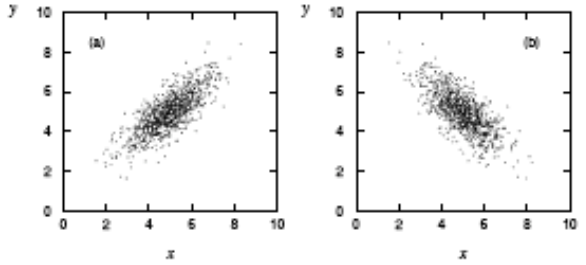
$$= \int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} a(\vec{x})b(\vec{x})f(\vec{x})dx_1\cdots dx_n - \mu_a\mu_a$$

Donde $g(a,b)$ es la pdf conjunta para $a(x)$ y $b(x)$ y $f(\boldsymbol{x})$ es la pdf conjunta para **x**.

NB: Por construcción V_{xy} es simétrica en a y b. Además los elementos diagonales $V_{xy}=\sigma^2$ son positivos.

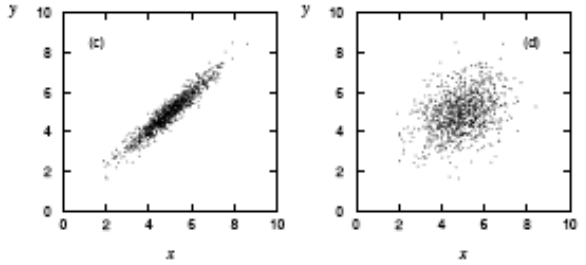
Coeficiente de correlación

$\rho = 0.75$



$\rho = -0.75$

$\rho = 0.95$



$\rho = 0.25$

ρ_{xy} proporciona una medida del nivel de correlación entre dos variables

$$\rho_{xy} = \frac{V_{xy}}{\sigma_x \sigma_y}, \quad -1 \le \rho_{xy} \le 1$$

Si x,y son independientes $f(x,y)=f_x(x)f_y(y)$

$$E[xy] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_x(x)f_y(y)dxdy = \mu_x\mu_y$$

$$V_{xy} = E[xy] - \mu_x\mu_y = 0$$

NB: El teorema recíproco no es cierto en general

Propagación de errores

Dado un vector $\boldsymbol{x} = (x_1,\ldots,x_n)$ de variables aleatorias distribuidas conforme a una pdf conjunta tales que:

La pdf no se conoce completamente

Los valores medios de los x_i, $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)$ se conocen (al menos se conoce una estimación de éstos)

La matriz de covarianza V_{ij} se conoce o ha sido estimada

Considerar una función de n variables $y(\boldsymbol{x})$:

Puesto que f(x) no se conoce completamente no es posible obtener la pdf de y

Si es posible encontrar el valor aproximado del valor esperado de y (o valor medio) y de la varianza $V[y]$ expandiendo $y(\boldsymbol{x})$ en primer orden entorno a los valores medios de x_i.

$$y(\vec{x}) = y(\vec{\mu}) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i)$$

El valor medio de y es, a primer orden: $E[y(\vec{x})] \approx y(\vec{\mu})$ ya que $E[x_i - \mu_i] = 0$

El valor esperado de y^2 se calcula fácilmente:

$$y^2(\vec{x}) = y^2(\vec{\mu}) + 2y(\vec{\mu}) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i) + \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i) \sum_{j=1}^{n} \left[\frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j)$$

$$E[y^2(\vec{x})] \approx y^2(\vec{\mu}) + y(\vec{\mu}) \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} E[(x_i - \mu_i)] + E\left[\sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} (x_i - \mu_i) \sum_{j=1}^{n} \left[\frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} (x_j - \mu_j) \right]$$

$$= y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} E[(x_i - \mu_i)(x_j - \mu_j)]$$

$$= y^2(\vec{\mu}) + \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

Y por lo tanto la varianza de y: $\sigma_y^2 = E[y^2] - (E[y])^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$

Análogamente, para m funciones $y_1(\boldsymbol{x}),\ldots,y_m(\boldsymbol{x})$, la matriz de covarianza es:

$$U_{kl} = \mathrm{cov}[y_k y_l] \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

En notación matricial:

$U = AVA^T$ donde la matriz de derivadas A es:

$$\mathrm{A}_{\mathrm{ij}} = \left[\frac{\partial y_i}{\partial x_j} \right]_{\vec{x}=\vec{\mu}}$$

La ecuación anterior es la ley de propagación de errores en el caso más general. Es decir, las varianzas, que se usan como medidas de las incertidumbres estadísticas, se propagan desde las variables x_i a las funciones $y_1, y_2 \ldots$etc.

Si las x_i no están correlacionadas entonces:

$$\sigma_y^2 = \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]^2_{\vec{x}=\vec{\mu}} \sigma_i^2$$

$$U_{kl} = \sum_{i=1}^{n} \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_i} \right]_{\vec{x}=\vec{\mu}} \sigma_i^2$$

Casos particulares:

Suma:

$$y = x_1 + x_2$$

$$\sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$$

Producto:

$$y = x_1 x_2$$

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{V_{12}}{x_1 x_2}$$

Hipótesis en el cálculo de la propagación de errores

Hipótesis:

Las medias y varianzas de las variables $x_1,\ldots,x_n$ (las medidas) se conocen (o existe una aproximación razonable)

Las funciones de $x_1,\ldots,x_n$ que estamos estudiando pueden aproximarse mediante una expansión de Taylor a primer orden entorno a los valores medios $m_1,\ldots,m_n$.

Esta hipótesis sólo es exacta cuando las funciones estudiadas son lineales y deja de ser válida cuando el comportamiento es fuertemente no lineal entorno a la media comparable a la desviación estándar.