Лабораторна робота №2
з непараметричної статистики

Горбунов Даніел Денисович
1 курс магістратури
група "Прикладна та теоретична статистика"
Варіант №4

22 лютого 2022 р.

Перша частина.

Вступ.

У даній роботі реалізована зміщена вибіркова процедура за відомою зміщуючою функцією $w(t) = (1 + \cos(t))/2$. На модельованих даних, отриманих за зміщеною та незміщеною процедурами, обчислені оцінки невідомої функції розподілу $F \sim Exp_{\lambda=1}$. Обсяги зміщеної та незміщеної вибірок рівні та дорівнюють $n = 300$. Зроблені висновки щодо якості використаних оцінок на модельованих даних.

Хід роботи.

Зміщена вибіркова процедура: реалізація в R.

Оскільки нам відома зміщуюча функція $w(t)$, то ми знаємо умовний розподіл індикатора потрапляння об'єкта O до вибірки з точністю до деякої сталої $c > 0$:

$$P(s(O) = 1 \mid \xi(O) = t) = c \cdot w(t)$$

Не втрачаючи загальності, будемо вважати, що $c = 1$. Насправді цього достатньо для того, що описати алгоритм моделювання зміщеної вибірки обсягу n за відомим розподілом $F(t)$ та зміщуючою функцією $w(t)$:

1. **Вхідні дані:** Обсяг вибірки n, функція розподілу $F(t)$, зміщуюча функція $w(t)$.
2. **Процедура:**
   (а) Створюємо порожній масив ξ^b.
   (б) **Цикл з умовою завершення:** $|\xi^b| = n$
      i. Моделюємо в.в. ξ з розподілу F.
      ii. Обчислюємо $w(\xi)$. Долучаємо ξ до ξ^b з ймовірністю $w(\xi)$.
3. **Результат:** Зміщена вибірка $\xi^b = (\xi^b_1, \ldots, \xi^b_n)$.

Далі наведена програмна реалізація вищенаведених викладок.

```
# Функцію, що генерує зміщену вибірку обсягу n з Exp(1) за
# відомою функцією зміщення w = w(t) та k > 0, де
# P( s(O) = 1 | xi(O) = t ) = k * w(t)
biased.sampling <- function(n, k, w)
{
  to.return <- c()
  while(length(to.return) != n)
  {
    d <- n - length(to.return)
    generated <- rexp(d, rate = lambda)
    p <- k * w(generated)
    m <- sapply(p, function(q) { sample(c(T,F), 1, prob = c(q, 1 - q)) })
    to.return <- c(to.return, generated[m])
  }
  to.return
}
```

Для простоти, у скрипті функція, що відповідає за незміщену процедуру, позначається як "unbiased.sampling". Простота тут лише в тому, щоб код було легше читати, ніякого іншого смислового навантаження тут немає.

Підготовчі відомості.

Як і завжди, емпірична функція розподілу має вигляд:

$$\hat{F}_{emp}(t) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}\{\xi_j < t\}$$

де $(\xi_1, \ldots, \xi_n)$ – кратна вибірка, отримана за незміщеною процедурою.

```
Femp <- function(x)
{
  f <- function(t) { sum(x < t) / length(x) }; f
}
```

Якщо припустити, що вона отримана за зміщеною процедурою, то використовувати "в лоб" $\hat{F}_{emp}(t)$ недоречно. Знаючи $w(t)$, краще скористатися оцінкою Горвіца-Томпсон:

$$\hat{F}_{HT}(t) = \frac{\sum_{j=1}^{n} \frac{\mathbb{1}\{\xi_j < t\}}{w(\xi_j)}}{\sum_{j=1}^{n} \frac{1}{w(\xi_j)}}$$

```
FHT <- function(x, w)
{
  w.x <- w(x)
  f <- function(t) { sum(as.numeric(x < t) / w.x) / sum(1 / w.x) }; f
}
```

Якщо маємо незміщену $\xi = (\xi_1, \ldots, \xi_{n_1})$ та зміщену $\eta = (\eta_1, \ldots, \eta_{n_2})$ незалежні вибірки, тоді можна розглянути опуклу комбінацію оцінок:

$$\hat{F}_{conv}(t) = \lambda \hat{F}_{emp}(t) + (1-\lambda)\hat{F}_{HT}(t),\ \xi \sim \hat{F}_{emp},\ \eta \sim \hat{F}_{HT}$$

де $\lambda \in [0,1]$ буде обиратися з припущення про мінімізацію $\mathbb{D}\left[\hat{F}_{conv}(t)\right]$. Теорія (крім того, програмна реалізація), пов'язана із визначенням такого оптимального λ, буде викладена далі. На завершення пропонується використати оцінку ЕМНВ, що приводить до оцінки Варді:

$$\hat{F}_V(t) = \sum_{j=1}^{n_1+n_2} p_j \mathbb{1}\{\zeta_j < t\}$$

де $\zeta = (\zeta_1, \ldots, \zeta_{n_1+n_2}) = (\xi_1, \ldots, \xi_{n_1}, \eta_1, \ldots, \eta_{n_2})$, а ваговий коефіцієнт p_j дорівнює

$$p_j = \frac{W}{\lambda + n_2 w(\zeta_j)},\ W = \left(\sum_{j=1}^{n_1+n_2} \frac{1}{\lambda + n_2 w(\zeta_j)} \right)^{-1},$$

де λ – додатній корінь рівняння

$$\sum_{j=1}^{n_1+n_2} \frac{w(\zeta_j)}{\lambda + n_2 w(\zeta_j)} = 1$$

Для кожної ситуації, побудуємо та порівняємо графіки теоретичної та емпіричної функцій розподілу у точках $t_j := Q^{Exp(1)}(0.01) \cdot \left(1 - \frac{j}{B}\right) + Q^{Exp(1)}(0.99) \cdot \frac{j}{B},\ j = \overline{0, B}$, де $B = 1000$. Зокрема у цих точках обчислимо абсолютні відхилення функцій розподілу, тобто величину:

$$\max_{0 \le j \le B} \left| \hat{F}(t_j) - F(t_j) \right|$$

```
library("nleqslv")
# Програмна реалізація суми, що використовується у рівнянні
fv.target <- function(lambda, wz, n2)
{
  sum(wz / (lambda + n2 * wz)) - 1
}

# Перші два аргументи - незміщена та зміщена вибірки,
# w - зміщуюча функція, x0 - початкове значення для знаходження кореня
# рівняння чисельними методами, plot.c - чи будувати графік функції,
# для якої знаходили додатний нуль.
FV <- function(unbiased, biased, w, x0 = 0, plot.c = F)
{
  n2 <- length(biased)
  z <- c(unbiased, biased)
  w.z <- w(z)
  to.solve <- function(l) { fv.target(l, w.z, n2) }
  solution <- nleqslv(x0, to.solve)
  lambda.opt <- solution$x
  ...
```

```
  ...
  if(plot.c)
  {
    t.val <- seq(0, (5/4 * lambda.opt), 0.01)
    plot(t.val, sapply(t.val, to.solve), type = 'l',
         xlab = "lambda", ylab = "sum")
    abline(h = 0, col = "black", lty = 2)
    abline(v = lambda.opt, col = "red", lty = 2)
    grid()
  }
  W <- 1 / (sum(1 / (lambda.opt + n2 * w.z)))
  p <- W / (lambda.opt + n2 * w.z)
  f <- function(t) { sum(p * (z < t)) }; f
}
```

Все необхідне для виконання наступних пунктів є, залишається змоделювати вибірки.

```
set.seed(0)

n <- 300
x.unbiased <- unbiased.sampling(n)
x.biased <- biased.sampling(n, 1, w)

q1 <- qexp(0.01, lambda)
q2 <- qexp(0.99, lambda)
B <- 1000
I <- q2 * (0:B)/B + q1 * (B:0)/B

Fx.emp <- Femp(x.unbiased)
Fx.HT <- FHT(x.biased, w)
Fx.V <- FV(x.unbiased, x.biased, w)
# Для опуклої комбінації - далі
```

Графічні результати. Класична оцінка та оцінки Горвіца-Томпсон, Варді.

Покажемо графіки емпіричної та теоретичної функцій розподілу.

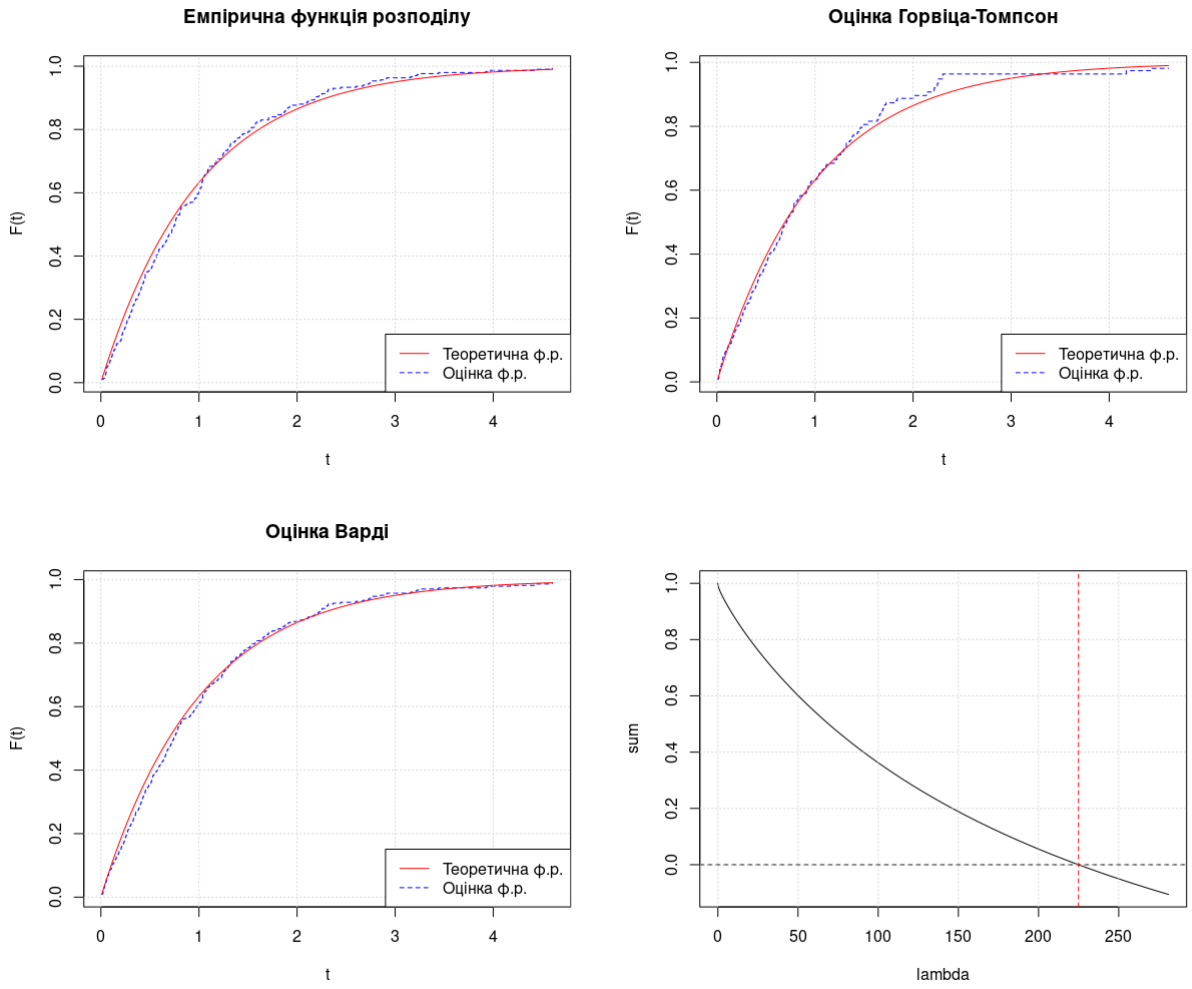

Рис. 1: Функції розподілу та графік $\varphi(\lambda) = \sum_{j=1}^{n_1+n_2} w(\zeta_j)/(\lambda + n_2 w(\zeta_j)) - 1$ (у правому нижньому куті). На графіку для $\varphi(\lambda)$ червоною лінією відмічено розв'язок рівняння $\varphi(\lambda) = 0$.

На око можна відмітити, що найменші відхилення від теоретичної функції розподілу спостерігаються саме для оцінки Варді. Можна висунути гіпотезу про те, що оцінка Варді є точнішою у порівнянні з іншими двома оцінками. З іншого боку, як на такій відносно невеликій вибірці, оцінка Горвіца-Томпсон спрацювала непогано.

```
[1] "Empirical: 0.0504254585849177"
[1] "Horvitz-Thompson: 0.0635525326877217"
[1] "Vardi: 0.041515579754284"
```

На змодельованих вибірках, найбільше відхилення на побудованому розбитті видно для оцінки Горвіца-Томпсон, а найменше - для оцінки Варді.

Опукла комбінація оцінок: Теорія.

Продублюємо задачу: нехай задані незміщена $\xi = (\xi_1, \ldots, \xi_{n_1})$ та зміщена $\eta = (\eta_1, \ldots, \eta_{n_2})$ незалежні вибірки, тоді можна розглянути опуклу комбінацію оцінок:

$$\hat{F}_{conv}(t) = \lambda \hat{F}_{emp}(t) + (1 - \lambda)\hat{F}_{HT}(t),\ \xi \sim \hat{F}_{emp},\ \eta \sim \hat{F}_{HT}$$

де $\lambda \in [0, 1]$ потрібно обирати з припущення про мінімізацію $\mathbb{D}\left[\hat{F}_{conv}(t)\right]$. Питання такого плану: як мінімізувати? Спочатку зробимо просту арифметику в силу незалежності двох вибірок:

$$\mathbb{D}\left[\hat{F}_{conv}(t)\right] = \lambda^2 \mathbb{D}\left[\hat{F}_{emp}(t)\right] + (1 - \lambda)^2 \mathbb{D}\left[\hat{F}_{HT}(t)\right]$$

Обирати λ будемо з міркувань зменшення асимптотичної дисперсії. ... *А далі – хто знає. Я ще не придумав.*

Опукла комбінація оцінок: Практика.

Розглянемо $\lambda = 1/2$. Тоді маємо такий результат:

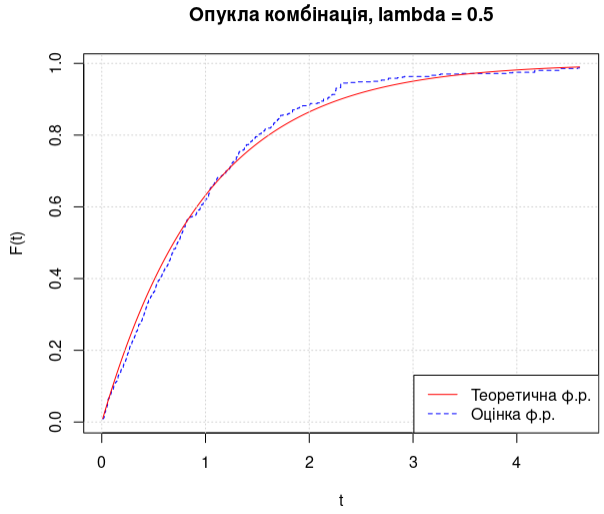

Рис. 2: Графік теоретичної функції розподілу та її оцінки.

```
[1] "Convex: 0.0437196128788422"
```

За однією вибіркою, звісно, мало чого можна сказати, однак видно, що при хорошому виборі вагового коефіцієнта в опуклій комбінації, то можна досягти непоганих результатів.

Висновки.

Оцінка емпіричної найбільшої вірогідності вийшла хорошою. Звісно, коли є дві вибірки, то це чудово (бо є можливість побудувати більш точну оцінку), але на практиці вже як вийде.

Друга частина.

Вступ.

У даній роботі описана програмна реалізація обчислення оцінок функціоналів для заданої функції розподілу $F(t)$, що взята з першої частини роботи (тобто $F \sim Exp_{\lambda=1}$). Якість оцінок перевірена з використанням техніки імітаційного моделювання. Додатково обчислені L_p-відстані (при $p = 1, 2, \infty$) між теоретичною функцією розподілу та її оцінкою.

Хід роботи.

Надалі

$$\hat{F}(t) = \sum_{j=1}^{n} p_j \mathbb{1}\{\xi_j < t\}$$

довільна оцінка функції розподілу $F(t)$ (емпірична, Горвіца-Томпсон, Варді тощо) за деякою вибіркою $(\xi_1, \ldots, \xi_n)$ з незалежних спостережень.

Математичне сподівання.

Для експоненційного розподілу відомо, що математичне сподівання дорівнює $\frac{1}{\lambda} \mid_{\lambda=1} = 1$. У формулу для математичного сподівання замість $F(t)$ підставимо $\hat{F}(t)$

$$\int_{-\infty}^{+\infty} t d\hat{F}(t) = \sum_{j=1}^{n} \xi_j p_j$$

Аналогічно, взагалі кажучи, можна спробувати оцінити функціональний момент для борелевої функції $h : \mathbb{R} \to \mathbb{R}$:

$$\int_{-\infty}^{+\infty} h(t) d\hat{F}(t) = \sum_{j=1}^{n} h(\xi_j) p_j$$

Доречно буде відсортувати спостережувані значення за зростанням. Остання оцінка є узагальненням для обчислення теоретичних моментів, тому варто саме її реалізувати.

```
# В аргументах функції: x - вибірка, Fx - функція, що реалізує оцінку ф.р. за x,
# h - деяка функція, за якою обчислюється теоретичний момент
expected.value <- function(x, Fx, h = function(t) {t})
{
  sx <- sort(x) # Відсортовуємо значення
  p <- diff(Fx(c(sx, Inf))) # Обчислюємо стрибки емпіричної функції розподілу
  sum(p * h(sx)) # Рахуємо функціональний момент
}
```

Така реалізація годиться у рамках цієї роботи. Покажемо приклад застосування функції.

```
# Математичне сподівання
m.estim.emp <- expected.value(x.unbiased, Fx.emp)
print(paste("Оцінка мат. сподівання за Femp:", m.estim.emp))
m.estim.HT <- expected.value(x.biased, Fx.HT)
print(paste("Оцінка мат. сподівання за FHT:", m.estim.HT))
m.estim.conv <- 0.5 * (m.estim.emp + m.estim.HT)
print(paste("Оцінка мат. сподівання за Fconv:", m.estim.conv))
m.estim.V <- expected.value(c(x.unbiased, x.biased), Fx.V)
print(paste("Оцінка мат. сподівання за FV:", m.estim.V))
```

Результат виконання цього блоку наступний:

```
[1] "Оцінка мат. сподівання за Femp: 0.990856053075742"
[1] "Оцінка мат. сподівання за FHT: 0.97659961925472"
[1] "Оцінка мат. сподівання за Fconv: 0.983727836165231"
[1] "Оцінка мат. сподівання за FV: 1.01742564623483"
```

Отримані значення оцінок близькі до справжнього значення математичного сподівання, тому можна очікувати, що вони оцінюють щось адекватне. За допомогою імітаційного моделювання дослідимо асимптотичний розподіл оцінок, зокрема наближено підраховуючи на кожному етапі зміщення та коефіцієнт розсіювання. Детальніше, згенеруємо по $B = 1000$ повторних вибірок $\zeta^{(b)}$, $b = \overline{1, B}$ обсягу n, підрахувавши на кожній оцінку математичного сподівання $\hat{\theta}^{(b)} = \hat{\theta}(\zeta^{(b)})$. Маючи вибірку зі значень $\vec{\theta}^{(B)} = (\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$, можна оцінити

1. **Зміщення:** $\mathrm{bias}(\hat{\theta}) \approx \sqrt{n}\left(\overline{\vec{\theta}^{(B)}} - \theta\right)$;
2. **Асимптотичну дисперсію:** $\mathrm{Var}_\infty(\hat{\theta}) \approx nS_0^2(\vec{\theta}^{(B)})$.

Нормуючий множник обрано з тих міркувань, як обирався множник для асимптотичної нормальності оцінок функції розподілу. Зауважимо, що таку процедуру ми будемо використовувати і для наступних оцінок функціоналів, тому далі дублювати цю схему не будемо. Покажемо програмну реалізацію попередніх викладок:

```
# Повертає таблицю зі значень оцінок на повторних вибірках для вказаних оцінок
imit.model.expectation <- function(n, k, w, h = function(t) {t}, B = 1000)
{
  exp.estim.tabl <- data.frame(empirical = double(),
                               horvitz.thompson = double(),
                               convex = double(),
                               vardi = double())
  for(b in 1:B)
  {
    x.repl.0 <- unbiased.sampling(n)
    x.repl.b <- biased.sampling(n, k, w)

    Fx.repl.emp <- Femp(x.repl.0)
    Fx.repl.HT <- FHT(x.repl.b, w)
    Fx.repl.V <- FV(x.repl.0, x.repl.b, w)
    ...
```

```
    ...
    m.repl.emp <- expected.value(x.repl.0, Fx.repl.emp)
    m.repl.HT <- expected.value(x.repl.b, Fx.repl.HT)
    m.repl.conv <- 0.5 * (m.repl.emp + m.repl.HT)
    m.repl.V <- expected.value(c(x.repl.0, x.repl.b), Fx.repl.V)

    exp.estim.tabl[nrow(exp.estim.tabl) + 1,] <- c(
      m.repl.emp, m.repl.HT, m.repl.conv, m.repl.V
    )
  }
  exp.estim.tabl
}
```

Код досить брудний, але робочий. Щоб довести це, покажемо результати наближеного оцінювання при значеннях вибірки $n = 100, 250, 500, 1000$.

```
N <- 100 # 250, 500, 1000
Ex.estim <- imit.model.expectation(N, 1, w)
Ex.estim.norm <- apply(Ex.estim, 2, function(v) { sqrt(n) * (v - Ex) })

bias.est <- apply(Ex.estim.norm, 2, mean)
var.est <- apply(Ex.estim.norm, 2, var)
for(j in 1:4)
{
  hist(Ex.estim.norm[,j],
       prob = T, xlab = "value",
       main = paste("Method:", colnames(Ex.estim.norm)[j], "for", "n:", n))
  curve(dnorm(x, bias.est[j], sqrt(var.est[j])), col = "red", add = T)
  grid()
}
```

Маємо такі наближені значення для нормованої оцінки математичного сподівання:

n	**empirical**	**horvitz.thompson**	**convex**	**vardi**
100	-0.04444327	-0.5049203	-0.2746818	-0.04363694
250	0.04184152	-0.5303153	-0.2442369	0.01091464
500	0.02688092	-0.4341112	-0.2036151	0.03486291
1000	0.01220264	-0.6452307	-0.316514	-0.002758411
n	**empirical**	**horvitz.thompson**	**convex**	**vardi**
100	0.98194876	5.0362879	1.4321515	0.65347910
250	0.98347273	8.0114344	2.2581525	0.68727182
500	1.02649025	13.2315687	3.5170948	0.68657671
1000	1.00576843	18.2227391	4.879119	0.719483998

Табл. 1: Оцінки зміщення та асимптотичної дисперсії.

Подивимося на гістограми. Наступні декілька сторінок присвячені саме гістограмам в залежності від обраної оцінки функції розподілу.

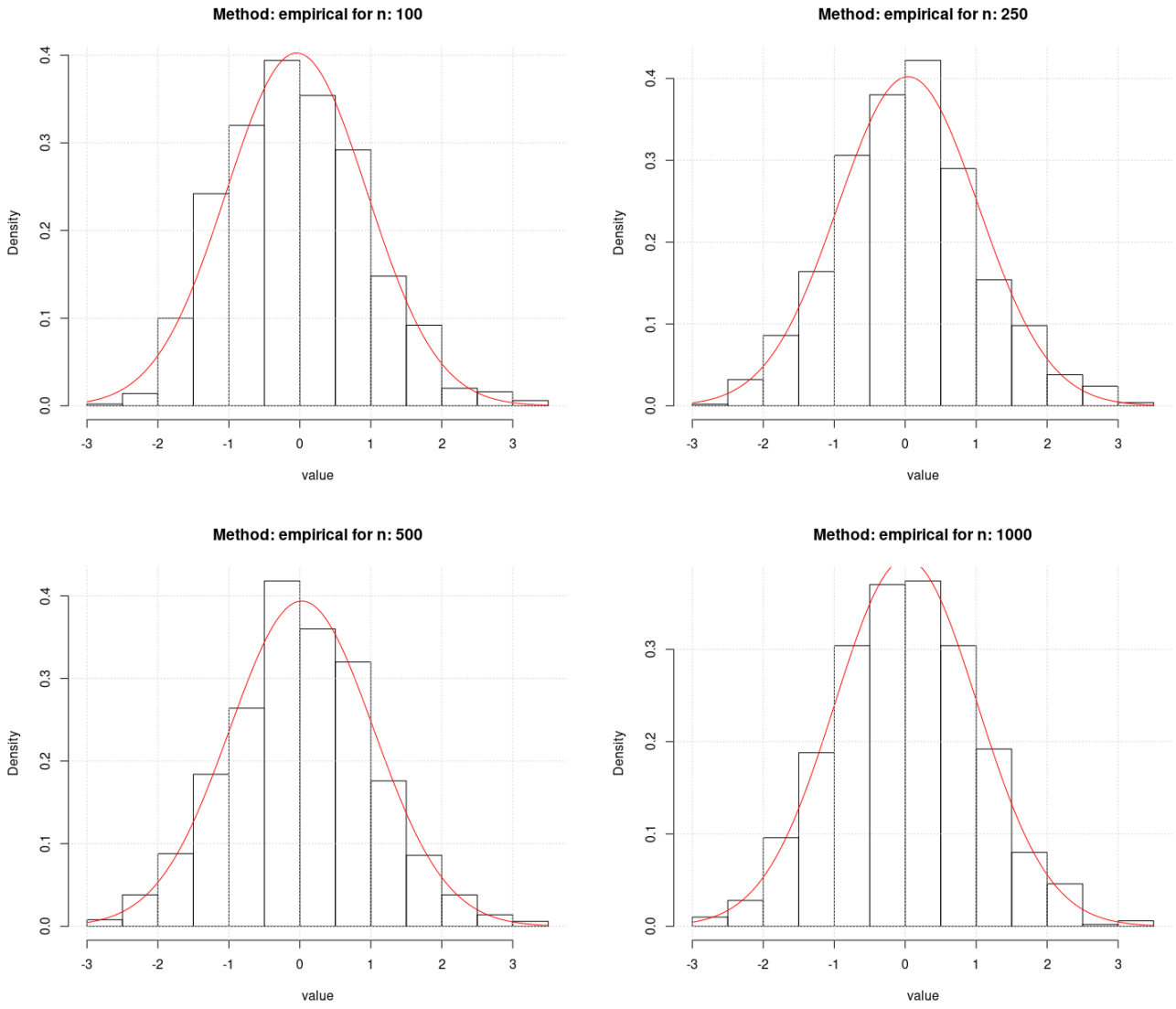

Рис. 3: Гістограми нормованої оцінки математичного сподівання за емпіричною функцією розподілу.

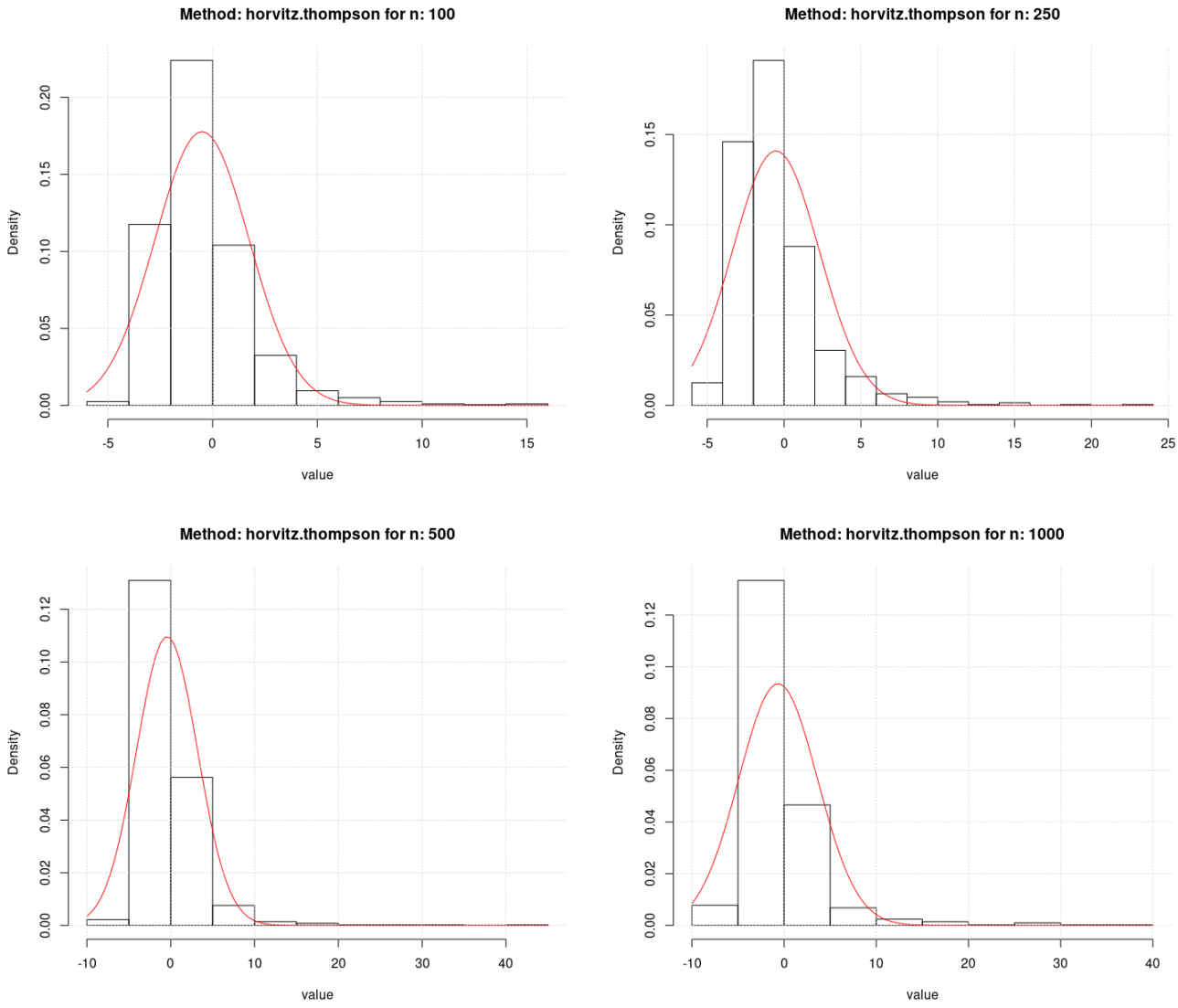

Рис. 4: Гістограми нормованої оцінки математичного сподівання за оцінкою Горвіца-Томпсон.

Дисперсія оцінки "летить" із збільшенням обсягу вибірки у стратосферу. Можливо, це можна пояснити тим, що дисперсія оцінки функції розподілу туди ж і летить.

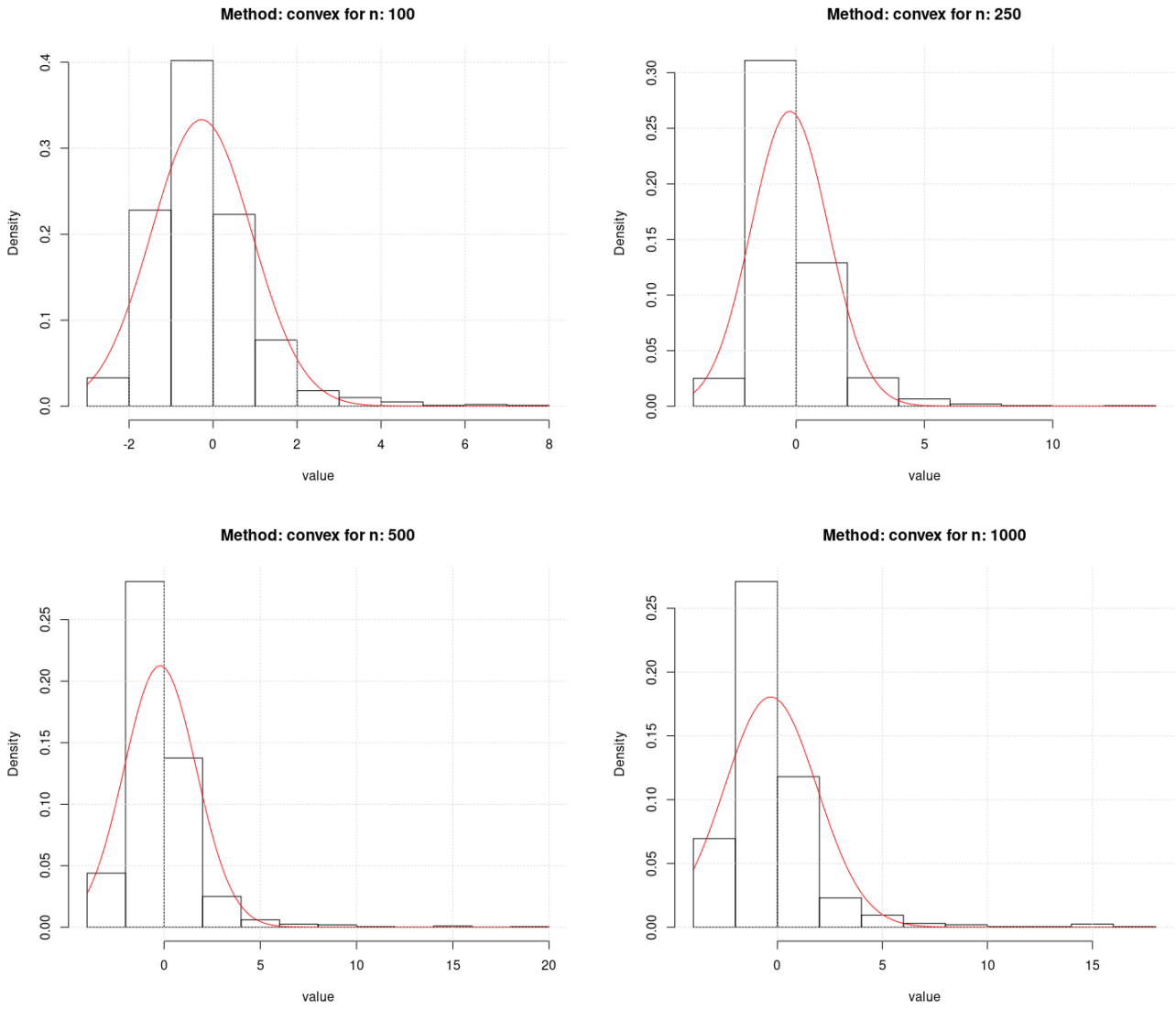

Рис. 5: Гістограми нормованої оцінки математичного сподівання за опуклою комбінацією оцінок.

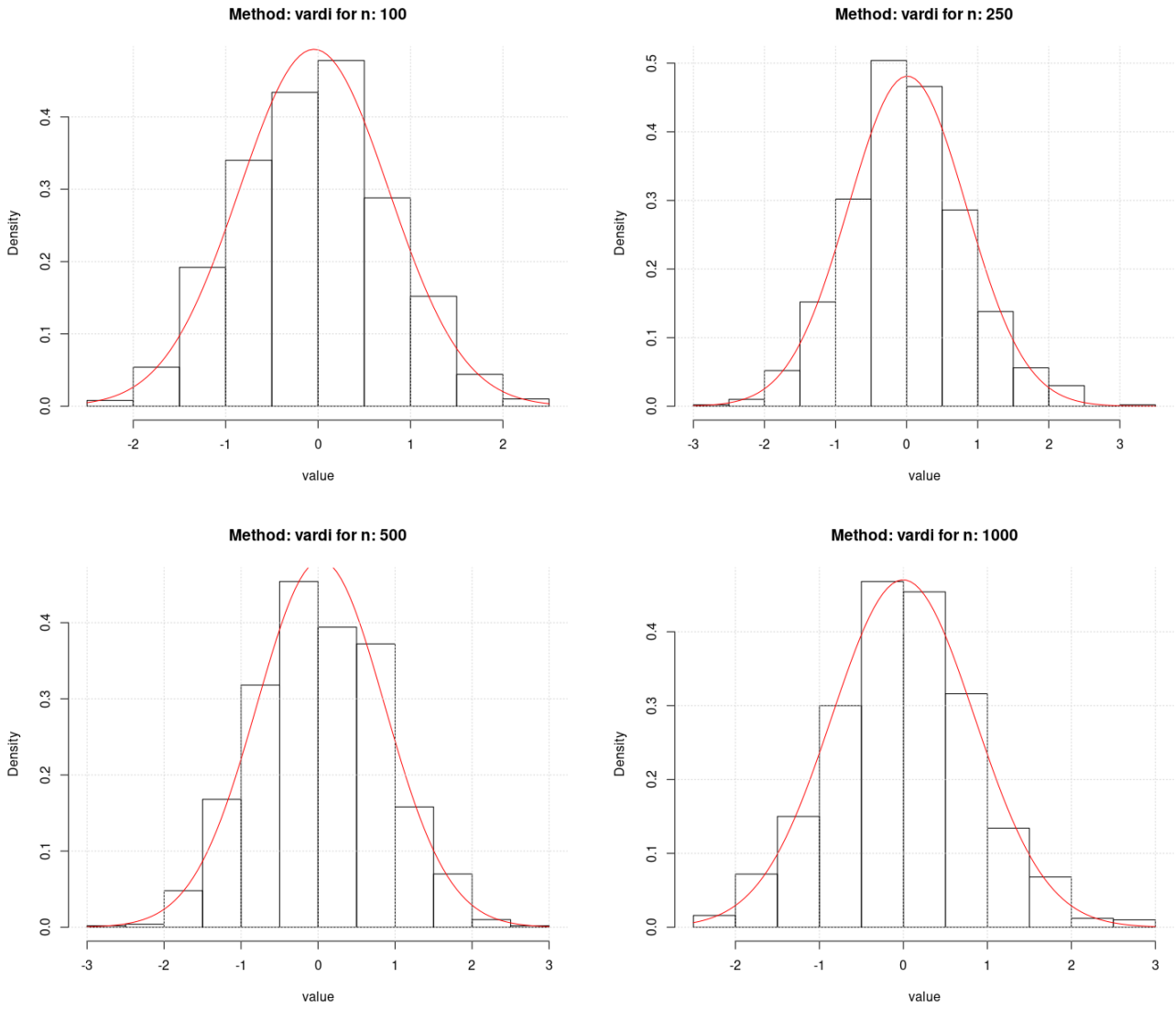

Рис. 6: Гістограми нормованої оцінки математичного сподівання за оцінкою Варді.

Загалом адекватна картина спостерігається у граничному розподілі оцінок математичного сподівання за емпіричною функцією розподілу та оцінкою Варді.

Рівномірна метрика (L_∞).

Опишемо алгоритм обчислення L_∞-метрики за неперервною теоретичною функцією розподілу $F(t)$ та її оцінкою $\hat{F}(t)$ за вибіркою $\{\zeta_1, \ldots, \zeta_n\}$:

1. Розглянути варіаційний ряд $\zeta_{(1)} \leq \ldots \leq \zeta_{(n)}$, обчислити $f_j = F(\zeta_{(j)})$ та $\hat{f}_j = \hat{F}(\zeta_{(j)})$.
2. Знайти $f^- = \max\limits_{1 \leq j \leq n} |f_j - \hat{f}_j|$, $f^+ = \max\limits_{1 \leq j \leq n} |f_j - \hat{f}_{j+1}|$, де $\hat{f}_{n+1} := \hat{F}(+\infty) = 1$.
3. Знайти $\max\{f^-, f^+\}$. Отримана величина буде шуканою відстанню.

Програмна реалізація алгоритму:

```
sup.dist <- function(sorted.x, Fest, Ftheor)
{
  fj <- Ftheor(sorted.x)
  hat.fj <- Fest(sorted.x)
  fplus <- max(abs(fj - hat.fj))
  fminus <- max(abs(fj - c(hat.fj[-1], 1)))
  max(fminus, fplus)
}
```

на попередньо згенерованих даних з нульовою зерниною маємо такі показники метрики:

```
> print(sup.dist(x.unbiased, Fx.emp, F.theor)) # Емпірична оцінка
[1] 0.05080491
> print(sup.dist(x.biased, Fx.HT, F.theor)) # Оцінка Горвіца-Томпсон
[1] 0.05384684
> print(sup.dist(c(x.unbiased, x.biased), Fx.conv, F.theor)) # Опукла комбінація
[1] 0.04454717
> print(sup.dist(c(x.unbiased, x.biased), Fx.V, F.theor)) # Оцінка Варді
[1] 0.0423566
```

Обчислимо середні значення відстаней, згенерувавши по $B = 1000$ повторних вибірок:

```
imit.model.distance <- function(n, k, w, B = 1000)
{
  dist.estim.tabl <- data.frame(empirical = double(),
                                horvitz.thompson = double(),
                                convex = double(),
                                vardi = double())
  for(b in 1:B)
  {
    x.repl.0 <- unbiased.sampling(n)
    x.repl.b <- biased.sampling(n, k, w)

    Fx.repl.emp <- Femp(x.repl.0)
    Fx.repl.HT <- FHT(x.repl.b, w)
    Fx.repl.conv <- function(t) { 0.5 * (Fx.repl.emp(t) + Fx.repl.HT(t)) }
    Fx.repl.V <- FV(x.repl.0, x.repl.b, w)
    ...
```

```
    ...
    d.emp <- sup.dist(x.repl.0, Fx.repl.emp, F.theor)
    d.HT <- sup.dist(x.repl.b, Fx.repl.HT, F.theor)
    d.conv <- sup.dist(c(x.repl.0, x.repl.b), Fx.repl.conv, F.theor)
    d.V <- sup.dist(c(x.repl.0, x.repl.b), Fx.repl.V, F.theor)

    dist.estim.tabl[nrow(dist.estim.tabl) + 1,] <- c(d.emp, d.HT, d.conv, d.V)
  }
  dist.estim.tabl
}

N <- 100
dist.estim.tabl.N <- imit.model.distance(N, 1, w, B = 1000)
print(apply(dist.estim.tabl.N, 2, mean))
```

Програму можна ще оптимізувати, однак для цієї роботи достатньо і такої реалізації. Маємо такі результати при $n = 100, 250, 500, 1000$:

n	**empirical**	**horvitz.thompson**	**convex**	**vardi**
100	0.08060948	0.11358188	0.07167572	0.06097499
250	0.05250551	0.08533162	0.05138241	0.03885491
500	0.03771669	0.06619005	0.03885326	0.02834517
1000	0.02690911	0.05616298	0.03194188	0.02014370

Табл. 2: Середні значення відстаней в залежності від обсягу вибірки n.

Висновки.