Análisis Multivariado

El análisis multivariado es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable o característica sobre una muestra de individuos.

Objetivos del análisis multivariado

1. Resumir los datos mediante un pequeño conjunto de nuevas variables, construidas como transformaciones de las originales, con la mínima pérdida de información

Ejemplos

- La calidad de una universidad se puede resumir en unos pocos indicadores de eficiencia
- El crecimiento económico de un pais puede explicarse mediante un numero reducido de variables, sin tener que tomar todas las variables originales.

Ventajas

- Se pueden representar gráficamente los individuos y visualizar posibles comportamientos entre ellos
- La interpretación de las nuevas variables ayuda al mejor conocimiento del fenomeno

Objetivos

2. Encontrar grupos en los datos, si existen
Ejemplo

- Si se mide un conjunto de observaciones que describen el funcionamiento de empresas, se espera que existan grupos de observaciones con propiedades similares, las empresas se pueden dividir en grupos de observaciones en función de su rentabilidad, su eficiencia comercial o su estructura productiva.
- En la mayoría de los casos, los grupos son desconocidos, entonces se requiere procedimientos objetivos para obtenerlos.
- El número de grupos necesarios para describir la estructura de los datos no está definida, existen métodos para determinarlos.

Objetivos

3. Clasificar nuevas observaciones en grupos definidos
Ejemplos

- Se desea clasificar a clientes que solicitan créditos, como clientes fiables o no fiables.
- Diseñar una maquina que identifique billetes falsos

4. Relacionar dos conjuntos de variables
Ejemplos

- Relacionar un conjunto de variables de capacidad intelectual con otro de resultados profesionales y determinar cuantas dimensiones tiene esta relación.
- Los dos conjuntos pueden corresponder a la misma variable medidas en dos momentos distintos o en espacios diferentes, y se quiere estudiar la relación entre los dos grupos.

Aplicaciones del análisis multivariado

- Administración de empresas: construir tipologías de clientes mediante la identificación de características homogéneas entre ellos
- Arqueología: clasificar restos arqueológicos cronológicamente de acuerdo a su formas, materiales, ubicación, etc
- Expansión de Negocios: identificar los factores que determinan el éxito de un negocio, utilizados con fines de expansión.
- Ciencias de la computación: diseñar algoritmos de clasificación automática
- Psicología: determinar los factores que componen la inteligencia humana
- Sociología y ciencias políticas: construir tipologías de los votantes de un partido

Enfoques del análisis multivariado: exploratorio e inferencial

Enfoque exploratorio: tiene como objetivo extraer la mayor información posible de los datos disponibles e identificar el modelo teorico que mejor represente a la población de donde provienen los datos muestrales

1. Descripción de datos multivariados mediante:
   - *Medidas de centralización y dispersión*: vector de medias, matriz de varianzas y covarianzas, varianza total, varianza generalizada.
   - *Medidas de dependencia lineal entre variables*:
     - *Dependencia por pares de variables:* la matriz de correlación
     - *Dependencia de cada variable* y el resto: regresión múltiple.
   - Representaciones gráficas: histogramas, diagramas de dispersión, representación mediante figuras (caras de chernoff,diagramas de cajas, etc)
   - Detección de datos atípicos mediante el uso de distancias.

Enfoque exploratorio

- **Métodos de reducción de dimensionalidad**
  - *Componentes principales*: A partir de un conjunto de variables p-dimensionales se construye un conjunto de variables en una dimensión menor a p, que resume la información original.
  - *Escalamiento multidimensional*: a partir de una matriz de disimilaridad se encuentra un conjunto de variables, que identifica la dimensión de las disimilaridades .
  - *Análisis de correspondencia*: a partir de datos cualitativos que se presentan en una tabla de contingencia, identificar las dimensiones subyacentes de los datos.
- **Métodos de agrupamiento y clasificación**
  - Técnicas de cluster como *K-medias* agrupan los elementos en clusters basados en criterios de homogeneidad y separación
  - *La clasificación de los elementos* en los grupos formados, basados en el concepto de distancia

Enfoque inferencial

Enfoque inferencial: tiene como objetivo obtener conclusiones sobre la población de donde provienen los datos, mediante un modelo que explique su generación y permita prever lo datos futuros

- Se requiere un modelo estadístico, basado en una distribución de probabilidad multivariada de la población
- Consiste en estimación los parámetros de la poblacion a partir de los datos de una muestra
- Probar hipótesis sobre las características de la población

Ventajas

- Se identifica el mecanismo generador de los datos
- Se pueden realizar predicciones respecto a datos no observados pero generados por el mismo sistema o distribución

Técnicas multivariadas: enfoque inferencial

- *Análisis factorial*: es una generalización de los componentes principales para reducir la dimensionalidad de los datos.
- *Análisis discriminante*: se asume que los datos provienen de dos o mas poblaciones que siguen una distribución conocida y se desea clasificar un nuevo dato en una de ellas
- *Regresión multivariada*: identifica la relación entre dos conjuntos de variables multivariadas. El primer conjunto incluye variables continua o discretas y se utilizan para explicar las variables continuas del segundo conjunto
- *Correlación canónica*: el objetivo es encontrar indicadores del primer conjunto que expliquen lo mejor posible los indicadores de las variables del segundo grupo. El número de relaciones independientes nos indica la dimensión de la relación

Contenido del curso

1. Datos multivariados
   - Descripción de las observaciones multivariadas
   - Medidas de localizacion y matriz de varianzas y covarianzas
   - Medidas de dependencia lineal: la matriz de correlaciones
2. Distribuciones multivariadas
   - Variables aletorias vectoriales
   - Densidad normal multivariada y sus propiedades
   - Distribuciones de Hotelling, Wishart y Wilks.
   - Estimación de los Parámetros de la distribución normal multivariada
   - Pruebas de hipótesis multivariadas y regiones de confianza
3. Modelos de regresión lineal
   - Modelo de regresion lineal clásico.
   - Regresión lineal Multivariada. Inferencia sobre los parámetros
   - Correlación Canónica

Contenido del curso

4. Análisis de factores
   - Modelo de Factores Ortogonales
   - Métodos de estimación de los parámetros
   - Determinación del número de factores
   - Rotación de factores
   - Relación con componente principales
   - Análisis de factores confirmatorio
5. Análisis de datos categóricos
   - Comparación de proporciones
   - Pruebas de independencia
   - Asociación en tablas de contingencia
   - Análisis de correspondencia

Bibliografía

- Johnson, R.A. y Wichern, D.W. (2007) . Applied multivariate statistical analysis, 6th Ed. Prentice – Hall.
- Daniel Peña (2002). Análisis de datos multivariantes. McGraw Hill.
- Izenman, J. (2008). Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer.
- Tinsley, H. and Brown, S. (2000). Handbook of Applied Multivariate Statistics and Mathematical Modeling. Academic Press

Bibliografía

- Everitt, Brian S. (2005) An R and S-Plus® companion to multivariate analysis. Springer.
- Agresti, A (2007). An introduction to categorical data analysis, 2nd Ed. Wiley.
- T. W. Anderson. (2003). An Introduction to Multivariate Statistical Analysis. Wiley

Evaluación del curso

- Dos examenes parciales: 50% (el primero a la mitad del curso y otro al final)
- Tareas:30%
- Proyecto final: 20%.
  - Consistirá en la aplicación de alguna de las técnicas vistas a un conjunto de datos reales
  - Una presentación de 30 minutos y un reporte por escrito.
  - El trabajo debe contener una introducción, objetivos, metodología, resultados, conclusiones y bibliografía.

Observaciones multivariadas

- Una observación multivariada es el resultado de observar p características en un elemento de la población
- Las obs. multivariadas son valores particulares de un conjunto de p variables que se denomina *variable aleatoria vectorial*
- Por ejemplo, si se observa la edad, peso, altura de los estudiantes de una universidad, tendremos un conjunto de obs multivariadas tomadas de una *variable aleatoria tridimensional*.
- Una *variable vectorial p-dimensional* es discreta cuando lo es cada una de las p variables escalares que la componen.
- Variable aleatoria es continua, cuando todos sus componentes son continuos.
- Variable aleatoria es mixta cuando algunos componentes son discretos y otros continuos.

Matriz de datos

- Los valores de las p variables escalares en cada uno de los n elementos se representan mediante una matriz $\mathbf{X}$, $(n \times p)$, llamada **matriz de datos** que se denota de la siguiente forma

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_i' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

$$\mathbf{X} = [\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(p)}]$$

Medidas de Centralización: el vector de medias

- El vector de medias $\overline{\mathbf{x}}$, es un vector de dimensión p cuyas componentes son las medias de cada una de las p variables

$$\overline{\mathbf{x}} = (\overline{x}_1, ..., \overline{x}_j, ..., \overline{x}_p)' \text{ donde } \overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

  O bien denotando por $\mathbf{x}_i$ a la i-esima fila de $\mathbf{X}$,

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

- $\overline{\mathbf{x}}$ se expresa a partir de la matriz de datos $\mathbf{X}$ como

$$\overline{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$$

  donde $\mathbf{1}$ representa un vector de **unos** de dimension n

Medidas de Centralización

El vector de medias se encuentra en el centro de los datos, en el sentido de que la suma de desviaciones es cero:

$$\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}) = 0$$

Problemática con otras medidas de Centralización

- La mediana que es otra medida de centralización basada en el orden de las observaciones no puede generalizarse al caso multivariado
- Por ejemplo podemos calcular el vector de medianas, pero este punto no tiene necesariamente una interpretación como centro de los datos.
- La dificultad se debe a que no existe un orden natural de los datos multivariados.

Covarianza entre dos variables

- En el caso univariado, la variabilidad respecto a la media se mide usualmente por la varianza o su raiz cuadrada (desviación típica):

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- La relación lineal entre dos variables se mide por la **covarianza**. La covarianza entre las variables x_j y x_k se calcula por:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)$$

- Mide la dependencia lineal entre las variables.
- La covarianza entre la misma variable (es decir, si $j = k$) es la varianza

La matriz de varianzas y covarianzas

Para una muestra de observaciones $\mathbf{X}$ que provienen de una variable multivariada se define la matriz de *varianzas y covarianzas* como

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$$

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} x_{i1} - \overline{x}_1 \\ \vdots \\ x_{ip} - \overline{x}_p \end{bmatrix}\begin{bmatrix} x_{i1} - \overline{x}_1, & \ldots, & x_{ip} - \overline{x}_p \end{bmatrix}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} (x_{i1} - \overline{x}_1)^2 & \cdots & (x_{i1} - \overline{x}_1)(x_{ip} - \overline{x}_p) \\ \vdots & & \vdots \\ (x_{ip} - \overline{x}_p)(x_{i1} - \overline{x}_1) & \cdots & (x_{ip} - \overline{x}_p)^2 \end{bmatrix}$$

La matriz de varianzas y covarianzas

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} (x_{i1}-\overline{x}_1)^2 & \dots & (x_{i1}-\overline{x}_1)(x_{ip}-\overline{x}_p) \\ \vdots & & \vdots \\ (x_{ip}-\overline{x}_p)(x_{i1}-\overline{x}_1) & \dots & (x_{ip}-\overline{x}_p)^2 \end{bmatrix}$$

- La expresion dentro de la sumatoria representa la matriz de cuadrados y productos cruzados de las p variables en el elemento i.
- Al sumarla para todos los elementos y dividir por n se obtienen las varianzas en la diagonal y las covarianzas fuera de ella.
- La matriz de *varianzas y covarianzas* es una matriz cuadrada y simétrica de orden p que contiene en la diagonal las varianzas y fuera de la diagonal las covarianzas entre las variables.

La matriz de varianzas y covarianzas

La matriz de *varianzas y covarianzas* o simplemente matriz de *covarianzas* se denota comunmente como

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

Matriz de datos centrados

- La matriz de *datos centrados* $\widetilde{\mathbf{X}}$, se define como $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}'$
- $\widetilde{\mathbf{X}}$ se puede reescribir como

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}' = \mathbf{X} - \mathbf{1}(\frac{1}{n}\mathbf{X}'\mathbf{1})' = \mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X} = (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X} = \mathbf{P}\mathbf{X},$$

donde $\mathbf{P} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$ es la **matriz de centrado**.

La matriz de varianzas y covarianzas

- La matriz de centrado **P** es util debido a que muchas características multivariadas se expresan mejor a partir de ella.
- La matriz P tiene algunas propiedades interesantes:
  - es simétrica:$P = P'$
  - es idempotente: $PP = P$
  - Tiene rango $(n-1)$, es decir, tiene $n-1$ columnas o renglones linealmente independientes
  - Sus valores propios son 1 o 0.

La matriz de varianzas y covarianzas

- La matriz $\mathbf{S}$ puede escribirse como:

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \frac{1}{n}\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \frac{1}{n}(\mathbf{PX})'\mathbf{PX} = \frac{1}{n}\mathbf{X}'\mathbf{PX}$$

Propiedades de la matriz de covarianzas. Al igual que la varianza es siempre un número no negativo, la matriz de covarianzas tiene una propiedad similar:

- La matriz de covarianzas es semidefinida positiva. Esto es, si $\mathbf{y}$ es cualquier vector, $\mathbf{y}'\mathbf{Sy} \geq 0$
- También la traza, el determinante y los valores propios de $\mathbf{S}$ son no negativos.

Medidas globales de variabilidad

- Existen otras medidas de variabilidad obtenidas a partir de **S**. Son muy utiles cuando queremos comparar conjuntos de variables. En particular, hay dos medidas de variabilidad que son muy utiles en el análisis multivariado: la **varianza total de los datos** y **la varianza generalizada**
- **Varianza total de los datos.** Resume la variabilidad de un conjunto de variables, y se define mediante la traza de **S**,

$$tr(\mathbf{S}) = \sum_{i=1}^{p} s_i^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

- **Varianza generalizada** se define como el determinante de **S**,

$$|\mathbf{S}| = \lambda_1 * \lambda_2 * \cdots * \lambda_p$$

Medidas globales de variabilidad

- **La varianza total de los datos**, es muy útil en técnicas como componentes principales y análisis de factores. Permite identificar el porcentaje de variabilidad total de los datos que es explicada por cada componente o factor, lo cual facilita la elección del numero de componentes con fines de reducción de dimensión.
- Cuando la **varianza generalizada** es cero, es decir $|\mathbf{S}| = 0$, esto indica que existe una o mas variables que son un combinación lineal de otra (son colineales).
- Estas variables no están aportando nada nuevo en el análisis de datos multivariados, y deben ser eliminadas del estudio.
- Por tanto la varianza generalizada nos permite determinar si existen variables correlacionadas.

Medida de dependencia lineal: la matriz de correlación

- La dependencia lineal entre dos variables se estudia mediante el **coeficiente de correlación lineal**. El **coeficiente de correlación** para las variables x_j y x_k se define como:

$$r_{jk} = \frac{s_{jk}}{s_j s_k}, \quad -1 \leq r_{jk} \leq 1$$

  donde s_{jk} es la covarianza entre las variables x_j y x_k y s_j y s_k son la raiz cuadrada de las varianzas de x_j y x_k, respec.
- Si existe una relación lineal positiva exacta entre las variables x_j y x_k, $r_{jk} = 1$
- Si existe una relación lineal negativa exacta entre x_j y x_k, $r_{jk} = -1$
- Si no existe relación lineal entre las variables x_j y x_k, $r_{jk} = 0$

Medida de dependencia lineal: la matriz de correlación

La dependencia lineal entre todos los pares de variables se mide por la *matriz de correlación*. La matriz de correlación **R** se denota por:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ \vdots & \vdots & \dots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

- Es una matriz cuadrada $p \times p$, simetrica, que contiene **unos** en la diagonal (lo cual indica la relación de una variable consigo misma) y los **coeficientes de correlacion** entre los pares de variables fuera de la diagonal
- Es positiva semidefinida

Relación de la matriz de covarianzas y la matriz de correlación

- Aunque el coeficiente de correlación y la covarianza son medidas de dependencia lineal, los coeficientes están estandarizados a diferencia de la covarianza.
- Cuando las variables estan medidas en unidades distintas es preferible usar $\mathbf{R}$ sobre $\mathbf{S}$ debido a que la estandarización de las variables elimina este sesgo
- Sea $\mathbf{D}=\mathbf{D}(\mathbf{S})$ la matriz diagonal de orden p formada por los elementos de la diagonal de $\mathbf{S}$. La matriz $\mathbf{D}^{1/2}$ contiene las desviaciones típicas.
- La matriz $\mathbf{R}$ está relacionada con la matriz de covarianzas $\mathbf{S}$ mediante

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2} \quad \text{que implica} \quad \mathbf{S} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$$

Variables aleatorias vectoriales

Los datos en el análisis multivariado suelen provenir de una población caracterizada por una distribución multivariada

.

- Sea $\mathbf{x} = (x_1, ..., x_p)$ una variable aleatoria vectorial o vector aleatorio.
- **La función de densidad (o de probabilidad) conjunta** de una variable aleatoria vectorial $\mathbf{x}$, queda definida cuando se especifica:
  1. El espacio muestral o conjunto de sus valores posibles. Cada valor representa un punto en el espacio de dimensión p.
  2. Las probabilidades de cada posible resultado del espacio muestral.

Distribución conjunta de una variable aleatoria vectorial

- La **función de distribución conjunta**, $F(\mathbf{x})$, de una variable aleatoria vectorial, $\mathbf{x}$, se define en un punto dado $\mathbf{x}^0 = (x_1^0, x_2^0, \ldots, x_p^0)$ mediante

$$F(\mathbf{x}^0) = P(\mathbf{x} \leq \mathbf{x}^0) = P(x_1 \leq x_1^0, x_2 \leq x_2^0, \ldots, x_p \leq x_p^0)$$

- $F(\mathbf{x}^0)$ acumula las probabilidades de todos los valores menores o iguales al punto considerado.
- Aunque $F(\mathbf{x}^0)$ tiene gran interés teórico, es mas cómodo en la práctica trabajar con la función de probabilidades o función de densidad para variables continuas.

Distribución conjunta de una variable aleatoria vectorial

- Se dice que una variable aleatoria vectorial es continua si existe una **función de densidad** $f(\mathbf{x})$ que satisface:

$$F(\mathbf{x}^0) = \int_{-\infty}^{x^0} f(\mathbf{x})d\mathbf{x} = \int_{-\infty}^{x_1^0} \int_{-\infty}^{x_2^0} \cdots \int_{-\infty}^{x_p^0} f(\mathbf{x})d\mathbf{x}$$

  donde $d\mathbf{x} = dx_1, \ldots, dx_p$
- La integral es una integral múltiple en dimensión p. La densidad $f(\mathbf{x})$ tiene la interpretación habitual: masa por unidad de volumen.
- Sus límites de integración están definidos por los valores que toma cada variable univariada x_i

Distribución conjunta de una variable aleatoria vectorial

1. La función de densidad $f(\mathbf{x})$ debe verificar las siguientes condiciones:
2. $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_p) \geq 0$
3. $\int_{-\infty}^{\infty} f(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{\infty}^{-\infty} f(x_1, x_2, \ldots, x_p) dx_1 \cdots dx_p = 1$

Ejercicio

- Sea (x, y) una variable aleatoria bidimensional continua con función de densidad conjunta:

$$f(x,y) = \begin{cases} xy & 0 \leq x \leq 2, 0 \leq y \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

- Obtener la funcion de distribución conjunta de (x, y)

Distribuciones marginales

- Dada una variable aleatoria vectorial $\mathbf{x} = (x_1, ..., x_p)$, llamaremos distribución marginal de cada componente x_i a la distribución univariada de dicho componente, considerado individualmente, e ignorando los valores del resto de variables.
- Por ejemplo, para variables bidimensionales continuas $\mathbf{x} = (x_1, x_2)$, las distribuciones marginales de x_1 y x_2 se obtienen como:

$$f(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2,$$

$$f(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1$$

donde $f(x_1)$ y $f(x_2)$ representan las funciones de densidad de cada variable

Ejercicio:

- Sea (x, y) una variable aleatoria bidimensional continua con función de densidad conjunta:

$$f(x,y) = \begin{cases} 3x(1-xy) & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

- Obtener las funciones marginales $f(x)$ y $f(y)$
- Obtener la funcion de distribución conjunta de (x, y)

Distribuciones condicionadas

- Si $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, donde $\mathbf{x}_1$ y $\mathbf{x}_2$ son a su vez variables vectoriales, se define la *distribución condicionada* de $\mathbf{x}_1$ para un valor concreto de la variable $\mathbf{x}_2$ por

$$f(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f(\mathbf{x}_2)} \quad \text{o bien} \quad f(\mathbf{x}_2|\mathbf{x}_1) = \frac{f(\mathbf{x}_1, \mathbf{x}_2)}{f(\mathbf{x}_1)}$$

De la última definición se deduce que

$$f(\mathbf{x}_1, \mathbf{x}_2) = f(\mathbf{x}_2|\mathbf{x}_1) f(\mathbf{x}_1).$$

La distribucion marginal de $\mathbf{x}_2$ se puede calcular como

$$f(\mathbf{x}_2) = \int_{-\infty}^{\infty} f(\mathbf{x}_1, \mathbf{x}_2) d\mathbf{x}_1 = \int_{-\infty}^{\infty} f(\mathbf{x}_2|\mathbf{x}_1) f(\mathbf{x}_1) d\mathbf{x}_1$$

Distribuciones condicionadas

- De los resultados anteriores, la distribución condicionada $f(\mathbf{x}_1|\mathbf{x}_2)$ se puede escribir como

$$f(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x}_2|\mathbf{x}_1)f(\mathbf{x}_1)}{\int_{-\infty}^{\infty} f(\mathbf{x}_2|\mathbf{x}_1)f(\mathbf{x}_1)d\mathbf{x}_1}$$

- Representa el teorema de bayes para funciones de densidad, y constituye la herramienta fundamental de la inferencia bayesiana.
- Para variables discretas los conceptos son similares, pero ahora las integrales se sustituyen por sumas.

Independencia de variables aleatorias

- Dos vectores aleatorios $\mathbf{x}_1$ y $\mathbf{x}_2$ son independientes si el conocimiento de uno de ellos no aporta información respecto a los valores del otro.
- Es decir, la distribución de valores de $\mathbf{x}_2$ no depende de $\mathbf{x}_1$ y es la misma cualquiera que sea el valor de $\mathbf{x}_1$
- Formalmente esto se expresa como $f(\mathbf{x}_2|\mathbf{x}_1) = f(\mathbf{x}_2)$
- De la relación $f(\mathbf{x}_1, \mathbf{x}_2) = f(\mathbf{x}_2|\mathbf{x}_1)f(\mathbf{x}_1)$ se puede obtener una definición equivalente de independencia entre dos vectores aleatorios $\mathbf{x}_1$ y $\mathbf{x}_2$
- Dos vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ son independientes si $f(\mathbf{x}_1, \mathbf{x}_2) = f(\mathbf{x}_1)f(\mathbf{x}_2)$
- Su distribución conjunta es el productos de las marginales

Independencia de variables aleatorias

- En general, se dice que las variables aleatorias $(x_1, ..., x_p)$ con densidad conjunta $f(x_1, ..., x_p)$ son independientes si se verifica:

$$f(x_1, ..., x_p) = f(x_1)f(x_2)\cdots f(x_p)$$

- Al ser $x_1, ..., x_p$ independientes, tambien lo sera cualquier subconjunto de variables $(x_1, ..., x_h)$ con $h \leq p$
- Al ser $x_1, ..., x_p$ independientes, también lo será cualquier conjunto de funciones de las variables individuales $g(x_1), g(x_2), ...g(x_p)$.
- Cuando los componentes de un vector aleatorio son independientes, no se gana nada si los estudiamos conjuntamente, y conviene estudiarlos de forma separada o univariada. De hecho, la independencia de variables facilita el calculo de probabilidades.

Esperanza de una variable aleatoria vectorial

$\bar{\mathbf{x}}$ y $\mathbf{S}$ son obtenidos a partir de una muestra de datos multivariados, organizados en la matriz de datos $\mathbf{X}$. Estos datos provienen de un vector aletorio $\mathbf{x}$ que sigue cierta distribucion. Asi, denotaremos por μ a la media o *esperanza* de $\mathbf{x}$ y como $\mathbf{\Sigma}$ a la matriz de varianzas y covarianzas.

- **La esperanza o valor esperado de una variable vectorial** $\mathbf{x} = (x_1, ..., x_p)$ se denota por

$$\mu = E[\mathbf{x}] = E[x_1, ..., x_p] = [E(x_1), ..., E(x_p)] = [\mu_1, ..., \mu_p].$$

- Si la variable $\mathbf{x}$ es continua:

$$\mu = E(\mathbf{x}) = \int \mathbf{x} f(\mathbf{x}) d\mathbf{x}$$

Propiedades de la esperanza de un vector

- La **esperanza es una función lineal**, para cualquier matriz $\mathbf{A}$, y vector $\mathbf{b}$, se tiene que $E(\mathbf{Ax}+\mathbf{b}) = \mathbf{A}E(\mathbf{x}) + \mathbf{b}$
- Si $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, tenemos también que para escalares a y b:

$$E(a\mathbf{x}_1 + b\mathbf{x}_2) = aE(\mathbf{x}_1) + bE(\mathbf{x}_2)$$

- Si $\mathbf{x}_1$ y $\mathbf{x}_2$ son independientes:

$$E(\mathbf{x}_1\mathbf{x}_2) = E(\mathbf{x}_1)E(\mathbf{x}_2)$$

- Si $\mathbf{x}_1$ y $\mathbf{x}_2$ son independientes y definimos $y_1 = g_1(\mathbf{x}_1)$ y $y_2 = g_2(\mathbf{x}_2)$ entonces

$$E(y_1 y_2) = E(g_1(\mathbf{x}_1))E(g_2(\mathbf{x}_2))$$

Matriz de covarianzas de una variable vectorial

- La **Matriz de varianzas y covarianzas** de un vector aleatorio $\mathbf{x} = (x_1, ..., x_p)$, con vector de medias $\mu = (\mu_1, ..., \mu_p)$ se define por:

$$\Sigma_{\mathbf{x}} = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)']$$

- $\Sigma_{\mathbf{x}}$ contiene en la diagonal las varianzas de los componentes, representados por σ_i^2, y fuera de ella las covarianzas entre los pares de variables, representadas por σ_{ij}.
- $\Sigma_{\mathbf{x}}$ es simétrica y semidefinida positiva, es decir para cualquier vector, ω se verifica que $\omega' \Sigma_{\mathbf{x}} \omega \geq 0$

Matriz de correlación de una variable vectorial

Matriz de correlación

- La matriz de correlación de un vector aleatorio $\mathbf{x}$, con matriz de covarianzas $\mathbf{\Sigma_x}$ se define por

$$\mathbf{R_x} = \mathbf{D}^{-1/2}\mathbf{\Sigma_x}\mathbf{D}^{-1/2}, \quad \text{donde} \quad \mathbf{D} = diag(\sigma_1^2, ..., \sigma_1^p)$$

- La matriz de correlación es cuadrada y simétrica, con **unos** en la diagonal y los coeficientes de correlación entre los pares de variables fuera de la diagonal.
- Los coeficientes de correlación simple o coeficientes de correlación lineal, vienen dados por

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

- la matriz de correlación es positiva semidefinida.

Esperanzas y varianzas de transformaciones lineales

- Sea $\mathbf{x}$ un vector aleatorio de dimensión p y definimos un nuevo vector aleatorio $\mathbf{y}$ de dimensión m ($m \leq p$) tal que

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

donde $\mathbf{A}$ es una matriz rectangular de dimensiones $m \times p$.

- Denotando por $\mu_{\mathbf{x}}$, $\mu_{\mathbf{y}}$ a sus vectores de medias y $\mathbf{\Sigma}_{\boldsymbol{x}}$, $\mathbf{\Sigma}_{\boldsymbol{y}}$ a las matrices de covarianzas, se verifican las relaciones:

$$\mu_{\mathbf{y}} = \mathbf{A}\mu_{\mathbf{x}} \quad \text{y} \quad \mathbf{\Sigma}_{\boldsymbol{y}} = \mathbf{A}\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{A}^{'}$$

Para verificar la segunda relación, usamos la definicion de covarianzas, así

$$\mathbf{\Sigma}_{\boldsymbol{y}} = E[(\mathbf{y} - \mu_{\mathbf{y}})(\mathbf{y} - \mu_{\mathbf{y}})^{'}] = E[\mathbf{A}(\mathbf{x} - \mu_{\mathbf{x}})(\mathbf{x} - \mu_{\mathbf{x}})^{'}\mathbf{A}^{'}] = \mathbf{A}\mathbf{\Sigma}_{\mathbf{x}}\mathbf{A}^{'}$$

Ejemplo.

Ejemplo.Las valoraciones de los clientes sobre la puntualidad (x_1), rapidez (x_2) y limpieza (x_3) de un servicio de transporte tienen unas medias, en una escala de cero a diez, de 7, 8 y 8.5, respectivamente, con una matriz de varianzas y covarianzas:

$$\mathbf{\Sigma_x} = \begin{pmatrix} 1 & ,5 & ,7 \\ ,5 & ,64 & ,6 \\ ,7 & ,6 & 1,44 \end{pmatrix}$$

- Se construyeron dos indicadores de la calidad del servicio. El primero, es el promedio de las tres puntuaciones y el segundo es la diferencia entre el promedio de la puntualidad y la rapidez, que indica la fiabilidad del servicio y la limpieza, que indica la comodidad del mismo.

Ejemplo.

- Calcular el vector de medias y la matriz de varianzas y covarianzas para estos dos indicadores.
- Calcular la matriz de correlaciones para los dos indicadores

Dependencia entre variables aleatorias

Esperanzas condicionadas

- La esperanza de un vector $\mathbf{x}_1$ condicionada a un valor concreto de otro vector $\mathbf{x}_2$ se define como

$$E(\mathbf{x}_1|\mathbf{x}_2) = \int \mathbf{x}_1 f(\mathbf{x}_1|\mathbf{x}_2) d\mathbf{x}_1$$

- En general, esta expresión sera función del valor de $\mathbf{x}_2$. Cuando $\mathbf{x}_2$ es un valor fijo, $E(\mathbf{x}_1|\mathbf{x}_2)$ será una constante. Si $\mathbf{x}_2$ es una variable aleatoria, $E(\mathbf{x}_1|\mathbf{x}_2)$ será también una variable aleatoria.
- La esperanza de un vector aleatorio $\mathbf{x}_1$ se puede calcular a partir de las esperanzas condicionadas de la siguiente forma:

$$E(\mathbf{x}_1) = E[E(\mathbf{x}_1|\mathbf{x}_2)]$$