PENDUGAAN GALAT BAKU DENGAN METODE BOOTSTRAP MENGGUNAKAN *R LANGUAGE*

Stanislaus S. Uyanto
Staf Pengajar Statistika – Fakultas Ekonomi
Universitas Katolik Atmajaya – Jakarta

e-mail: ss.uyanto@lpa.atmajaya.ac.id

ABSTRACT

A summary statistics such as $\hat{\theta} = t(\hat{F})$ are often the first outputs of a data analysis. The next thing we want to know is the accuracy of $\hat{\theta}$. The most common measure of an estimator's accuracy is the standard error. The bootstrap method is a computer based method for assigning measures of accuracy to statistical estimates. In this writing we will show how to obtain the standard error of a statistic using the bootstrap method. Some practical examples using the R language are also given.

Kata kunci: statistic, bootstrap, estimate, standard error, galat baku

PENDAHULUAN

Di dalam statistika kita menganalisis data berdasarkan sampel acak $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ dari suatu distribusi probabilitas F yang tidak diketahui, misalkan kita ingin menduga suatu parameter $\theta = t(F)$. Untuk menduga parameter θ ini kita menghitung statistik $\hat{\theta} = s(\mathbf{x})$ berdasarkan sampel acak $\mathbf{x}$ tersebut. Tentu kita ingin mengetahui seberapa akurat dugaan $\hat{\theta}$ ini? Ukuran keakuratan suatu dugaan statistik yang paling sering digunakan adalah *galat baku* (*standard error*). Metode bootstrap yang diperkenalkan Efron(1979) (lihat pula Davison and Hinkley, 1997; Efron 1981, 1987, 1993) sebagai metode yang berdasarkan komputer akan digunakan untuk memperoleh galat baku (*standard error*) dari suatu nilai dugaan $\hat{\theta}$. Keunggulan metode bootstrap dalam memperoleh galat baku adalah tidak memerlukan kalkulasi teoritis dan selalu dapat digunakan baik untuk pendugaan yang secara matematis sangat rumit maupun yang sederhana (Efron, 1993). Metode bootstrap juga dapat digunakan untuk distribusi normal maupun distribusi non-normal (Karian and Dudewicz, 2000; Shao and Tu, 1995). Dengan metode bootstrap kita dapat menggunakan data yang ada secara efisien tanpa memerlukan sampel data yang besar, mahal, dan memakan waktu lama untuk mengumpulkannya.

METODE BOOTSTRAP

Permasalahan dalam statistika inferensi kerapkali melibatkan pendugaan beberapa aspek dari suatu distribusi probabilitas *F* berdasarkan sampel acak yang diambil dari distribusi *F* tersebut. Fungsi distribusi empiris, yang diberi notasi $\hat{F}$, merupakan dugaan sederhana dari keseluruhan distribusi *F*. Cara paling mudah untuk menduga beberapa aspek yang menarik dari *F*, seperti purata (*mean*) atau median atau korelasi, adalah menggunakan aspek padanan dari $\hat{F}$.

Metode boostrap tergantung pada *sampel bootstrap*. Misalkan $\hat{F}$ merupakan suatu distribusi empiris diskrit, yang didefinisikan dengan probabilitas $\frac{1}{n}$ untuk setiap nilai $x_i, i = 1,2,\ldots,n$ yang diamati. Sampel bootstrap didefinisikan sebagai suatu sampel acak

sebesar n yang diambil dari $\hat{F}$, misalkan $\mathbf{x}^* = \{x_1^*, x_2^*, \ldots, x_n^*\}$. Tanda *asterisk* pada $\mathbf{x}^*$ menunjukkan bahwa $\mathbf{x}^*$ bukan merupakan data aktual dari $\mathbf{x}$, tetapi merupakan sampel acak dari $\mathbf{x}$ sebesar n *dengan pemulihan* (*with replacement*).

Berpadanan dengan data bootstrap $\mathbf{x}^*$ adalah replikasi bootstrap dari $\hat{\theta}$, yaitu:

$$\hat{\theta}^* = s(\mathbf{x}^*) \qquad (1)$$

Dugaan galat baku dari statistik $\hat{\theta}$ dengan metode bootstrap didefinisikan sebagai:

$$\mathrm{se}_{\hat{F}}\left(\hat{\theta}^*\right) \qquad (2)$$

Persamaan (2) disebut *dugaan galat baku ideal bootstrap*. Tetapi dalam praktek tidak selalu tersedia persamaan untuk menghitung $\mathrm{se}_{\hat{F}}\left(\hat{\theta}^*\right)$. Sebagai solusi alternatif untuk mengatasi hal tersebut dapat digunakan metode bootstrap. Algoritma metode bootstrap (Efron, 1979; Efron, 1993) untuk memperoleh galat baku suatu dugaan statistik adalah sebagai berikut:

1) Pilih B sampel bootstrap independen $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \ldots, \mathbf{x}^{*B}$ yang masing-masing terdiri dari n data yang diambil dengan pemulihan dari $\mathbf{x}$.
2) Hitung replikasi bootstrap yang berpadanan dengan setiap sampel bootstrap,

$$\hat{\theta}^*(b) = s\left(\mathbf{x}^{*b}\right), \quad b = 1,2,\ldots,B \qquad (3)$$

3) Hitung dugaan galat baku $\mathrm{se}_F\left(\hat{\theta}\right)$ menggunakan persamaan simpangan baku sampel dari replikasi B:

$$\hat{se}_B = \sqrt{\frac{\sum_{b=1}^{B}\left[\hat{\theta}^*(b) - \hat{\theta}^*(\bullet)\right]^2}{B-1}} \qquad (4)$$

di mana:

$$\hat{\theta}^*(\bullet) = \sum_{b=1}^{B}\frac{\hat{\theta}^*(b)}{B} \qquad (5)$$

Limit dari $\hat{se}_B$ bila $B \to \infty$ adalah *dugaan galat baku ideal bootstrap* $\mathrm{se}_F\left(\hat{\theta}\right)$, yaitu:

$$\lim_{B\to\infty} \hat{se}_B = se_{\hat{F}} = se_{\hat{F}}\left(\hat{\theta}^*\right) \qquad (6)$$

Hal ini dapat dijelaskan karena simpangan baku sampel selalu menghampiri simpangan baku populasi bila besarnya replikasi B menuju tak berhingga (Walpole et al., 2002).

BESAR REPLIKASI BOOTSTRAP *B*

Berapa besar nilai replikasi B untuk mengevaluasi $\hat{se}_B$? Dugaan galat baku ideal bootstrap memerlukan $B \to \infty$ (lihat Persamaan (6)). Karena waktu yang dibutuhkan komputer untuk menghitung Persamaan (3) bertambah secara linier sejalan dengan bertambahnya B, maka kita harus memilih nilai B yang tidak terlalu besar, terutama bila $\hat{\theta} = s(\mathbf{x})$ merupakan fungsi $\mathbf{x}$ yang sangat rumit. Pendekatan yang cukup memuaskan untuk menentukan nilai B adalah menggunakan *koefisien variasi* (*coefficient of variation*) dari $\hat{se}_B$, yang merupakan rasio simpangan baku dari $\hat{se}_B$ terhadap nilai harapannya (Walpole et. al., 2002), yaitu:

$$\mathrm{cv}\left(\hat{se}_B\right) = \frac{\sqrt{\mathrm{var}\left(\hat{se}_B\right)}}{E\left(\hat{se}_B\right)} \qquad (7)$$

Bila diketahui data $\mathbf{x}$, maka varians dari $\hat{se}_B$ adalah (Efron, 1993):

$$\text{var}\left(\hat{se}_B\right) = \text{var}\left[E\left(\hat{se}_B|\mathbf{x}\right)\right] + E\left[\text{var}\left(\hat{se}_B|\mathbf{x}\right)\right] \quad (8)$$

Misalkan $\hat{m}_i$ adalah momen ke-*i* dari distribusi bootstrap $s(\mathbf{x}^*)$ dan $\hat{\Delta} = \frac{\hat{m}_4}{\hat{m}_2^2} - 3$, yang merupakan kurtosis dari distribusi bootstrap $s(\mathbf{x}^*)$. Di sini $\hat{m}_i$ dan $\hat{\Delta}$ merupakan merupakan fungsi dari $\mathbf{x}$. Dengan menggunakan rumus untuk nilai harapan dan varians dari suatu simpangan baku sampel didapat (Efron, 1993; Venables and Smith, 2004):

$$\text{var}\left(\hat{se}_B\right) \approx \text{var}\left[\hat{m}_2^{1/2}\right] + E\left[\frac{\hat{m}_2}{4B}\left(\hat{\Delta}+2\right)\right] \quad (9)$$

Jika kita membagi Persamaan (9) dengan $\hat{m}_2$ dan menarik akar kuadrat, kita mendapatkan persamaan koefisien variasi dari $\hat{se}_B$ sebagai berikut (Efron, 1993):

$$\text{cv}\left(\hat{se}_B\right) = \sqrt{\text{cv}\left(\hat{se}_\infty\right)^2 + \frac{E\left(\hat{\Delta}\right)+2}{4B}} \quad (10)$$

karena $\hat{se}_\infty = \hat{m}_2^{1/2}$. Nilai $\hat{\Delta} = 0$ bila distribusi normal. Rentang dari nilai $\hat{\Delta}$ adalah antara -2 untuk ujung distribusi terpendek (*shortest-tailed distribution*) sampai dengan tak berhingga untuk distribusi yang berujung panjang (*long-tailed distribution*). Dalam praktek, nilai $\hat{\Delta}$ tidak lebih dari 10. Notasi $\hat{se}_\infty$ diatas melambangkan *dugaan galat baku ideal bootstrap*.

Tabel 1 membandingkan $\text{cv}\left(\hat{se}_B\right)$ dengan $\text{cv}\left(\hat{se}_\infty\right)$ untuk berbagai variasi nilai *B*, dengan asumsi $\hat{\Delta}$=0 berdasarkan Persamaan (10) (lihat Efron, 1993).

Tabel 1: Koefisien variasi $\text{cv}\left(\hat{se}_B\right)$ *dari Persamaan (10) sebagai fungsi* $\text{cv}\left(\hat{se}_\infty\right)$

		B →				
		25	50	100	200	∞
$\text{cv}\left(\hat{se}_\infty\right)$	0,25	0,29	0,27	0,26	0,25	0,25
↓	0,20	0,24	0,22	0,21	0,21	0,20
	0,15	0,21	0,18	0,17	0,16	0,15
	0,10	0,17	0,14	0,12	0,11	0,10
	0,05	0,15	0,11	0,09	0,07	0,05
	0,00	0,14	0,10	0,07	0,05	0,00

Dari Tabel 1 terlihat bahwa untuk menduga galat baku, pada nilai replikasi $B = 200$ koefisien variasi relatif sudah stabil dan tidak banyak berubah. Dengan demikian penggunaan replikasi *B* =200 sudah dianggap memadai, meskipun dalam praktek digunakan *B* > 200 (kebanyakan *software* menggunakan nilai *B* = 1000).

IMPLEMENTASI DENGAN *R LANGUAGE*

Untuk melakukan pendugaan galat baku dengan metode bootstrap akan digunakan bantuan komputer yang menjalankan *R language* (Maindonald, 2001; Venables and Smith, 2004). *R language* adalah suatu software gratis untuk komputasi statistis dan grafik yang dapat dijalankan di Linux, Windows, dan MacOS. *R language* dapat diperoleh dari situs http://www.r-project.org.

Sebagai contoh akan digunakan *distribusi seragam* (*uniform distribution*) $U(0,\theta)$.

Misalkan kita memiliki data $X_1, X_2, \ldots, X_n$ dari suatu distribusi seragam $U(0,\theta)$, di mana $\theta > 0$ dan θ merupakan parameter yang akan diduga. Bentuk fungsi kerapatan probabilitas (*probability density function*) dari distribusi seragam adalah:

$$f_x(X) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{lainnya} \end{cases} \quad (11)$$

dan fungsi *likelihood* (*likelihood function*) dari distribusi seragam adalah:

$$L(\theta) = \begin{cases} \frac{1}{\theta^n} & \text{bila } \theta \ge \max x_i \\ 0 & \text{lainnya} \end{cases} \quad (12)$$

Maka *maximum likelihood estimator* untuk θ adalah:

$$\hat{\theta} = \max_{i=1,\ldots,n} X_i = X_{(n)} \quad (13)$$

Kita akan menganalisis bentuk distribusi dari $\hat{\theta}$ sebagai berikut:

Kita tahu bahwa,

$$P[\theta - c < X_{(n)} < \theta] = 1 - P[X_{(n)} < \theta - c]$$
$$= 1 - \left(\frac{\theta - c}{\theta}\right)^n \quad (14)$$

sehingga

$$X_{(n)} \xrightarrow{P} \theta \quad (15)$$

Dan

$$P[n(\theta - X_{(n)}) \le x] = \left(1 - \frac{x}{n\theta}\right)^n \quad (16)$$

dan untuk $n \to \infty$ didapat

$$H(x) = 1 - e^{-\frac{x}{\theta}} \quad (17)$$

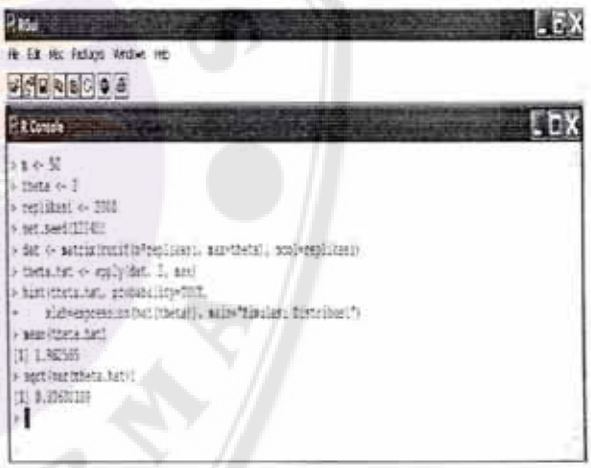
sehingga distribusi *sampling* dari $\hat{\theta} = \max_{i=1,\ldots,n} X_i = X_{(n)}$ cenderung menjadi distribusi *exponential* yang tergantung pada parameter θ yang tidak diketahui.

Kita lakukan simulasi *Monte Carlo* untuk distribusi seragam θ =2, *n* = 50, dan replikasi = 2000 dengan *R language* untuk menduga $\hat{\theta}$ dan galat baku dari $\hat{\theta}$ sebagai berikut:

```
n <- 50
theta <- 2
replikasi <- 2000
set.seed(12345)
dat <- matrix(runif(n*replikasi, max=theta),
ncol=replikasi)
theta.hat <- apply(dat, 2, max)
hist(theta.hat, probability=TRUE,
  xlab=expression(hat(theta)), main="Simulasi
Distribusi")
mean(theta.hat)
sqrt(var(theta.hat))
```

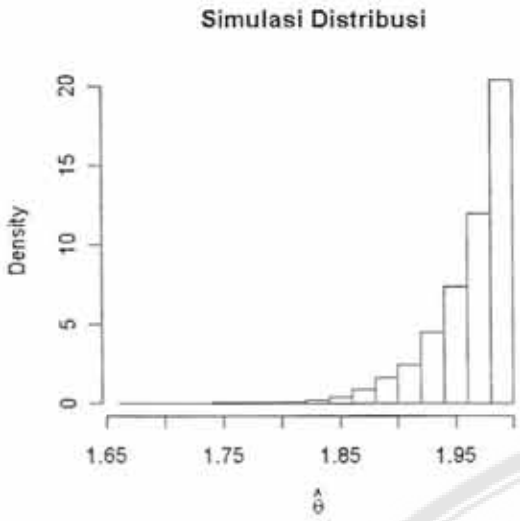
***Listing* 1:** Program simulasi Monte Carlo distribusi seragam θ =2 dalam *R Language*.

Output dari *R language* sebagai berikut:



Gambar 1: Tampilan R Language untuk simulasi distribusi seragam θ =2.

Dari tampilan *R language* pada Gambar 1 terlihat bahwa $\hat{\theta} = 1{,}962585$ dan galat baku dari $\hat{\theta}$ adalah 0,03630109.



Gambar 2: *Hasil bentuk simulasi distribusi seragam dengan* θ *=2.*

Meskipun sampel tidak terlalu besar dari Gambar 2 terlihat bahwa bentuk distribusi mendekati distribusi *exponential.*

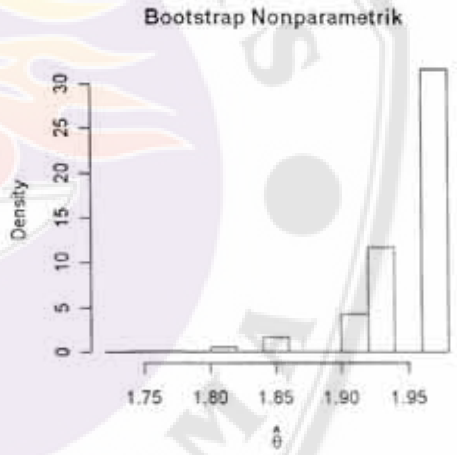
Bila kita menggunakan metode **bootstrap nonparametrik** dengan cara mengambil sampel ulang (*resampling*) dari suatu sampel data distribusi seragam dengan *n* = 50 dan replikasi= 2000 diperoleh hasil sebagai berikut:

```
n <- 50
theta <- 2
set.seed(12345)
dat <- runif(n, max=theta)
library(boot)
nboot <- 2000
stat <- function(x, i) max(x[i])
res <- boot(dat, stat, nboot)
hist(res$t, probability=TRUE,
   xlab=expression(hat(theta)),
   main=" Bootstrap Nonparametrik ")
mean(res$t)
sqrt(var(res$t))
```

Listing 2: Metode bootstrap nonparametrik dalam *R language*.

Gambar 3: Tampilan hasil metode bootstrap nonparametrik distribusi seragam θ *=2.*

Dari tampilan *R language* untuk **metode bootstrap nonparametrik** pada Gambar 3 terlihat bahwa $\hat{\theta} = 1{,}954604$ dan galat baku dari $\hat{\theta}$ adalah 0,03868347.



Gambar 4: Bentuk distribusi yang dihasilkan oleh metode bootstrap nonparametrik.

Gambar 4 menunjukkan penumpukan titik pada titik sampel maksimum $\hat{\theta} = 1{,}954604$. Hal ini disebabkan oleh:

$$1 - P\left(X_{(*)}^{*} = X_{(n)}\right) = 1 - P\left(\left(X_{(n)} \in \mathbf{X}_{*}^{*}\right) = 1 - \left(1 - \frac{1}{n}\right)^{n}\right) \quad (18)$$

$$= 1 - e^{-1}$$

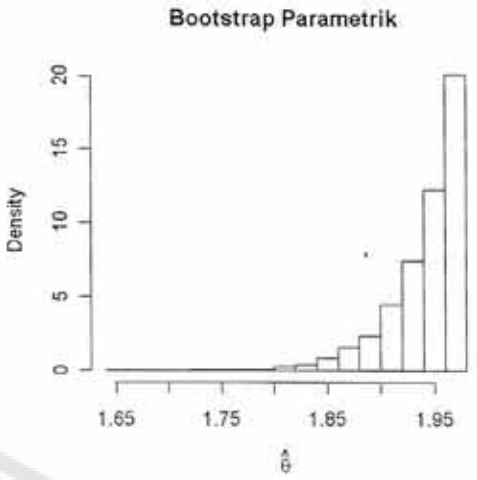
Bila kita menggunakan metode **bootstrap parametrik** dengan cara mengambil sampel ulang (*resampling*) dari distribusi seragam $U\left(0,\hat{\theta}\right)$ dengan *n* = 50 dan replikasi= 2000 diperoleh hasil sebagai berikut:

```
mle <- max(dat)
stat <- function(x) max(x)
set.seed(12345)
dat.gen <- function(dat, mle) runif(length(dat),
max=mle)
res <- boot(dat, stat, nboot, sim="parametric",
ran.gen=dat.gen, mle=mle)
hist(res$t, probability=TRUE,
   xlab=expression(hat(theta)),
    main="Bootstrap Parametrik")
mean(res$t)
sqrt(var(res$t))
```

Listing 3: Metode bootstrap parametrik dalam *R language*.

Gambar 5: Tampilan hasil metode bootstrap parametrik distribusi seragam θ *=2.*

Dari tampilan *R language* untuk **metode bootstrap parametrik** pada Gambar 5 terlihat bahwa $\hat{\theta} = 1{,}942443$ dan galat baku dari $\hat{\theta}$ adalah 0.03592853.



Gambar 6: Bentuk distribusi yang dihasilkan oleh bootstrap parametrik.

Gambar 6 memperlihatkan bentuk distribusi yang sangat mirip dengan distribusi *exponential*, hal ini disebabkan karena kita mengambil sampel ulang dari distribusi seragam $U\left(0,\hat{\theta}\right)$.

Tabel 2: Ringkasan hasil maximum likelihood estimator distribusi seragam θ *=2.*

Metode	$\hat{\theta}$	Galat baku dari $\hat{\theta}$
Simulasi Monte Carlo	1,962585	0,03630109
Bootstrap nonparametrik	1,954604	0,03868347
Bootstrap parametrik	1,942443	0.03592853

Dari Tabel 2 terlihat bahwa hasil pendugaan untuk distribusi seragam dengan θ =2, yakni $\hat{\theta}$ berkisar antara 1,942 sampai dengan 1,963; sedangkan pendugaan galat baku dari $\hat{\theta}$ berkisar antara 0,0359 sampai dengan 0,0387. Dengan kata lain dapat dikatakan bahwa tingkat keakuratan dari metode bootstrap cukup baik untuk sampel yang tidak terlalu besar tersebut.

PENUTUP

- Dengan menggunakan metode bootstrap kita tidak memerlukan asumsi normalitas dari distribusi populasi data seperti yang biasa disyaratkan dalam buku-buku teks statistika. Dengan kata lain kita bisa menggunakan metode bootstrap untuk menentukan galat baku suatu statistik dari suatu distribusi populasi non-normal.
- Implementasi *R language* (yang tersedia gratis) ke dalam metode bootstrap mempermudah evaluasi galat baku, sebagai ukuran keakuratan statistik.

DAFTAR PUSTAKA

[1] Davison, A.C. and Hinkley, D.V. *Bootstrap Methods and their Application.* Cambridge University Press, New York, 1997.

[2] Efron, B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7: 1-26, 1979.

[3] Efron, B. Nonparametric estimates of standard error: the jackknife, the bootstrap, and other methods. *Biometrika*, 68: 589-599, 1981.

[4] Efron, B. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82: 171-200, 1987.

[5] Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993.

[6] Karian, Z. A. and Dudewicz, E. J. *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods.* Chapman & Hall/CRC, Boca Raton Florida, 2000.

[7] Maindonald, J.M. *Using R for Data Analysis and Graphics: An Introduction.* Statistical Consulting Unit of the Graduate School, Australian National University, 2001.

[8] Shao, J. and Tu, D. *The Jakcknife and Bootstrap.* Springer-Verlag, New York, 1995.

[9] Venables, W. N. and Smith, D. M. *An Introduction to R.* R Development Core Team, 2004 (http://cran.r-project.org/manuals.html).

[10] Walpole, R.E., et. al. *Probability & Statistics for Engineers & Scientists.* 7th Edition. Prentice-Hall, Inc., Upper Saddle River, New Jersey, 2002.