Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

STATYSTYKA MATEMATYCZNA

z pakietem R

IV. Estymacja punktowa i przedziałowa

Przemysław Grzegorzewski
Konstancja Bobecka-Wesołowska
Marek Gągolewski

Spis treści

1 Wprowadzenie

1.1 Dystrybuanta empiryczna

Definicja 1. *Dystrybuantą empiryczną* z próby $(X_1, \ldots, X_n)$ nazywamy funkcję

$$F_n(t; X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, t]}(X_i), \quad t \in \mathbb{R}. \tag{1}$$

Dla ustalonej wartości próby $(x_1, \ldots, x_n)$ funkcja $F_n(\cdot; x_1, \ldots, x_n)$ jest dystrybuantą schodkową, mającą skoki wielkości $1/n$ w punktach $x_1, \ldots, x_n$. Do narysowania wykresu dystrybuanty empirycznej można użyć w programie `R` funkcji `ecdf()`. Wektor `x`, zawierający wartości próbki $x_1, \ldots, x_n$, podajemy jako argument tej funkcji: `ecdf(x)`.

1.2 Przedziały ufności

W pakiecie `R` mamy do dyspozycji funkcje pozwalające wyznaczać przedziały ufności (ang. *confidence intervals*) dla wartości oczekiwanej w modelu normalnym z nieznanym odchyleniem standardowym oraz dla wskaźnika struktury (proporcji) w modelu dwupunktowym.

Niech $(X_1, \ldots, X_n)$ będzie próbą z rozkładu normalnego $\mathrm{N}(\mu, \sigma)$ o nieznanych parametrach μ i σ. Do wyznaczenia przedziału ufności dla wartości oczekiwanej μ można użyć funkcji `t.test()`. Pierwszym argumentem tej funkcji jest wektor zawierający wartości próbki, na podstawie których szacujemy μ. Poziom ufności (ang. *confidence level*) podajemy jako drugi argument tej funkcji, np. `conf.level=0.9` (domyślnie wybierany jest 0,95).

Niech $(X_1, \ldots, X_n)$ będzie próbą z rozkładu dwupunktowego $\mathrm{Bern}(p)$. Do wyznaczenia przedziału ufności dla wskaźnika struktury (proporcji) p można użyć funkcji `binom.test()` lub `prop.test()`, przy czym w drugim przypadku dostajemy asymptotyczny przedział ufności. Pierwszym argumentem obu tych funkcji jest liczba jedynek w naszej próbie (odpowiadająca liczbie elementów posiadających interesującą nas cechę), a drugim — liczność próby n. Poziom ufności podajemy jako kolejny argument, np. `conf.level=0.9` (domyślnie wybierany jest 0,95).

Inne przedziały ufności wyznaczamy sami, pisząc odpowiedni program.

2 Zadania rozwiązane

Zadanie 4.1. Wygeneruj dwie próby losowe z rozkładu standardowego normalnego: 20 i 100 elementową. Narysuj dla obu prób dystrybuanty empiryczne i porównaj je z odpowiednią dystrybuantą teoretyczną.

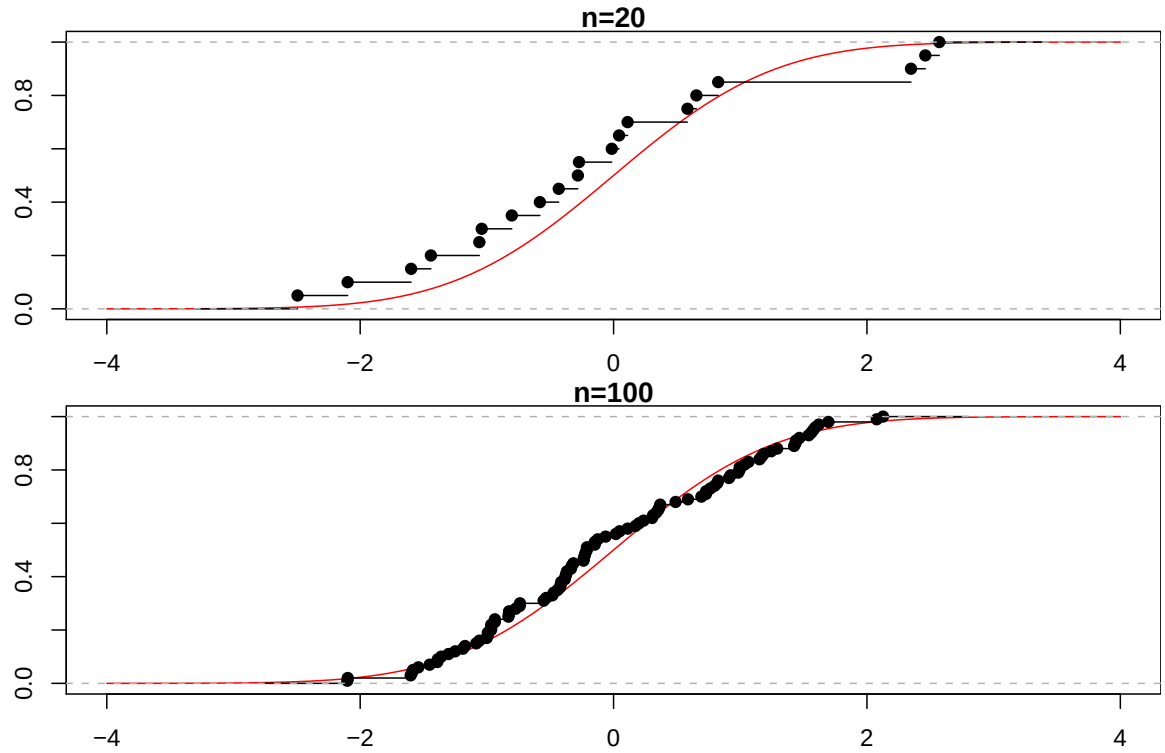
Rozwiązanie.

Generujemy próby losowe: 20 i 100 elementową z rozkładu $\mathrm{N}(0, 1)$.

```
> y20  <- rnorm(20)
> y100 <- rnorm(100)
```

Dla każdej z próbek rysujemy wykres dystrybuanty rozkładu $\mathrm{N}(0, 1)$ i nakładamy na niego wykres dystrybuanty empirycznej:

```
> par(mfrow=c(2,1))# tworzymy miejsce do narysowania 2 wykresów na jednym rysunku
> curve(pnorm(x),from=-4,to=4,col="red",main="n=20");   plot(ecdf(y20),  add=T)
> curve(pnorm(x),from=-4,to=4,col="red",main="n=100");  plot(ecdf(y100), add=T)
```



□

Zadanie 4.2. Wygeneruj $n = 500$-elementową próbę $(Y_1, \ldots, Y_n)$ z rozkładu standardowego Cauchy'ego (z parametrem położenia $a = 0$ i parametrem rozproszenia $b = 1$).

a) Dla każdej podpróbki zawierającej i początkowych elementów próbki wyjściowej, tzn. dla $\mathbf{X}_i = (Y_1, \ldots, Y_i)$, gdzie $i = 1, \ldots, n$, oblicz średnią $\bar{\mathbf{X}}_i$ oraz medianę Med_i Następnie przedstaw na wspólnym wykresie zbiory $\left\{\bar{\mathbf{X}}_i : i = 1, \ldots, n\right\}$ oraz $\{\mathrm{Med}_i : i = 1, \ldots, n\}$. Przeanalizuj wpływ liczności próby na zachowanie się średniej oraz mediany z próby. Czy statystyki te wydają się być sensownymi estymatorami parametru położenia a w tym modelu?

b) Dla każdej podpróbki zawierającej $i = 2, \ldots, n$ początkowych elementów próbki wyjściowej oblicz odchylenie standardowe s_i oraz odchylenie ćwiartkowe $r_i = \mathrm{IQR}(\mathbf{X}_i)/2$ (czyli rozstęp międzykwartylowy podzielony przez 2). Następnie przedstaw na wspólnym wykresie zbiory $\{s_i : i = 2, \ldots, n\}$ oraz $\{r_i : i = 2, \ldots, n\}$. Przeanalizuj wpływ liczności próby na zachowanie się s_i oraz r_i. Czy statystyki te wydają się być sensownymi estymatorami parametru rozproszenia b w tym modelu?

Rozwiązanie.
Porównanie estymatorów parametru położenia $a = 0$ w rozkładzie Cauchy'ego C(0, 1).

Do rozwiązania tego przykładu konieczne będzie użycie jakiegoś mechanizmu, który pozwoli nam na rozpatrzenie dla różnych i próbek $\mathbf{X}_i$ i policzenie dla każdej z nich odpowiednich statystyk.

Sposób postępowania powinien być następujący. Dla każdego i ze zbioru $1, 2, \ldots, n$ chcemy policzyć średnią i medianę dla ciągu $\mathbf{X}_i$ złożonego z elementów $(Y_1, \ldots, Y_i)$. Informacje te winny być przechowywane jako elementy ciągów wyjściowych $\bar{\mathbf{X}}_i$ oraz Med_i.

Do zapisania powyższego algorytmu użyjemy pętli `for`.

Uwaga

Pętla `for` w języku `R` służy do cyklicznego wykonywania ciągu instrukcji. Jej składnia jest następująca: Pętla `for`

```
for (Zmienna in WektorWartosci)
{
   ... InstrukcjeDoWykonania ...
}
```

`InsktrukcjeDoWykonania` wykonają się `length(WektorWartosci)` razy, z tym że w każdej kolejnym powtórzeniu `Zmienna` będzie przyjmować kolejną wartość z ciągu `WektorWartosci`, czyli w porządku: `WektorWartosci[1]`, `WektorWartosci[2]`, ...

Najprościej jest pokazać tę zasadę na przykładzie:

```
> for (i in 1:5)          # dla każdego i=1,2,3,4,5
  {                       # wykonaj:
     print(i);            # wypisz i
  }                       # koniec

[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
```

Jeżeli jest tylko jedna instrukcja do wykonania, można pominąć nawiasy klamrowe `{·}` — grupują one wiele poleceń.

```
> for (i in 1:5) print(2^i);

[1] 2
[1] 4
[1] 8
[1] 16
[1] 32
```

W wielu (naprawdę wielu!) zastosowaniach użycie pętli w języku `R` jest nieuzasadnione. Należy, gdzie się da (i gdzie się potrafi), starać się używać innych mechanizmów — zastosowanie `for` jest często niewydajne. Trzeba zatem porzucić nawyki z języka `C`. Np. powyższy przykład lepiej jest zaimplementować tak:

```
> print(2^(1:5)) # tylko działania na wektorach

[1]  2  4  8 16 32
```

Dla porównania, spróbujmy wyznaczyć wektor liczb $a_1, \ldots, a_n$, gdzie $a_i = \left(1 + \frac{1}{i}\right)^i$ (kolejne przybliżenie liczby e) — za pomocą zarówno pętli `for`, jak i operacji na wektorach. Interesować nas będą czasy wykonania tych operacji: zwraca je funkcja `system.time()`.

```
> n <- 1000000;
> a1 <- numeric(n); # pusty wektor o rozmiarze n;
> system.time( {     for (i in 1:n) a1[i] <- (1+1/i)^i;      } ); # sposób I
> system.time( {     a2 <- (1+1/(1:n))^(1:n);               } ); # sposób II
```

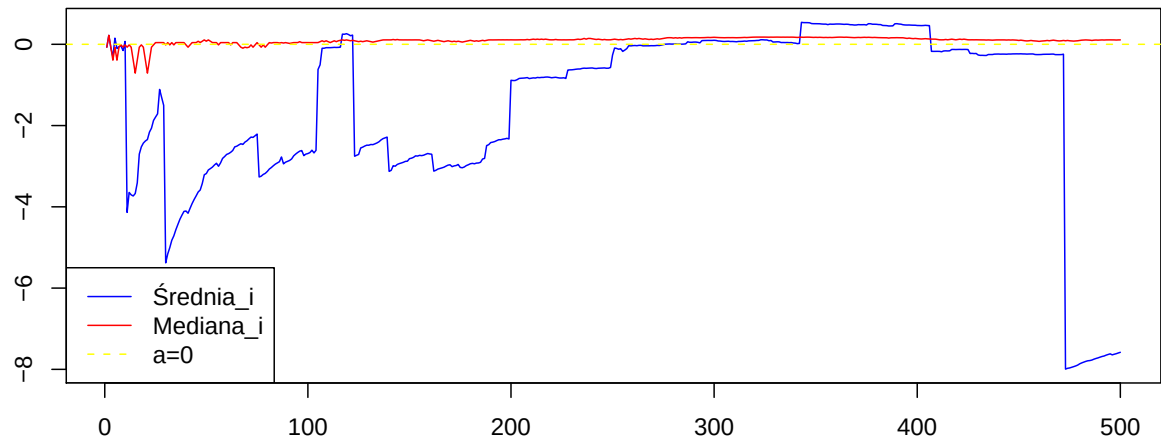
Wyniki pomiarów wynoszą na naszym komputerze odpowiednio (kolumna `user`) [s]:

```
# sposób I:
   user  system elapsed
  5.045   0.007   5.081
# sposób II:
   user  system elapsed
  0.232   0.013   0.248
```

Wnioski pozostawiamy Czytelnikowi.

Rozwiązanie sformułowanego problemu można więc przedstawić w sposób następujący.

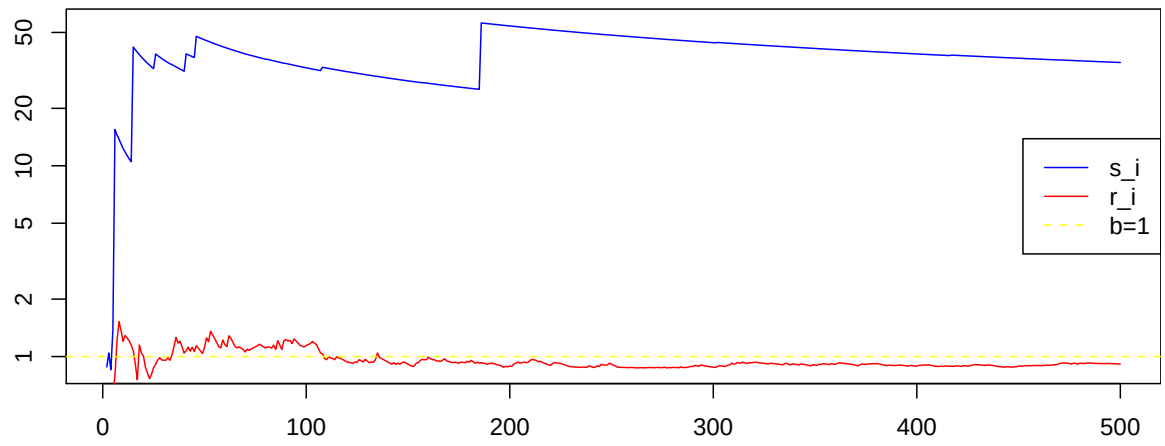
```
> n <- 500;
> X <- rcauchy(n);  # n-elementowa próba ze standaryzowanego rozkładu Cauchy'ego
> mn <- numeric(n); # tu będziemy przechowywać średnie z podpróbek
> md <- numeric(n); # tu będziemy przechowywać mediany z podpróbek
> for (i in 1:n)
  {
     mn[i] <- mean(X[1:i])
     md[i] <- median(X[1:i])
  }
> plot (1:i, mn, type="l", col="blue", xlab="", ylab="", lty=1);
> lines(1:i, md, col="red", lty=1);
> abline(h=0, col="yellow", lty=2);
> # lines(c(1,i), c(0,0), col="yellow"); # można i tak
> legend("bottomleft", c("Średnia_i", "Mediana_i", "a=0"),
     col=c("blue", "red", "yellow"), lty=c(1,1,2));
```



Rzecz jasna, rozpatrywanie kolejnej próby losowej daje za każdym razem w wyniku całkiem inny wykres. Warto więc samemu przyjrzeć się kilku różnym przykładom.

Porównanie estymatorów parametru skali $b = 1$ w rozkładzie standaryzowanym Cauchy'ego.

```
> n <- 500;
> X <- rcauchy(n);
> s <- numeric(n-1);
> r <- numeric(n-1);
> for (i in 2:n)
  {
     s[i-1] <- sd(X[1:i]);
     r[i-1] <- IQR(X[1:i])/2;
  }
> plot (2:i, s, type="l", col="blue", xlab="", ylab="", log="y");
> lines(2:i, r, col="red");
> abline(h=1, col="yellow", lty=2);
> legend("right", c("s_i", "r_i", "b=1"),
     col=c("blue", "red", "yellow"), lty=c(1,1,2));
```



Uwaga

Zwróćmy uwagę na zastosowanie skali logarytmicznej na osi pionowej powyższego wykresu.

⊡

Zadanie 4.3. Wygeneruj $m = 10000$ n-elementowych próbek ($n = 20$) z rozkładu jednostajnego na odcinku jednostkowym. Porównaj empirycznie obciążenie i błąd średniokwadratowy estymatora momentów i estymatora największej wiarogodności parametru θ w rozkładzie jednostajnym $U([0, \theta])$.

Rozwiązanie.
Niech $\mathbf{X} = (X_1, \ldots, X_n)$ będzie próbą z rozkładu jednostajnego $U([0, \theta])$. Można pokazać, że estymatory parametru θ mają postać:

– Estymator otrzymany metodą momentów (EMM):

$$\hat{\theta}_1 = 2\bar{\mathbf{X}}, \tag{2}$$

– Estymator największej wiarogodności (ENW):

$$\hat{\theta}_2 = X_{n:n}, \tag{3}$$

– Estymator nieobciążony o minimalnej wariancji (ENMW):

$$\hat{\theta}_3 = \frac{n+1}{n} X_{n:n} = \frac{n+1}{n} \hat{\theta}_2. \tag{4}$$

Obciążenia

$$b\left(\hat{\theta}\right) = \mathbb{E}\left(\hat{\theta} - \theta\right) = \mathbb{E}\left(\hat{\theta}\right) - \theta \tag{5}$$

oraz błędy średniokwadratowe

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathbb{E}\left(\hat{\theta} - \theta\right)^2 = \mathrm{Var}\left(\hat{\theta}\right) + \left[b\left(\hat{\theta}\right)\right]^2 \tag{6}$$

tych estymatorów wynoszą:

i	$b\left(\hat{\theta}_i\right)$	$\mathrm{MSE}\left(\hat{\theta}_i\right)$
1	0	$\dfrac{\theta^2}{3n}$
2	$-\dfrac{\theta}{n+1}$	$\dfrac{2\theta^2}{(n+1)(n+2)}$
3	0	$\dfrac{\theta^2}{n(n+2)}$

Niech $\theta = 1$. Generujemy 10000 20-elementowych próbek z rozkładu $U([0, 1])$ i porównujemy estymatory parametru θ. Wyniki zapiszemy w macierzy o rozmiarze $m \times 3$, w której każdy wiersz będzie odpowiadał innemu estymatorowi.

```
> m <- 10000;
> n <- 20;
> theta <- 1

> wyniki <- matrix(nrow=3, ncol=m,          # tu będą przechowywane wyniki
   dimnames=list(c("emm", "enw", "enmw"))) # nadajemy nazwy wierszom macierzy

> for (k in 1:m)  # :-(
  {
     X <- runif(n, 0, theta); # w każdej iteracji pętli nowa próbka
     wyniki[1, k] <- 2*mean(X);
     wyniki[2, k] <- max(X);
     wyniki[3, k] <- max(X)*(n+1)/n;
  }
```

Rozwiązanie tego problemu za pomocą pętli `for` nie jest najszczęśliwsze. Zobaczmy, że przebiega ono według schematu:

```
k razy
{
   wykonaj eksperyment losowy
   zapisz wynik
}
```

Uwaga

Wydajna implementacja powyższej metody może być stworzona za pomocą funkcji `replicate()`. Służy ona do wielokrotnego przeprowadzania pewnego eksperymentu losowego i zapisywania wyników w wektorze bądź macierzy wyjściowej. Składnia:

Funkcja replicate()

```
replicate(IleRazy,
{
   ... różne operacje, np. losowanie próby, działania arytmetyczne itp. ...
   wyznacz wynik eksperymentu jako wektor (też: pojedyncza liczba)
}
```

Lepiej jest więc zastosować konstrukcję:

```
> wyniki <- replicate(m,
  {
     X <- runif(n, 0, theta);
     c(2*mean(X), max(X), max(X)*(n+1)/n); # wynik pojedynczego eksperymentu
  });
> wyniki[,1:5] # zobaczmy wyniki 5 pierwszych eksperymentów

          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.9433351 0.9599519 0.8277112 0.8667324 1.1048221
[2,] 0.9692459 0.8932497 0.9134830 0.8910931 0.9813953
[3,] 1.0177082 0.9379122 0.9591571 0.9356477 1.0304651
```

Dzięki temu rozwiązanie liczy się nieco szybciej (dla $m = 100000, n = 20$ otrzymaliśmy czas 3,5 miast 4,3 s). Ponadto, co chyba ważniejsze, kod jest bardziej zwięzły i zrozumiały.

Szacujemy obciążenia:

```
> mean(wyniki[1, ])  - theta

[1] 0.000892507


> mean(wyniki[2, ])  - theta

[1] -0.0482485


> mean(wyniki[3, ]) - theta

[1] -0.0006609281
```

Szacujemy błędy średniokwadratowe:

```
> var(wyniki[1, ]) + (mean(wyniki[1, ]) - theta)^2
```

```
[1] 0.01640478
```

```
> var(wyniki[2, ]) + (mean(wyniki[2, ]) - theta)^2
```

```
[1] 0.004486981
```

```
> var(wyniki[3, ]) + (mean(wyniki[3, ]) - theta)^2
```

```
[1] 0.002380804
```

Teoretyczne wartości obciążeń i błędów średniokwadratowych:

- Obciążenie $b\left(\hat{\theta}_2\right)$:

  ```
  > -theta/(n+1)
  [1] -0.04761905
  ```

- MSE $\left(\hat{\theta}_1\right)$:

  ```
  > (theta^2)/(3*n)
  [1] 0.01666667
  ```

- MSE $\left(\hat{\theta}_2\right)$:

  ```
  > 2*(theta^2)/(n+1)/(n+2)
  [1] 0.004329004
  ```

- MSE $\left(\hat{\theta}_3\right)$:

  ```
  > (theta^2)/n/(n+2)
  [1] 0.002272727
  ```

□

Zadanie 4.4. Wygeneruj $m = 50$ n-elementowych próbek ($n = 10$) z rozkładu normalnego $\mathrm{N}(1, 2)$. Przedstaw na jednym wykresie przedziały ufności dla wartości oczekiwanej μ na poziomie ufności 0,95. Ile z nich powinno zawierać wartość $\mu = 1$?

Rozwiązanie.
Dla próby: $\mathbf{X} = (X_1, \ldots, X_n)$ z rozkładu normalnego $N(\mu, \sigma)$, o nieznanych parametrach μ, σ, przedział ufności dla μ na poziomie ufności $1 - \alpha$ ma postać

$$\left(\bar{\mathbf{X}} - t_{1-\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}, \bar{\mathbf{X}} + t_{1-\alpha/2}^{[n-1]} \frac{s}{\sqrt{n}}\right), \tag{7}$$

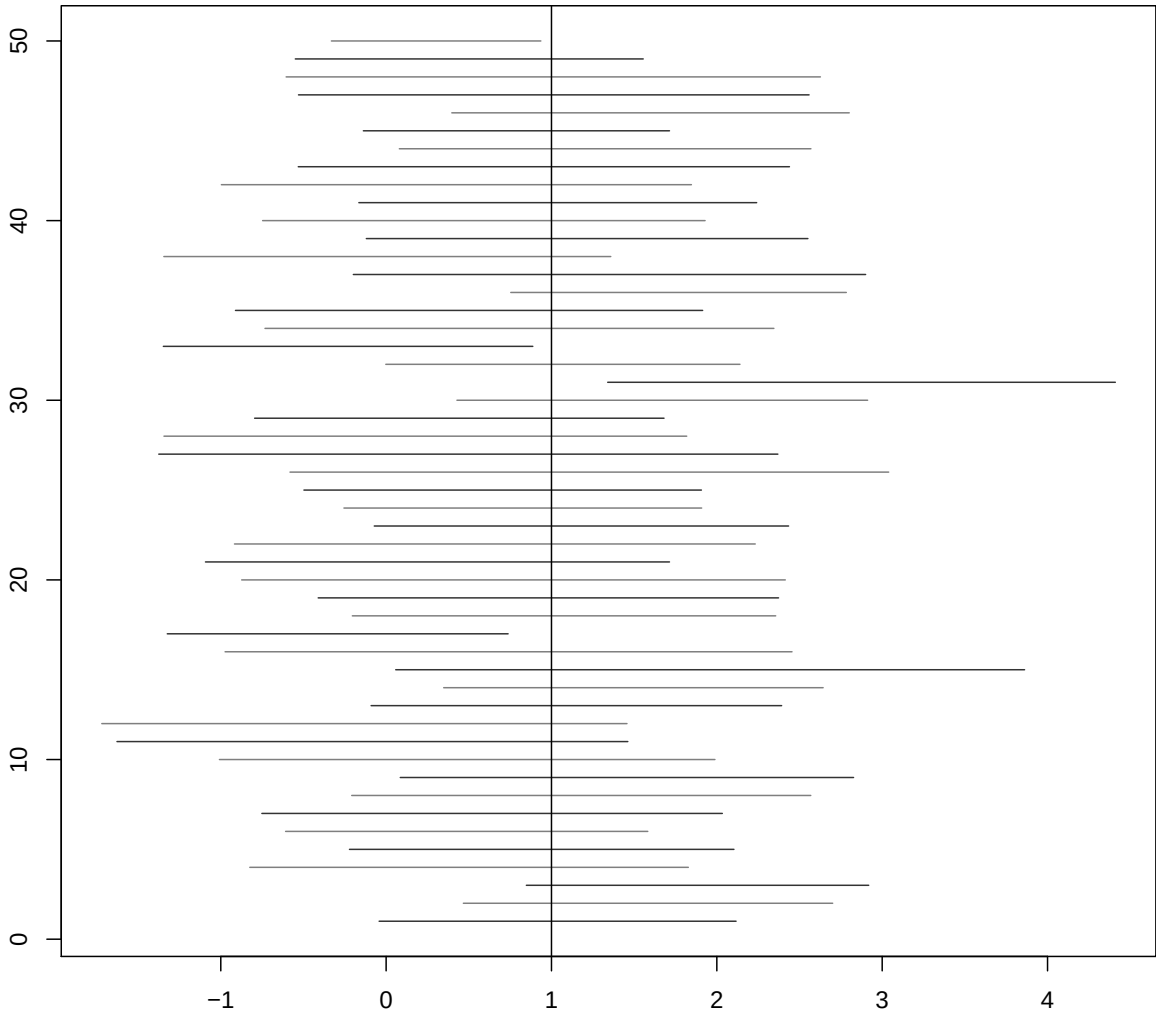
gdzie $t_{1-\alpha/2}^{[n-1]}$ oznacza kwantyl rzędu $1-\alpha/2$ rozkładu t-Studenta z $n-1$ stopniami swobody.

Generujemy m próbek o liczności n z rozkładu normalnego $\mathrm{N}(1, 2)$ i tworzymy dla nich wektor `mn` zawierający średnie $\bar{X}$ oraz wektor `d` zawierający wartości $s/\sqrt{n}$:

```
> m  <- 50;
> n  <- 10;
> mi <- 1;
> sigma <- 2;
> wyniki <- replicate(m,
  {
     X <- rnorm(n, mi, sigma);
     c(mean(X), sd(X)/sqrt(n));
  });
> mn <- wyniki[1,];
> d  <- wyniki[2,];
```

Teraz konstruujemy przedziały ufności i zaznaczamy je na jednym wykresie:

```
> alfa <- 0.05;
> q <- qt(1-alfa/2,n-1);
> matplot(rbind(mn-q*d,mn+q*d), rbind(1:m,1:m), type="l", lty=1,
     col=c("gray20", "gray50")); # wydaj polecenie: ?matplot
> abline(v=mi);
```



⊡

Zadanie 4.5. Wygeneruj $m = 10000$ próbek n-elementowych ($n = 10$) z rozkładu normalnego. Następnie zakładając, iż o próbkach wiemy tylko tyle, że pochodzą one z rozkładu normalnego o nieznanych parametrach, wyznacz dla każdej próbki przedział ufności dla wartości oczekiwanej na poziomie ufności 0,95. Porównaj frakcję pokryć przez przedział ufności faktycznej wartości oczekiwanej z założonym poziomem ufności.

Rozwiązanie.

```
> m <- 10000;
> n <- 10;
> alpha <- 0.05;
> q <- qt(0.975,9)
```

Generujemy próbki i „sprawdzamy" poziom ufności:

```
> ileWpada <- replicate(m,
  {
     X  <- rnorm(n);
     mn <- mean(X);
     s  <- sd(X);
     (mn-q*s/sqrt(n) < 0) & (mn+q*s/sqrt(n) > 0) # to jest wynik eksperymentu -
                          # albo TRUE (mi wpada do przedziału ufności) albo FALSE
  });
> ileWpada[1:10] # pierwsze 10 wyników

 [1]  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

Sprawdźmy frakcję pokryć liczby $\mu = 0$ przez wygenerowane przedziały ufności:

```
> sum(ileWpada)/m

[1] 0.9516
```

□

Zadanie 4.6. Średnia cena 50 losowo wybranych podręczników akademickich wyniosła 28,40 zł. Wiadomo, że odchylenie standardowe cen podręczników wynosi 4,75 zł. Wyznacz 95% przedział ufności dla średniej ceny podręcznika akademickiego zakładając, że rozkład cen jest rozkładem normalnym.

Rozwiązanie.
Dla próby: $\mathbf{X} = (X_1, \ldots, X_n)$ z rozkładu normalnego $\mathrm{N}(\mu, \sigma)$, o znanym parametrze σ, przedział ufności dla μ na poziomie ufności $1 - \alpha$ ma postać:

$$\left(\bar{\mathbf{X}} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{\mathbf{X}} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \tag{8}$$

gdzie $z_{1-\alpha/2}$ oznacza kwantyl rzędu $1 - \alpha/2$ rozkładu $\mathrm{N}(0, 1)$.

W naszym zadaniu: $\sigma = 4.75$, $n = 50$, $\bar{\mathbf{X}} = 28{,}4$, $1-\alpha = 0{,}95$. Zatem krańce szukanego przedziału ufności wynoszą:

```
> 28.4-qnorm(0.975)*4.75/sqrt(50)
```

```
[1] 27.08339
```

```
> 28.4+qnorm(0.975)*4.75/sqrt(50)
```

```
[1] 29.71661
```

□

Zadanie 4.7. Przeprowadzono 18 niezależnych pomiarów temperatury topnienia ołowiu i otrzymano następujące wyniki (w stopniach Celsjusza):

$$330.0, 322.0, 345.0, 328.6, 331.0, 342.0,$$
$$342.4, 340.4, 329.7, 334.0, 326.5, 325.8,$$
$$337.5, 327.3, 322.6, 341.0, 340.0, 333.0.$$

Zakładamy, że temperatura topnienia ołowiu ma rozkład normalny. Wyznacz dwustronny przedział ufności dla wartości oczekiwanej i odchylenia standardowego temperatury topnienia ołowiu na poziomie ufności 0,95.

Rozwiązanie.
Nasza próba $(X_1, \ldots, X_n)$ pochodzi z rozkładu normalnego $\mathrm{N}(\mu, \sigma)$ o nieznanych parametrach μ i σ, zatem do wyznaczenia przedziału ufności dla wartości oczekwianej μ można użyć funkcji `t.test`.

```
> x <- c(330.0, 322.0, 345.0, 328.6, 331.0, 342.0,
  342.4, 340.4, 329.7, 334.0, 326.5, 325.8,
  337.5, 327.3, 322.6, 341.0, 340.0, 333.0);
> mean(x);
```

```
[1] 333.2667
```

```
> t.test(x, conf.level=0.95)$conf.int
```

```
[1] 329.6482 336.8851
attr(,"conf.level")
[1] 0.95
```

Uwaga

Szczegóły dotyczące funkcji `t.test` poznamy w części dotyczącej weryfikacji hipotez.

Porównajmy wynik z wartością wyznaczoną wg wzoru (7):

```
> mean(x)-qt(0.975, length(x)-1)*sd(x)/sqrt(length(x));
```

```
[1] 329.6482
```

```
> mean(x)+qt(0.975, length(x)-1)*sd(x)/sqrt(length(x));
```

```
[1] 336.8851
```

Przedział ufności dla odchylenia standardowego σ na poziomie ufności $1-\alpha$ liczymy ze wzoru:

$$\left(\sqrt{\frac{(n-1)\,s^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)\,s^2}{\chi^2_{\alpha/2,n-1}}}\right), \tag{9}$$

Zatem dolny kraniec szukanego przedziału ufności to:

```
> sqrt((length(x)-1)*var(x) / qchisq(0.975, (length(x)-1)))

[1] 5.460114
```

a górny to:

```
> sqrt((length(x)-1)*var(x) / qchisq(0.025, (length(x)-1)))

[1] 10.90836
```

⊡

Zadanie 4.8. Wygeneruj $m = 10$ $n = 100$-elementowych próbek z rozkładu dwupunktowego Bern(p). Przedstaw na jednym wykresie przedziały ufności dla parametru $p = 0{,}5$ na poziomie ufności 0,9. Ile z nich powinno zawierać wartość $p = 0{,}5$?

Rozwiązanie.
Dla próby o dużej liczności: $(X_1, \ldots, X_n)$ z rozkładu dwupunktowego Bern(p) o nieznanym parametrze p, przedział ufności dla p na poziomie ufności $1-\alpha$ ma postać

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right), \tag{10}$$

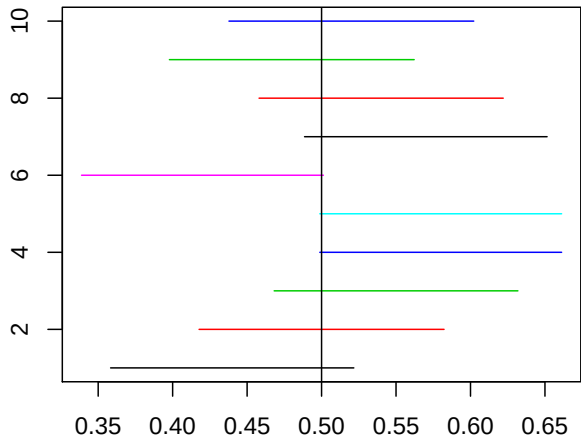
gdzie $z_{1-\alpha/2}$ oznacza kwantyl rzędu $1-\alpha/2$ rozkładu $\mathrm{N}(0,1)$.

Tworzymy wektor `pp`, zawierający wartości $\hat{p}$ wyliczone dla m próbek, oraz wektor `d`, zawierający wartości $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ wyliczone dla m próbek z rozkładu Bern(p) o licznościach n.

```
> m <- 10;
> n <- 100;
> p <- 0.5;
> alfa <- 0.1
> z <- qnorm(1-alfa/2)
> pp <- rbinom(m, n, p)/n
> d  <- sqrt(pp*(1-pp)/n)
```

Zaznaczamy otrzymane przedziały na jednym wykresie:

```
> matplot(rbind(pp-z*d,pp+z*d), rbind(1:m,1:m), type="l", lty=1)
> abline(v=p)
```



□

Zadanie 4.9. W sondażu przeprowadzonym przez magazyn „Time" („Time", 22 czerwca 1987) 578 spośród 1014 dorosłych respondentów stwierdziło, że dla dobra dzieci lepiej jest, gdy matka nie pracuje poza domem. Wyznacz 95% przedział ufności dla odsetka dorosłych podzielających tę opinię.

Rozwiązanie.
Dla próby o dużej liczności: $(X_1, \ldots, X_n)$ z rozkładu dwupunktowego Bern(p) o nieznanym parametrze p, przedział ufności dla p na poziomie ufności $1-\alpha$ ma postać:

$$\left(\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right), \tag{11}$$

gdzie $z_{1-\alpha/2}$ oznacza kwantyl rzędu $1-\alpha/2$ rozkładu $N(0,1)$.

Zastosujmy powyższy wzór do stosownych obliczeń:

```
> p <- 578/1014
> n <- 1014;
> p + c(-1,1)*qnorm(0.975)*sqrt(p*(1-p)/n)

[1] 0.5395479 0.6004915
```

Wyznaczmy przedział ufności dla p, używając funkcji `prop.test`:

```
> prop.test(578, 1014, conf.level=0.95)$conf.int

[1] 0.5388446 0.6006578
attr(,"conf.level")
[1] 0.95
```

Wyniki różnią się. Spowodowane jest to stosowaniem przez R-a korekty na ciągłość. Można ją oczywiście wyłączyć z obliczeń:

```
> prop.test(578, 1014, conf.level=0.95, correct=F)$conf.int
```

```
[1] 0.5393401 0.6001709
attr(,"conf.level")
[1] 0.95
```

Wygląda więc na to, że w naszym programie zaimplementowany jest estymator przedziału ufności dla proporcji niebazujący na przybliżeniu rozkładem normalnym wg tzw. centralnego twierdzenia granicznego. Rzeczywiście, stosowana jest tutaj poprawka zaproponowana przez Wilsona (1927). Zainteresowanego Czytelnika odsyłamy do literatury.
⊡

Zadanie 4.10. Na 12 oddanych niezależnie rzutów kostką otrzymano 3 „szóstki". Wyznacz 95% przedział ufności dla prawdopodobieństwa otrzymania „szóstki" w pojedynczym rzucie kostką.

Rozwiązanie.
Mamy próbę $(X_1, \ldots, X_n)$ z rozkładu dwupunktowego Bern (p), o nieznanym parametrze p i małej liczności. Nie możemy więc skorzystać z przybliżenia tego modelu rozkładem normalnym. Do wyznaczenia przedziału ufności dla p użyjemy funkcji `binom.test`:

```
> # prop.test(3, 12, conf.level=0.95)$conf.int  # ... ???
> binom.test(3, 12, conf.level=0.95)$conf.int

[1] 0.05486064 0.57185846
attr(,"conf.level")
[1] 0.95
```

⊡

3 Zadania do rozwiązania

Zadanie 4.11. Wygeneruj $n = 500$-elementową próbę $(Y_1, \ldots, Y_n)$ z rozkładu normalnego standardowego.

a) Dla każdej podpróbki zawierającej i początkowych elementów próbki wyjściowej, tzn. dla $\mathbf{X}_i = (Y_1, \ldots, Y_i)$, gdzie $i = 1, \ldots, n$, oblicz średnią $\bar{\mathbf{X}}_i$ oraz medianę Med_i Następnie przedstaw na wspólnym wykresie zbiory $\left\{\bar{\mathbf{X}}_i : i = 1, \ldots, n\right\}$ oraz $\{\mathrm{Med}_i : i = 1, \ldots, n\}$. Przeanalizuj wpływ liczności próby na zachowanie się średniej oraz mediany z próby. Czy statystyki te wydają się być sensownymi estymatorami parametru wartości oczekiwanej w tym modelu?

b) Dla każdej podpróbki zawierającej $i = 2, \ldots, n$ początkowych elementów próbki wyjściowej oblicz odchylenie standardowe s_i oraz $d_i = \mathrm{IQR}(\mathbf{X}_i)/1{,}35$ (czyli rozstęp międzykwartylowy podzielony przez 1,35). Następnie przedstaw na wspólnym wykresie zbiory $\{s_i : i = 2, \ldots, n\}$ oraz $\{d_i : i = 2, \ldots, n\}$. Przeanalizuj wpływ liczności próby na zachowanie się s_i oraz d_i. Czy statystyki te wydają się być sensownymi estymatorami odchylenia standardowego w tym modelu?

Zadanie 4.12. Na podstawie danych zawartych w pliku `samochody.csv` oszacuj przedziałowo średnie zużycie paliwa i odchylenie standardowe zużycia paliwa samochów o przyspieszeniu mniejszym niż 20 $\mathrm{m/s^2}$ (wykorzystaj zmienne `mpg` i `przysp`). Załóż, że badana cecha ma rozkład normalny. Przyjmij poziom ufności 0,95.

Zadanie 4.13. Biolog, badający pewien gatunek ryb, pobrał losową próbę 15 ryb i zmierzył ich długość. Otrzymał następujące wyniki (w mm):

$$92, 88, 85, 82, 89, 86, 81, 66, 75, 61, 78, 76, 91, 82, 82.$$

Zakładając, że rozkład długości ryb badanego gatunku jest normalny, oszacuj przedziałowo średnią długość ryb badanego gatunku na poziomie ufności 0,95.

Zadanie 4.14. Przeprowadzono sondaż opinii publicznej i okazało się, że 57% spośród 1000 ankietowanych Polaków uważa, że Polska skorzystała na wejściu do UE. Wyznacz 90% przedział ufności dla odsetka Polaków podzielających ten pogląd.

$\star$ **Zadanie 4.15.** Niech $(X_1, \ldots, X_n)$ oznacza ciąg obserwacji czasów poprawnego działania n urządzeń pracujących niezależnie. Zakładamy, że czas poprawnej pracy każdego urządzenia ma rozkład wykładniczy z nieznanym parametrem θ. Urządzenia te nie są obserwowane w sposób ciągły, lecz kontrola dokonywana jest w dyskretnych chwilach $1, 2, \ldots, k$. Stąd też, de facto, obserwujemy jedynie $Y_1, \ldots, Y_n$, gdzie

$$Y_j = \begin{cases} i & \text{gdy} \quad i-1 < X_j \leqslant i, \quad \text{dla pewnego } i = 1, \ldots, k, \\ k+1 & \text{gdy} \quad X_j > k, \end{cases}$$

przy czym $j = 1, \ldots, n$. Niech $N_i = \#\{j : Y_j = i\}$, $i = 1, \ldots, k+1$. Wyznacz estymator największej wiarogodności parametru θ. Dokonaj obliczeń dla przypadku, gdy $n = 10$, $k = 2$ oraz $N_1 = 5$, $N_2 = 2$ i $N_3 = 3$.

4 Wskazówki i odpowiedzi

Odpowiedź do zadania 4.12. $\mu \in (27{,}30534, 29{,}69748)$, $\sigma \in (6{,}45733, 8{,}16175)$.

Odpowiedź do zadania 4.13. $\mu \in (76{,}08535, 85{,}78131)$.

Odpowiedź do zadania 4.14. $p \in (0{,}54359, 0{,}59602)$.