Penentuan Distribusi Statistik Nisbah Kemungkinan G^2 dalam Analisis Tabel Kontingensi Dimensi Tiga dengan Metode Bootstrap Parametrik

Adi Setiawan
Jurusan Matematika Fakultas Sains dan Matematika UKSW
Jl. Diponegoro 52-60 Salatiga 50711

Abstrak

Tabel kontingensi awal dianggap berasal dari distribusi tertentu yang tergantung pada parameter yang tidak diketahui. Parameter ditaksir berdasarkan tabel kontingensi awal. Tabel kontingensi baru disimulasikan dari distribusi anggapan dengan parameter hasil taksiran. Berdasarkan tabel kontingensi baru dihitung nilai statistik nisbah kemungkinan (*likelihood ratio statistics*) G^2 dalam analisis tabel kontingensi. Metode bootstrap parametrik digunakan dalam pendekatan distribusi G^2.

1. Pendahuluan

Dalam analisis tabel kontingensi (dimensi tiga) biasanya dilakukan analisis untuk melihat ketergantungan antara peubah (*variable*) satu dengan peubah yang lain. Statistik Pearson X^2 dapat digunakan untuk menguji apakah ada ketergantungan antar peubah (Setiawan, 1999). Statistik nisbah kemungkinan G^2 lebih berguna yaitu disamping untuk menguji apakan ada ketergantungan antar peubah juga untuk membandingkan antara dua model mana yang lebih sesuai. Dalam tulisan ini akan dijelaskan bagaimana menggunakan metode bootstrap parametrik dalam penentuan distribusi G^2 yang digunakan dalam analisis tabel kontingensi dimensi tiga.

2. Dasar Teori

Dalam tabel kontingensi dimensi 3 mempunyai tiga peubah yaitu *A*, *B*, dan *C* dengan peubah A mempunyai I kategori, peubah B mempunyai J kategori dan peubah C mempunyai K kategori. Frekuensi sel pengamatan yang termasuk kategori *i* peubah *A* kategori *j* peubah *B* dan kategori *k* peubah *C* dinyatakan dengan Y_{ijk} .

Jika ketiga peubah saling bebas maka peluang terjadinya kejadian pada sel ke-*ijk* dapat dirumuskan di bawah ini:

P_{ijk} = *P*(peubah *A* kategori *i*, peubah *B* kategori *j*, dan peubah *C* kategori *k*).
= *P*(peubah *A* kategori *i*) *P*(peubah *B* kategori *j*) *P*(peubah *C* kategori *k*).
= $P_{i..}P_{.j.}P_{..k}$.

dengan $p_{i..}$ merupakan peluang munculnya kategori i , $p_{.j.}$ merupakan peluang munculnya kategori j dan $p_{..k}$ merupakan peluang munculnya kategori k.

Taksiran nilai harapan untuk sel (*i,j,k*) di bawah hipotesis bahwa ketiga peubah saling bebas adalah

$$\hat{m}_{ijk} = n\,\hat{P}_{i..}\,\hat{P}_{.j.}\,\hat{P}_{..k}$$

dengan, $\hat{P}_{i..} = \dfrac{y_{i..}}{n}$ $\hat{P}_{.j.} = \dfrac{y_{.j.}}{n}$ dan $\hat{P}_{..k} = \dfrac{y_{..k}}{n}$. Dalam hal ini

$$y_{i..} = \sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}\ ;\ y_{.j.} = \sum_{i=1}^{I}\sum_{k=1}^{K} y_{ijk}\ ;\ y_{..k} = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ijk}$$

Statistik nisbah kemungkinan dirumuskan dengan

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk} \ln\left(\frac{y_{ijk}}{\hat{m}_{ijk}}\right)$$

Bila ukuran sampel n besar dan tingkat kepercayaan (*level of significance*) α, hipotesis nol ditolak jika nilai-*p* = $P_{H0}(G^2 > G^2_{hit}) < \alpha$.

3. Penggunaan Metode Boostrap Parametrik dalam Penentuan Distribusi.

Prinsip penggunaan metode bootstrap parametrik dalam pembuatan tabel kontingensi baru dapat dijelaskan berikut ini. Tabel kontingensi dianggap berasal dari populasi yang berdistribusi multinomial dengan parameter n dan p_{ijk} untuk $i = 1,2,\ldots,I$; $j = 1,2,\ldots,J$; $k = 1,2,\ldots,K$. Bila diinginkan untuk menguji hipotesis bahwa setiap peubah saling bebas satu sama lain atau secara matematik dapat dinyatakan sebagai

$$P_{ijk} = P_{i..}P_{.j.}P_{..k}\,.$$

Penaksir kemungkinan maksimum untuk parameter p_{ijk} adalah

$$\hat{P}_{ijk} = \hat{P}_{i..}\,\hat{P}_{.j.}\,\hat{P}_{..k}$$

(Agresti, 1990). Berdasarkan pada anggapan tersebut penggunaan metode bootstrap parametrik dalam penentuan distribusi G^2 dilakukan dengan algoritma berikut :

1. Cuplikan acak (*random sample*) ukuran n dibangkitkan (dari distribusi seragam (*uniform*) pada (0,1)) dan misalkan diperoleh $u_1, u_2, \ldots, u_n$.
2. Interval (0,1) dibagi menjadi IJK interval bagian dan interval bagian yang bersesuaian dengan sel ke-*ijk* mempunyai panjang interval sebesar $\hat{P}_{ijk}$ untuk $i = 1,2,\ldots\ldots,I$; $j = 1,2,..,J$; $k = 1,2,..,K$.
3. Jika terdapat sebanyak Y_{ijk} anggota cuplikan acak $u_1, u_2, \ldots, u_n$ terletak pada interval yang bersesuaian dengan sel ke-*ijk* maka frekuensi sel untuk sel ke-*ijk* pada tabel kontingensi baru adalah Y_{ijk}.
4. Berdasarkan pada tabel kontingensi baru dihitung nilai statistik G^2.
5. Bila prosedur di atas diulang sebanyak bilangan besar B kali maka akan diperoleh nilai-nilai statistik $G^2_1, G^2_2, \ldots, G^2_B$.

   Distribusi dari statistik G^2 dapat didekati dengan histogram dari $G^2_1, G^2_2, \ldots, G^2_B$ dan nilai-p dari suatu hipotesis dapat ditentukan. Algoritma tersebut dapat dinyatakan dalam bahasa yang digunakan dalam paket program statistik Splus versi 3.2 seperti pada Lampiran.

4. Kasus

Untuk memberikan gambaran bagaimana metode bootstrap parametrik dapat digunakan dalam penentuan distribusi G^2 berikut ini diberikan suatu kasus. Misalkan dimiliki tabel kontingensi dimensi tiga seperti pada Tabel 1 dan akan diuji hipotesis yang menyatakan bahwa ketiga peubah A, B dan C saling bebas dengan menggunakan tingkat kepercayaan $\alpha = 0{,}05$.

Tabel 1. Suatu kasus tabel kontingensi dimensi tiga.

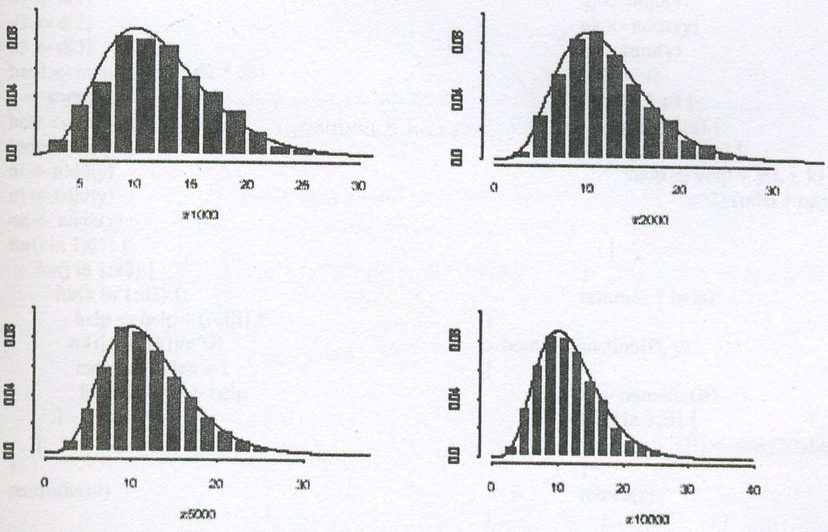
A	B	C	
		C1	C2
A1	B1	15	10
	B2	23	14
	B3	14	8
A2	B1	85	58
	B2	102	60
	B3	49	21
A3	B1	42	51
	B2	81	61
	B3	27	35

Nilai statistik G^2 untuk tabel kontingensi tersebut adalah 25,951. Tabel 2 menyatakan nilai-p bila digunakan pendekatan distribusi G^2 dengan penggunaan metode bootstrap parametrik untuk ulangan sebesar B = 1.000, 2.000 , 5.000 dan 10.000.

Tabel 2. Nilai-p untuk beberapa ulangan B.

B	Nilai-p	Mean nilai-nilai bootstrap G^2	Variansi nilai-nilai bootstrap G^2
1.000	0,008	12,220	23,274
2.000	0,014	12,317	24,794
5.000	0,010	12,120	24,989
10.000	0,012	12,025	24,892

Distribusi G^2 secara analitik dapat ditentukan yaitu G^2 berdistribusi χ^2 dengan derajat bebas 12 (Fienberg, 1991). Gambar 1 menjelaskan perbadingan antara histogram nilai-nilai bootstrap dari dengan distribusi χ^2 dengan derajat bebas 12. Dengan melihat Gambar 1 maka dapat dikatakan bahwa metode bootstrap parametrik berhasil mendekati distribusi eksak untuk statistik G^2.



Gambar 1. Histogram nilai-nilai bootstrap statistik G2 dengan B = 1000, 2000, 5000 dan 10000.

Berdasarkan pada nilai-p maka dengan menggunakan tingkat kepercayaan sebesar 0,05 dapat disimpulkan bahwa hipotesis yang menyatakan bahwa peubah A, B dan C saling bebas akan ditolak. Bila digunakan distribusi eksak dari G^2 maka diperoleh nilai-p sebesar 0,0109 sehingga kesimpulan yang sama juga akan diperoleh bila digunakan distribusi eksak.

5. Kesimpulan

Distribusi statistik nisbah kemungkinan G^2 dapat ditentukan dengan menggunakan metode bootstrap parametrik. Histogram dari nilai-nilai bootstrap statistik G^2 yang diperoleh merupakan pendekatan dari distribusi G^2 sehingga dengan berdasarkan pada nilai-nilai bootstrap ini nilai-p dari suatu uji hipotesis dapat ditentukan.

Daftar Pustaka

[1] A. Agresti , *Categorical Data Analysis*, John Wiley & Sons, Inc, New York (1990)

[2] S.E. Fienberg, The analysis of Cross-Classified Categorical Data, 2nd, The MIT Press, Cambridge (1991)

[3] A. Setiawan, *Penggunaan Metode Bootstrap Parametrik dalam Analisis Tabel Kontingensi Dimensi Dua*, makalah diseminarkan pada SPMIPA di Undip, 13 Nopember 1999.

Lampiran

```
> nidd <- function(y)
{
        d <- dim(y)
        d1 <- d[1]
        d2 <- d[2]
        d3 <- d[3]
        hasil <- numeric(d1)
        for(i in 1:d1) {
                hasil[i] <- sum(y[i, , ])
        }
        return(hasil)
}
> ndjd <- function(y)
{
        d <- dim(y)
        d1 <- d[1]
        d2 <- d[2]
        d3 <- d[3]
        hasil <- numeric(d2)
        for(j in 1:d2) {
                hasil[j] <- sum(y[, j, ])
        }
        return(hasil)
}
> nddk <- function(y)
{
        d <- dim(y)
        d1 <- d[1]
        d2 <- d[2]
        d3 <- d[3]
        hasil <- numeric(d3)
        for(k in 1:d3) {
                hasil[k] <- sum(y[, , k])
        }
        return(hasil)
}
> proba <- function(y)
{
        d <- dim(y)
        d1 <- d[1]
        d2 <- d[2]
        d3 <- d[3]
        hasil <- numeric(d1 * d2 * d3)
        n <- sum(y)
        help <- 0
        index <- 0
        ni <- nidd(y)
        nj <- ndjd(y)
        nk <- nddk(y)
        for(i in 1:d1) {
          for(j in 1:d2) {
            for(k in 1:d3) {
              help <- help + ((ni[i] *
              nj[j] * nk[k])/n^3)
              index <- index + 1
              hasil[index] <- help
            }
          }
        }
        return(hasil)
}
> letak <- function(p, u)
{
        n <- length(u)
        f <- numeric(length(p))
        for(i in 1:n) {
           k <- sum(p < u[i]) + 1
           f[k] <- f[k] + 1
        }
        return(f)
}
> tabelpqr <- function(y)
{
        pr <- proba(y)
        d <- dim(y)
        p <- d[1]
        q <- d[2]
        r <- d[3]
        z <- array(1:(p * q * r), c(p, q, r))
        n <- sum(y)
        u <- runif(n)
        let <- letak(pr, u)
        index <- p * q * r
        for(i in 1:p) {
          for(j in 1:q) {
            for(k in 1:r) {
              z[p - i + 1, q - j + 1, r - k + 1] <- let[index]
        index <- index - 1
            }
          }
        }
        return(z)
}
> statg2 <- function(y)
{
        d <- dim(y)
        p <- d[1]
        q <- d[2]
        r <- d[3]
        ni <- nidd(y)
        nj <- ndjd(y)
        nk <- nddk(y)
        n <- sum(y)
        help <- 0
        for(i in 1:p) {
          for(j in 1:q) {
            for(k in 1:r) {
              help <- help + y[i, j, k] *log((y[i, j, k] *
                   n^2)/(ni[i] * nj[j] *nk[k]))
            }
          }
        }
        return(2 * help)
}
> boot <- function(B, y)
{
        z <- numeric(B)
        for(i in 1:B) {
                z[i] <- statg2(tabelpqr(y))
        }
        return(z)
}
```