

## BAB II

### METODE PERMUTASI DAN BOOTSTRAP UNTUK SAMPLING

Dalam bab berikut ini akan disajikan konsep mengenai pengambilan sampel dalam pengujian hipotesis dengan metode permutasi dan bootstrap yang akan mendukung untuk pembahasan selanjutnya mengenai uji permutasi ASL.

#### 2.1 Uji Hipotesis

Kebenaran atau ketidakbenaran suatu hipotesis statistik tidak pernah diketahui dengan pasti kecuali bila seluruh populasi diamati. Hal ini tentunya tidak praktis dalam banyak keadaan. Karena itu diambil sampel acak dari populasi yang ingin diselidiki, dengan menggunakan informasi yang terkandung dalam sampel itu, diputuskan apakah hipotesis tersebut wajarnya benar atau salah.

Petunjuk dari sampel yang tidak sesuai dengan hipotesis menjurus kepada penolakan hipotesis, sedangkan petunjuk yang mendukung hipotesis menjurus kepada penerimaannya.

Hipotesis yang dirumuskan dengan harapan untuk ditolak disebut hipotesis nol dan dinyatakan dengan  $H_0$ . Penolakan  $H_0$  menjurus pada penerimaan suatu hipotesis alternatif yang dinyatakan dengan  $H_1$ . Sehingga jika  $H_0$  menyatakan hipotesis nol  $p = 0,5$  untuk populasi binomial, hipotesis alternatif  $H_1$  mungkin salah satu dari  $p = 0,75$ ;  $p > 0,5$  atau  $p \neq 0,5$ .

Prosedur membuat keputusan tentang kebenaran atau kesalahan hipotesis disebut pengujian hipotesis.

## 2.2 Metode Permutasi

*Permutasi* adalah penyusunan objek-objek sejumlah  $n$  yang tiap-tiap kali diambil sejumlah  $r$  dengan memperhatikan tata susunannya. Di dalam istilah matematis, apabila ada suatu populasi yang terdiri dari  $n$  objek, sedangkan dari gugus tersebut akan dipilih  $r < n$  buah objek untuk dibentuk menjadi suatu susunan, maka tempat pertama susunan tersebut dapat diduduki oleh salah satu dari  $n$  objek tersebut, kemudian tempat kedua oleh salah satu dari  $(n-1)$  objek yang belum terpilih, dan seterusnya tempat ke  $r$  dapat diduduki oleh salah satu dari  $(n-r+1)$  objek yang tersisakan. Oleh karena itu jumlah permutasi tingkat  $r$  dari  $n$  objek dapat dinyatakan dengan rumus sebagai berikut :

$$P_{(n,r)} = n (n-1) (n-2) \dots (n-r+1) \text{ atau} \dots\dots\dots(2.1)$$

$$P_{(n,r)} = \frac{n!}{(n-r)!} \dots\dots\dots(2.2)$$

Keterangan :  $n! = 1 \cdot 2 \cdot 3 \dots n$  atau

$$n (n-1) (n-2) \dots 2 \cdot 1$$

$$0 ! = 1$$

Dengan demikian dapat diketahui dengan jelas bahwa permutasi dilakukan dengan pengambilan suatu objek sebagai sampel tanpa melakukan pengembalian kembali (*without replacement*).

Selanjutnya untuk menghitung suatu permutasi dengan himpunan ganda dapat digunakan suatu perhitungan sebagai berikut :

Misal  $S$  suatu himpunan ganda dengan faktor pengulangan  $n_1$ , untuk unsur pertama,  $n_2$ , untuk unsur yang kedua,  $\dots$ ,  $n_k$  untuk unsur yang ke  $k$ . Bila jumlah

semua unsur dari  $S$  adalah  $n = n_1 + n_2 + \dots + n_k$ , maka jumlah permutasi dari  $S$  adalah

$$\frac{n!}{n_1! n_2! \dots n_k!}$$

Himpunan ganda  $S$  dapat dituliskan sebagai  $\{ n_1 \cdot a_1, n_2 \cdot a_2, \dots, n_k \cdot a_k \}$ , dimana  $a_1$  adalah unsur pertama dari  $S$ ,  $a_2$  adalah unsur kedua dan seterusnya. Pertama-tama misalkan bahwa setiap unsur dapat dibedakan, termasuk unsur-unsur yang dianggap sama seperti  $n_1$  buah  $a_1$ . Maka untuk  $n$  buah unsur dapat dibuat sebanyak  $n!$  permutasi  $n$ . Tetapi perhitungan  $n!$  permutasi ini memisalkan bahwa semua unsur berbeda, sedang pada kenyataannya tidak demikian. Untuk setiap indeks  $i = 1, \dots, k$  ada sebanyak  $n_i$  unsur  $a_i$  yang sama, ini berarti sebanyak  $n_i!$  permutasi dari setiap unsur  $a_i$  juga dianggap sama dan bukan merupakan permutasi yang berbeda. Hal ini berlaku untuk setiap  $i = 1, 2, \dots, k$ . Jadi jumlah semua permutasi yang berbeda adalah

$$\frac{n!}{n_1! n_2! \dots n_k!}$$

Sehingga bila hanya ada dua unsur yang berbeda yaitu  $a_1$  dan  $a_2$ , masing-masing dengan faktor pengulangan  $n_1$  dan  $n_2$ , dimana  $n = n_1 + n_2$ , maka rumus permutasi dari  $n$  unsur yang ada dalam  $S$  menjadi

$$\frac{n!}{n_1! n_2!} = \frac{n!}{n_1! (n - n_1)!} = \binom{n}{n_1}$$

Jadi besaran  $\binom{n}{n_1}$  dapat dikatakan sebagai kombinasi  $n_1$  dari suatu himpunan dari  $n$  unsur atau juga dapat dikatakan sebagai permutasi  $n$  dari suatu himpunan ganda dengan faktor pengulangan  $n$  dan  $n - n_1$ .

Beranjak dari hal tersebut, maka dapat digunakan suatu teknik perhitungan statistik yang digunakan dalam pengamatan sampel, yang disebut sebagai uji permutasi.

### 2.3 Metode Bootstrap

Metode Bootstrap diperkenalkan oleh *Bradley Efron* dalam *Annals of Mathematical Statistics*, 1979. Metode Bootstrap ini cenderung digunakan untuk menaksir distribusi suatu statistik. 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.

Suatu data  $x_1, \dots, x_n$  yang masing-masing bersifat bebas dari distribusi  $F(x)$ . Sebuah parameter  $\theta = g(F(\cdot))$  mempunyai estimator  $\hat{\theta} = \hat{\theta}(x_1, \dots, x_n)$ . Bootstrap akan melakukan imitasi dari permasalahan yang sesungguhnya yaitu dengan  $F$  yang digantikan oleh distribusi empiris  $F_n(x)$  sebagai berikut :

$$F_n(x) = \frac{\#\{x_i \leq x\}}{n},$$

sehingga parameter dari distribusi empiris tersebut adalah :

$$\tilde{\theta} = g(F_n(\cdot)).$$

Simulasi dari  $F_n(x)$  disebut sebagai sampel bootstrap adalah

$$X_1^*, X_2^*, \dots, X_n^* \text{ , } X_i^* \in F_n(x).$$

Penghitungan estimasi parameternya

$$\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \dots, X_n^*).$$

Mengulangi simulasi di atas akan menunjukkan keakurasian  $\hat{\theta}^*$  yang mengestimasi  $\hat{\theta}$  pada permasalahan bootstrap. Pengambilan sampel bootstrap dilakukan dengan secara random sebanyak  $n$  kali, dengan pengembalian, dari titik – titik data yang orisinil  $x_1, x_2, \dots, x_n$ . Sebagai contoh, dengan  $n = 7$  maka salah satu kemungkinan sampel bootstrap yang akan muncul adalah  $x^* = (x_5, x_7, x_5, x_4, x_7, x_3, x_1)$ .

Untuk selanjutnya  $\text{Var}^*(\hat{\theta}^*)$  adalah  $E^*[\hat{\theta}^* - \hat{\theta}] = \text{Bias}^*$ . Dan distribusi erornya adalah

$$P^*(\hat{\theta}^* - \hat{\theta} \leq u),$$

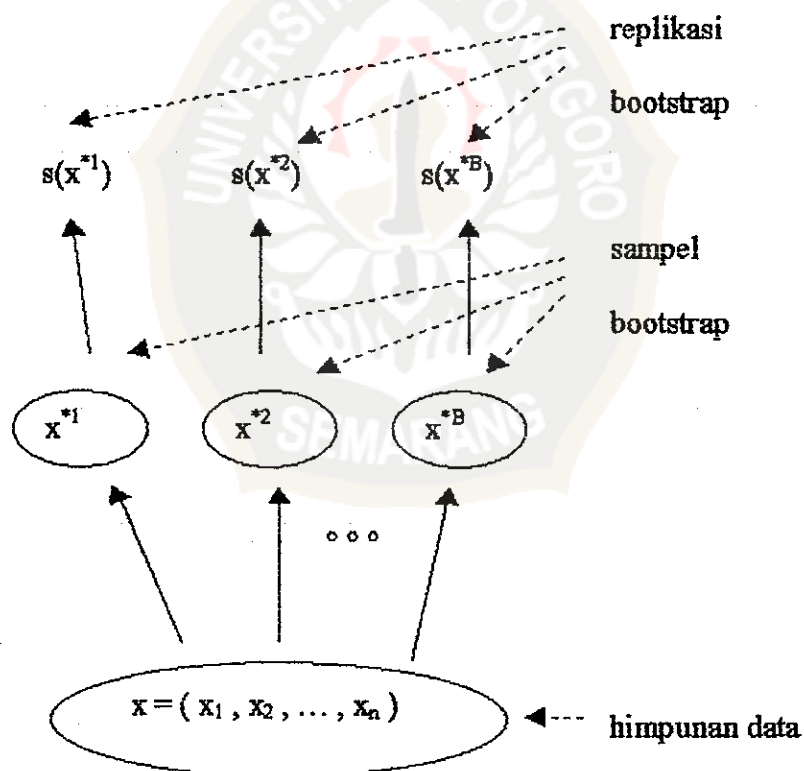
Indeks  $*$  berarti nilai yang diharapkan. Karena terdapat relasi satu-satu antara himpunan dari nilai yang diamati dan distribusi empiris  $F_n(x)$ , maka

$$E^*[\hat{\theta}^* - \hat{\theta}] = E[\hat{\theta}^* - \hat{\theta} | F_n(x)]$$

Sehingga properti estimasi dari masalah bootstrap dapat digunakan untuk menentukan properti dari permasalahan estimasi yang sesungguhnya. Dengan demikian, variabilitas dari  $\hat{\theta}$  dapat diestimasi dengan  $\sigma^* = \sqrt{\text{Var}^*(\hat{\theta}^*)}$  atau dengan histogram atau gambar lain dari distribusi  $\hat{\theta}^* - \hat{\theta}$ .

Berikut ini skema dari proses bootstrap.

Gambar 2.1. skema proses bootstrap untuk estimasi standar error dari sebuah statistik  $s(x)$ . B sampel bootstrap dibangkitkan dari himpunan data yang orisinal. Masing-masing sampel bootstrap mempunyai  $n$  elemen, yang dibangkitkan dengan melakukan sampling dengan pengembalian sebanyak  $n$  kali. Replikasi bootstrap  $s(x^{*1})$ ,  $s(x^{*2})$ , ...  $s(x^{*B})$  diperoleh dengan menghitung nilai dari statistik  $s(x)$  pada masing-masing sampel bootstrap. Standar deviasi dari nilai  $s(x^{*1})$ ,  $s(x^{*2})$ , ...  $s(x^{*B})$  adalah estimasi dari standar error  $s(x)$ .



### 2.3.1 Simulasi Bootstrap menggunakan Median

Misalkan terdapat sebelas data terukur sebagai berikut : 5700, 36300, 12400, 28000, 19300, 21500, 12900, 4100, 91400, 7600, 1600. Nilai data tersebut merupakan observasi independen  $x_i$  pada distribusi kontinu  $F(x)$  dengan median  $\hat{\theta}$ .

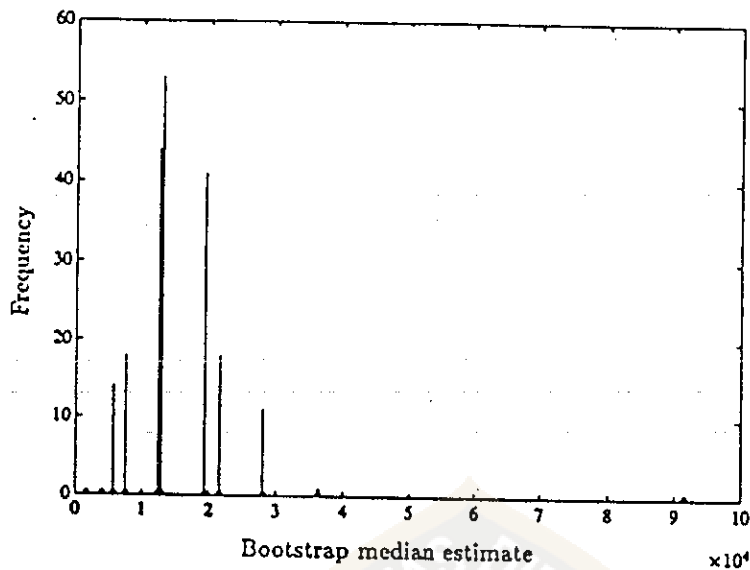
Estimasi  $\hat{\theta}$  dengan nilai median dari data diperoleh

$$\hat{\theta} = x_{(6)} = 12900$$

Sampel bootstrap dari  $F_n(x)$  dapat digambarkan dengan masing-masing  $x_i^*$  secara random di antara nilai-nilai yang diamati  $x_1, \dots, x_{11}$ . Karena  $x_i^*$  bebas dan observasi dilakukan dengan pengembalian, maka nilai-nilai yang sama dapat terambil lebih dari sekali. Pada Tabel 2.1 nilai-nilai yang diambil pada lima simulasi yang pertama diindikasikan dengan tanda +, dan estimasi parameter yang berhubungan dihitung. Nilai parameter yang benar dalam  $F_n(x)$  adalah  $\theta = 12900$ .

Tabel 2.1. Data yang diambil pada lima sampel bootstrap yang pertama.

Data orisinal terurut	Sampel bootstrap ke -				
	1	2	3	4	5
1600			+		++
4100	+++	++	+	+	
5700	+	+	+	+++	+
7600				++	++
12400	+	+	+		+
12900	+		++		
19300	+	+	+	++	
21500		+++	+	+	+
28000	+	++	+	++	+
36300	+	+			+
91400	++		++		++
$\hat{\theta}^*$	12900	21500	12900	7600	12400



Kesimpulan dari 200 simulasi

Dari permasalahan bootstrap di atas maka dari estimasi parameter orisinal  $\hat{\theta} = 12900$  mempunyai :

- Sebuah estimasi bias dari 1943 unit
- Sebuah estimasi standar deviasi  $\hat{\sigma}_{\hat{\theta}} = 5737$
- Sebuah distribusi yang menyerupai diagram di atas (tetapi dalam bentuk kontinu).
- Sebuah estimasi bias diberikan  $\hat{\theta}_1 = 12900 - 1943 = 10957$ .

Meskipun hasil rata-rata bootstrap berada di atas 12900, kesimpulannya estimasinya berada di bawah 12900. Jika estimasi bootstrap terhadap rata-rata terlalu tinggi, maka estimasi nyatanya mungkin terlalu tinggi pula.



### 2.3.2 Simulasi Bootstrap menggunakan Rata-rata

Selanjutnya akan digunakan rata-rata  $\bar{x}$  sebagai estimator untuk nilai tengah dan kemudian diaplikasikan pada teknik bootstrap. Motivasi untuk estimator ini dijelaskan sebagai berikut :

Misal :

Jika distribusi  $F(x)$  simetri di sekitar nilai tengah  $\theta$ , dan nilai yang diharapkan

$$\mu = \int_{-\infty}^{\infty} x dF(x) ,$$

maka  $\mu$  dan nilai tengah  $\theta$  adalah identik. Nilai yang diharapkan biasanya diestimasi oleh rata-rata  $\bar{x}$  dari observasi. Karena  $\theta = \mu$ , rata-rata menjadi optimal untuk estimasi  $\theta$  dalam kasus ini dan kasus yang serupa. Ketika  $F(x)$  non simetri, rata-rata biasanya bias dan tidak berperan lagi sebagai estimator nilai tengah.

Definisi dari parameter  $\tilde{\theta}$  dalam distribusi empiris tidak terpengaruh oleh perubahan estimator ini.  $\theta$  tetap didefinisikan sebagai nilai tengah dari  $F(x)$  dan  $\tilde{\theta}$  sebagai nilai tengah dari  $F_n(x)$  yang berarti bahwa  $\tilde{\theta} = x_{(n)}$ .

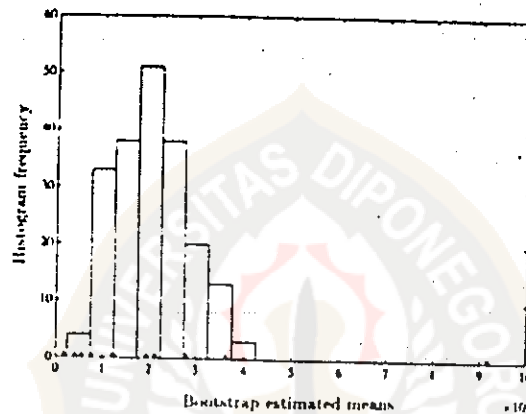
Masih menggunakan data yang sama, ambil  $\theta$  sebagai median dari distribusi teoritik  $F(x)$ . Sehingga estimatornya menjadi

$$\hat{\theta}_b = \bar{x} = 21891$$

Nilai tersebut jauh dari estimasi awal. Permasalahan bootstrap mulai dengan didahului oleh distribusi empiris  $F_n(x)$  dan mediannya  $\tilde{\theta} = 12900$ . Pada masing-masing simulasi, sebelas observasi independen diambil dari  $F_n(x)$ . Estimasi bootstrap

$$\hat{\theta}_b^* = \frac{X_1^* + \dots + X_{11}^*}{11}$$

kemudian akan dihitung. Hasil dari 200 simulasi digambarkan sebagai histogram berikut:



Simulasi bootstrap dari  $\hat{\theta}_b^*$  memberikan hasil sebagai berikut :

rata-rata            22608  
 standar deviasi    7539  
 estimasi bias        22608 – 12900 = 9708

Nilai-nilai di atas tentu saja hanya merupakan estimasi. Estimasi  $\hat{\theta}^*$  ditentukan untuk memiliki sebuah distribusi dengan pendekatan bentuk yang sama seperti pada histogram tersebut. Akhirnya karena bias diestimasi sangat besar maka dihitung estimasi bias yang dikoreksi ( untuk permasalahan orisinal tanpa \* )

$$\hat{\theta}_2 = 21891 - 9708 = 12183.$$

$$E[\hat{\theta}_b^*] = E[X_1^*] = \frac{1600 + \dots + 91400}{11} = 21891$$

merupakan rata-rata dari data orisinal, dan

$$\text{Var}_*(\hat{\theta}_b^*) = \text{Var}_*(\bar{X}^*) = \frac{1}{11} \text{Var}(x_1^*)$$

$$= \frac{1}{11} \sum_{i=1}^{11} (x_i - 21891)^2 \cdot \frac{1}{11},$$

di mana  $x_i$  merupakan data orisinal sehingga memberikan

$$\text{Var}_*(\hat{\theta}_b^*) = 53290000$$

$$\sigma_*(\hat{\theta}_b^*) = 7300$$

$$\text{Bias}^* = 21891 - 8991 = 12900.$$

Dari berbagai pemisalan yang telah diuraikan di atas maka dapat diketahui bahwa bootstrap sangat penting dalam masalah keakurasian dari suatu kesimpulan distribusi dalam populasi. Teknik pengambilan sampel dalam bootstrap adalah pengambilan sampel tanpa pengembalian (*with replacement*) sehingga pada sampel bootstrap,  $x_i$  bisa muncul nol kali, bisa satu kali dan maksimal  $n$  kali.