Piotr Zieleń

Zadanie na ocenę celującą

9 lipca 2022

1. Wstęp

Celem pracy jest symulacyjne porównanie testu Cochrana, dotyczącego jednorodności prawdopodobieństw sukcesu niezależnych zmiennych losowych z rozkładu dwumianowego, z innymi z zaproponowanych testów, w artykule „Park2019”.

W drugim rozdziale zaprezentuję problem, jakiego dotyczy praca. W trzeciej części przedstawię procedurę testowania jednorodności k zmiennych losowych z rozkładu Bernoulliego, dla klasycznej wersji testu Cochrana oraz moją implementację testu w pakiecie R. W rozdziale czwartym — procedurę testowania dla modyfikacji testu Cochrana wraz z implementacją. W rozdziale piątym — procedurę testowania dla nowego testu. Następnie w przedostatniej części pracy — symulacje, dla której porównam test Cochrana w jego klasycznej wersji, z modyfikacją testu Cochrana i nowym testem, a w ostatniej części — wnioski.

2. Problem

Problem sprowadza się do testowania jednorodności prawdopodobieństwa sukcesu dla k niezależnych zmiennych losowych z rozkładu dwumianowego ($X \sim \mathcal{B}(n, \pi)$). Liczba prób oraz prawdopodobieństwa sukcesu mogą być różne dla rozważanych zmiennych losowych.
Niech $X_i \sim \mathcal{B}(n_i, \pi_i)$, gdzie $1 \leqslant i \leqslant k$, wtedy hipoteza zerowa dla k niezależnych zmiennych losowych jest wyrażona następująco:

$$H_0 : \pi_1 = \pi_2 = \cdots = \pi_k \equiv \pi \tag{1}$$

3. Test Cochrana

Statystyka testowa w teście Cochrana wyrażona jest wzorem:

$$T_s = \sum_{i=1}^{k} \frac{(X_i - n_i\hat{\pi})^2}{n_i\hat{\pi}\,(1-\hat{\pi})}, \tag{2}$$

gdzie $\hat{\pi} = \frac{\sum_{i=1}^{k} X_i}{\sum_{i=1}^{k} n_i}$. Dla prawdziwej hipotezy zerowej, statystyka testowa T_s ma w przybliżeniu rozkład chi-kwadrat z $(k-1)$ stopniami swobody. Hipotezę zerową H_0 odrzucamy wtedy, gdy $T_s > \chi^2_{1-\alpha,k-1}$, gdzie $\chi^2_{1-\alpha,k-1}$, to kwantyl rzędu $1-\alpha$ rozkładu chi-kwadrat z $(k-1)$ stopniami swobody. Autorzy artykułu zwrócili jeszcze uwagę, że gdy k jest stosunkowo duże, to statystykę $\frac{T_s - k}{\sqrt{2k}}$ można przybliżać rozkładem $\mathcal{N}(0,1)$.

3.1. Implementacja w pakiecie R

Poniżej przedstawiłem swoją implementację testu klasycznej wersji testu Cochrana, w pakiecie R:

```
cochran_test <- function(Xi, ni, alfa=0.05){
        k <- length(Xi)
        pi_est <- sum(Xi)/sum(ni)
        Ts <- sum((Xi-ni*pi_est)^2/(ni*pi_est*(1 - pi_est)))
        Ts < qchisq(1-alfa, k-1)
}
```

Funkcja `cochran_test` przyjmuje jako argumenty:

- `Xi` — wektor realizacji zmiennych losowych z rozkładu dwumianowego;
- `ni` — wektor ilości prób w wyżej wymienionych realizacjach zmiennych losowych (parametry n_i w zmiennych losowych $X_i \sim \mathcal{B}(n_i, \pi_i)$);
- `alfa` — poziom istotności (argument opcjonalny, domyślnie wartość 0.05)

4. Modyfikacja testu Cochrana

Statystyka testowa dana jest wzorem (2). Hipotezę zerową odrzucamy wtedy, gdy:

$$\frac{T_s - k}{\sqrt{\hat{\mathcal{B}}_{0k}}} > z_{1-\alpha}, \tag{3}$$

gdzie:

- $z_{1-\alpha}$ to kwantyl rzędu $1-\alpha$, rozkładu standardowego normalnego;
- $\hat{\mathcal{B}}_{0k} = \sum_{i=1}^{k}\left(2 - \frac{6}{n_i} + \frac{1}{n_i \hat{\overline{\pi}}(1-\hat{\overline{\pi}})}\right)$;
- $\hat{\overline{\pi}} = \frac{\sum_{i=1}^{k} n_i \hat{\pi}_i}{N}$
- $\hat{\pi} = \frac{\sum_{i=1}^{k} X_i}{\sum_{i=1}^{k} n_i}$
- $N = \sum_{i=1}^{k} n_i$

4.1. Implemetacja w pakiecie R

Poniżej przedstawiłem swoją implemetację testu Cochrana w wersji zmodyfikowanej, w pakiecie R:

```
modified_cochran <- function(Xi, ni, alfa=0.05){
        k <- length(Xi)
        pi_est <- sum(Xi)/sum(ni)
        Ts <- (Xi-ni*pi_est)^2/(ni*pi_est*(1 - pi_est))
        overline_pi <- 1/sum(ni)*sum(ni * pi_est)
        B0k <- sum(2 - 6/ni + 1/(ni*overline_pi*(1-overline_pi)))
        ETs <- mean(Ts)
        VarTs <- var(Ts)
        T0 <- (sum(Ts) - k)/sqrt(B0k)
        T0 < qnorm(1-alfa)
}
```

Funkcja `modified_cochran` przyjmuje dokładnie takie same argumenty jak funkcja `cochran_test` (opisane w pozrozdziale (3.1)).

5. Nowy test

W artykule został zaproponowany jeszcze nowy test w dwóch wersjach — oparty na dwóch różnych estymatorach. Hipotezę zerową dla tego testu odrzucamy, gdy:

$$T_{\text{new1}}(\text{lub } T_{\text{new2}}) > z_{1-\alpha} \tag{4}$$

gdzie $z_{1-\alpha}$ to kwantyl rzędu $1-\alpha$ rozkładu $\mathcal{N}(0,1)$. Poniżej przedstawiłem w punktach obliczenia jakie należy wykonać, aby wyznaczyć statystyki T_{new1} i T_{new2}:

- $T_{\text{new1}} = \frac{T}{\sqrt{\hat{\mathcal{V}}_1}}$, $T_{\text{new2}} = \frac{T}{\sqrt{\hat{\mathcal{V}}_{1,*}}}$
- $\hat{\mathcal{V}}_1 = \sum_{i=1}^{k}\sum_{l=1}^{4} = a_{li}\hat{\eta}_{li}$, gdzie $\hat{\eta}_{li} = \frac{n_i^l}{\prod_{j=0}^{l-1}(n_i-j)}\prod_{j=0}^{l-1}\left(\hat{\pi}_i - \frac{j}{n_i}\right)$ dla $l = 1,2,3,4$ oraz $a_{1i} = \mathcal{A}_{2i}$, $a_{2i} = \mathcal{A}_{1i} - \mathcal{A}_{2i}$, $a_{3i} = -2\mathcal{A}_{1i}$, $a_{4i} = \mathcal{A}_{1i}$
  $\hat{\mathcal{V}}_{1*} = \sum_{i=1}^{k}\sum_{l=1}^{4} = a_{li}\hat{\eta}_l$, gdzie $\hat{\eta}_l = \frac{N^l}{\prod_{j=0}^{l}(N-j)}\prod_{j=0}^{l-1}\left(\hat{\pi} - \frac{j}{N}\right)$ dla $l = 1,2,3,4$ oraz a_{li} jak wyżej
- $T = \sum_{i=1}^{k} n_i(\hat{\pi}_i - \hat{\overline{\pi}})^2 - \sum_{i=1}^{k} d_i\hat{\pi}_i(1-\hat{\pi}_i)$
- $d_i = \frac{n_i c_i}{n_i - 1}$
- $c_i = \left(1 - \frac{n_i}{N}\right)$

5.1. Implementacja testu T_{new2} w pakiecie R

```
c_i <- function(ni, index){
        1 - ni[index]/sum(ni)
}

d_i <- function(ni, index){
        (ni[index]*c_i(ni, index))/(ni[index] - 1)
}

hat_pi_i <- function(data, ni, index){
        data[index]/ni[index]
}

hat_overline_pi2 <- function(data, ni){
        hat_pi <- data/ni
        sum(ni * hat_pi)/sum(ni)
}

T <- function(data, ni){
        len <- 1:length(data)
        sum(ni * (hat_pi_i(data, ni, len) -
        hat_overline_pi2(data, ni))^2) -
        sum(d_i(ni, len) * hat_pi_i(data, ni, len) *
        (rep(1, length(data)) - hat_pi_i(data, ni, len)))
}

A_1i <- function(ni, index){
        2 - 6/ni[index] - di(ni, index)^2/ni[index] +
```

```
        (8*di(ni, index)^2)/ni[index]^2 - (6*di(ni, index)^2)/ni[index]^3 +
        12*di(ni, index)*(ni[index] - 1)/ni[index]^2
}

di <- function(ni, index){
        ni[index]/(ni[index] - 1) * (1 - ni[index]/sum(ni))
}

A_2i <- function(ni, index){
        ni[index]/sum(ni)^2
}

a_li <- function(l, ni, index){
        if (l==1){
                A_2i(ni, index)
        } else if (l==2){
                A_1i(ni, index) - A_2i(ni, index)
        } else if (l==3){
                -2*A_1i(ni, index)
        } else if (l==4){
                A_1i(ni, index)
        }
}

eta2 <- function(data, ni, l){
        N <- sum(ni)
        hat_pi <- 1/N * sum(ni * hat_pi_i(data, ni, 1:length(data)))
        N^l/(prod(rep(N, l) - 0:(l-1))) * prod(rep(hat_pi, l) - 0:(l-1)/N)
}

V1_gwiazdka <- function(data, ni){
        k <- length(data)
        suma1 <- 0
        sum(sapply(1:k, function(i){
                sum(sapply(1:4, function(j){
                        a_li(j, ni, i) * eta2(data, ni, j)
                }))
        }))
}

T_new2 <- function(data, ni){
        T(data, ni)/sqrt(V1_gwiazdka(data, ni)) < qnorm(0.95)
}
```

6. Symulacje

W tym rozdziale przedstawię przeprowadzone symulacje. Przeprowadziłem wszystkie dziesięć symulacji z artykułu dla testu Cochrana, jego zmodyfikowanej wersji i testu T_{new2} dla 10^4 powtórzeń. Kod do symulacji przedstawiłem tylko w przypadku pierwszej symulacji. W kolejnych przypadkach kody wyglądały analogicznie.

6.1. Setup 1

$\pi_i = 0.001$ dla $1 \leqslant i \leqslant k$, $\pi_k = 0.001 + \delta$, k=8
$\mathbf{n}_8 = 20(2, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8)$:

```
n8 <- 20*c(2, 2^2, 2^3, 2^4, 2^5, 2^6, 2^7, 2^8)
result_cochran <- c()
result_mod_cochran <- c()
result_tnew2 <- c()
for (j in 1:10){
        Y <- NA
        while (is.na(mean(Y))){
        Y <- sapply(1:10000, function(...){
                pi <- c(0.001, 0.001, 0.001, 0.001,
                0.001, 0.001, 0.001, 0.001 * j)
        random_variables <- c()
        for (i in 1:8){
                random_variables <- c(random_variables,
                rbinom(1, n8[i], pi[i]))
        }
        c(cochran_test(random_variables, n8),
        modified_cochran(random_variables, n8),
        T_new2(random_variables, n8))
        })
        }
        result_cochran <- c(result_cochran, 1 - mean(Y[1,]))
        result_mod_cochran <- c(result_mod_cochran, 1 - mean(Y[2,]))
        result_tnew2 <- c(result_tnew2, 1 - mean(Y[3,]))
}


df <- data.frame(seq(0.000, 0.009, by=0.001),
result_cochran, result_mod_cochran, result_tnew2)
colnames(df) <- c("delta", "Test Cochrana", "Zmodyfikowany test Cochrana",
"Test Tnew2")
df

##    delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1  0.000        0.1159                      0.0568     0.0363
## 2  0.001        0.0684                      0.0310     0.0759
## 3  0.002        0.1184                      0.0362     0.3168
## 4  0.003        0.3298                      0.1429     0.6441
## 5  0.004        0.6236                      0.3962     0.8617
## 6  0.005        0.8409                      0.6823     0.9627
## 7  0.006        0.9492                      0.8793     0.9925
## 8  0.007        0.9873                      0.9621     0.9982
## 9  0.008        0.9988                      0.9924     0.9999
## 10 0.009        0.9991                      0.9980     0.9998
```

6.2. Setup 2

(w tym przypadku, w porównaniu do artykułu, zrobiłem mały wyjątek — δ zwiększam o 0.001, zamiast o 0.01)
$\pi_i = 0.001 + \delta$, dla $k = 1$, $\pi_i = 0.001$, dla $2 \leqslant i \leqslant k$, dla $k = 8$ i $\mathbf{n}_8$ — tak jak w (6.1):

```
##    delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1  0.000         0.1134                        0.0532     0.0355
## 2  0.001         0.1460                        0.0839     0.0389
## 3  0.002         0.1762                        0.1154     0.0426
## 4  0.003         0.2143                        0.1409     0.0456
## 5  0.004         0.2480                        0.1802     0.0521
## 6  0.005         0.2671                        0.2035     0.0596
## 7  0.006         0.3066                        0.2192     0.0675
## 8  0.007         0.3317                        0.2616     0.0715
## 9  0.008         0.3548                        0.2792     0.0771
## 10 0.009         0.3909                        0.3030     0.0955
```

6.3. Setup 3

$\pi_1 = 0.001 + \delta$ i $\pi_i = 0.001$ dla $2 \leqslant i \leqslant 8$, $k = 8$, $n_i = 2560$ dla $1 \leqslant i \leqslant 8$:

```
##     delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1  0.0000         0.0444                        0.0288     0.0551
## 2  0.0005         0.0638                        0.0456     0.0820
## 3  0.0010         0.1372                        0.1099     0.1558
## 4  0.0015         0.2449                        0.2087     0.2817
## 5  0.0020         0.4004                        0.3537     0.4316
## 6  0.0025         0.5599                        0.5145     0.5856
## 7  0.0030         0.6896                        0.6474     0.7123
## 8  0.0035         0.7900                        0.7634     0.8180
## 9  0.0040         0.8741                        0.8482     0.8867
## 10 0.0045         0.9259                        0.9143     0.9342
```

6.4. Setup 4

$\pi_i = 0.001$ dla $1 \leqslant i \leqslant k-1$ and $\pi_k = 0.001 + \delta$, dla $k = 40$
$\mathbf{n}_{40} = 20(\mathbf{n}_1^*, \mathbf{n}_2^*, \mathbf{n}_3^*, \mathbf{n}_4^*, \mathbf{n}_5^*, \mathbf{n}_6^*, \mathbf{n}_7^*, \mathbf{n}_8^*)$, gdzie $\mathbf{n}_m^* = (2^m, 2^m, 2^m, 2^m, 2^m)$ to 5-wymiarowy wektor

```
##    delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1  0.000         0.1473                        0.0546     0.0510
## 2  0.001         0.1601                        0.0531     0.1212
## 3  0.002         0.3282                        0.1301     0.4551
## 4  0.003         0.6297                        0.3830     0.8043
## 5  0.004         0.8712                        0.7119     0.9631
## 6  0.005         0.9755                        0.9155     0.9947
## 7  0.006         0.9959                        0.9827     0.9992
## 8  0.007         0.9994                        0.9973     0.9999
## 9  0.008         1.0000                        0.9999     1.0000
## 10 0.009         1.0000                        1.0000     1.0000
```

6.5. Setup 5

$\pi_i = 0.001 + \delta$, dla $i = 1$ i $\pi_i = 0.001$ dla $2 \leqslant i \leqslant k$ dla $k = 40$ i $\mathbf{n}_{40}$ — tak jak w (6.4):

```
##    delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1  0.000        0.1571                      0.0526     0.0513
## 2  0.001        0.1715                      0.0630     0.0557
## 3  0.002        0.1839                      0.0666     0.0559
## 4  0.003        0.2060                      0.0778     0.0595
## 5  0.004        0.2185                      0.0887     0.0677
## 6  0.005        0.2392                      0.1020     0.0729
## 7  0.006        0.2559                      0.1111     0.0775
## 8  0.007        0.2799                      0.1245     0.0872
## 9  0.008        0.2889                      0.1388     0.0959
## 10 0.009        0.3046                      0.1547     0.1068
```

6.6. Setup 6

$\pi_i = 0.001 + \delta$, dla $k = 1$ i $\pi_i = 0.001$ dla $2 \leqslant i \leqslant k$ dla $k = 40$ i $n_i = 2560$ dla $1 \leqslant i \leqslant k$:

```
##   delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1 0.000        0.0486                      0.0365     0.0616
## 2 0.001        0.0923                      0.0766     0.1111
## 3 0.002        0.2767                      0.2525     0.2928
## 4 0.003        0.5510                      0.5068     0.5717
## 5 0.004        0.7634                      0.7518     0.7860
## 6 0.005        0.9089                      0.8984     0.9208
## 7 0.006        0.9667                      0.9601     0.9727
```

6.7. Setup 7

$\pi_i = 0.001(1 + \epsilon_i)$, $k = 40$, $n_i = 2560$ dla $1 \leqslant i \leqslant 40$, gdzie ϵ_i jest równomiernie rozłożony na odcinku $[-\delta, \delta]$:

```
##   delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1  0.00        0.0471                      0.0391     0.0594
## 2  0.25        0.0841                      0.0660     0.0999
## 3  0.50        0.2425                      0.2066     0.2692
## 4  0.75        0.6094                      0.5624     0.6336
## 5  1.00        0.9330                      0.9150     0.9421
```

6.8. Setup 8

$\pi_i = 0.01(1 + \epsilon_i)$, $k = 40$ i $\mathbf{n}_{40}$ — jak w (6.4) oraz ϵ_i jest równomiernie rozłożony na odcinku $[-\delta, \delta]$:

```
##   delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1  0.00        0.0704                      0.0492     0.0599
## 2  0.25        0.0263                      0.0174     0.0889
## 3  0.50        0.0611                      0.0417     0.3415
## 4  0.75        0.2787                      0.2329     0.7924
## 5  1.00        0.7240                      0.6662     0.9842
```

6.9. Setup 9

$\pi_i = 0.01(1+\epsilon_i)$, $k = 2000$, $n_i = 100$ dla $1 \leqslant i \leqslant 2000$, gdzie ϵ_i jest równomiernie rozłożony na odcinku $[-\delta, \delta]$:

```
##   delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1   0.0          0.0474                        0.0220     0.0538
## 2   0.2          0.1084                        0.0569     0.1160
## 3   0.4          0.5039                        0.3755     0.5031
## 4   0.6          0.9716                        0.9453     0.9709
## 5   0.8          1.0000                        1.0000     1.0000
```

6.10. Setup 10

$\pi_i = 0.01(1+\epsilon_i)$, $k = 2000$, $\mathbf{n} = (\mathbf{n}_{1,250}, \mathbf{n}_{2,250}, \ldots, \mathbf{n}_{8,250})$, gdzie $\mathbf{n}_{m,250} = (2^m, 2^m, \ldots, 2^m)$ jest 250-cio wymiarowym wektorem, ϵ_i jest równomiernie rozłożony na odcinku $[-\delta, \delta]$

```
##   delta Test Cochrana Zmodyfikowany test Cochrana Test Tnew2
## 1   0.0          0.2458                        0.0503     0.0433
## 2   0.2          0.0100                        0.0006     0.0281
## 3   0.4          0.0000                        0.0000     0.0276
## 4   0.6          0.0000                        0.0000     0.0407
## 5   0.8          0.0000                        0.0000     0.0871
## 6   1.0          0.0000                        0.0000     0.2036
```

7. Wnioski

Celem pracy było symulacyjne porównanie testu Cochrana, sprawdzającego jednorodność prawdopodobieństw sukcesu n niezależnych zmiennych losowych z rozkładu dwumianowego. Liczba prób dla zmiennych losowych (n w $\mathcal{B}(n, \pi)$) może być różna. W pracy skupiłem się na symulacyjnym porównaniu testu Cochrana z jego zmodyfikowaną wersją i jednym z nowych testów. W tym celu zaimplementowałem testy opracowane w artykule i przeprowadziłem 10 symulacji z artykułu (uzyskane wartości są wyznaczone na podstawie symulacji Monte-Carlo, na podstawie 10^4 powtórzeń).
Symulacje pokazały, jak różnie reagują rozważane testy, na zmieniające się parametry. Pierwszy wiersz, w każdej przeprowadzonej symulacji opisuje prawdopodobieństwo popełnienia błędu pierwszego rodzaju — czyli prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest ona prawdziwa. Wartość ta powinna wskazywać liczbę w przybliżeniu 0.05, ponieważ wykonywane testy były robione na poziomie istotności $\alpha = 0.05$. Na podstawie przeprowadzonych symulacji, nie można jednoznacznie stwierdzić, który z testów jest najlepszy, każdy ma w zależności od rozważanych sytuacji swoje „mocniejsze” i „słabsze” strony.