Самостійна робота №1
з дисципліни "асимптотична статистика"
Варіант №4

Горбунова Даніела Денисовича
4 курс бакалаврату
група "комп'ютерна статистика"

4 лютого 2021 р.

1 Вступ.

У даній роботі проаналізовано дані про спроби несанкціонованого доступу до інформації в інтернеті. Більш точно, використано критерій χ^2 для перевірки складної гіпотези про розподіл часових інтервалів між послідовними спробами несанкціонованого доступу з рівнем значущості $\alpha = 0.05$.

Зауваження. Надалі $\xi = (\xi_j)_{j=1}^n$ - початкова вибірка, $n = 120$.

2 Переформулювання статистичних гіпотез.

У рамках даної роботи, сформульовані такі гіпотези, які потрібно було перевірити:

$$H_0 : \text{Хакерської атаки не було;}$$
$$H_1 : \text{Має місце хакерська атака.}$$

Якщо вважати, що хакерської атаки не було, тоді можна припустити, що спроби несанкціонованого доступу незалежні між собою. Тому інтервали між ними можна інтерпретувати як час до першого успіху (наступний несанкціонований доступ). З цих міркувань виходимо на те, що при виконанні основної гіпотези розподіл інтервалів є експоненційним.

$$H_0 : \exists \lambda > 0 : \xi_1 \sim Exp(\lambda)$$
$$H_1 : \forall \lambda > 0 : \text{Спостереження розподілені інакше}$$

3 Групування даних.

Зауважимо, що мінімальне та максимальне значення ξ дорівнюють:

$$\min_{1 \leqslant j \leqslant n} \xi_j = 0.0001437491, \ \max_{1 \leqslant j \leqslant n} \xi_j = 9.935669$$

Тому всі спостереження можна помістити у відрізок $\mathcal{P} = [0, 10]$. Далі $\mathcal{P}$ розбили на $K = 5$ інтервалів $\mathcal{P}_j$ однакової довжини ($\mathcal{P}_1 = [t_0, t_1]$, $\mathcal{P}_j = (t_{j-1}, t_j], j \in \overline{2, K}$; $t_j = 2j$). Для кожного $\mathcal{P}_j$ були обчислені емпіричні частоти ν_j, тобто:

$$\nu_j = \#\{1 \leqslant i \leqslant n : \xi_i \in \mathcal{P}_j\},\ j \in \overline{1, K}$$

Нижче наведена таблиця емпіричних частот:

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_5$
ν_j	72	26	13	5	4

Видно, що лише одна частота (для крайнього інтервалу) менша 5. У нашому випадку, частка тих частот, що перевищує 5, становить 80%. Якщо довіряти застереженню в підручнику Карташова, то можна більш-менш довіритися результатам, які отримаємо за допомогою тесту χ^2.

4 Оцінювання невідомих параметрів розподілу.

В експоненційному розподілі необхідно оцінити єдиний параметр інтенсивності $\lambda > 0$ (кількість невідомих параметрів позначимо через $d = 1$). Для кратної вибірки відома така оцінка найбільної вірогідності (замінивши вибіркове середнє на альтернативу для групованих даних):

$$\hat{\lambda}_n = \left(\overline{\xi}_n^{group}\right)^{-1},\ \overline{\xi}_n^{group} = \frac{\sum_{j=1}^{K} \nu_j x_j}{\sum_{j=1}^{K} \nu_j},\ x_j = \frac{t_j - t_{j-1}}{2},\ j = \overline{1, K}$$

На основі розбиття та частот, отриманих у попередньому розділі, побудуємо гістограму відносних частот. Щільність експоненційного можна побудувати, маючи оцінку найбільшої вірогідності (у даному випадку $\hat{\lambda}_n = 0.4195804$) Форма гістограми за початковими даними

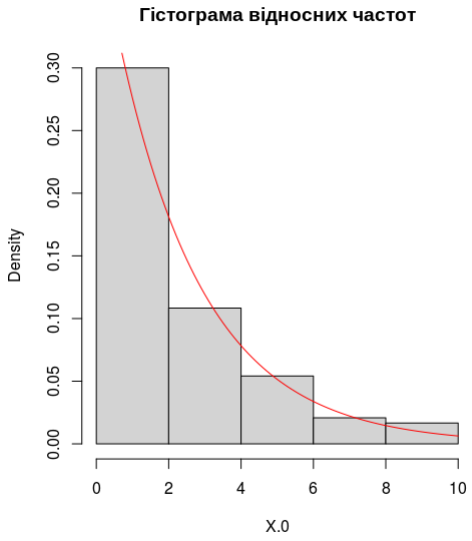

Рис. 1: Гістограма відносних частот для початкових даних. Червоною кривою відмічено графік щільності експоненційного розподілу.

дещо узгоджується з експоненційним розподілом, суттєвих відхилень при заданному розбитті не видно.

5 Обчислення χ^2 статистики. Результати тесту.

Нагадаємо, що статистика тесту χ^2 для перевірки складної гіпотези про розподіл даних має вигляд:

$$\widehat{\chi}^2_{emp} = \sum_{j=1}^{K} \frac{(\nu_j - p_j n)^2}{p_j n},$$

де $p_j = \mathbb{P}_{H_0}(\xi_1 \in \mathcal{P}_j) = \mathbb{P}_{H_0}(\xi_1 < t_j) - \mathbb{P}_{H_0}(\xi_1 < t_{j-1}), j = \overline{1, K-1}$; $p_K = 1 - \sum_{i=1}^{K-1} p_i$. Порогове значення тесту - квантиль рівня $1-\alpha$ розподілу χ^2 з $K - d - 1 = 3$ ступенями вільності (позначимо через χ^2_{th}).

Після нескладних обчислень маємо значення статистики $\widehat{\chi}^2_{emp} = 0.6796512$ і квантиль $\chi^2_{th} = Q^{\chi^2_3}(1-\alpha) = 7.814728$.

6 Висновки.

За результатами тесту χ^2, альтернативна гіпотеза, що полягає у наявності хакерської атаки, відхиляється при $\alpha = 0.05$.

7 Програмна реалізація.

```
workspace_setup <- function()
{
  workspace <- dirname(sys.frame(1)$ofile)
  setwd(workspace)
  print(getwd())
}

workspace_setup()

alpha <- 0.05
d <- 1

X <- read.table("haker.txt", header = T)
X.0 <- data.matrix(X[colnames(X)[1]])
N <- length(X.0)

# H0: Хакерської атаки немає | Exp
# H1: Була хакерська атака | not Exp

h <- hist(X.0, probability = T, breaks = 5)

K <- length(h$breaks)
O <- h$counts
delta <- diff(h$breaks)[1]

halves <- h$breaks[-K] + delta/2
X.0.mean <- sum(O * halves)/sum(O)

l.est <- 1/X.0.mean

curve(dexp(x, l.est), add = T)
p.boundaries <- pexp(h$breaks, rate = l.est)
p.diff <- diff(p.boundaries[-K])
p <- c(p.diff, 1 - p.boundaries[K-1])

df.fix <- K - d - 1

E <- N * p
chisq.stat <- sum((O - E)^2/E)
chisq.quan <- qchisq(1 - alpha, df = df.fix)
print(paste("chisq stat: ", chisq.stat))
print(paste("chisq.quan: ", chisq.quan))
print(1 - pchisq(chisq.stat, df = df.fix))
```