

- Variables
- Categoricos o cualitativos (sin orden)
 - Cuasi cuantitativos o ordinales (sin distancia)
 - cuantitativos discretos (con orden)
 - cuantitativos continuos (con orden)

Medidas descriptivas

Tendencia central → promedio muestral \bar{x} → no robusto

- mediana \tilde{x} → robusta
- moda
- media L -poda

Posición o estadísticos de orden

Cuartiles

Cuartiles

De dispersión

Rango muestral $x^{(n)} - x^{(1)}$

Variante muestral s_x^2

Desviación estandar muestral s_x

Cociente de variación $\frac{s_x}{\bar{x}}$ → relatives

Rango intercuartil: $Q_3 - Q_1$

MAD

Otras medidas

coeficiente de similitud muestral de fisher

coef. sim. Pearson

coef. sim. Bowley

coef. curtois muestral

Info Multivariada \rightarrow estudios estadísticos de varios variables. Clasificación aislada x los relaciones definidas entre ellos

Análisis exploratorio \rightarrow conocer datos

- ↳ descubrir regularidades
- ↳ existencia de estructuras ocultas
- ↳ fracciones descubiertas
- ↳ resumir info
- ↳ relaciones entre var.
- ↳ detectar anomalías

Vectores de medias muestral

$$\bar{x} = (\bar{x}_1, \dots, \bar{x}_p) \in \mathbb{R}^p$$

Matriz de covarianza x covarianza muestral

$$\Sigma = \frac{1}{n} (X - \bar{X})^T (X - \bar{X})$$

$$\Sigma \in \mathbb{R}^{p \times p}$$

Σ es semidef \Rightarrow autovalores ≥ 0

- ↳ diagonal son las varianzas muestrales de cada una de las variables
- ↳ fuera diagonal son las covarianzas

Transformaciones para variables

↳ variales aleatorias estandarizadas

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j^2}}$$

Si \bar{x} y s^2 son buenas estimaciones, sino usar otros

Transformaciones para intervalos

$$T(x) = \begin{cases} \frac{x - \bar{x}}{\bar{x}_{max} - \bar{x}} & x > \bar{x} \\ \frac{x - \bar{x}}{\bar{x} - \bar{x}_{min}} & x < \bar{x} \end{cases}$$

Convertir variables

$$\Sigma = \begin{pmatrix} s_{11} & \dots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \dots & s_{nn} \end{pmatrix}$$

$$s_{ik} = \begin{cases} > 0 & \rightarrow \text{esoc lineal positiva} \\ < 0 & \rightarrow \text{esoc lineal negativa} \\ = 0 & \rightarrow \text{no hay esoc lineal} \end{cases}$$

$$s_{ik} = \frac{1}{n} \left(\sum (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \right)$$

Correlación

$$R = \begin{pmatrix} 1 & \dots & r_{1n} \\ r_{21} & \dots & r_{2n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \dots & 1 \end{pmatrix}$$

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}} \sqrt{s_{kk}}}$$

$$r_{ik} = \begin{cases} = 1 & \rightarrow \text{línea recta pendiente positiva} \\ -1 & \rightarrow \text{línea recta pendiente negativa} \\ > 0, < 1 & \rightarrow \text{alrededor linea recta} \\ & \text{pendiente positiva} \\ & \text{> -1 y < 0} \rightarrow \text{alrededor linea recta} \\ & \text{pendiente negativa} \\ = 0 & \rightarrow \text{no hay assoc. linea} \end{cases}$$

Trazo \rightarrow suma de los diagonales
 \rightarrow mitad de cov \rightarrow importancia del problema

Alternativos robustos

- ↳ Evitan: efecto enmascaramiento: outliers engañan otros outliers
- efecto de agrupación: uno obs. es outlier en presencia de otro.

- Robustos Methodos
- Vector de mediciones
- MVE (Min Volume Ellipsoid)
- MCD (Min Covariance Determinant)

Análisis de componentes principales

Técnica exploratoria → siempre puede aplicarse

[] Técnica descriptiva

[] Libre de distorsión

[] Puede hacer combinaciones

[] Reducción de la varianza

Todos los V.O.R tienen el mismo papel

Reducir dimensiones, descartar info redundante

Visualizar info multa

Exploración vrs. latentes

Matriz de correlación x^T no usa las medidas

Detectar outliers

Si hay alto correlación entre los V.O.R

[] Puede aplicarse

$$X_i = \sum_{j=1}^p \alpha_{ij} X_j = \alpha_{i1} X_1 + \dots + \alpha_{ip} X_p$$

$\alpha = (\alpha_{11}, \dots, \alpha_{1p}) \in \mathbb{R}^p$, se busca $\|\alpha\|=1$ y ρ_{oc}

Y si se busca var máx. Los α_{ij} son loadings

$\Rightarrow \alpha_1$ es el autovector asociado al mayor autovalor de la matriz de covarianza y covariance

$$\alpha_1 \text{ con } \|\alpha_1\|=1, \text{ Cov}(X_1, X_2) = \text{Cov}(\alpha_1 X_1, \alpha_2 X_2)$$

$$= \alpha_1 \Sigma \alpha_2^T = \alpha_1 \lambda_1 \alpha_2^T = \lambda_1 \alpha_1 \alpha_2^T = 0$$

α_2 t.p. Var(X_2) sea máx. con $\alpha_2 \alpha_1^T = 0$

$$\alpha_2 \alpha_2^T = 1$$

Variabilidad de cí CP

Cada var X_k tiene variabilidad $\text{Var}(X_k) = \sum_{kk}$

$\Rightarrow \text{tr}(\Sigma) = \sum_{11} + \sum_{22} + \dots + \sum_{pp}$ variabilidad total

$$\text{tr}(\Sigma) = \sum_{i=1}^p \lambda_i, \lambda_i \text{ autovalores de } \Sigma.$$

$$\Rightarrow \lambda_i = \text{Var}(X_i) \quad \lambda_1 > \lambda_2 > \dots > \lambda_p$$

Variabilidad de cí componente es

$$\frac{\lambda_i}{\text{tr}(\Sigma)}$$

Contenido de componentes principales

Criterio 1: $\phi\%$ o variabilidad explicada

$$\frac{\lambda_1 + \dots + \lambda_{kk}}{\text{tr}(\Sigma)} \geq \frac{\phi}{100} \times \frac{\lambda_1 + \dots + \lambda_{K-1}}{\text{tr}(\Sigma)} < \frac{\phi}{100}$$

Criterio 2: n primeros componentes tf

$$\lambda_1, \dots, \lambda_n \geq 1 \text{ y } \lambda_{n+1} < 1$$

(algunos recomiendan 0.7 en lugar de ≈ 1)

Criterio 3: horizonte reflejo \Rightarrow profundo de sedimentación
o screen plot

Criterio 4: prueba de esterilidad

Corregir x los dengos

{ Si lo correg de una CP es \Rightarrow correlacion + entre la var y CP

\Rightarrow lo correg de una CP es \Rightarrow bajar se correlacion - con la CP

Los CP se calculan a partir de los mat. de correlación xf no dependen de las escales de las medidas

Biplot → interpretación de las distancias entre los obs.

→ obs que no tienen patrones

→ explícito con utilizadas correlaciones

→ representación profusa de datos multivariados

→ óvalos chicas \rightarrow muy correlacionadas
ortogonales \rightarrow no correlacionadas

verid

Matriza de vect. propios V define combio de base

Los prim. col. de V proyección de los puntos en \mathbb{R}^P sobre el q -dimensional

Los elem. de V son coseno de los ~~entre~~ ángulos

Los nuevos coord. $Vx^t = yt$

→ son scores

ACP usa redundancia para reducir dimensión

Los CP son no correlacionados y c_i no comparten info independiente

La varianza de la i -esima CP es λ_i

Componentes Principales robustas

→ ante la presencia de outliers univariados o multivariados
o usar MCD

PCA → 1^o comp → mismo signo entre correlaciónes de los datos

2^o → signos → de forma

Contraste de Hipótesis

H_0 es la que se contrasta

H_0 nunca probada → pero puede ser rechazada por la muestra

Hay evidencia, o no, contra H_0

Hipótesis parámetricas

Hipótesis libre de obs

Estadístico de contraste

- proporciona info empírica sobre la afirmación H_0
- posee una dist. muestral conocida

Región de rechazo \rightarrow si H_0 es V. \Rightarrow el estadístico (crítico)

→ poco probable esté en
lo de la mitad/las el/los valores
críticos

→ prob. de la reg. crítica es el nivel de
significación del test α

rechazo H_0 si el est. toma un valor $\in -\text{la RR}$

no rechazo H_0 si el est. toma un valor $\notin -\text{la RR}$

$$\alpha = 0.01, 0.05, 0.10$$

Error tipo 1: rechazar H_0 cuando H_0 es verd.

$$P(\text{rech } H_0 \mid H_0 \text{ es verd}) = \alpha$$

Error tipo 2: no rechazar H_0 , cuando H_0 no es falso

$$P(\text{no rech } H_0 \mid H_0 \text{ es falso}) = \beta$$

Nivel de significación (α)

→ más prob. de rech. H_0 cuando H_0 es V.

→ más riesgo admisible → elige el investigador

Error tipo 2 depende → H_1 se considera V.

→ el valor de β

→ el tamaño del error

p-valor → prob. sup. que H_0 es V, de obtener una muestra como la obtenida ó más alejada aún, en el sentido de H_1

→ menor p-valor → mayor seguridad con que rech. H_0

→ cuantifica la seguridad del rechazo del H_0

Contraste de independencia

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

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ con } P(B) > 0$$

→ mismo problema con sujetos

→ variables categóricas

→ test χ^2

Contraste de homogeneidad

↳ una variable

- ↳ si sigue la misma dist. en τ subgrupos
- ↳ τ muestras de tamaño n_j de una muestra
- ↳ ver si tiene la misma dist en los τ subgrupos
- ↳ 1 problo con τ subproblos con una una variable y K categorías
- ↳ los prob de X es ind. del muestra

Mismo este distico de contraste con $\chi^2_{(k-1)(\tau-1)}$

→ grados de libertad en función de la cont de filas y columnas

Re Re

→ unilateral o derecha

→ cuando el este distico es grande

χ^2 es asintotico y vale cuando

los frecuencias esperadas > 1 y se

sumo el 20%inf = 5

→ cuando no se puede \Rightarrow test exacto de Fisher

Independencia

Dos variable

Una población

RR. una o dos categorías

Rechaza grandes de F

$$\text{Muestra } e_{ij} = \frac{n_{i,j} - n_i \cdot n_j}{n_{..}}$$

Homogeneidad
Una variable

Al menos 2 subpopulaciones

RR. una o dos categorías

Rechaza grandes de F

$$\hat{e}_{ij} = \frac{n_{i,j} - n_i \cdot n_j}{n_{..}}$$

Test exacto Fisher

→ valores esperados de al menos 80% > 5

→ 2 métodos para el cálculo del p-value

→ para independencia

Análisis de correspondencia

- variables categóricas u ordinales
- se tiene sentido cuando las variables son dependientes

perfil fila = Frecuencias relativas de las filas

perfil columna = Frecuencia relativa de las columnas

Lectura

- coocurrencia entre el estadístico χ^2 y el tamaño muestral
- def r variables binarias para las filas y k variables binarias para columnas
- (1) Calcular freq. relativos condicionales y considerarlos como puntos en el espacio (la norma es 1)
 - Distancia χ^2 entre esos puntos
 - Proyectamos los puntos en dirección de los autovectores de $Z^T Z$
- $$Z_{ij} = \frac{f_{ij}}{\sqrt{F_i f_j}}$$
- para filas y columnas es similar

Cuanto se separan los datos del supuesto de independencia (total)

Biplot simétrico

- col cercanas al origen reflejan cat. sim
- la col promedio (col = filas)
- col cercanas entre sí reflejan cat de similar perfil en términos col. (col = filas)
- las col. cercanas y lejanas al origen reflejan alto o soc. positivo entre las categorías representadas (col = filas)
- ejes vsp Factores ocultos
- en el eje el gde de interés

Test χ^2 que reduce solo var. no son independientes \Rightarrow son dependientes pero no indica en qué sentido.

- se puede evaluar los perfiles
- estructura residuals del modelo
 - topofícales corregidas
 - ~~morfológicas~~ o ajustadas
- son en $N(0,1)$
- ortotópicos

Estadístico de Pearson e intervalo

Intervalo Σ de distancias entre perfiles fila
y perfil fila promedio ponderado
por cant. de obs.

$$\rightarrow \text{total} = \Sigma \text{ autovalores de } Z^T Z = Z^T Z$$

\Rightarrow análisis de fila o col es simétrico

\rightarrow dist $N^2 \rightarrow$ principio de equivalencia
distribucional



2 filas con misma estr.

se agrupan en una nueva.
los dist. entre los restantes

Filas permanecen invariantes
(colm col.)

\rightarrow método de variancabilidad total
de la tabla, independiente del
tamaño.

Análisis de correspondencia multiple

\rightarrow similar al ~~análisis~~ análisis simple

\rightarrow matriz de Burt.

Residuo es alto $\Leftrightarrow \geq 2$