Geometría de la Muestra y Muestreo Aleatorio

Vectores aleatorios, Geometría de la Muestra

Santiago Alférez

Agosto de 2020

MACC
Universidad del Rosario

Contenidos

Geometría de la Muestra

Muestras Aleatorias

Valores Esperados de la Media Muestral y la Matriz de Covarianza

Geometría de la Muestra

Observación multivariable

Se miden p variables en n observaciones :

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- Cada fila de $\mathbf{X}$ representa una observación multivariable.
- Dado que el conjunto de medidas es a menudo una realización particular de lo que podría haberse observado, decimos que los datos son una muestra de tamaño n de una población de p-variables.
- La muestra entonces consta de n mediciones, cada una de las cuales tiene componentes p.

Geometría de la muestra

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \begin{array}{l} \leftarrow \text{primera observación} \\ \\ \\ \leftarrow n\text{-ésima observación} \end{array}$$

Formas de graficar

- Los datos se pueden graficar de dos formas:
  - * n puntos en el espacio p-dimensional.
  - * p vectores en el espacio n-dimensional.

Geometría de la muestra

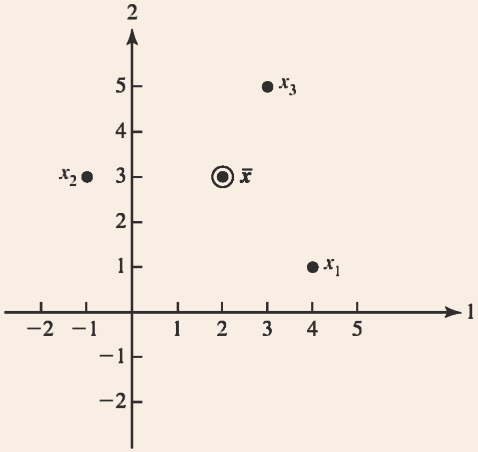
N puntos en el espacio p-dimensional

- El diagrama de dispersión de n puntos en el espacio p-dimensional proporciona información sobre las ubicaciones y la variabilidad de los puntos.
- Si los puntos se consideran esferas sólidas, el vector de media muestral $\overline{\mathbf{x}}$, es el centro de equilibrio.
- La variabilidad ocurre en más de una dirección, y se cuantifica mediante la matriz de varianza-covarianza de la muestra $\mathbf{S}_n$.
- Cuando p es mayor que 3, esta representación del diagrama de dispersión no se puede representar gráficamente.
- La consideración de los datos como n puntos en p dimensiones proporciona conocimientos que no se ven en las expresiones algebraicas.
- los conceptos ilustrados para $p = 2$ o $p = 3$ siguen siendo válidos para los demás casos.

Ejemplo

Graficar el vector media para los siguientes datos tomando $n = 3$ puntos en un espacio de $p = 2$ dimensiones.

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$



Geometría de la muestra

p vectores en el espacio n-dimensional

- La idea es tomar los elementos de las columnas de la matriz como las coordenadas de los vectores:

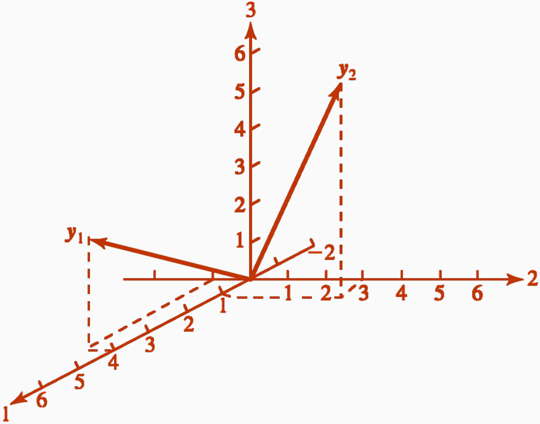
$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \left[\mathbf{y}_1 \mid \mathbf{y}_2 \mid \cdots \mid \mathbf{y}_p\right]$$

- El i-ésimo punto $\mathbf{y}_i' = \left[\; x_{1i}, x_{2i}, \ldots, x_{ni} \;\right]$ está determinado por la n-tupla de todas las medidas en la i-ésima variable.
- En esta representación geométrica, representamos $\mathbf{y}_1, \ldots, \mathbf{y}_p$ como vectores en lugar de puntos, en el espacio de dimensión n.

Ejemplo

Graficar los siguientes datos como $p = 2$ vectores en un espacio de $n = 3$ dimensiones.

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$



Interpretación geométrica de la media muestral

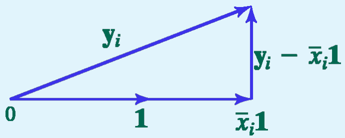
- Sea el vector $\mathbf{1}$ de $\mathbb{R}^n$ definido por $\mathbf{1}_n' = [1, 1, \ldots, 1]$.
- El vector $\mathbf{1}$ forma ángulos iguales con cada uno de los ejes de coordenadas n, entonces el vector $(1/\sqrt{n})\mathbf{1}$ tiene magnitud de una unidad en la misma dirección.
- Considere el vector $\mathbf{y}_i' = [x_{1i}, x_{2i}, \ldots, x_{ni}]$. La proyección de $\mathbf{y}_i$ sobre el vector unitario $(1/\sqrt{n})\mathbf{1}$ es,

$$\mathbf{y}_i' \left(\frac{1}{\sqrt{n}} \mathbf{1} \right) \frac{1}{\sqrt{n}} \mathbf{1} = \frac{x_{1i} + x_{2i} + \cdots + x_{ni}}{n} \mathbf{1} = \bar{x}_i \mathbf{1}$$

- Así, la media de la muestra $\bar{x}_i = (x_{1i} + x_{2i} + \cdots + x_{ni})/n = \mathbf{y}_i' \mathbf{1}/n$ corresponde al múltiplo de $\mathbf{1}$ requerido para dar la proyección de y_i en la línea determinada por $\mathbf{1}$.

Geometría de la muestra

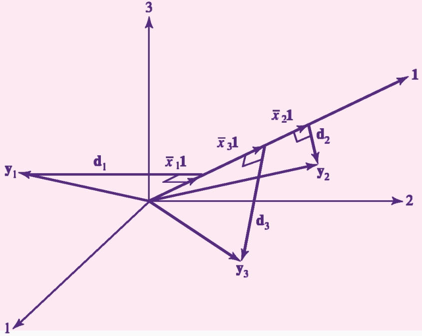
Descomposición de $\mathbf{y}_i$



Desviación o media corregida:

$$\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1} = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix}$$

Descomposición de $\mathbf{y}_i$ en la componente media $\bar{x}_i\mathbf{1}$ y la desviación $\mathbf{d}_i$



Ejemplo

Para los siguientes datos, descomponga cada vector columna en las componentes de vector media y de desviación.

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

$$L_x = \sqrt{\mathbf{x}'\mathbf{x}}$$

$$\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i \mathbf{1} = \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix}$$

Relación entre la desviación y la desviación estándar

$$L_{\mathbf{d}_i}^2 = \mathbf{d}_i'\mathbf{d}_i = \sum_{j=1}^n \left(x_{ji} - \bar{x}_i\right)^2$$

$$\mathbf{d}_i'\mathbf{d}_k = \sum_{j=1}^n \left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right)$$

$$\sum_{j=1}^n \left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right) = \sqrt{\sum_{j=1}^n \left(x_{ji} - \bar{x}_i\right)^2}\sqrt{\sum_{j=1}^n \left(x_{jk} - \bar{x}_k\right)^2}\cos\left(\theta_{ik}\right)$$

$$\cos\theta = \frac{\mathbf{x}'\mathbf{y}}{L_{\mathbf{x}}L_{\mathbf{y}}}$$

El coseno del ángulo es el coeficiente de correlación

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos\left(\theta_{ik}\right)$$

$$\mathbf{d}_i'\mathbf{d}_k = L_{\mathbf{d}_i}L_{\mathbf{d}_k}\cos\left(\theta_{ik}\right)$$

Ejemplo

Para los siguientes datos, calcule la matriz de covarianza muestral $\mathbf{S}_n$ y la matriz de correlación muestral $\mathbf{R}$ usando las desviaciones y los conceptos geométricos.

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

Resumen sobre la interpretación geométrica de la muestra

1. La proyección de una columna y_i de la matriz de datos X sobre el vector $\mathbf{1}$ es el vector $\bar{x}_i\mathbf{1}$. El vector $\bar{x}_i\mathbf{1}$ tiene una longitud $\sqrt{n}\,|\bar{x}_i|$. Por lo tanto, la i-ésima media muestral, $\bar{x}_i$, está relacionada con la longitud de la proyección de $\mathbf{y}_i$ en $\mathbf{1}$.
2. La información que comprende S_n se obtiene de los vectores de desviación $\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i\mathbf{1} = [x_{1i} - \bar{x}_i, \bar{x}_{2i} - \bar{x}_i, \ldots, x_{ni} - \bar{x}_i]'$. El cuadrado de la longitud de $\mathbf{d}_i$ es ns_{ii}, y el producto (interno) entre $\mathbf{d}_i$ y $\mathbf{d}_k$ es ns_{ik}.
3. La correlación de muestra r_{ik} es el coseno del ángulo entre $\mathbf{d}_i$ y $\mathbf{d}_k$

Muestras Aleatorias

Muestras aleatorias

Consideraciones

- Para estudiar la variabilidad muestral de la estadística, por ejemplo de $\bar{\mathbf{x}}$ y $\mathbf{S}_n$, con el objetivo de realizar inferencias, es necesario hacer supuestos acerca de las variables cuyos valores observados forman el conjunto de datos $\mathbf{X}$.
- Supongamos que los datos no se han observado, pero intentamos recolectar n conjuntos de medidas sobre p variables.
- Dado que las medidas no se pueden predecir exactamente antes de que se realicen, las tratamos (las medidas) cómo variables aleatorias.

Definiendo una muestra aleatoria

- Sea X_{jk} una variable aleatoria que representa la entrada (j,k)-ésima en la matriz de datos.
- Cada conjunto de medidas $\mathbf{X}_j$ con p variables es un vector aleatorio, y tenemos la matriz aleatoria

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix}$$

Muestras aleatorias

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix}$$

Definiendo una muestra aleatoria

- Si los vectores de fila $\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n'$ representan observaciones independientes de una distribución conjunta común con función de densidad $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_p)$ entonces $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ se dice que forman una muestra aleatoria de $f(\mathbf{x})$.
- $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ forman una muestra aleatoria si su función de densidad conjunta está dada por el producto $f(\mathbf{x}_1) f(\mathbf{x}_2) \cdots f(\mathbf{x}_n)$, donde $f(\mathbf{x}_j) = f(x_{j1}, x_{j2}, \ldots, x_{jp})$ es la función de densidad para el j-ésimo vector fila.

Muestra aleatoria

Consideraciones de la definición de muestra aleatoria

- Las medidas de las variables p en una sola prueba, como $\mathbf{X}'_j = [X_{j1}, X_{j2}, \ldots, X_{jp}]$, normalmente estarán correlacionadas. Por otro lado, las mediciones de diferentes ensayos deben ser independientes.
- La independencia de las mediciones de una prueba a otra puede no ser válida cuando es probable que las variables se cambien con el tiempo, como ocurre con los conjuntos de precios de acciones p o indicadores económicos p.
- Las violaciones del supuesto tentativo de independencia pueden tener un impacto grave en la calidad de las inferencias estadísticas.

Valores Esperados de la Media Muestral y la Matriz de Covarianza

Valores esperados de la media y la covarianza muestrales

Consideraciones de los valores esperados

- Si las n componentes no son independientes o las distribuciones marginales no son idénticas, la influencia de las medidas individuales (coordenadas) sobre la localización es asimétrica.
- Entonces, podríamos considerar usar una distancia en la cuál las coordenadas fueran ponderadas de forma desigual, cómo en la distancia estadística o la forma cuadrática.
- Se puede ver cómo $\overline{\mathbf{X}}$ y $\mathbf{S}_n$ se comportan como estimadores puntuales del vector de media poblacional correspondiente $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$.

Valores esperados de la media y la covarianza muestrales

Estimador para μ

Sea $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ una muestra aleatoria con una distribución conjunta que tiene vector media $\boldsymbol{\mu}$ y matriz de covarianza $\boldsymbol{\Sigma}$. Entonces, $\overline{\mathbf{X}}$ es un estimador insesgado de $\boldsymbol{\mu}$ y su matriz de covarianza es:

$$\frac{1}{n}\boldsymbol{\Sigma}.$$

$$E(\overline{\mathbf{X}}) = \boldsymbol{\mu} \qquad \text{(vector de media poblacional)}$$

$$\mathrm{Cov}(\overline{\mathbf{X}}) = \tfrac{1}{n}\boldsymbol{\Sigma} \qquad \left(\begin{array}{c}\text{matriz de covarianza poblacional} \\ \text{dividida por el tamaño de la muestra}\end{array}\right)$$

Valores esperados de la media y la covarianza muestrales

Estimador para $\boldsymbol{\Sigma}$

Para la matriz de covarianza $\mathbf{S}_n$,

$$E\left(\mathbf{S}_n\right) = \frac{n-1}{n}\boldsymbol{\Sigma} = \boldsymbol{\Sigma} - \frac{1}{n}\boldsymbol{\Sigma} \quad \text{o,} \quad E\left(\frac{n}{n-1}\mathbf{S}_n\right) = \boldsymbol{\Sigma}$$

Entonces, $[n/(n-1)]\mathbf{S}_n$ es un estimador insesgado de $\boldsymbol{\Sigma}$, mientras que $\mathbf{S}_n$ es un estimador sesgado con $sesgo = E\left(\mathbf{S}_n\right) - \boldsymbol{\Sigma} = -(1/n)\boldsymbol{\Sigma}$

Matriz de covarianza muestral insesgada

$$\mathbf{S} = \left(\frac{n}{n-1}\right)\mathbf{S}_n = \frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{X}_j - \overline{\mathbf{X}}\right)\left(\mathbf{X}_j - \overline{\mathbf{X}}\right)'$$

Puesto que $\mathbf{S}$ se usa generalmente para pruebas multivariables, $\mathbf{S}_n$ será sustituida por $\mathbf{S}$.