Raport 5 - Porównanie metody Agrestiego-Coulla z metodą klasyczną

Alicja Wiączkowska

2023-05-20

Cel raportu

Raport skupia się na przedstawieniu różnic w wyznaczaniu przedziałów ufności między metodami Agrestiego-Coulla a klasyczną oraz porównaniu którą z nich można uzyskać bardziej adekwatne wyniki.

Część teoretyczna

Problem: konstrukcja przedziału ufności dla średniej zmiennej losowej o rozkładzie zero-jedynkowym.

Założenia:

- badana zmienna losowa to X o średniej $\mathbb{E}(X)$=p
- X_i to i-te powtórzenie doświadczenia opisywanego zmienną losową X (doświadczenia te są niezależne)
- ilość doświadczeń n jest ustalona (n jest rozmiarem próby losowej)
- Zmienna losowa $Y = \sum_{i=1}^{n} X_i$ ma rozkład Bermoulliego $Y \sim B(n,p)$ - można go więc przybliżać rozkładem normalnym $Y \sim N(\ np,\ \sqrt{np(1-p)}\ )$

Metoda Klasyczna

Parametrem przybliżającym wartość średnią p jest $\hat{p} = \frac{Y}{n}$ o rozkładzie normalnym $\hat{p} \sim N(\ p,\ \sqrt{\frac{p(1-p)}{n}}\ )$.
Standardowy błąd to: $SE_{\hat{p}} = \sqrt{n\hat{p}(1-\hat{p})}$.
$Z_{\alpha/2}$ - kwantyl rozkładu normalnego

Przybliżony przedział ufności na poziomie istotności α wyznaczony metodą klasyczną to

$$[\hat{p} - Z_{\alpha/2} \cdot SE_{\hat{p}},\ \hat{p} + Z_{\alpha/2} \cdot SE_{\hat{p}}]$$

Metoda Agrestiego-Coulla

Średnia p jest przybliżana parametrem $\tilde{p} = \frac{Y + \frac{1}{2}(Z_{\alpha/2})^2}{n + (Z_{\alpha/2})^2}$.
Natomiast błąd standardowy zadaje wzór: $SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+(Z_{\alpha/2})^2}}$

Przybliżony przedział ufności na poziomie istotności α wyznaczony metodą Agrestiego-Coulla to

$$[\tilde{p} - Z_{\alpha/2} \cdot SE_{\tilde{p}},\ \tilde{p} + Z_{\alpha/2} \cdot SE_{\tilde{p}}]$$

Symulacje

W celu porównania metod należy wykonać doświadczenia lub przeprowadzić symulacje komputerowe i wyciągnąć wnioski o skuteczności obu metod w poszczególnych sytuacjach.

Podłoże programistyczne - niezbędny kod

Pierwszym krokiem przeprowadzenia syulacji jest napisanie funkcji poprawnie obliczających przedziały ufności wartości średniej o zadanym poziomie istotności według metody klasycznej oraz metody Agrestiego-Coulla. Przydatna może być też funkcja spradzająca czy rzeczywisa wartość średnia mieści się w obliczonym przedziale.

```
PU_klas<-function(wek, alpha=0.05){
  'obliczenie PU o poziomie istotności alpha dla wektora wek metodą klasyczną'
  n=length(wek)
  Y=sum(wek)
  Z=qnorm(alpha/2,lower.tail = F)

  p_hat = Y/n
  SE=sqrt(p_hat*(1-p_hat) / n)

  lew<-p_hat-Z*SE
  praw<-p_hat+Z*SE

  return(c(lew,praw))
}

PU_AC<-function(wek, alpha=0.05){
  'obliczenie PU o poziomie istotności alpha dla wektora wek metodą
  Agrestiego-Coulla'
  n=length(wek)
  Y=sum(wek)
  Z=qnorm(alpha/2,lower.tail = F)

  p_til = (Y + 0.5 * Z**2 ) / (n+ Z**2)
  SE=sqrt(p_til*(1-p_til) / (n+ Z**2) )

  lew<-p_til-Z*SE
  praw<-p_til+Z*SE

  return(c(lew,praw))
}

IN <-function(p, PU){
  #funkcja sprawdza czy wartość p mieści się w podanym przedziale PU
  lew=PU[1]
  praw=PU[2]
  if(p<= praw & p>=lew) {return(T)}
  else {return(F) }
}
```

W kolejnym kroku należy stworzyć funkcje przeprowadzające serie symulcji, obliczające przybliżone przedziały ufności i sprawdzające jak często rzeczywista wartość średnia do nich należy w zależności od wartości parametru p.

```
sym_klas<-function(n,p,alpha=0.05,m=100){
  'n - ilosc powtórzeń doświadczenia;
   p - wektor wartości średnich E(X)
   m - wielkość próby w jednym doświadczeniu
  funkcja zwraca wektor poziomów ufności wyznaczonych przedziałów'
  p_len<-length(p)
  ufnosc<-numeric(p_len)

  for(k in 1:p_len){
    dosw<-numeric(n)
    for(j in 1:n){
      x<-sample( c(1,0),m,T,prob=c( p[k],1-p[k] ) )
      PU<-PU_klas(x,alpha)
      dosw[j]<-IN(p[k],PU)
    }
    ufnosc[k]<-mean(dosw)
  }
  return(ufnosc)
}


sym_AC<-function(n,p,alpha=0.05,m=100){
  'n - ilosc powtórzeń doświadczenia;
   p - wektor wartości średnich E(X)
   m - wielkość próby w jednym doświadczeniu
  funkcja zwraca wektor poziomów ufności wyznaczonych przedziałów'
  p_len<-length(p)
  ufnosc<-numeric(p_len)

  for(k in 1:p_len){
    dosw<-numeric(n)
    for(j in 1:n){
      x<-sample( c(1,0),m,T,prob=c( p[k],1-p[k] ) )
      PU<-PU_AC(x,alpha)
      dosw[j]<-IN(p[k],PU)
    }
    ufnosc[k]<-mean(dosw)
  }
  return(ufnosc)
}
```

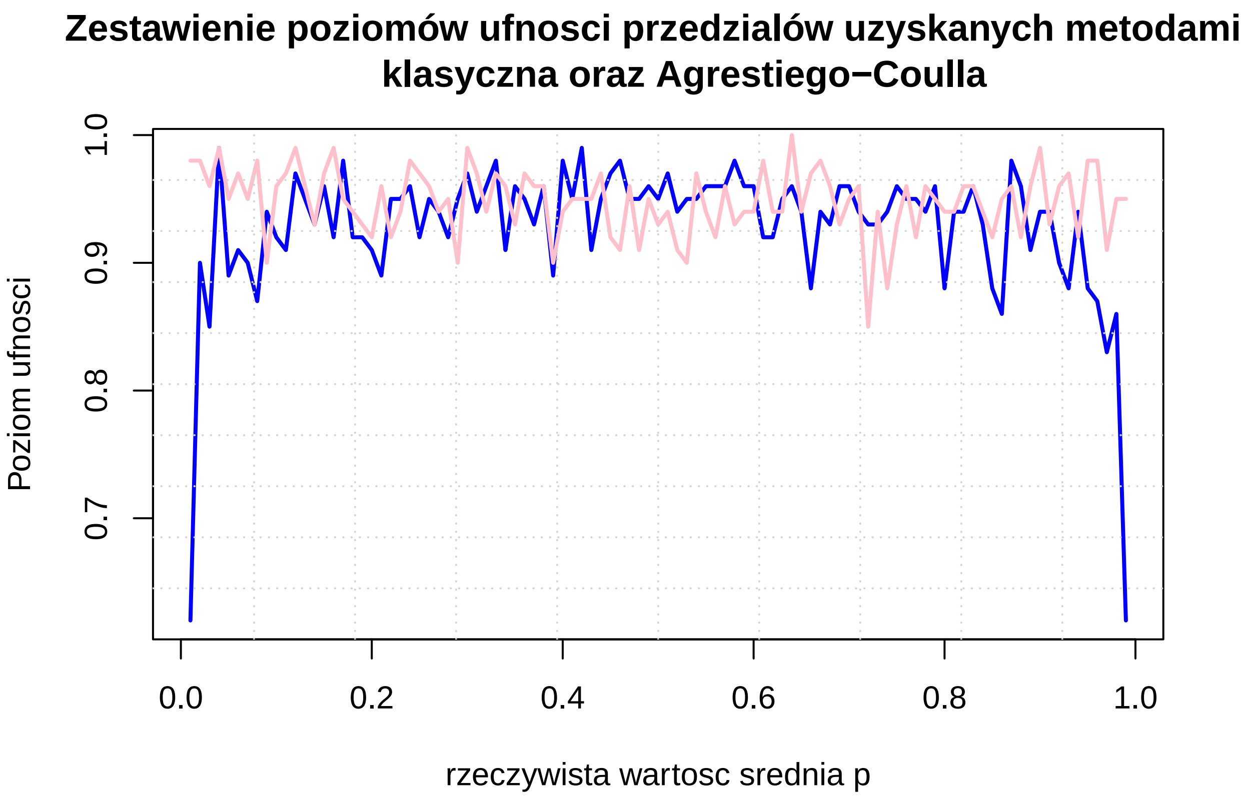
Zestawienie danych

Z użyciem powyższych funkcji przeprowadzono symulacje z parametrami: n=1000, m=100, $\alpha = 0.05$. Doświadczalnie sprawdzony poziom ufności dla zadanych wartości p przedstawiono w tabeli:

p	klasyczna	A-C
0.01	0.631	0.975
0.05	0.877	0.969
0.10	0.930	0.972

p	klasyczna	A-C
0.20	0.948	0.950
0.50	0.929	0.937
0.80	0.931	0.935
0.90	0.928	0.971
0.95	0.897	0.976
0.99	0.642	0.985

Dla parametrów n=m=100, $\alpha = 0.05$ narysowano wykres zależności poziomu ufności od wartości średniej p. Kolorem niebieskim oznaczono wynki uzuskane metodą klasyczną, a różowym metodą Agrestiego-Coulla.



Wnioski

Z obu powyższych zestawień wynika, że niezależnie od wartosci p metoda Agrestiego-Coulla zwraca przedział z w zybliżeniu 95%-owym poziomem ufności - czyli zgodnym z oczekiwaniami. Natomiast metoda klasyczna zwraca znacznie mniej dokładne wyniki dla wartości średniej p leżącej w ogonach. Metody tej nie można zatem stosować, kiedy p jest bliskie 0 lub 1. W pozostałych przypadkach obie metody są porównywalne.