Лабораторна робота №1
з непараметричної статистики

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

2 квітня 2022 р.

1 Частина перша.

1.1 Вступ.

У першій частині роботи побудована емпірична функція розподілу за кратною вибіркою з експоненційного розподілу з параметром інтенсивності $\lambda = 1$. Побудовано асимптотичні довірчі інтервали для справжніх значень теоретичної функції розподілу та перевірено за допомогою імітаційного експерименту, що побудовані інтервали доставляють задану точність α. Процедура була проведена для вибірок обсягу $n \in \{10, 50, 100, 500, 1000\}$.

1.2 Хід роботи.

Позначимо кратну вибірку через $X = (X_1, \ldots, X_n)$, а $F(t) = \mathbb{1}_{t>0} \cdot \left(1 - e^{-\lambda t}\right)$. Тоді емпірична функція розподілу має вигляд:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{X_j < t},\ t \in \mathbb{R}$$

Програмна реалізація обчислення значення $\hat{F}_n$ у точці $t \in \mathbb{R}$ зовсім проста:

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

Нагадаємо теорему Глівенко-Кантеллі: емпірична функція розподілу $\hat{F}_n(t)$ є рівномірно строго конзистентною оцінкою функції розподілу $F(t)$, тобто справедлива збіжність вигляду:

$$\sup_{t \in \mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \to^{P1} 0,\ n \to +\infty$$

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

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

Покажемо графіки при $n \in \{10, 50, 100, 500, 1000\}$:

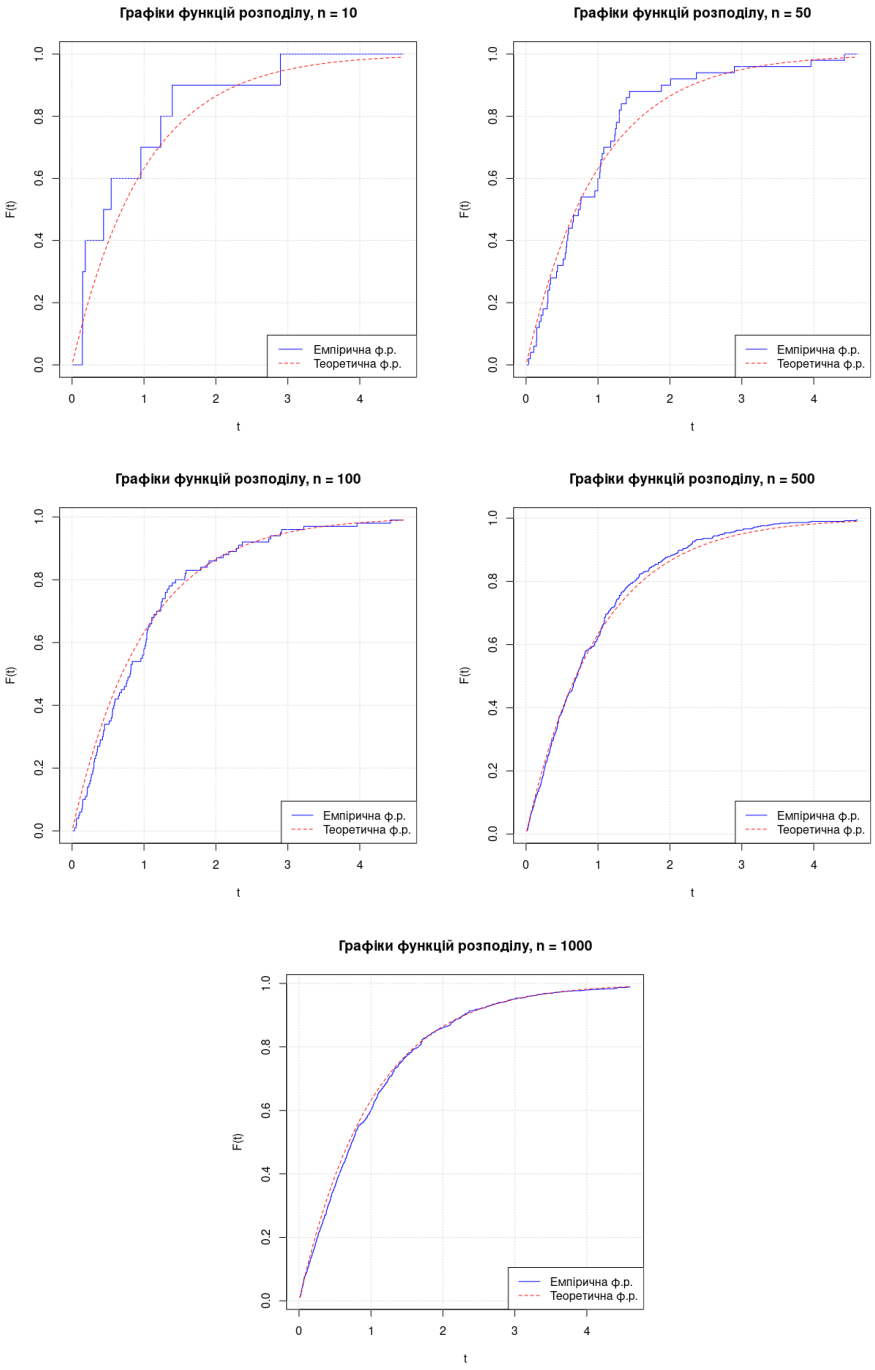


Рис. 1: Графіки $\hat{F}_n(t)$ та $F(t)$. Відхилення зменшуються при збільшенні обсягу вибірки.

На рисунках можна побачити, що сходинки емпіричної функції розподілу "лягають" на графік теоретичного розподілу для вибірок більших розмірів. Зокрема "затухає" величина найбільшого за модулем відхилення функцій (на змодельованих вибірках):

n	$\max\left\lvert\hat{F}_n - F\right\rvert$
10	0.2319
50	0.1180
100	0.0658
500	0.0315
1000	0.0377

Табл. 1: Показник абсолютного відхилення функцій розподілу в залежності від обсягу вибірки. Округлено до 4 знаків після коми.

З попередніх результатів можна вважати, що теорема виконується. Далі, побудуємо асимптотичний довірчий інтервал для значень функцій розподілу $F(t)$ використовуючи поточкову асимптотичну нормальність $\hat{F}(t)$:

$$\forall t \in \mathbb{R}: \ \sqrt{n} \cdot \frac{\hat{F}_n(t) - F(t)}{\sqrt{\sigma^2(t)}} \to^W_{n\to+\infty} \xi \sim N(0,1), \ \sigma^2(t) = F(t)(1-F(t))$$

Оскільки припускається, що дослідник не знає справжній вигляд $F(t)$, тому замість $\sigma^2(t)$ розглядається її оцінка:

$$\hat{\sigma}^2(t) = \hat{F}_n(t)(1-\hat{F}_n(t))$$

А в силу теореми Слуцького слабка збіжність до нормального розподілу виконується навіть при такій заміні.

Отже, асимптотичний довірчий інтервал рівня $1-\alpha$ матиме вигляд:

$$\hat{F}_n(t) - Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right) \cdot \sqrt{\frac{\hat{\sigma}^2(t)}{n}} \le F(t) \le \hat{F}_n(t) + Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right) \cdot \sqrt{\frac{\hat{\sigma}^2(t)}{n}}$$

Програмна реалізація така:

```
Fconf <- function(t, x, alpha = 0.05)
{
  F.emp.at.t <- Femp(t, x)
  z.alpha <- qnorm(1 - alpha / 2)
  var.at.t <- F.emp.at.t * (1 - F.emp.at.t) / length(x)
  ci.at.t <- F.emp.at.t + z.alpha * sqrt(var.at.t) * c(-1,1)
  ci.at.t
}
```

За допомогою імітаційного моделювання, для $Q_1 = Q^{Exp(\lambda)}(1/3)$ та $Q_2 = Q^{Exp(\lambda)}(2/3)$ перевіримо, чи будуть доставляти відповідні асимптотичні інтервали для $F(Q_j)$ доставляти точність $\alpha = 0.05$:

$F(x)$ \ n	10	50	100	500	1000
$F(Q_1)$	0.133	0.054	0.051	0.056	0.047
$F(Q_2)$	0.147	0.06	0.05	0.056	0.041

Табл. 2: Оцінки точності інтервалів для відповідних значень, в залежності від обсягу вибірки.

Асимптотичний довірчий інтервал доставляє потрібну точність при збільшенні вибірки.

1.3 Висновки.

Теорія узгоджується з практикою, непогані оцінки функції розподілу маємо вже за вибіркою зі ста елементів.

2 Частина друга.

2.1 Вступ.

У другій частині роботи використано класичну оцінку Каплана-Мейєра для оцінювання невідомої функції розподілу $F(t)$ справжніх значень кратної цензурованої справа вибірки $(X_j, \delta_j)_{j=1}^n$. $F(t)$ – функція логнормального розподілу з нульовим математичним сподіванням логарифма та одиничною дисперсією логарифма, а функція розподілу цензора $G(t)$ є функція χ^2-розподілу з трьома ступенями вільності. За допомогою імітаційного моделювання перевірені асимптотичні властивості оцінки. Побудовано асимптотичні довірчі інтервали для справжніх значень теоретичної функції розподілу та перевірено за допомогою моделювання, що побудовані інтервали доставляють задану точність α. Процедура була проведена для вибірок обсягу $n \in \{10, 50, 100, 500, 1000\}$.

2.2 Хід роботи.

Припустимо що вибірка $(X_j, \delta_j)_{j=1}^n$ не має однакових значень X_j. Тоді оцінку Каплана-Мейєра для $F(t)$ можна подати у вигляді:

$$\hat{F}_n^{KM}(t) = 1 - \prod_{X_{[j]} \le t} \left(1 - \frac{\delta_{[j]}}{n - j + 1}\right)$$

де $X_{[j]}$ – це j-ий елемент варіаційного ряду за X_j, а $\delta_{[j]}$ - j-ий індикатор відсутності цензурування у впорядкованому наборі індикторів за попереднім варіаційним рядом. Можлива наступна програмна реалізація обчислення значення $\hat{F}_n^{KM}(t)$:

```
FKM <- function(t, x, d)
{
  n <- length(x)
  x.v <- sort(x)
  d.v <- d[order(x)]
  m <- x.v <= t
  if(!sum(m))
  {
    return(0)
  }
  idx <- (1:n)[m]
  mults <- sapply(idx, function(j) {
    1 - d.v[j] / (n - j + 1)
  })
  1 - prod(mults)
}
```

Також доведеться генерувати цензуровані вибірки, тому опишемо відповідну функцію:

```
gencens <- function(N)
{
  t.vect <- rlnorm(N, meanlog = 0, sdlog = 1)
  c.vect <- rchisq(N, df = 3)
  d <- t.vect < c.vect
  z <- ifelse(d, t.vect, c.vect)
  list(dat = z, ind = d)
}
```

Побудуємо графіки $F(t)$, $\hat{F}_n^{KM}$ у точках $t_j := Q^{LN(0,1)}(0.01) \cdot \left(1 - \frac{j}{B}\right) + Q^{LN(0,1)}(0.99) \cdot \frac{j}{B}$, де $j = \overline{0, B}$ та $B = 1000$, для кожного $n \in \{10, 50, 100, 500, 1000\}$.

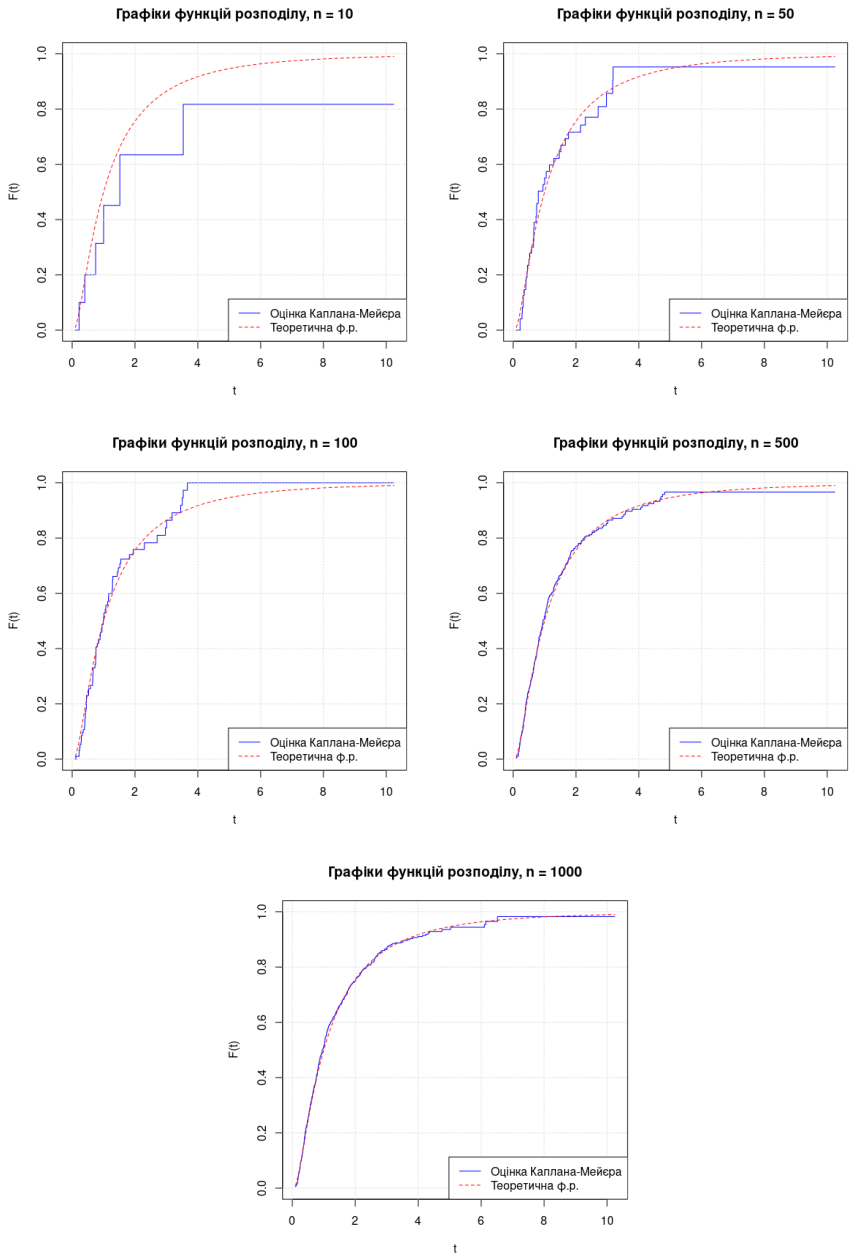


Рис. 2: Графіки $\hat{F}_n^{KM}(t)$ та $F(t)$. Для великих n ефект "вибування" зникає.

Далі дослідимо за допомогою імітаційного експерименту асимптотичний розподіл нормованої оцінки Каплана-Мейєра у точці $t_0 = Q^{LN(0,1)}\left(\frac{1}{2}\right)$. Розглянемо таке нормування $\hat{F}_n^{KM}(t_0)$:

$$\frac{\sqrt{n}\left(\hat{F}_n^{KM}(t_0) - F(t_0)\right)}{\sqrt{n \cdot V_n(t_0)}} = \frac{\hat{F}_n^{KM}(t_0) - 1/2}{\sqrt{V_n(t_0)}},\ n \geq 1 \tag{1}$$

Де $V_n(t_0)$ – формула Грінвуда для обчислення оцінки асимптотичної дисперсії $\hat{F}_n^{KM}(t_0)$:

$$V_n(t_0) = (1 - \hat{F}_n^{KM}(t_0))^2 \sum_{X_{[j]} \leq t_0} \frac{\delta_{[j]}}{(n-j+1)(n-j+1-\delta_{[j]})}$$

Оскільки $\sqrt{n}\left(\hat{F}_n^{KM}(t_0) - F(t_0)\right) \to^W \xi \sim N(0, \sigma^2(t_0))$ та $nV_n(t_0) \to^P \sigma^2(t_0)$, то за теоремою Слуцького:

$$\frac{\hat{F}_n^{KM}(t_0) - 1/2}{\sqrt{V_n(t_0)}} \to \eta \sim N(0,1),\ n \to +\infty$$

В імітаційному експерименті згенеруємо по $m = 1000$ повторних цензурованих вибірок обсягу n та за кожною з них обчислимо (1). Отримавши набір з нормованих значень оцінки Каплана-Мейєра у точці t_0, побудуємо гістограму абсолютних частот та нарисуємо графік нормованої щільності стандартного нормального розподілу. Для достатньо великих n можна очікувати, що емпіричний розподіл нормованої оцінки узгоджується зі стандартним нормальним.

Програмна реалізація підрахунку $V_n(t)$:

```
FKM.var.estim <- function(t, x, d, FKM.at.t)
{
  n <- length(x)
  x.v <- sort(x)
  d.v <- d[order(x)]
  m <- x.v <= t
  if(!sum(m)) { return(0) }
  idx <- (1:n)[m]
  sums <- sapply(idx, function(j) {
    d.v[j] / ((n - j + 1) * (n - j + 1 - d.v[j]))
  })
  (1 - FKM.at.t)^2 * sum(sums)
}
```

Генерування вибірки зі значень нормованої оцінки $\hat{F}_n^{KM}(t_0)$:

```
p <- 0.5
t0 <- qlnorm(p, meanlog = m.log, sdlog = s.log)

m <- 1000
FKM.boot.cor <- replicate(m, {
  u.boot <- gencens(n)
  Y.boot <- u.boot$dat
  d.boot <- u.boot$ind
  FKM.at.t0 <- FKM(t0, Y.boot, d.boot)
  s.est <- sqrt(FKM.var.estim(t0, Y.boot, d.boot, FKM.at.t0))
  (FKM.at.t0 - p) / s.est
})
```

Покажемо гістограми з нормованими щільностями для $n \in \{10, 50, 100, 500, 1000\}$.

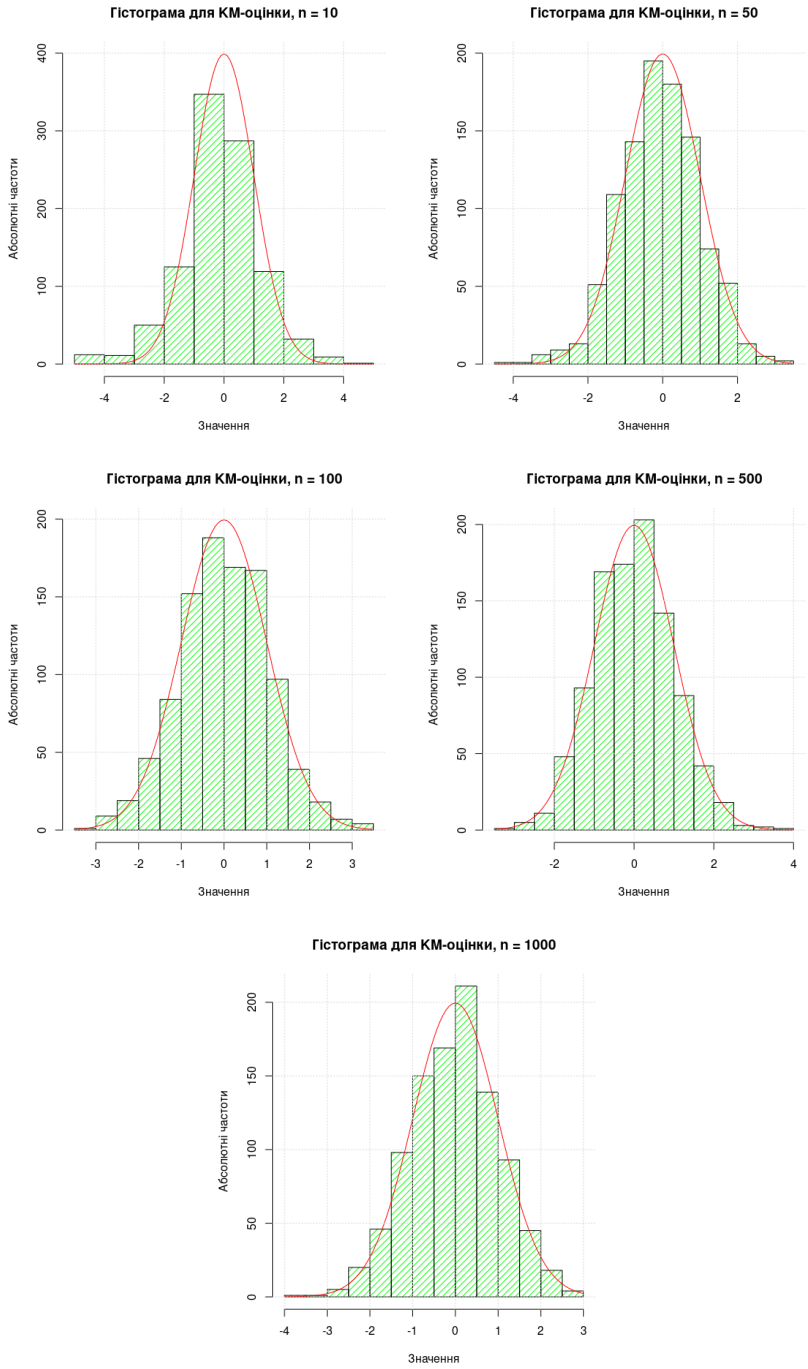


Рис. 3: Гістограми абсолютних частот за вибіркою зі значень нормованої оцінки $\hat{F}_n^{KM}(t_0)$.

Можна припускати, що розподіл нормованої оцінки Каплана-Мейєра добре наближається нормальним починаючи з $n = 75 \pm 25$. Використовуючи формулу Грінвуда, можна побудувати асимптотичний довірчий інтервал для $F(t_0)$ рівня $1 - \alpha$, де $\alpha \in (0, 1)$:

$$\hat{F}_n^{KM}(t_0) - Q^{N(0,1)}\left(1 - \frac{\alpha}{2}\right)\sqrt{V_n(t_0)} \leq F(t_0) \leq \hat{F}_n^{KM}(t_0) + Q^{N(0,1)}\left(1 - \frac{\alpha}{2}\right)\sqrt{V_n(t_0)} \qquad (2)$$

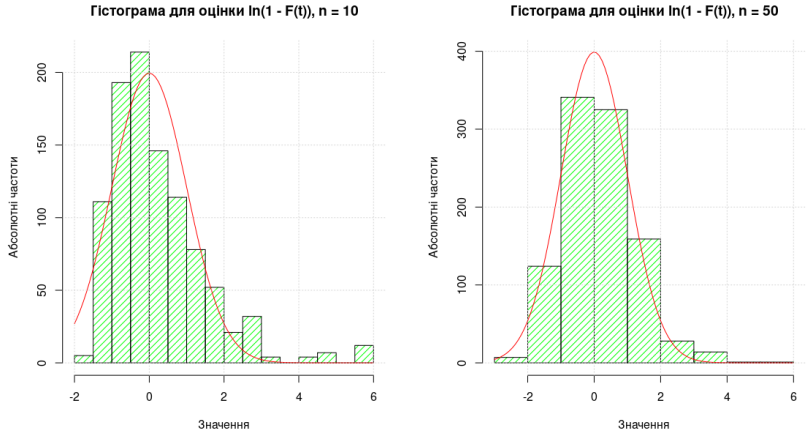
Інтервали будемо будувати пізніше при $\alpha = 0.05$, зараз дослідимо асимптотичний розподіл $\ln\left(1 - \hat{F}_n^{KM}(t_0)\right)$ як оцінки для $\ln\left(1 - F(t_0)\right)$. Застосуємо аналогічний підхід з нормуванням.

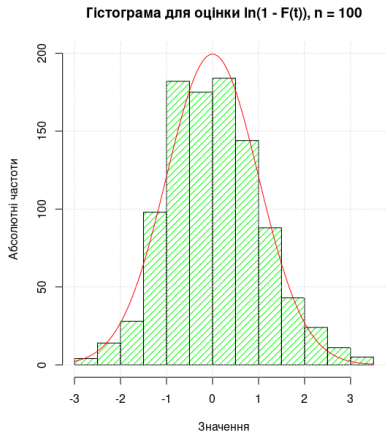
Оскільки $g(t) = \ln(1-t)$ – неперервно диференційовна по $t \in (-\infty, 1)$, то маємо:

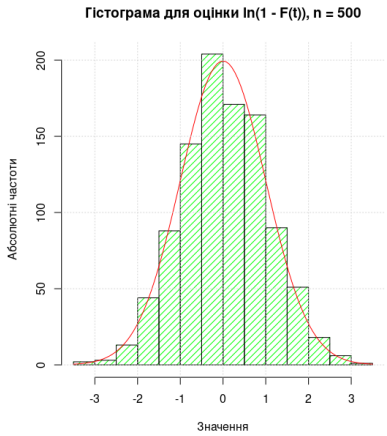
$$\sqrt{n}\left(g(\hat{F}_n^{KM}(t_0)) - g(F(t_0))\right) \to^W \xi \sim N(0, (g'(F(t_0)))^2 \cdot \sigma^2(t_0)),\ n \to +\infty$$

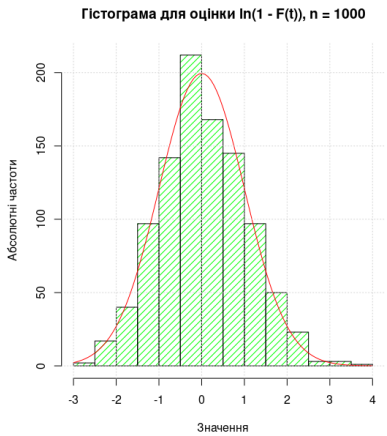
де $g'(t) = -1/(1-t) \neq 0$. У нашій роботі $G(x) < 1$, тому $\hat{F}_n^{KM}(t_0) \to^P F(t_0)$, а звідси $g'(\hat{F}_n^{KM}(t_0)) \to^P g'(F(t_0))$ в силу неперервності похідної $g(t)$. За теоремою Слуцького:

$$\frac{g(\hat{F}_n^{KM}(t_0)) - g(F(t_0))}{|g'(\hat{F}_n^{KM}(t_0))| \cdot \sqrt{V_n(t_0)}} \to \eta \sim N(0,1),\ n \to +\infty$$









У цьому разі можна звернути на те, що на вибірках де $n < 100$ апроксимація недоречна. Кращі наближення маємо починаючи з $n \gg 100$. Теоретичний факт узгодився з реальністю.

Асимптотичний довірчий інтервал рівня $1-\alpha$ для $\ln(1-F(t_0))$ можна будувати використавши попередні результати про асимптотичну нормальність:

$$g(\hat{F}_n^{KM}(t_0)) - Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right)\sqrt{S_n(t_0)} \le g(F(t_0)) \le g(\hat{F}_n^{KM}(t_0)) + Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right)\sqrt{S_n(t_0)} \quad (3)$$

де $S_n(t_0) = (g'(\hat{F}_n^{KM}(t_0)))^2 \cdot V_n(t_0)$. Для $g(t)$ існує обернена $g^{-1}(t) = 1 - e^t$. На кінці інтверали спробуємо подіяти $g^{-1}(t)$, щоб отримати інший довірчий інтервал для $F(t_0)$:

$$g^{-1}\left(g(\hat{F}_n^{KM}(t_0)) + Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right)\sqrt{S_n(t_0)}\right) \le F(t_0) \le g^{-1}\left(g(\hat{F}_n^{KM}(t_0)) - Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right)\sqrt{S_n(t_0)}\right)$$

Зауважимо, що

$$g^{-1}\left(g(\hat{F}_n^{KM}(t_0)) \pm Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right)\sqrt{S_n(t_0)}\right) = 1 - (1-\hat{F}_n^{KM}(t_0)) \cdot e^{\pm Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right)\sqrt{S_n(t_0)}} \quad (4)$$

Тоді (3) має спрощений вигляд:

$$1 - (1-\hat{F}_n^{KM}(t_0)) \cdot e^{Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right)\sqrt{S_n(t_0)}} \le F(t_0) \le 1 - (1-\hat{F}_n^{KM}(t_0)) \cdot e^{-Q^{N(0,1)}\left(1-\frac{\alpha}{2}\right)\sqrt{S_n(t_0)}}$$

Перевіримо коли асимптотичні властивості справджуються для інтервалів (2) та (4) за допомогою імітаційного моделювання:

n	**FKM**	**gFKM**
10	*0.121*	*0.086*
50	0.054	0.059
100	0.061	0.063
500	0.047	0.048
1000	0.051	0.052

Табл. 3: Оцінена точність асимптотичних довірчих інтервалів в залежності від обсягу вибірки.

Для $n = 10$ оцінка зрізана, бо в експерименті згенеровано такі цензуровані вибірки, де цензурування не припадало на останній момент часу, що і повернуло некоретне значення для дисперсії (NaN):

```
> Y.boot
 [1] 0.69805441 1.25634045 0.58654182 0.79954436 0.90837815 0.44817786
 [7] 0.73290753 0.01994765 0.28506201 0.30462092
> d.boot
 [1] FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
> d.boot[order(Y.boot, decreasing = T)[1]] # В останній момент цензуров. немає
[1] TRUE
> Y.boot[order(Y.boot, decreasing = T)[1]] # Зафіксований час
[1] 1.25634
> mean(na.omit(fkm.ci.m))
[1] 0.1214859
> mean(na.omit(gfkm.ci.m))
[1] 0.08634538
```

Насправді така проблема була і на початку, бо зернина не змінювалася для генерування псевдовипадкових чисел, а тому можна точно знати для яких вибірок отримали некоректні результати (зрозуміло що у hist по замовчуванню значення NaN не враховуються, тому одразу цього не можна помітити). Потрібно залучити до вибірок додатковий момент часу, більший за максимальний серед зафіксованих: $X_{n+1} := +\infty$ з $\delta_{n+1} = 1$. Для цього виправимо програмну реалізацію генерування цензурованої вибірки:

```
gencens <- function(N)
{
  t.vect <- rlnorm(N, meanlog = 0, sdlog = 1)
  c.vect <- rchisq(N, df = 3)
  d <- t.vect < c.vect
  z <- ifelse(d, t.vect, c.vect)
  # Додатковий момент часу
  z <- c(z, Inf)
  d <- c(d, 1)
  list(dat = z, ind = d)
}
```

Покажемо графіки функцій розподілу, гістограм нормованих оцінок використовуючи виправлений генератор. Насправді графіки оцінки Каплана-Мейєра не зовсім цікавлять, бо там незначна різниця з попередньою версією (лише $\hat{F}_n^{KM}(t) = 1$ не досягається коли вводиться нескінченно віддалений момент часу).

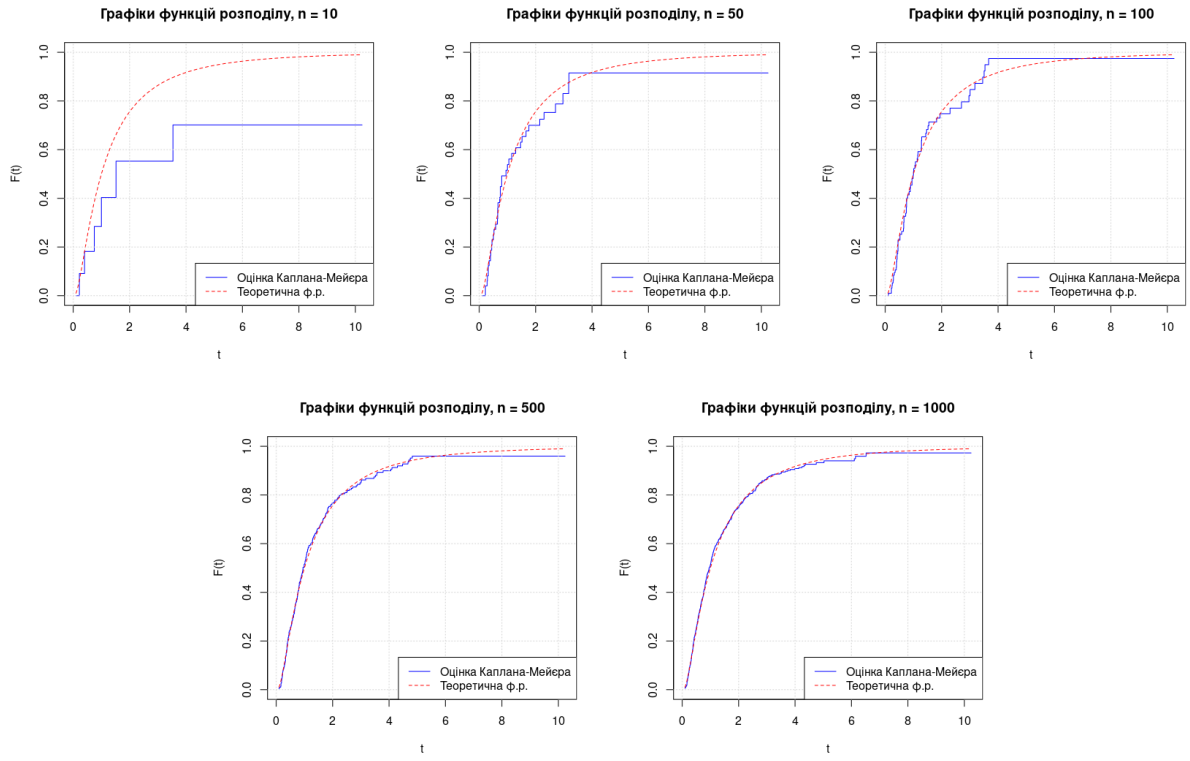


Рис. 4: Графіки $\hat{F}_n^{KM}(t)$ та $F(t)$ після виправлення.

Далі покажемо емпіричний розподіл нормованої оцінки $\hat{F}_n^{KM}(t_0)$.

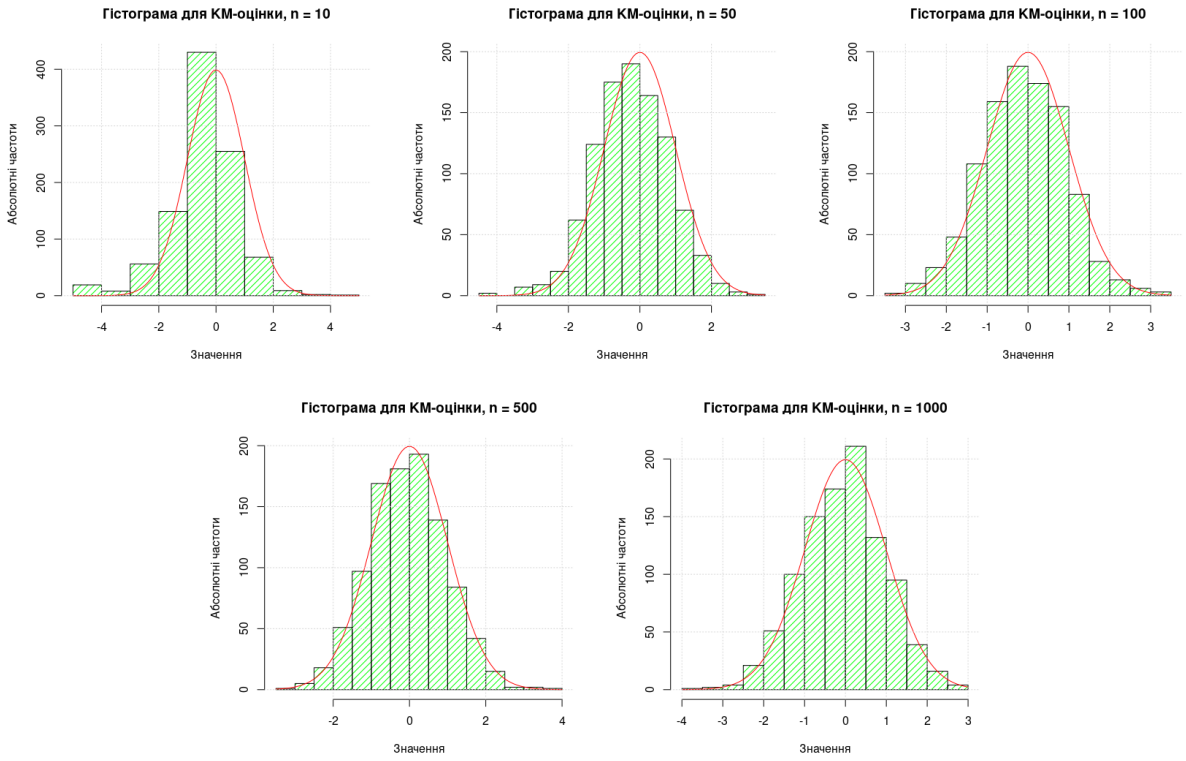


Рис. 5: Гістограми для нормованої оцінки $\hat{F}_n^{KM}(t_0)$ після виправлення.

Попередні висновки щодо адекватності використання кормальної апроксимації для розподілу $\hat{F}_n^{KM}(t_0)$ залишається без змін.

n	**FKM**
10	0.101
50	0.055
100	0.060
500	0.048
1000	0.052

Табл. 4: Оцінена точність асимптотичних довірчих інтервалів в залежності від обсягу вибірки за формулою Грінвуда після виправлення.

Вже при $n \gg 50$ асимптотичний довірчий інтервал може доставляти задачу точність $\alpha = 0.05$.

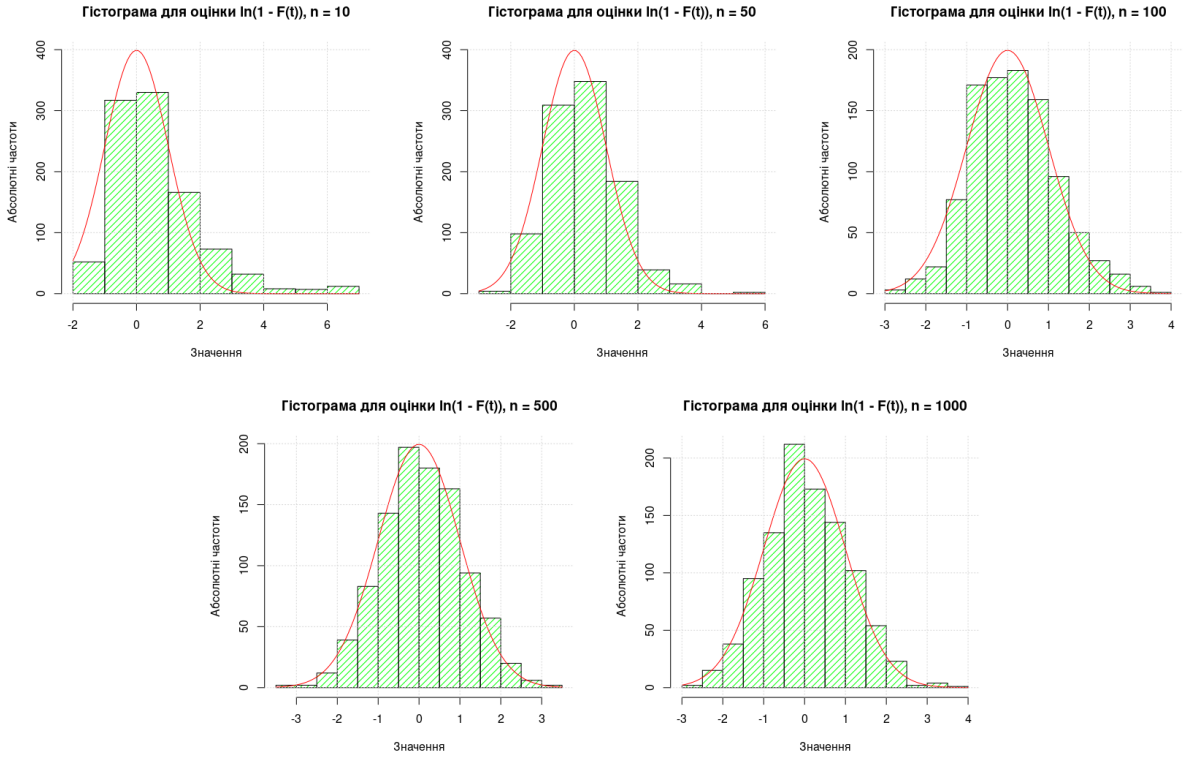


Рис. 6: Гістограми для нормованої оцінки $\ln(1 - \hat{F}_n^{KM}(t_0))$ після виправлення.

Попередні висновки щодо адекватності використання кормальної апроксимації для розподілу $\ln(1 - \hat{F}_n^{KM}(t_0))$ залишається без змін. Зокрема можна побачити які важкі хвости формуються за вибірками обсягу меншого за 100. "Нормальні" форми маємо для вибірок з більшою кількістю спостережень.

n	**gFKM**
10	0.139
50	0.065
100	0.068
500	0.047
1000	0.053

Табл. 5: Оцінена точність аисмптотичних довірчих інтервалів в залежності від обсягу вибірки за лог-трансформацією після виправлення.

Повільніше доставляється задана точність. У випадку асимптотичних довірчих інтервалів з лог-трансформацією задана точність α досягається при $n \gg 100$.

2.3 Висновки.

Розподіл $\hat{F}_n^{KM}(t)$ добре наближується нормальним за вибіркою з меншої кількості спостережень, ніж для розподілу $\ln(1-\hat{F}_n^{KM}(t))$. Інколи краще не морочити голову та використовувати простіші інтервали, які забезпечують кращі асимптотичні властивості (як краща швидкість забезпечення заданої точності).