Nama : Isda Sari

NIM : 0607093

3.3 Sampel Acak dan Nilai Harapan dari Rerata Sampel dan Matrik Kovarians

Kita asumsikan bahwa nilai variabel pengamatan berdasarkan himpunan data . Andaikan data bukan hasil pengamatan, tetapi kita berniat mengumpulkan n himpunan dari pegukuran p variabel. Sebelum pengukuran dibuat, nilai-nilai tersebut tidak ada, oleh karena itu diprediksi. Konsekuensinya, diproses sebagai variabel acak.

Didalam konteks ini, misal entri (i,j) di dalam data matrik menjadi variabel acak . Masing-masing himpunan berukuran pada p variabel adalah sebuah vektor acak dan kita mempunyai matrik acak:

$$\underset{(pxn)}{\mathbf{X}} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{bmatrix} = [X_1, X_2, \ldots, X_n] \qquad (3.8)$$

Sebuah sampel acak didefinisikan:

> Jika vektor kolom $X_1, X_2, \ldots\ldots, X_n$ pada (3.8) muncul pengamatan independen dari distribusi gabungan dengan fungsi kepadatan $f(x) = f(x_1, x_2, \ldots\ldots, x_p)$, maka $X_1, X_2, \ldots\ldots, X_n$ dikatakan bentuk suatu sampel acak dari $f(x)$. Secara matematis, hasil $f(x_1) f(x_2) \ldots f(x_n)$, dimana $f(x_j) = f(x_{1j}, x_{2j}, \ldots, x_{pj})$ adalah fungsi kepadatan dari kolom vektor ke-j.

Perhatikan , dua *point* yang berhubungan dengan definisi sampel acak:

1. Pengukuran p variabel pada sebuah percobaan tunggal, seperti $X'_j = [X_{1j}, X_{2j}, \ldots, X_{pj}]$, biasanya berhubungan. Tentu saja, kita harapkan ini menjadi sebuah kasus. Ukuran dari percobaan yang berbeda , bagaimanapun harus menjadi independen.

2. Kebebasan pengukuran dari percobaan ke percobaan tidak mungkin tetap manakala variabel mungkin menyimpang dari waktu ke waktu, seperti himpunan p harga bursa atau p indikator ekonomi.

Jarak Euclid, jika komponen-komponennya suatu vektor bebas dan mempunyai beberapa varians. Misalkan letak baris ke-i $Y_i' = [\mathrm{X}_{i1}, \mathrm{X}_{i2}, \ldots, \mathrm{X}_{in}]$ dari X dipandang sebagai titik pada n dimensi.

Letak dari titik ini ditentukan oleh peluang distribusi gabungan $f(y_i) = f(x_{i1}, x_{i2}, \ldots, x_{in})$ ketika $x_{i1}, x_{i2}, \ldots, x_{in}$ adalah sampel acak, $f(y_i) = f(x_{i1}, x_{i2}, \ldots, x_{in}) = f_i(x_{i1}) f_i(x_{i2}) \ldots f_i(x_{in})$ dan akibatnya, setiap penambahan koordinat x_{ij} sama dengan letak distribusi marginal identik $f_i(x_{ij})$. Jika n komponen tidak bebas atau distribusi marginal tidak identik, kegagalan dari pengukuran (koordinat) individual pada letak asimetris.

Kesimpulannya, jangkauan distribusi sampling untuk $\bar{\mathrm{X}}$ dan S_n dengan tidak membuat asumsi mengenai variabel pokok distribusi gabungan. Sehingga kita dapat melihat bagaimana $\bar{\mathrm{X}}$ dan S_n menyatakan titik estimasi dari populasi korespondensi rerata vektor μ dan matrik kovarians Σ.

Akibat 3.1

Misalkan $\mathrm{X}_1, \mathrm{X}_2, \ldots\ldots, \mathrm{X}_n$ adalah sampel acak dari distribusi gabungan dengan rerata vektor μ dan matrik kovarians Σ. Maka $\bar{\mathrm{X}}$ adalah estimasi tak bias pada μ dan matrik kovariansnya adalah $\frac{1}{n}\Sigma$, dimana

$$E(\bar{\mathrm{X}}) = \mu \qquad (\textit{populasi rerata vektor})$$

$$Cov(\bar{\mathrm{X}}) = \frac{1}{n}\Sigma \quad \text{(populasi matrik varians-kovarians dibagi oleh ukuran sampel)}$$

untuk matrik kovarians S_n, $E(S_n) = \frac{n-1}{n}\Sigma = \Sigma - \frac{1}{n}\Sigma$ maka $E\left(\frac{n}{n-1}S_n\right) = \Sigma$

Jadi $\left[\frac{n}{(n-1)}\right]S_n$ merupakan estimasi tak bias pada Σ. Sedangkan S_n merupakan estimasi bias dengan $(bias) = E(S_n) - \Sigma = -\left(\frac{1}{n}\right)\Sigma$

Bukti:

Diketahui $\bar{\mathrm{X}} = (\mathrm{X}_1 + \mathrm{X}_2 + \ldots + \mathrm{X}_n)/n$, dengan menggunakan pernyataan (2-24)

$$E(\mathrm{X} + Y) = E(\mathrm{X}) + E(Y)$$

$$E(A\mathrm{X}B) = AE(\mathrm{X})B$$

Diperoleh:

$$
\begin{aligned}
E(\bar{\mathrm{X}}) &= E\left(\frac{1}{n}\mathrm{X}_1+\frac{1}{n}\mathrm{X}_2+\ldots+\frac{1}{n}\mathrm{X}_n\right)\\
&= E\left(\frac{1}{n}\mathrm{X}_1\right)+E\left(\frac{1}{n}\mathrm{X}_2\right)+\ldots+E\left(\frac{1}{n}\mathrm{X}_n\right)\\
&= \frac{1}{n}E(\mathrm{X}_1)+\frac{1}{n}E(\mathrm{X}_2)+\ldots+\frac{1}{n}E(\mathrm{X}_n)\\
&= \frac{1}{n}\mu+\frac{1}{n}\mu+\ldots+\frac{1}{n}\mu\\
&= \mu
\end{aligned}
$$

Kemudian $\left(\bar{\mathrm{X}}-\mu\right)\left(\bar{\mathrm{X}}-\mu\right)' = \left(\frac{1}{n}\sum_{j=1}^{n}\left(\mathrm{X}_j-\mu\right)\right)\left(\frac{1}{n}\sum_{l=1}^{n}\left(\mathrm{X}_l-\mu\right)\right)'$

$$= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{l=1}^{n}\left(\mathrm{X}_j-\mu\right)\left(\mathrm{X}_l-\mu\right)'$$

maka $Cov(\bar{\mathrm{X}}) = E\left(\bar{\mathrm{X}}-\mu\right)\left(\bar{\mathrm{X}}-\mu\right)' = \frac{1}{n^2}\left(\sum_{j=1}^{n}\sum_{l=1}^{n}E\left(\mathrm{X}_j-\mu\right)\left(\mathrm{X}_l-\mu\right)'\right)$

untuk $j \neq l$, setiap entri di $E\left(\mathrm{X}_j-\mu\right)\left(\mathrm{X}_l-\mu\right)'$ adalah nol.
Sebab entri-entrinya adalah kovarians antara komponen dari X_j dan komponen X_l dan independen.
Pada pernyataan (2-29) yaitu: $Cov\left(\mathrm{X}_i,\mathrm{X}_k\right)=0,$ *jika* X_i *dan* X_k *independen*

Karena itu, $Cov\left(\bar{\mathrm{X}}\right) = \frac{1}{n^2}\left(\sum_{j=1}^{n}E\left(\mathrm{X}_j-\mu\right)\left(\mathrm{X}_j-\mu\right)'\right)$.

Karena $\Sigma = E(\mathrm{X}_j-\mu)(\mathrm{X}_j-\mu)'$ adalah populasi umum matrik kovarians untuk setiap X_j , kita peroleh

$$Cov(\bar{\mathrm{X}}) = \frac{1}{n^2}\left(\sum_{j=1}^{n}E\left(\mathrm{X}_j-\mu\right)\left(\mathrm{X}_j-\mu\right)'\right) = \frac{1}{n^2}\underbrace{\left(\Sigma+\Sigma+\ldots+\Sigma\right)}_{n\ suku} = \frac{1}{n^2}(n\Sigma) = \left(\frac{1}{n}\right)\Sigma$$

Berlaku pada nilai harapan S_n dengan catatan $\left(\mathrm{X}_{ij}-\bar{\mathrm{X}}_i\right)\left(\mathrm{X}_{kj}-\bar{\mathrm{X}}_k\right)$
Merupakan elemen (i,k) pada $\left(\mathrm{X}_i-\bar{\mathrm{X}}\right)\left(\mathrm{X}_j-\bar{\mathrm{X}}\right)'$

Mewakili penjumlahan dari matrik persegi dan perkalian dapat dituliskan sebagai berikut:

$$\sum_{j=1}^{n}\left(\mathrm{X}_j-\bar{\mathrm{X}}\right)\left(\mathrm{X}_j-\bar{\mathrm{X}}\right)'=\sum_{j=1}^{n}\left(\mathrm{X}_j-\bar{\mathrm{X}}\right)\mathrm{X}_j'+\left(\sum_{j=1}^{n}\left(\mathrm{X}_j-\bar{\mathrm{X}}\right)\right)\left(-\bar{\mathrm{X}}\right)'=\sum_{j-1}^{n}\mathrm{X}_j\mathrm{X}_j'-n\bar{\mathrm{X}}\bar{\mathrm{X}}'$$

karena $\sum_{j-1}^{n}(\mathrm{X}_j-\bar{\mathrm{X}})=0$ dan $n\bar{\mathrm{X}}'=\sum_{j=1}^{n}\mathrm{X}_j'$. Sehingga nilai harapannya adalah

$$E\left(\sum_{j-1}^{n}\mathrm{X}_j\mathrm{X}_j'-n\bar{\mathrm{X}}\bar{\mathrm{X}}'\right)=\sum_{j-1}^{n}E(\mathrm{X}_j\mathrm{X}_j')-nE(\bar{\mathrm{X}}\bar{\mathrm{X}}')$$

Untuk beberapa vektor V dengan $E(V)=\mu_v$ dan $Cov(V)=\Sigma_v$, kita mempunyai $E(VV')=\Sigma_v+\mu_v\mu_v'$. Konsekuensinya $E(\mathrm{X}_j\mathrm{X}_j')=\Sigma+\mu\mu'$ dan $E(\bar{\mathrm{X}}\bar{\mathrm{X}}')=\frac{1}{n}\Sigma+\mu\mu'$

Gunakan hasil ini, $\sum_{j=1}^{n}E(\mathrm{X}_j\mathrm{X}_j')-nE(\bar{\mathrm{X}}\bar{\mathrm{X}}')=n\Sigma+n\mu\mu'-n\left(\frac{1}{n}\Sigma+\mu\mu'\right)=(n-1)\Sigma$

karena $S_n=\left(\frac{1}{n}\right)\left(\sum_{j=1}^{n}\mathrm{X}_j\mathrm{X}_j'-n\bar{\mathrm{X}}\bar{\mathrm{X}}'\right)$ sedemikian sehingga $E(S_n)=\frac{(n-1)}{n}\Sigma$.

Akibat 3.1

Ditunjukkan bahwa entri (i,k) dari $\left[\frac{n}{(n-1)}\right]S_n$ adalah estimasi takbias pada σ_{ik}

Bagaimanapun, sampel individu standar deviasi $\sqrt{S_{ii}}$ yang bukan estimasi takbias dari koresponden banyak populasi $\sqrt{S_{ii}}$, dihitung dengan salah satu n atau n-1 sebagai pembagi. Selain itu, koefisien korelasi r_{ik} yang bukan estimasi takbias dari banyak populasi ρ_{ik} . Meskipun, bias $E(\sqrt{S_{ii}})-\sqrt{\sigma_{ii}}$ atau $E(r_{ik})-\rho_{ik}$, biasanya dapat diabaikan jika ukuran n sampel cukup besar.

Pertimbangan pada bias mendorong suatu perubahan definisi dari matrik varians-kovarians sampel. Akibat 3.1 menyediakan kepada kita dengan suatu estimasi takbias S untuk Σ . (Takbias) Matrik varians-kovarians sample

$$S=\left(\frac{n}{n-1}\right)S_n=\frac{1}{n-1}\sum_{j=1}^{n}\left(\mathrm{X}_j-\bar{\mathrm{X}}\right)\left(\mathrm{X}_j-\bar{\mathrm{X}}\right)' \quad \text{.... (3-11)}$$

Disini S mempunyai entri (i,k) , $(n-1)^{-1}\sum\left(\mathrm{X}_{ij}-\bar{\mathrm{X}}_i\right)\left(\mathrm{X}_{kj}-\bar{\mathrm{X}}_k\right)$. Definisi dari sampel kovarians ini biasa dipakai dalam tes statistik multivariat. Oleh karena itu, akan digantikan S_n sebagai matrik kovarians sampel dalam berbagai materi.

3.4 Varians Umum

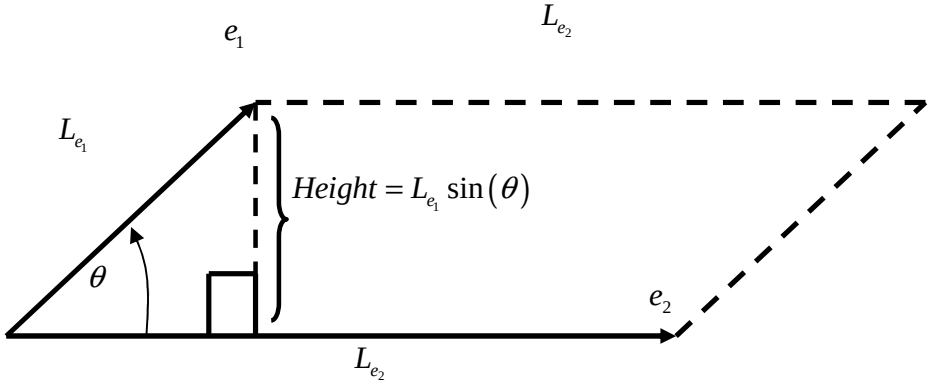
Untuk variabel tunggal biasanya varians sampel digunakan untuk menggambarkan banyaknya variasi pengukuran pada variabel itu. Ketika p variabel merupakan pengamatan pada masing-masing unit, variasi digambarkan oleh matrik varians-kovarians sampel

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1p} & s_{2p} & \cdots & s_{pp} \end{bmatrix} = \left\{ s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} \left(x_{ij} - \overline{x}_i \right) \left(x_{kj} - \overline{x}_k \right) \right\}$$

Matrik kovarians sampel memuat p varians dan $\frac{1}{2} p(p-1)$ kemungkinan kovarians yang lain. Kadang-kadang itu yang diperlukan untuk menentukan suatu nilai numerik tunggal pada variasi yang dinyatakan oleh S. Satu pilihan untuk nilai determinan dari S, yang mereduksi varians sampel biasa dari suatu karakteristik tunggal p = 1. Determinan ini disebut varians sampel umum.

$$\textit{Varians sampel umum} = |S| \quad \text{.... (3-12)}$$

Ketika p>1, di dalam prosesnya beberapa informasi sampel hilang. Penjelasan secara geometrik dari $|S|$ akan membantu kita mendeskripsikannya. Anggap dua vektor deviasi dalam bidang $e_1 = y_1 - \overline{x}_1 1 \; dan \; e_2 = y_2 - \overline{x}_2 1$. Misalkan L_{e_1} merupakan panjang dari e_1 dan L_{e_2} merupakan panjang dari e_2 . Geometrinya,



dan luas trapesium di gambar ini adalah $\left(L_{e_1} \sin(\theta) \right) L_{e_2}$.

Karena $\cos^2(\theta) + \sin^2(\theta) = 1$ kita dapat menggambarkan luas ini sebagai

$$\textit{Luas} = L_{e_1} L_{e_2} \sqrt{1 - \cos^2(\theta)}$$

Dari (3-5) dan (3-7)

$$L_{e_i}{}^2 = e_i' e_i = \sum_{j=1}^{n} \left(x_{ij} - \bar{x}_i \right)^2 \qquad (3-5)$$

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \cos\left(\theta_{ik}\right) \qquad (3-7)$$

Sehingga $L_{e_1} = \sqrt{\sum_{j=1}^{n} \left(x_{1j} - \bar{x}_1 \right)^2} = \sqrt{(n-1)s_{11}}$, $L_{e_2} = \sqrt{\sum_{j=1}^{n} \left(x_{2j} - \bar{x}_2 \right)^2} = \sqrt{(n-1)s_{22}}$ dan $\cos(\theta) = r_{12}$

Oleh karena itu, $Luas = (n-1)\sqrt{s_{11}}\sqrt{s_{22}}\sqrt{1-r_{12}^2} = (n-1)\sqrt{s_{11}s_{22}\left(1-r_{12}^2\right)}$ ... (3-13)

Juga,

$$|S| = \left| \begin{bmatrix} s_{11} & s_{21} \\ s_{12} & s_{22} \end{bmatrix} \right| = \left| \begin{bmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}}r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}}r_{12} & s_{22} \end{bmatrix} \right| = s_{11}s_{22} - s_{11}s_{22}r_{12}^2 = s_{11}s_{22}\left(1-r_{12}^2\right) \quad \text{....(3-14)}$$

Jika kita dibanding (3-14) dan (3-13), terlihat bahwa $|S| = \frac{(luas)^2}{(n-1)^2}$

Asumsikan bahwa $|S| = (n-1)^{-(p-1)}(volume)^2$ menyatakan volume dengan n space oleh p-1 vektor deviasi $e_1, e_2, \ldots, e_{p-1}$ kita dapat menentukan hasil untuk p vektor deviasi dengan induksi

$$Varians\ sampel\ umum = |S| = (n-1)^{-p}(volume)^2 \qquad \text{..... (3-15)}$$

dengan vektor deviasi $e_1 = y_1 - \bar{x}_1 1, e_2 = y_2 - \bar{x}_2 1, \ldots, e_p = y_p - \bar{x}_p 1$

Varians umum digambarkan ke dalam p-space titik sebaran yang mewaliki data. Seringnya menggambarkan secara intuisi menyangkut sebaran titik rerata sample $\bar{x}' = \left[\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p \right]$. Anggap pengukuran jarak

$$0 < (jarak)^2 = x'Ax \quad jika\ x \neq 0 \quad \text{..... (2-19)}$$

dengan $\bar{x}$ berperan pada titik tertentu μ dan S^{-1} berperan pada A. Dengan pilihan ini, koordinat $x' = \left[x_1, x_2, \ldots, x_p \right]$ pada titik suatu jarak konstan c dari $\bar{x}$ yang memenuhi

$$\left(x - \bar{x}\right)' S^{-1}\left(x - \bar{x}\right) = c^2 \quad \text{..... (3-16)}$$

Ketika p=1 $\left(x - \bar{x}\right)' S^{-1}\left(x - \bar{x}\right) = \frac{\left(x_1 - \bar{x}_1\right)^2}{s_{11}}$ merupakan jarak persegi dari x_1 ke $\bar{x}_1$ dalam satuan standar deviasi.

Persamaan (3-16) menetapkan pusat hiperelipsoid di $\bar{x}$ (suatu elips jika p=2) . Ini dapat ditunjukkan menggunakan kalkulus integral bahwa volume pada hiperelipsoid berhubungan dengan $|S|$. Yaitu,

$$\left\{x:\left(x-\bar{x}\right)' S^{-1}\left(x-\bar{x}\right)\le c^2\right\}=k_p\left|S\right|^{1/2}c^p$$

atau

$$\left(volume\ elipsoid\right)^2=\left(kons\tan\right)\left(\text{var}\,iansi\ sampel\ umum\right)$$

dimana $k_p=\dfrac{2\pi^{p/2}}{p\Gamma\left(p/2\right)}$; dengan $\Gamma(z)$ merupakan notasi fungsi gamma.

Walaupun variansi umum mempunyai beberapa intuisi yang memuaskan penafsiran geometris, inilah kelemahannya dalam menggambarkan matrik kovarians sampel S.

Ilustrasi:

Misalkan tiga matrik kovarians sampel dan diperoleh koefisien kovarians sebagai berikut:

$$S=\begin{bmatrix}5 & 4\\ 4 & 5\end{bmatrix} \qquad S=\begin{bmatrix}5 & -4\\ -4 & 5\end{bmatrix} \qquad S=\begin{bmatrix}3 & 0\\ 0 & 3\end{bmatrix}$$

$$r_{12}=\frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}}=\frac{4}{\sqrt{5}\sqrt{5}}=8 \qquad r_{12}=\frac{-4}{\sqrt{5}\sqrt{5}}=-8 \qquad r_{12}=\frac{0}{\sqrt{3}\sqrt{3}}=0$$

Setiap matrik kovarians mempunyai beberapa varians umum yaitu $\left|S\right|=9$, namun jelas berbeda struktur korelasi (kovarians). Perbedaan struktur korelasi tidak ditemukan oleh $\left|S\right|$.