ISSN 1907-6452

Centre

Civil & Electrical Engineering Journal

Volume 8 | Number 2 | Desember 2013

Centre
Civil & Electrical Engineering Journal
ISSN: 1907-6452
Volume 8, Number 2, December 2013

CENTRE Journal is a scientific journal publised by Universitas Internasional Batam two times year (June and December). Submission of the article on this journal is addressed to editorial office above. All articles will be subjected to editorial review and no charges for submission.

1907 6454

Centre

Civil & Electrical Engineering Journal

ISSN: 1907-6452

Volume 8, Number 2, December 2013

Interval Konfidensi Klasik dan Interval Konfidensi *Bootstrap* dari Simulasi Data N(100,23) menggunakan *Software* R

Yayuk Setyaning Astutik[1)]
Program Studi Teknik Sipil, Universitas Internasional Batam
Email : [1)] *yayuk@uib.ac.id*, *yayuksetyaninga@gmail.com*

Abstrak

Bootstrap adalah teknik resampling nonparametrik yang bertujuan untuk menentukan estimasi standar error dan interval konfidensi dari parameter populasi seperti *mean*, rasio, median, proporsi, koefisien korelasi atau koefisien regresi tanpa menggunakan asumsi distribusi. Algoritma resampling untuk interval konfidensi klasik maupun *bootstrap* menggunakan *software* R dengan membangkitkan data dari variabel random berdistribusi normal.

Kata Kunci: Interval Konfidensi, *Bootstrap*, Simulasi.

1. Pendahuluan

Estimasi interval adalah suatu interval tertentu yang memuat parameter dengan probabilitas tertentu. Misalkan $X_1, X_2, \cdots, X_n$ adalah sampel random yang diambil dari populasi dengan parameter θ, maka interval konfidensi $(1-\alpha)$ untuk θ adalah $P(a \le \theta \le b) = 1 - \alpha$.

Metode *bootstrap* merupakan metode yang digunakan untuk mengestimasi suatu distribusi populasi yang tidak diketahui dengan distribusi empiris yang diperoleh dari proses penyampelan ulang (Efron dan Tibshirani, 1993:5). Teknik penarikan sampel metode *bootstrap* adalah dengan pengembalian dari sebuah sampel asli. Sampel asli merupakan sampel yang diperoleh dari hasil observasi yang diperlakukan seolah-olah sebagai populasi.

Tujuan utama penggunaan *bootstrap* adalah untuk memperoleh estimasi parameter berdasarkan data yang minimal dengan bantuan komputer. Dalam statistika, data yang minimal dapat diartikan sebagai data yang sedikit, data yang menyimpang dari asumsi tertentu, atau bahkan data yang tidak memiliki asumsi apapun tentang distribusinya.

Sedangkan tujuan dari penelitian ini adalah simulasi data dari distribusi normal dengan nilai parameter tertentu, kemudian membandingkan hasilnya dengan menggunakan interval konfidensi biasa dengan interval konfidensi *bootstrap*.

2. Kerangka Teoritis

2.1. Konsep Dasar Interval Konfidensi

Dalam teori estimasi pada statistik klasik, parameter θ dianggap sebagai suatu konstanta yang akan diestimasi dari n sejumlah sampel pengamatan yang ada. Pendekatan yang paling biasa digunakan untuk interval konfidensi (*confidence interval*) adalah pendekatan normal terhadap Binomial. Sehingga persamaan interval konfidensi $(1-\alpha)$ dinyatakan dengan :

$$P\left(\bar{X}_n - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha \quad (1)$$

Jika diketahui $\sigma^2 = \dfrac{\hat{\theta}(1-\hat{\theta})}{n}$ dan $\bar{X} = \hat{\theta}$

maka interval konfidensi untuk *mean* $\hat{\theta}$ adalah :

$$\left[\hat{\theta} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right]$$

Dimana $\hat{\theta}$ adalah estimator parameter proporsi distribusi Binomial yang diestimasi dari statistik sampel, $Z_{\frac{\alpha}{2}}$ adalah persentil $\frac{\alpha}{2}$ dari distribusi normal standar dan adalah ukuran sampel. (Bain dan Engelhardt, 1992).

2.2. Konsep Dasar Interval Konfidensi *Bootstrap*

Untuk memperoleh interval konfidensi pada *bootstrap* terdapat beberapa metode yang dapat digunakan, salah satunya adalah *bootstrap* *percentile interval* atau interval persentil *bootstrap*. Misalkan dibangkitkan suatu data set x^*, kemudian dihitung replikasi *bootstrap* $\widehat{\theta}^* = s(x^*)$. Jika $\widehat{G}$ adalah fungsi distribusi kumulatif dari $\widehat{\theta}^*$, interval persentil ke $(1-2\alpha)$ didefinisikan dengan persentil ke-α dan ke-$(1-\alpha)$ dari $\widehat{G}$:

$$\left[\widehat{\theta}_{\%.lo}, \widehat{\theta}_{\%.up}\right] = \left[\widehat{G}^{-1}(\alpha), \widehat{G}^{-1}(1-\alpha)\right] \qquad (2)$$

Dengan definisi bahwa $\widehat{G}^{-1}(\alpha) = \widehat{\theta}^{*(\alpha)}$ adalah persentil ke-α distribusi *bootstrap*, maka interval persentile dapat ditulis :

$$\left[\widehat{\theta}_{\%.lo}, \widehat{\theta}_{\%.up}\right] = \left[\widehat{\theta}^{*(\alpha)}, \widehat{\theta}^{*(1-\alpha)}\right] \qquad (3)$$

Bentuk diatas sama dengan keadaaan *bootstrap* yang ideal, dimana jumlah replikasi *bootstrap*-nya tidak terbatas. Langkah pertama yang dilakukan untuk mengestimasi interval konfidensi *bootstrap* adalah membentuk sejumah *b* data set independen yaitu :

$$x^{*1}, x^{*2}, \cdots, x^{*B}$$

Kemudian menghitung replikasi bootstrap :.

$$\widehat{\theta}^*(b) = s(x^{*b}), b = 1, 2, \cdots, B \qquad (4)$$

Dengan persentil empiris ke-100α dari nilai $\widehat{\theta}^*(b)$, yaitu nilai ke-$B\alpha$ dalam urutan replikasi *B* dari $\widehat{\theta}^*$. Jadi jika nilai B = 2000 dan $\alpha = 0.05$, maka $\widehat{\theta}_B^{*(\alpha)}$ adalah nilai ke-100 dari urutan replikasinya. Demikian juga $\widehat{\theta}_B^{*(1-\alpha)}$ adalah persentil empiris ke-$100(1-\alpha)$. Pendekatan interval presentil $(1-2\alpha)$ adalah :

$$\left[\widehat{\theta}_{\%.lo}, \widehat{\theta}_{\%.up}\right] \approx \left[\widehat{\theta}_B^{*(\alpha)}, \widehat{\theta}_B^{*(1-\alpha)}\right] \qquad (5)$$

Interval ini tidak mengasumsikan data berdistribusi normal, namum interval ini tidak memberikan hasil yang baik kecuali dengan perulangan *bootstrap* paling sedikit 1000 kali.

3. Metodologi Penelitian

Penelitian ini bersifat eksperimen dengan memahami teori untuk maksud dan tujuan. Eksperimen ini dimulai dengan tinjauan pustaka, mengambil kasus yang dapat diselesaikan dengan metode statistik. Penyelesaian kasus dengan simulasi data dan menganalisa output yang dihasilkan dengan *software* R.

3.1. Interval Konfidensi Klasik

Diberikan sampel random $X_1, X_2, \cdots, X_n$ yang diambil dari populasi yang berdistribusi $N\left(\mu, \sigma^2\right)$ dengan σ^2 diketahui. Untuk mendapatkan interval konfidensi $(1-\alpha)$ untuk parameter μ dapat menggunakan *Pivotal Quantity* :

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \qquad (6)$$

Apabila ditentukan nilai $(1-\alpha) = 95\%$, maka interval konfidensi 95% untuk parameter μ diberikan oleh :

$$P(-1,96 \le Z \le 1,96) = 95\%$$

$$\Leftrightarrow P\left(-1,96 \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \le 1,96\right) = 95\%$$

$$\Leftrightarrow P\left(\overline{X} - 1,96.\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 1,96.\frac{\sigma}{\sqrt{n}}\right) = 95\%$$

Jika σ tidak diketahui, diganti dengan s yaitu standar deviasi sampel. Dengan demikian, untuk dapat menentukan interval konfidensi $(1-\alpha)$ untuk μ digunakan *Pivotal Quantity* :

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)} \qquad (7)$$

Dengan demikian, interval konfidensi $(1-\alpha)$ dapat diperoleh dari :

$$P\left(-t_{\left(n-1;\frac{\alpha}{2}\right)} \le T \le t_{\left(n-1;\frac{\alpha}{2}\right)}\right) = 1-\alpha$$

$$\Leftrightarrow P\left(\overline{X} - t_{\left(n-1;\frac{\alpha}{2}\right)}\frac{s}{\sqrt{n}} \le \mu \le \overline{X} + t_{\left(n-1;\frac{\alpha}{2}\right)}\frac{s}{\sqrt{n}}\right) \qquad (8)$$

$$= 1-\alpha.$$

Misalkan bahwa L dan U adalah suatu statistik, katakan : $L = l(X_1, X_2, \cdots, X_n)$ dan $U = u(X_1, X_2, \cdots, X_n)$. Jika data hasil pengamatan $x_1, x_2, \cdots, x_n$, maka sebuah interval $(l(x_1, x_2, \cdots, x_n), u(x_1, x_2, \cdots, x_n))$ disebut interval konfidensi $(1-\alpha)100\%$ untuk θ, jika :

$$P\{l(X_1, X_2, \cdots, X_n) < \theta < u(X_1, X_2, \cdots, X_n)\} = 1-\alpha$$

untuk $0 < \alpha < 1$. Nilai $(1-\alpha)$ disebut koefisien konfidensi atau tingkat konfidensi. $L = l(x_1, x_2, \cdots, x_n)$ dan $U = u(x_1, x_2, \cdots, x_n)$ disebut batas bawah (lower) dan batas atas (*upper*) limit konfidensi yang nilainya dapat dihitung.

3.2. Jumlah Resplikasi *Bootstrap*

Efron dan Tibshirani (1993) menyatakan terdapat beberapa hal yang perlu diperhatikan mengenai jumlah replikasi *bootstrap*, yaitu :

1) Meskipun jumlah replikasi *bootstrap* kecil biasanya sudah cukup informatif. Tetapi jika jumlah replikasinya diperbanyak maka akan sangat cukup untuk memberikan estimasi $se_F(\widehat{\theta})$ yang akurat.
2) Jumlah replikasi *bootstrap* yang besar misal B = 200, biasanya tidak perlu dilakukan dalam mengestimasi standar error (jumlah replikasi *bootstrap* yang besar diperlukan dalam interval konfidensi *bootstrap*).

3.3. Asumsi Metode *Bootstrap*

Ada beberapa asumsi dalam metode *bootstrap*, yaitu :

1) Sampel yang dimiliki merupakan sampel yang sesuai untuk mewakili populasi.
2) Metode *bootstrap* adalah metode yang digunakan untuk mengestimasi suatu distribusi populasi yang tidak diketahui dengan distribusi empiris yang diperoleh dari proses penyempelan ulang. Setiap sampel *bootstrap* berdistribusi sama satu dengan lainnya, atau dapat diasumsikan bahwa sampel *bootstrap* berasal dari distribusi populasi yang sama, tetapi setiap sampel *bootstrap* salin independen.

3.4. Bias Bootstrap

Efron dan Tibshirani (1993) menyatakan bahwa metode *bootstrap* dapat digunakan untuk mengestimasi bias dari estimator $\widehat{\theta} = s(x)$. Estimator *bootstrap* untuk bias didefinisikan sebagai $Bias_B^*$, yaitu :

$$Bias_B^* = B^{-1}\sum_{b=1}^{B}\widehat{\theta}^*(b) - \widehat{\theta} \tag{9}$$

dengan B adalah banyaknya replikasi *bootstrap.*

3.5. Standar Error *Bootstrap*

Efron dan Tibshirani (1993) menyetakan bahwa untuk $se_F(\widehat{\theta})$ adalah standar error dari statistik $\widehat{\theta}$ adalah estimasi *plug-in* yang digunakan pada fungsi distribusi empiris $\widehat{F}$ pada distribusi F yang tidak diketahui. Estimator *bootstrap* untuk $se_F(\widehat{\theta})$ didefinisikan sebagai berikut :

$$se_{\widehat{F}}(\widehat{\theta}^*) = \left[E_{\widehat{F}}\left\{(\widehat{\theta}^*) - E_{\widehat{F}}(\widehat{\theta}^*)\right\}^2\right]^{\frac{1}{2}} \tag{10}$$

Dengan kata lain, estimator *bootstrap* $se_F(\widehat{\theta})$ adalah standar error dari $\widehat{\theta}$ untuk set data pada sampel random berukuran n dari $\widehat{F}$.

Estimator standar error *bootstrap* mungkin tidak akan mudah diselesaikan. Oleh karena itu digunakan algoritma bootstrap untuk mendekati $se_F(\widehat{\theta})$ secara numerik. Algoritma *bootstrap* merupakan cara komputasi untuk mendapatkan pendekatan yang baik terhadap nilai dari $se_{\widehat{F}}(\widehat{\theta}^*)$ dengan langkah-langkah sebagai berikut :

1) Pilih B sampel independen *bootstrap* $x^{*1}, x^{*2}, \cdots, x^{*B}$ dengan tiap-tiap sampel beranggotakan n data yang ditarik dengan pengembalian dari X.
2) Evaluasi replikasi *bootstrap* yang bersesuaian pada setiap sampel *bootstrap*

$$\widehat{\theta}^*(b) = s\left(x^{*b}\right), b = 1, 2, \cdots, B$$

Sehingga

$$\widehat{se}_B = \left\{ \sum_{b=1}^{B} \frac{\left[\widehat{\theta}^*(b) - \widehat{\theta}^*(\cdot)\right]^2}{B-1} \right\}^{-\frac{1}{2}} \tag{11}$$

dengan $\widehat{\theta}^*(.) = \sum_{b=1}^{B} \frac{\widehat{\theta}^*(b)}{B}$

Untuk mengestimasi standar error dari $\widehat{\theta} = s(x)$. Masing-masing sampel *bootstrap* adalah sampel random yang independen berukuran n dari $\widehat{F}$. Banyaknya replikasi *bootstrap* B untuk mengestimasi standar error biasanya antara 25-200. Jika $B \to \infty$, $\widehat{se}_B$ mendekati estimator *plug-in* dari $se_F(\widehat{\theta})$.

Limit $\widehat{se}_B$ dengan B mendekati tak hingga adalah estimator *bootstrap* yang sesuai untuk $se_F\left(\widehat{\theta}\right)$:

$$\lim_{B\to\infty} \widehat{se}_B = se_{\widehat{F}} = se_{\widehat{F}}\left(\widehat{\theta}^*\right) \tag{12}$$

Kenyataan bahwa $\widehat{se}_B$ mendekati $se_{\widehat{F}}$ dengan B menuju tak hingga dapat menunjukkan bahwa *standar error* empiris mendekati *standar error* populasi, jika replikasinya semakin banyak. Estimator *bootstrap* untuk $se_{\widehat{F}}(\widehat{\theta}^*)$ dan pendekatannya, $\widehat{se}_B$, disebut estimasi nonparametrik *bootstrap* karena estimasinya berdasarkan $\widehat{F}$.

3.6. Prinsip Metode *Bootstrap*

Ada beberapa alasan mengapa metode *bootstrap* diperlukan. Misalnya karena ukuran sampel n kecil dan asumsi normalitas tidak terpenuhi. Misalkan kita memiliki data sampel : $\mathbf{X} = \left(X_1, X_2, \cdots, X_3\right)$ yang diperoleh dengan cara sampling acak dari distribus tak diketahui F. Sampel *bootstrap* $\mathbf{X}^* = \left(X_1^*, X_2^*, \cdots, X_3^*\right)$ diperoleh dengan cara sampling acak berukuran n dengan pengembalian dari data asal $\mathbf{X}$.

Misalkan $\widehat{F}$ adalah distribusi empirik untuk distribusi F, yang didefinisikan sebagai : $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{x_i \le x\}$ dengan $I\{A\}$ adalah fungsi indikator dari himpunan A. Selanjutnya estimasi parameter statistik θ yang merupakan fungsional t, tepatnya :

$$\theta = t\left(X_1, X_2, \cdots, X_n; F\right).$$

Dengan menggunakan prinsip *plug-in* digunakan estimator *bootstrap* :

$\widehat{\theta}^* = t\left(X_1^*, X_2^*, \cdots, X_n^*; \widehat{F}\right)$.Untuk mengukur keakurasian, maka diperlukan estimator variansi *bootstrap*:

$$v_{BOOT} = \int\left(t_n(x) - \int t_n(y)\, d\prod_{i=1}^{n}\widehat{F}(y_i)\right)^2 d\prod_{i=1}^{n}\widehat{F}(x_i)$$

$$= \operatorname{var}_*\left(t_n\left(X_1^*, X_2^*, \cdots, X_n^*\right) \middle| X_1, X_2, \cdots, X_n\right)$$

Notasi $= \operatorname{var}_*\left(. \middle| X_1, X_2, \cdots, X_n\right)$menyatakan variansi bersyarat $X_1, X_2, \cdots, X_n$. Akar kuadrat dari v_{BOOT} merupakan estimasi standar error versi *bootstrap*. Estimasi *bootstrap* dari $se_F\left(\widehat{\theta}\right)$, standar error dari statistik $\widehat{\theta}$, adalah estimasi *plug-in* yang menggunakan distribusi empirik $\widehat{F}$ untuk menggantidistribusi tak diketahui F. Standar error ini mengukur keakurasian dari estimator $\widehat{\theta}^*$. Limit dari $\widehat{se}_B$ untuk $B \to \infty$ adalah estimasi *bootstrap* ideal dari $se_{\widehat{F}}\left(\widehat{\theta}^*\right)$.

Berikut adalah algorima *bootstrap* untuk mencari estimasi standar *error* :

1) Kita pilih B sampel *bootstrap* independen $X^{*1}, X^{*2}, \cdots, X^{*B}$, masing-masing berukuran n yang diambil secara acak tanpa pengembalian dari data asal $\mathbf{X}$.
2) Dievaluasi replikasi *bootstrap* berkaitan dengan masing-masing sampel :
$$\widehat{\theta}^*(b) = t\left(X^{*b}\right),\ b = 1, 2, \cdots, B.$$
3) Standar error $se_F\left(\widehat{\theta}\right)$diestimasi dengan standar deviasi B sampel replikasi

$$\widehat{se} = \left(\frac{\sum_{b=1}^{B} \left[\widehat{\theta}^*(b) - \widehat{\theta}^*(.)^2 \right]}{B-1} \right)^{\frac{1}{2}} \quad (13)$$

dimana $\widehat{\theta}^*(.) = \frac{\sum_{b=1}^{B} \widehat{\theta}^*(b)}{B}$.

4. Analisa dan Pembahasan

4.1. Interval Konfidensi Klasik

Sintax Interval Konfidensi Klasik :

```
Int.normal=function(miu,sigma,n,r,alfa)
{
ZAl2=qnorm(1-alfa/2)
y=0
IC=NULL
        for (i in 1:r)
        {
            data=rnorm(n,miu,sigma)
            xbar=mean(data)
            lower=xbar-ZAl2*sigma/sqrt(n)
            upper=xbar+ZAl2*sigma/sqrt(n)
            if ((lower<miu)&&(miu<upper))
                y=y+1
            else y=y
            IC=c(IC,lower,upper)
        }
matrixIC=matrix(IC,ncol=2,byrow=T)
cov=y/r
cat("Interval Normal Standar \n")
print(matrixIC)
cat("Coverage Probability \n")
print(cov)
}
Int.normal(100,32,100,100,0.01)
```

4.2 Interval Konfidensi *Bootstrap*

Sintax Interval Konfidensi Bootstrap :

```
Int.normal=function(mu,sigma,n,r,alpha)
{
data=rnorm(n,mu,sigma)
boot=replicate(r,sample(data,replace=TR
UE,s
    ize=n))
Zal2=qnorm(1-alpha/2)
y=0
IK=NULL
        for (i in 1:r)
        {
            xbar=mean(boot[,i])
            lo=xbar-Zal2*sigma/sqrt(n)
            up=xbar+Zal2*sigma/sqrt(n)
            if ((lo<mu)&&(mu<up))
                y=y+1
            else y=y
            IK=c(IK,lo,up)
        }
mtxIK=matrix(IK,ncol=2,byrow=T)
cov=y/r
print(mtxIK)
print(cov)
}
Int.normal(100,32,100,100,0.05)
```

Pada simulasi diatas terlihat bahwa untuk Interval Konfidensi Klasik dengan mengambil simulasi data yang berasal distribusi Gaussian Normal dengan jumlah data 100 dan parameternya 32 dimana *n* dan *r* berjumlah 100 dan $\alpha = 0.01$. Begitu juga dengan simulasi Interval Konfidensi *Bootstrap* mengambil distribusi Normal dengan jumlah data 100 dan parameternya 31 dimana n dan r berjumlah 100 dan $\alpha = 0.05$.

5. Kesimpulan

Berdasarkan uraian di atas, dapat diambil kesimpulan bahwa untuk Interval Konfidensi Klasik diambil dengan membangkitkan data yang berasal dari distribusi Gaussian Normal dengan parameter $\left(1-\frac{\alpha}{2}\right)$. Tetapi untuk Interval Konfidensi *Bootstrap* diambil dari data yang dibangkitkan berasal dari distribui Random Normal dengan parameter (n, μ, σ). Kemudian akan direplikasi dan nantinya data setelah direplikasi berdistribusi Gaussian Normal. Metode *bootstrap* ini lebih baik digunakan jika dimiliki data yang jumlahnya kecil. *Bootstrap* dapat digunakan untuk mengatasi permasalahan dalam statistika dengan jumlah data yang sedikit, data yang menyimpang dari asumsinya maupun data yang tidak memiliki asumsi distribusinya.

6. Daftar Pustaka

[1] Bain, L. J & Engelhardt, M. (1991). *Introduction to Probability and Mathematical Statistics.* Belmont : Duxbury.

[2] Efron, B. and Tinshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York : Chapman and Hall.

[3] Shao, J and D. Tu. (1995). *The Jackknife and Bootstrap*. New York : Springer Verlag Inc.

[4] SPSS Inc. (2010). *IBM SPSS to Bootstrapping 19*. New York : SPSS Inc.

[5] Versani, John. (2005). *Using R for Introductory Statistics*. Boca Raton : Chapman & Hall/CRC.