BAB II

MATERI PENUNJANG

Dalam bagian ini akan disajikan beberapa teori-teori dasar statistik klasik, khususnya mengenai statistika inferensi dan beberapa teorema yang penting untuk mendukung pembahasan berikutnya. Setelah itu akan digunakan metode bootstrap untuk menaksir parameter. Metode bootstrap yang akan dibahas nanti adalah metode bootstrap dari *Efron*.

2.1 Kekonvergenan dari Variabel Random dalam Statistika

Misalkan $\{X_n\}$ adalah penulisan dari sebuah barisan dari variabel random $(X_1, X_2, ..., X_n)$ dengan distribusi yang sama dan didefenisikan pada ruang sampel Ω yang sama. Misalkan X variabel random yang didefenisikan pada ruang yang sama.

Defenisi 2.1.1: $\{X_n\}$ disebut konvergen secara probalitive ke suatu variabel random X (ditulis, $X_n \xrightarrow{p} X$), $\forall \varepsilon > 0$. Untuk $n \to \infty$ berlaku $P[|X_n - X| \geq \varepsilon] \to 0$.

Teorema 2.1.1: Misalkan $X_n \xrightarrow{p} X$ dan $Y_n \xrightarrow{p} Y$, maka

$$X_n + Y_n \xrightarrow{p} X + Y$$

Bukti: Teorema ini dibuktikan menurut *defenisi 2.1.1* yaitu sebagai berikut:

$X_n \xrightarrow{p} X$ maka dengan *defenisi 2.1.1* $P[|X_n - X| \geq \varepsilon] \to 0$ untuk $n \to \infty$ $Y_n \xrightarrow{p} Y$, maka dengan *defenisi 2.1.1* $P[|Y_n - Y| \geq \varepsilon] \to 0$ untuk $n \to \infty$.

$$P[|(X_n+Y_n)-(X+Y)| \geq \varepsilon] = P[|(X_n-X)+(Y_n-Y)| \geq \varepsilon]$$

$$\leq P[|X_n-X|+|Y_n-Y| \geq \varepsilon]$$

$$\leq P[|X_n-X| \geq \varepsilon/2]+|Y_n-Y| \geq \varepsilon/2]$$

Dua suku terakhir untuk $n\to\infty$ masing-masing akan mencapai nol. Jadi sebagai hasil akhirnya:

$P[|(X_n+Y_n)-(X-Y)| \geq \varepsilon] \to 0$ jika $n\to\infty$.

Teorema 2.1.2: Jika $X_n \xrightarrow{p} X$ dan a konstan maka $aX_n \xrightarrow{p} aX$

Bukti: Jika a = 0 maka dengan sendirinya teorema terbukti karena $0 \xrightarrow{p} 0$.

Anggap $a \neq 0$, Untuk setiap $\varepsilon > 0$,

$$P[|aX_n-X| \geq \varepsilon] = P[|a||X_n-X| \geq \varepsilon]$$

$$= P[|X_n-X| \geq \varepsilon/|a|] \to 0 \text{ jika } n\to\infty.$$

Defenisi 2.1.2: X_n konvergen ke X dalam distribusi (ditulis, $X_n \xrightarrow{d} X$). Untuk ini fungsi distribusi dari X_n dan X masing-masing dinyatakan dengan $F_n(x)$ dan $F(x)$. Bila $F_n(x) \to F(x)$, $n \to \infty$ untuk semua titik x dimana $F(x)$ kontinu, maka dinyatakan bahwa barisan $X_n \xrightarrow{d} X$.

Dalam kasus seperti ini, dikatakan bahwa $F_n(x) = F_{X_n}(x)$ konvergen lengkap ke $F(x) = F_x(x)$, ditulis $F_n \to F$.

2.2 Model Regresi Linier

Dalam regresi linier dibedakan dua jenis variabel yaitu X yang merupakan variabel bebas (prediktor) sedangkan y merupakan variabel terikat (respon).

Sebuah model regresi linier yang mencakup lebih dari satu variabel bebas disebut *model regresi berganda*. Model umum regresi linier adalah

$$y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i , \qquad (2.2.1)$$

dengan: $\beta_0, \beta_1, \ldots, \beta_k$ parameter-parameter .

Pada umumnya model tersebut dapat ditulis dengan :

$$y = \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \quad \text{dengan } X_{i0} = 1 \text{ , untuk i} = 1$$

(2.2.2)

Model linier pada (2.2.1) dapat ditulis sebagai berikut:

$$y_i = x_i \beta + \varepsilon_i \qquad \text{untuk i} = 1,2,\ldots,n \qquad (2.2.3)$$

ε_i adalah galat acak yang tak teramati. Atau dalam bentuk matriks model (2.2.3) dapat ditulis sebagai berikut:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon \qquad (2.2.4)$$

untuk k buah peubah bebas $X_1, X_2, \ldots, X_i$, dimana:

$$\underset{nx1}{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad \underset{nxk}{\boldsymbol{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix} \quad \underset{kx1}{\beta} = \begin{bmatrix} \beta_0 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \quad \text{dan} \quad \underset{kx1}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_k \end{bmatrix}$$

Agar parameter β mudah dibuat taksirannya maka diasumsikan ε_i memenuhi kondisi Gauss-Markov (G – M), yakni:

$$\begin{aligned} &(1) \quad E_F(\varepsilon_i) = 0 \\ &(2) \quad E_F(\varepsilon_i^2) = \sigma^2 \\ &(3) \quad E_F(\varepsilon_i \varepsilon_j) = 0 \ , \qquad i \neq j \end{aligned} \qquad (2.2.5)$$

untuk semua i,j = 1,2,….,n. Apabila kondisi (G – M) dipenuhi maka:

$$\mu_i = E(y_i | X_i) = E(X_i\beta + \varepsilon_i | X_i) = E(X_i\beta | X_i) + E(\varepsilon_i | X_i) = X_i\beta \qquad (2.2.6)$$

Selanjutnya akan ditaksir parameter regresi β dari observasi data $((X_1, y_1)$, ..., $(X_n, y_n))$. Katakanlah $\hat{\beta} = \theta$, suatu taksiran untuk β_i dimana $\theta = (\theta_0, \theta_1, \ldots, \theta_k$.

Dari model (2.2.3) didapatkan $\varepsilon_i = y_i - X_i\beta$, maka kedua ruas dikuadratkan dan sekaligus dijumlahkan, . Ini disebut dengan Kuadrat Galat Sisa (KGS). Jadi, KGS $(\beta) = \sum_{i=1}^{n} [y_i - x_i\beta]^2$ atau dalam bentuk matriks sebagai berikut:

$$\begin{aligned} KGS(\beta) = (y - X\beta)'(y - X\beta) &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &= y'y - \beta'X'y - \beta'X'y + \beta'X'X\beta \\ &= y'y - 2\beta'X'y + \beta'X'X\beta \end{aligned}$$

karena $\beta X\dot{y}$ adalah matrik skalar maka tranpose $(\beta X\dot{y})' = \beta' X' y$. Suatu penaksir kuadrat terkecil (PKT), misalkan $\hat{\beta}$, didapatkan dengan cara meminimumkan KGS(β). Untuk meminimumkannya maka KGS(β) diturunkan terhadap masing-masing β_i dan turunan ini sama dengan nol, diperoleh

$$\partial JKG(\beta)/\partial\beta = -2\, X'y + 2X'X\beta = 0.$$

Dengan meletakkan θ pada β dan PKT $(\hat{\beta}) = \theta$ diberikan oleh persamaan $X'X\theta = X'y$ dan dari penyelesaian persamaan normal didapat

$$\theta = (X'X)^{-1} X'y \qquad (2.2.7)$$

Dengan menggantikan β dengan θ dan ε dengan **e** maka model prediksi (2.2.4) menjadi : $y = X\theta + \mathbf{e}$, dimana **e** adalah suatu taksiran untuk ε.

Sekarang diperkirakan model regresi tersebut adalah : $\hat{\boldsymbol{y}} = \boldsymbol{X}\theta$ dan didefenisikan vektor sisa $\mathbf{e} = \boldsymbol{y}_i - \hat{\boldsymbol{y}}$, yakni suatu penyimpangan prediksi dari nilai sebenarnya, dimana

$$\mathbf{e} = (e_1, e_2, \ldots, e_n), \text{ dan } \hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \boldsymbol{X}\theta = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

2.3 Rataan dan Variansi untuk Penaksir Parameter θ

Dibawah kondisi G-M maka diperoleh:

$$E(\theta) = E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E[\boldsymbol{y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta = \beta \qquad (2.3.1)$$

Hal ini menunjukkan bahwa θ suatu penaksir yang tak bias untuk β. Kemudian

$$\text{Cov}(\boldsymbol{y}) = E[(\boldsymbol{y} - \boldsymbol{X}\beta)(\boldsymbol{y} - \boldsymbol{X}\beta)'] = E(\varepsilon\varepsilon') = \sigma^2\mathbf{I} \qquad (2.3.2)$$

Misalkan ,A = $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ maka (2.2.6) menjadi

$$\theta = A\boldsymbol{y} \qquad (2.3.3)$$

Selanjutnya, menurut kondisi G-M,

$$\text{Cov}(\theta) = \text{Cov}(A\boldsymbol{y}) = A\,\text{Cov}(\boldsymbol{y})A' = A\,\sigma^2\mathbf{I}\,A' = A\,A'\sigma^2$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \qquad (2.3.4)$$

Supaya Cov (θ) dapat ditaksir , maka akan ditentukan lebih dahulu taksiran untuk σ^2 (yang mana biasanya tidak diketahui) yaitu,

$$S^2 = \sum_{i=1}^{n} e_i^2 (n-k-1)^{-1} = JKS/_{n-k-1} \qquad (2.3.5)$$

Sekarang dapat ditentukan taksiran untuk Cov (θ), yaitu:

$$\text{Cov}(\theta) = S^2 (\boldsymbol{X'X})^{-1} = S^2 G^{ij}.$$

Misalkan G^{ij} adalah elemen diagonal ke-j dari matriks $(\boldsymbol{X'X})^{-1}$ yang merupakan var(θ_j), variansi penaksir θ_j . Dan standar error (s.e.) dari θ_j adalah s.e.(θ_j) = $S\sqrt{G^{jj}}$.

2.4 Selang Kepercayaan untuk β pada Regresi Linier

Dalam hal ini diasumsikan kondisi G-M dipenuhi dan juga diasumsikan bahwa $\varepsilon \sim N(0,\sigma^2 I_n)$, yakni normal n-variate. Karena $\boldsymbol{X}$ bukan acak (matriks konstan), maka

$$\boldsymbol{y} \sim N(\boldsymbol{X}\,\beta, \sigma^2 \mathbf{I}). \qquad (2.4.1)$$

Dari (2.3.3) diperoleh $\theta = A\,\boldsymbol{y}$, dimana A suatu matriks konstan maka didapatkan:

$$\theta \sim N(A\boldsymbol{X}\beta\,,\, A\sigma^2 A') \text{ atau}$$

$$\theta \sim N(\beta, \sigma^2 (\boldsymbol{X'X})^{-1}). \qquad (2.4.2)$$

Ini berarti bahwa

$$\theta \xrightarrow{d} \beta\,, \qquad (2.4.3)$$

untuk parameter ke-j

$$\theta_j \sim N(\beta_j, \sigma^2 G^{jj}). \qquad (2.4.4)$$

Jika dibakukan diperoleh:

$$\frac{\theta_j - \beta_j}{\sigma\sqrt{G^{jj}}} \sim N(0,1) . \qquad (2.4.5)$$

Telah diketahui $S\sqrt{G^{jj}}$ = s.e.(θ_j) maka, menurut teori peluang, dari persamaan (2.3.5) didapatkan:

$$(n-k-1)\, S^2/\sigma^2 \quad \sim \chi^2_{(n-k-1)} . \qquad (2.4.6)$$

Berdasarkan (2.4.5) dan (2.4.6) diperoleh:

$$\frac{\left(\frac{\theta_j - \beta_j}{\sigma\sqrt{G^{jj}}}\right)}{\sqrt{\frac{(n-k-1)S^2/\sigma^2}{(n-k-1)}}} = (\theta_j - \beta_j)/\left(S\sqrt{G^{jj}}\right)$$

$$= [(\theta_j - \beta_j) / \text{s.e.}(\theta_j)] \sim t_{(n-k-1)} \qquad (2.4.7)$$

Selanjutnya berdasarkan (2.4.7) dapat dibuat selang kepercayaan (1-2α) 100% untuk parameter β_j sebagai berikut:

$P(L < \beta_j < U) = (1-2\alpha)$, asalkan $P(L \geq \beta_j) = \alpha$, dan $P(U \leq \beta_j)\, \alpha$.

Disini **L** adalah batas bawah selang dan **U** batas atas selang.

Jadi, selang kepercayaan (1-2α) 100% untuk β_j adalah:

$$\theta_j + \text{s.e.}(\theta_j)\, t_{(n-k-1),\alpha} \;,\; \theta_j - \text{s.e.}(\theta_j)\, t_{(n-k-1),\alpha} \qquad (2.4.8)$$

2.5 Model Distribusi Empiris dan Prinsip Penggantian (*Plug-in*)

Pandang $X_1, X_2, \ldots, X_n$ sampel random berukuran n dari populasi yang berdistribusi F yang tidak diketahui dan misalkan $X_i \overset{iid}{\sim} F$. Parameter sebagai fungsi dari F dinotasikan

$$\theta = t(F),$$

yang merupakan nilai sebenaranya dari parameter yang akan ditaksir. Parameter θ dapat dipandang sebagai nilai fungsi sebenarnya dari F. Distribusi sebenarnya F tidak diketahui, akan tetapi berdasarkan sampel $X_1, X_2, \ldots, X_n$ maka F dapat ditaksir dengan memakai *fungsi distribusi empiris* $\hat{F}_n$.

Fungsi distribusi empiris $\hat{F}_n$ dari $x_1, x_2, \ldots, x_n$ didefenisikan sebagai berikut:

$$\hat{F}_n(x) = \frac{1}{n}\{\#\{x_i \le x\}, 1 \le i \le n\}, \text{untuk} -\infty < x < \infty. \qquad (2.5.1)$$

Dengan kata lain $\hat{F}_n$ adalah suatu fungsi distribusi empiris dengan memberikan bobot peluang 1/n pada setiap observasi x_i , i = 1,2,...,n. $\hat{F}_n$ adalah penaksir yang baik dari F *karena* tidak bias, artinya $E(\hat{F}_n(x)) = F$. Dan menurut teorema 2.1.1 berarti

$$\hat{F}_n \to F \qquad (2.5.2)$$

Prinsip penggantian ***(Plug-in Principle)*** adalah metode sederhana penaksiran parameter-parameter dari sampel. Sebuah penaksir pengganti dari

parameter θ $t(F)$ akan ditaksir kembali dengan prisip penggantian yang didefenisikan sebagai:

$$\hat{\theta} = t\left(\hat{F}\right).$$

Dengan kata lain, penaksir fungsi $\theta = t(F)$ dari distribusi peluang F sama saja dengan menaksir parameter dari fungsi distribusi empiris $\hat{F}_n$, atau $\hat{\theta} = t\left(\hat{F}\right)$. Untuk selanjutnya istilah penaksir pengganti dapat disebut sebagai penaksir saja, dengan lambang " ^ ".

2.6 Metode bootstrap

Metode bootstrap diperkenalkan dalam statistika oleh *Bradley Efron* pada tahun 1977 sebagai suatu metode untuk menaksir variabilitas dari suatu statistik, seperti galat baku atau dengan membentuk selang kepercayaan. Dapat juga digunakan untuk menaksir distribusi suatu statistik (disebut distribusi bootstrap dari statistik tersebut). Distribusi ini diperoleh dengan menggantikan distribusi populasi yang tidak diketahui dengan distribusi empiris berdasarkan data sampel, kemudian melakukan resampel kembali dari distribusi empiris untuk dipergunakan dalam mencari penaksir bootstrap.

Misalkan $x_1, x_2, \ldots, x_n$ sampel random berukuran n dari suatu populasi dengan fungsi distribusi F tidak diketahui. Misalkan juga $\theta = t(F)$ parameter populasi yang menjadi perhatian dan yang akan ditaksir. Penaksir dari θ, $\hat{\theta}$, adalah suatu fungsi dari $x_1, x_2, \ldots, x_n$, misalkan kuantitas statistiknya

$$K_n = u(X_1, X_2, \ldots, X_n; F) \tag{2.6.1}$$

yang merupakan fungsi dari $(X_1, X_2, \ldots, X_n)$ dan F . Yang ingin dicari adalah distribusi dari K_n, dan misalkan fungsi distribusinya adalah

$$G_n(x) = P(K_n \leq x) \tag{2.6.2}$$

untuk $-\infty \leq x \leq \infty$. Jelas G_n tidak diketahui karena F tidak diketahui.

Berdasarkan $(X_1, X_2, \ldots, X_n)$, yang mempunyai distribusi empiris $\hat{F}_n$ yang memberi peluang 1/n pada setiap observasi dari X_i , i = 1,2,…,n:

$$\hat{F}_n(x) = \frac{1}{n} \{ \text{banyaknya } X_i \leq x,\ 1 \leq i \leq n \tag{2.6.3}$$

untuk $-\infty \leq x \leq \infty$. $\hat{F}_n$ adalah penaksir yang baik dari F karena tidak bias. ***Sampel bootstrap*** didefenisikan sebagai sampel random berukuran *n* yang diambil dari $\hat{F}_n$ dengan pengembalian, ditulis $X_1^*, X_2^*, \ldots, X_n^*$. Jadi ada n^n kombinasi yang mungkin untuk sampel bootstrap, bisa saja didapatkan $X_i^* = X_j^*$, untuk $i \neq j$. Pada sampel bootstrap, X_i bisa muncul nol kali, bisa satu kali dan maksimal *n* kali. Dengan kata lain ada *n* cara untuk memperoleh masing-masing $X_1^*, X_2^*, \ldots, X_n^*$. Jika X_i , i =1,2,…,n, ***n*** buah nilai berbeda, maka ada $\binom{2n-1}{n}$ sampel bootstrap yang berbeda.

Setiap sampel bootstrap berkorespondensi dengan satu *replikasi bootstrap* untuk K_n yang didefenisikan sebagai

$$K_n^* = u\left(X_1^*, \ldots, X_n^*; \hat{F}_n\right). \tag{2.6.4}$$

Sedangkan *penaksir bootstrap* untuk fungsi distribusi dari K_n didefenisikan sebagai

$$G_n^* = P^*\left(K_n^* \le x\right). \tag{2.6.5}$$

dimana P^* adalah peluang yang berkorespondensi dengan $\hat{F}_n$. Penaksir bootstrap G_n^* tidak lain adalah fungsi distribusi *bersyarat* dari K_n dengan syarat $\hat{F}_n$ diberikan, yaitu dengan observasi dari $(x_1, x_2, \ldots, x_n)$ yang diberikan. Untuk menegaskan kebersyaratan tersebut (2.6.5) dapat dituliskan sebagai

$$G_n^* = P^*\left(K_n^* \le x | x_1, \ldots, x_n\right). \tag{2.6.6}$$

Namun untuk memudahkan penulisan, dipakai (2.6.5).

Dalam hal ini penaksir bootstrap $(x_1^*, x_2^*, \ldots, x_n^*)$ pengganti dari $x_1, x_2, \ldots, x_n$ dan $\hat{F}_n$ penaksir penggantidari F. Dengan kata lain dalam metode bootstrap $\hat{F}_n$ dipandang sebagai populasi.

Tanda "*" menunjukkan bahwa x^* bukan data sesungguhnya x melainkan bentuk random dari x, atau hasil *resampling* dari data x secara random dengan pengembalian.

Pembahasan berikut adalah penjabaran tentang langkah-langkah dasar untuk metode bootstrap menurut *Efron*.

1. Menentukan distribusi empiris dari F, yaitu $\hat{F}_n(x)$ dengan memberi peluang 1/n untuk masing-masing titik $x_1, x_2, \ldots, x_n$.
2. Menentukan sampel bootstrap $(x_1^*, x_2^*, \ldots, x_n^*)$.

Menurut $\hat{F}_n(x)$ yang telah ditentukan, diambil sampel bootstrap ukuran n secara random dari $x_1, x_2, ..., x_n$ dengan pengembalian, sebut nilai sampelnya $(x_1^*, x_2^*, ..., x_n^*)$.

3. Menentukan statistik bootstrap $\hat{\theta}^*$.

   Dari $x_1^*, x_2^*, ..., x_n^*$ yang diperoleh pada langkah ke-2, selanjutnya hitung statistik bootstrap untuk menghasilkan $\hat{\theta}^*$.

4. Menentukan $K_n^* = u\left(x_1^*, ..., x_n^*; \hat{F}_n\right)$.

   Yaitu dengan menghitung $K_n^* = \sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right)$.

5. Menentukan distribusi K_n^*.

   Kita pelajari sifat-sifat dari K_n^* , misalnya untuk menentukan rata-rata dari K_n^*.

6. Ulangi langkah 2,3,4 dan 5 sebanyak B kali, untuk B yang cukup besar.

7. Berikan sebaran peluang darB $\hat{\theta}^*$ dengan menempatkan peluang bagi masing-masing $\hat{\theta}^*_1$, $\hat{\theta}^*_2$,..., $\hat{\theta}^*_B$. Sebaran ini adalah estimasi bootstrap untuk sebaran sampling $\hat{\theta}$; $F^*(\hat{\theta}^*)$.

Langkah-langkah inilah yang disebut sebagai dasar untuk metode bootstrap. Secara teoritis, kemungkinan $x_1^*, x_2^*, ..., x_n^*$ yang muncul adalah terhitung, yaitu sebanyak n^n kemungkinan. Hal ini dapat dilihat pada contoh dibawah ini.

Misalkan $X = (X_1, X_2, X_3)$ sampel random berukuran $n = 3$ dari suatu F dan $x = (x_1, x_2, x_3) = (3,6,9)$ adalah hasil pengamatan. Selanjutnya akan ditaksir distribusi sampling dari $K_n = u(x_1, x_2, ..., x_n; F) = \sqrt{n}\left(\hat{\theta} - \theta\right)$, maka yang harus dilakukan adalah :

1. $\hat{F}_n(x)$ memberi peluang 1/3 untuk setiap (3,6,9).

2. Menurut ketentuan dari $\hat{F}_n(x)$ diambil sampel bootstrap berukuran $n = 3$, maka x^* yang mungkin adalah:

   {(3,3,3),(3,3,6),(3,3,9),(3,6,3),(3,9,3)(6,3,3)(9,3,3)(3,6,6),(3,9,9),(6,6,6),(6,6,3),(6,6,9),(6,3,6),(6,9,6),(3,6,6),(9,6,6),(6,3,3),(6,9,9),(9,9,9),(9,9,3),(9,9,6),(9,3,9),(9,6,9),(3,9,9),(6,9,9),(9,3,3),(9,6,6) }

3. Tentukan $\theta\left(\hat{F}\right)$ dari $\bar{x} = \sum_{i=1}^{3} x_i / 3$, yaitu: $\hat{\theta} = (3+6+9) / 3 = 6$

4. Dari x^* , ditentukan $\hat{\theta}^*$.

   Untuk setiap pengambilan sampel bootstrap akan dihitung $\hat{\theta}_n^* = \sum_{i=1}^{3} x_i / n$,

   untuk n = 1,2,…,27, yaitu:

   $\hat{\theta}_n^* = (3+3+3) / 3$

   $\hat{\theta}_n^* = (3+3+6) / 3$

   $\hat{\theta}_n^* = (3+3+9) / 3$

   .

   .

   .

   $\hat{\theta}_{27}^* = (9+6+6) / 3$

5. Menentukan K_n^*, yaitu $K_n^* = u\left(x_1^*,...,x_n^*;\hat{F}_n\right) = \sqrt{n}\left(\hat{\theta}_n^* - \hat{\theta}\right)$, dimana $\hat{\theta}_n^*$ adalah rataan dari salah satu (3^3) kemungkinan x^* yang mungkin.

$K_1^* = \sqrt{3}$ x (3-6) = -5.1961524

$K_2^* = \sqrt{3}$ x (4-6) = -3.4641016

$K_3^* = \sqrt{3}$ x (5-6) = -1.7320508

.

.

.

$K_{27}^* = \sqrt{3}$ x (7-6) = 1.7320508

6. Hitung distribusi bootstrap dari $K_n^* = u\left(x_1^*,...,x_n^*;\hat{F}_n\right)$.

Untuk lebih jelasnya akan diterapkan hasil perhitungan dalam tabel dibawah ini:

Tabel 2.6.1 Sebaran frekuensi rataan resampling untuk (X_1, X_2, X_3) = (3,6,9)

i	$\hat{\theta}_{3i}^*$	$\sqrt{n}\left(\hat{\theta}_{3i}^* - \hat{\theta}\right)$ (K_{ni})	Frekwensi (f_i)	Frekwensi Relatif (P_i)	Frekwensi Relatif Kumulatip
1	3	-5.1961524	1	1/27	1/27
2	4	-3.4641016	4	4/27	5/27
3	5	-1.7320508	8	8/27	12/27
4	6	0	1	1/27	14/27
5	7	1.7320508	8	8/27	22/27
6	8	3.4641016	4	4/27	26/27
7	9	5.1961524	1	1/27	1
		Jumlah	$27 = 3^3$	1	

Resampling dilakukan dengan pengembalian untuk semua kemungkinan sampel.

Rataan setiap sampel bootstrap $(\hat{\theta}_{3i}^*)$ menyebar seperti pada tabel 2.6.1. Penaksir

bootstrap $G_{3i}^{*}(x) = P^{*}(\sqrt{3}\left(\hat{\theta}_{3i}^{*} - \hat{\theta}\right) < x)$ untuk setiap nilai x bagi $\sqrt{3}\left(\hat{\theta} - \theta\right)$ adalah $G_3^{*}(x) = \sum_{\substack{i=1 \\ K_n \le x}}^{9} P_i$.

Semakin besar n semakin baik (dalam arti, ini dikaitkan apabila grafik histogram dari hasil K_n sudah cukup menyerupai grafik distribusi normal). Meskipun secara prinsip dapat dihitung, tapi bisa dikatakan mustahil dalam prakteknya karena membutuhkan waktu yang lama sekali, apalagi jika bentuk K_n nya rumit. Ini menjadi tidak efektif lagi, sehingga digunakan pendekatan ***simulasi Monte-Carlo***, dimana semua kemungkinan sampel bootstrap dibuat menjadi sejumlah B yang cukup besar (tetapi cukup kecil jika dibandingkan dengan jumlah sampel bootstrap ideal, akan tetapi cukup menunjukkan grafik histogram daripada K_n^{*} untuk mendekati grafik distribusi normal). B dinamakan *umuran bootstrap*. Dengan bantuan paket komputer dalam bahasa ***pascal*** hal tersebut dapat dilakukan dengan mudah.

Simulasi Monte-carlo dapat digunakan, karena dengan teorema limit pusat akan menjamin kekonsistenan dari G_n^{*B} ke G_n^{*}.

Distribusi normal merupakan jenis distribusi dengan variabel random kontinyu. Jika X variabel random kontinyu mempunyai fungsi dentitas $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]}$ untuk $-\infty < x < \infty$, maka dikatakan variabel X berdistribusi normal.

Teorema limit pusat (central limit pusat)

Jika $\{ X_n \}$ adalah barisan dari variabel random yang independen dan didistribusikan secara identik dengan kesamaan nilai harapan dan varian yang terbatas, maka variabel random $Y_n = \frac{\sum X_i - n.\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ mempunyai limit untuk $n \to \infty$ sama dengan distribusi N(0,1).

Bukti:

Dalam hal ini dianggap, bahwa barisan $\{ X_n \}$ mempunyai fungsi pembangkit momen $m_X(t) = E\left[e^{tX}\right]$ $-h < t < h$. $\{ X_n \}$ barisan dari sampel random $X_1, X_2, \ldots, X_n$ (adalah independen).

Karena X_i merupakan barisan yang independen maka untuk fungsi Y_n juga merupakan barisan yang independen ($i = 1,2,\ldots,n$).

Sehingga berdasarkan defenisi fungsi pembangkit momen dari suatu variabel random maka setiap fungsi dari variabel random mempunyai fungsi pembangkit yang sama, berlaku

$$m_{y_n}(t) = E\left[e^{ty_n}\right]$$

$$= E\left[e^{t\left(\frac{\sum X_i - n\mu}{\sigma\sqrt{n}}\right)}\right] \quad ; \text{karena } X_1, X_2, \ldots, X_n \text{ independen}$$

$$= E\left[e^{t\left(\frac{X_1-\mu}{\sigma\sqrt{n}}\right)}\right].E\left[e^{t\left(\frac{X_2-\mu}{\sigma\sqrt{n}}\right)}\right]. \ldots . E\left[e^{t\left(\frac{X_n-\mu}{\sigma\sqrt{n}}\right)}\right]$$

$$= E\left[e^{t\left(\frac{\chi-\mu}{\sigma\sqrt{n}}\right)}\right]^n$$

$$= E\left[m_z\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \text{ dimana } z = X-\mu \text{ dan } -h < \frac{t}{\sigma\sqrt{n}} < h.$$

Dengan menggunakan deret taylor :

$$m_z(s) = m_z(0) + m_z'(0)s + \frac{m_z^2(s_0)s^2}{2} + \ldots \; ; 0<s_0<s \text{ dimana } s = \frac{t}{\sigma\sqrt{n}}$$

dari sifat fungsi pembangkit momen dimana variabel random $z = X-\mu$, maka $m_z(0) = 1$, $m_z'(0) = E[X-\mu] = 0$ dan $m_z^2(0) = \sigma^2$, maka deret taylor dapat dituliskan:

$$m_z(s) = 1 + \frac{m_z^2(s_0)}{2}.s^2 + \ldots$$

$$m_z\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + \frac{m_z^2(s_0)t^2}{2n\sigma^2} + \ldots$$

$$= 1 + \frac{t^2}{2n} + \frac{\left(\left(m_z^2(s_0)-\sigma^2\right)t^2\right)}{2n\sigma^2} + \ldots$$

$$m_{Y_n}(t) = \left[m_z\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

$$= \left[1 + \frac{t^2}{2n} + \frac{\left(\left(m_z^2(s_0) - \sigma^2 \right) t^2 \right)}{2n\sigma^2} + \ldots \right]^n$$

maka $\lim_{x \to \infty} it \; m_z^2\left(\frac{s_0}{\sigma\sqrt{n}} \right) = \sigma^2$, sehingga harga $m_{Y_n}(t) \to e^{\frac{t^2}{2}}$ jika $n \to \infty$.

Kesimpulannya $m(t) = e^{\frac{t^2}{2}}$ atau $F_{Y_n}(x) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$, dimana $n \to \infty$, adalah fungsi pembangkit momen dari N(0,1) terbukti.

2.7 Selang Kepercayaan Bootstrap Parameter θ

Pada bagian ini akan dibuat selang kepercayaan untuk parameter $\theta = t(F)$ .Misalkan $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n; F)$ adalah suatu statistik penaksir dari parameter θ, berdasarkan data $(x_1, x_2, \ldots, x_n)$. Didefenisikan suatu statistik:

$$K_n = \sqrt{n}\left(\hat{\theta} - \theta \right) \qquad (2.7.1)$$

Pendefinisian statistik ini adalah idenya dari teorema limit pusat. Mengatakan bahwa jumlah dari peubah acak dikurangi ekspektasinya mempunyai distribusi normal simetris disekitar nol, untuk *n* yang cukup besar . Disini diasumsikan statistik tersebut mempunyai distribusi asimptotis normal disekitar nol . Dan diasumsikan pula $E(\hat{\theta}) = \theta$,. Jadi merupakan suatu penaksir yang tak bias. Dalam hal ini $\hat{\theta}$ adalah suatu *penaksir kuadrat terkecil (PKT)* untuk parameter regresi β,

dimana $\hat{\theta}$ tersebut berdistribusi normal dengan rataan $E(\hat{\theta})=\theta$. Jadi jumlah hasil dari $n^{1/2}(\hat{\theta}-\theta)$ akan menghasilkan suatu distribusi normal simetris disekitar nol.

Dengan menggunakan prinsip penggantian dalam metode bootstrap diperoleh

$$K_n^* = \sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right) \qquad (2.7.2)$$

yakni dengan mengganti K_n dengan K_n^*, $\hat{\theta}^*$ dengan $\hat{\theta}$ dan $\hat{\theta}$ merupakan penaksir dari θ. Selanjutnya menentukan fungsi distribusi dari K_n yaitu:

$$G_n(x) = P(K_n(x_1,\ldots,x_n;F) \le x) = P_F\left(\sqrt{n}\left(\hat{\theta}-\theta\right) \le x\right) \qquad (2.7.3)$$

dan fungsi distribusi dari K_n^* adalah

$$G_n^*(x) = P^*\left(K_n^*\left(x_1^*,\ldots,x_n^*;\hat{F}_n\right) \le x\right) = P^*_{\hat{F}_n}\left(\sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right) \le x\right). \qquad (2.7.4)$$

Oleh karena $x_1^*,\ldots,x_n^*$ adalah sampel bootstrap yang diambil dari $x_1, x_2,\ldots,x_n$ bedasarkan $\hat{F}_n$ (distribusi empiris dari $x_1, x_2,\ldots,x_n$), maka dengan sendirinya G_n^* dapat diketahui karena didefenisikan berdasarkan $\hat{F}_n$. Jadi G_n^* merupakan *penaksir bootstrap* dari G_n.

G_n^* dapat dipandang sebagai penaksir pengganti (*plug-in estimate)* dari G_n dengan meletakkan $\hat{F}_n$ ke-i dan x_i^* ke x_i. Perhitungan G_n^* secara eksak berdasarkan semua kemungkinan sampel bootstrap inilah yang disebut "ideal bootstrap". Dengan menggunakan simulasi Monte-Carlo didapat penaksiran G_n^* yang cukup baik, artinya cukup dekat ke G_n.

Telah ditulis bahwa $\hat{F}_n \rightarrow F$. Selanjutnya karena G_n^* didefenisikan berdasarkan $\hat{F}_n$ dan G_n didefenisikan berdasarkan F. Maka bisa dikatakan bahwa $G_n \approx G_n^*$, untuk pengulangan n yang cukup besar, dalam arti jika $a_n = b_n$ maka a_n dan b_n dapat diganti . Berdasarkan ini, akan dibangun selang kepercayaan untuk parameter θ . Ide didasarnya adalah menempatkan kuantil (yang tak diketahui) dari G_n dengan menggunakan kuantil (diketahui) dari G_n^* . Kuantil (α) dan kuantil $(1-\alpha)$ dari G_n didefenisikan sebagai berikut:

$$c_{n\alpha} = G_n^{-1}(\alpha), \quad c_{n1-\alpha} = G_n^{-1}(1-\alpha) \tag{2.7.5}$$

dengan cara yang sama diperoleh kuantil (α) dan kuantil $(1-\alpha)$ dari G_n^*,

$$c_{n\alpha}^* = G_n^{*-1}(\alpha), \quad c_{n1-\alpha}^* = G_n^{*-1}(1-\alpha). \tag{2.7.6}$$

Karena $G_n^* \approx G_n$, untuk n yang cukup besar maka didapatkan $c_{n\alpha} \approx c_{n\alpha}^*$ dan $c_{n1-\alpha} \approx c_{n1-\alpha}^*$. Selang kepercayaan $(1\text{-}2\alpha)100\%$ untuk θ dibentuk sebagai berikut:

$$P\left(c_{n\alpha}^* \le \sqrt{n}\left(\hat{\theta} - \theta\right) \le c_{n1\ \alpha}^*\right).$$

Dalam hal ini, $P\left(c_{n\alpha}^* \le \sqrt{n}\left(\hat{\theta} - \theta\right) \le c_{n1-\alpha}^*\right) \approx P\left(c_{n\alpha} \le \sqrt{n}\left(\hat{\theta} - \theta\right) \le c_{n1-\alpha}\right)$.

Selang kepercayaan $(1\text{-}2\alpha)100\%$ diperoleh sebagai berikut:

$$P\left(c_{n\alpha} \le \sqrt{n}\left(\hat{\theta} - \theta\right) \le c_{n1-\alpha}\right) = P\left(\sqrt{n}\left(\hat{\theta} - \theta\right) \le c_{n1-\alpha}\right) - P\left(\sqrt{n}\left(\hat{\theta} - \theta\right) \le c_{n\alpha}\right)$$

$$= G_n(c_{n1-\alpha}) - G_n(c_{n\alpha})$$

$$= 1 - (\alpha) - (\alpha)$$

$$= 1 - 2\alpha$$

Sehingga diperoleh

$$c^*_{n\alpha} \le \sqrt{n}\left(\hat{\theta} - \theta\right) \le c^*_{n1-\alpha} \qquad (2.7.7)$$

yang ekivalen dengan

$$\hat{\theta} - n^{-\frac{1}{2}} c^*_{n1-\alpha} \le \theta \le \hat{\theta} - n^{-\frac{1}{2}} c^*_{n\alpha}. \qquad (2.7.8)$$

Sehingga dapat diperoleh selang

$$\hat{\theta} - n^{-\frac{1}{2}} c^*_{n1-\alpha} \;,\; \hat{\theta}_n - n^{-\frac{1}{2}} c^*_{n\alpha} \qquad (2.7.9)$$

sebagai selang kepercayaan bootstrap untuk parameter θ, dengan taraf kepercayaan mendekati $(1-2\alpha)$.

Selanjutnya pendefinisian kuantil (α) dan kuantil $(1-\alpha)$ dari distribusi (bersyarat) $\hat{\theta}^*$ dinyatakan $\hat{\theta}^{*(\alpha)}$ dan $\hat{\theta}^{*(1-\alpha)}$ Jadi, cukup dapat dimengerti jika ditulis $\sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right) = c^*_{np}$, untuk suatu nilai $p \in (0,1)$. Nilai p akan diambil sebesar (α) dan $(1-\alpha)$. Dimana bahwa $\hat{\theta}^*$ bertindak sebagai 'konstanta' dalam lingkungan bootstrap. Dengan demikian diperoleh selang kepercayaan bootstrap sebagai berikut:

$$\hat{\theta} - \sqrt{n}\left(\hat{\theta}^{*(1-\alpha)} - \hat{\theta}\right), \hat{\theta} - \sqrt{n}\left(\hat{\theta}^{*(\alpha)} - \hat{\theta}\right)$$

atau

$$\left(2\hat{\theta} - \hat{\theta}^{*(1-\alpha)}, 2\hat{\theta} - \hat{\theta}^{*(\alpha)}\right) \qquad (2.7.10)$$

dengan $\hat{\theta}_p^*$ adalah kuantil ke-100p dari distribusi empiris (bersyarat, berdasarkan $\hat{F}_n$) dari $\hat{\theta}^*$. Dari (2.7.10) didapatkan

$$\left(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)}\right) \tag{2.7.11}$$

Sekali lagi bahwa $\hat{\theta}^{*(\alpha)}$ dan $\hat{\theta}^{*(1-\alpha)}$ menyatakan kuantil (α) dan kuantil ($1 - \alpha$) dari distribusi (bersyarat) $\hat{\theta}^*$ .Tentu saja $\hat{\theta}^{*(\alpha)}$ dan $\hat{\theta}^{*(1-\alpha)}$ dihitung melalui aproksimasi simulasi Monte Carlo. Sehingga diurutkan $\hat{\theta}_j^*$; j = 1....,**B**. Harga **B** merupakan replikasi semu bootstrap dari $\hat{\theta}^*$. Setelah itu diambil nilai statistik terurut yang ke (α).**B** dan ke ($1-\alpha$)**B** sebagai aproksimasi Monte Carlo dari $\hat{\theta}^{*(\alpha)}$ dan $\hat{\theta}^{*(1-\alpha)}$. Selang kepercayaan Bootstrap seperti pada hasil 2.4.14 oleh Efron disebut *metoda persentil bootstrap.*