Лабораторна робота №2
з дисципліни ”комп’ютерна статистика”
Варіант №4

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

17 вересня 2020 р.

1 Вступ.

У даній роботі вказана інформація про отримані результати під час виконання роботи №2. Використані візуальні методи дослідження розподілу спостережень на прикладі гістограми відносних частот та P-P, Q-Q діаграм. Для вхідних даних вдалося підібрати найбільш слушний розподіл та оцінили його параметри методом моментів. Додатково були реалізовані функції для зображення прогнозних інтервалів для Q-Q діаграми та відносних частот на гістограмі.

1.1 Теоретичні відомості.

Гістограма - це один з методів відображення розподілу числових спостережень. Розрізняють гістограми абсолютних та відносних частот. В загальному, спочатку проводять групування даних, а на її основі виконується візуалізація частот для відповідних підмножин. Принцип групування описано нижче.

Нехай у нас задана кратна вибірка $\Xi = (\xi_1, \ldots, \xi_N)$ та оберемо множину $\mathcal{P} = [a, b]$, на якій зосереджені всі спостереження. Беремо деяке число K, що відповідає за кількість інтервалів на множині $\mathcal{P}$. Тому розділимо множину $\mathcal{P}$ на K інтервалів $P_1, \ldots, P_K$ однакової ширини $h = (b - a)/K$. Інтервали P_j можна визначити деклькома способами (всюди $t_j = a + jh$):

- Відкритий справа інтервал: $P_j = [t_j, t_{j+1}),\ j \in \overline{1, K-1},\ P_K = [t_K, t_{K+1}]$;
- Відкритий зліва інтервал: $P_j = (t_j, t_{j+1}],\ j > 1,\ P_1 = [t_1, t_2]$;
- Відрізок: $P_j = [t_j, t_{j+1}],\ j \in \overline{1, K}$. Якщо спостереження попадає на сумісну границю інтервалів, тоді частоту збільшують для двох інтервалів, але по 0.5.

У рамках цієї роботи обмежимося лише другим методом побудови інтервалів. Суттєвих відмінностей для значної кількості даних немає, але цікаво буде самостійно подивитися різницю на вибірці малого обсягу.

Для кожного P_j, $(j \in \overline{1,K})$ обчислимо кількість спостережень, які належать заданому інтервалу: $\zeta_j = \sum_{s=1}^{N} \mathbb{1}_{\xi_s \in P_j}$. Тоді $\zeta = (\zeta_1, \ldots, \zeta_K)$ - вектор абсолютних частот спостережень з вибірки Ξ. Маючи ζ, для кожної її координати будується стовпчик такої висоти, яка відповідає значенню частоти відповідного інтервалу. Ширина стовпчика визначається розмахом інтервалу з $\mathcal{P}$. Таким чином отримаємо гістограму абсолютних частот.

Кількість розбиттів K множини $\mathcal{P}$ або ширина h інтервалів P_j по факту залежить від вхідних даних, тому це число частіше підбирають у процесі їхньої обробки. Однак можна навести декілька емпіричні правила підбору K для конкретних даних:

- **Мінімаксне правило.** Для фіксованої ширини h, кількість розбиттів можна обчислити за формулою: $K = \lceil \mathrm{Range}(\Xi)/h \rceil$;
- **Формула Стургеса (Sturges' formula).** Отримана з властивостей біноміального розподілу. Вважається, що дані мають розподіл близький до нормального: $K = \lceil \log_2 N \rceil + 1$;
- **Правило Фрідмена-Дайконіса (Freedman–Diaconis' choice).** Для спостережень з нормального розподілу ширину інтервалів можна підібрати таким чином: $h = 2(\mathrm{IQR}(\Xi)/\sqrt[3]{n})$. Такий вибір є оптимальним в сенсі стійкості характеристики до викидів внаслідок робастності інтерквартильного розмаху.

Побудова гістограми відносих частот дещо відрізняється, але опирається на значення ζ та обране розбиття P_j, $j \in \overline{1,K}$. Нехай $\nu_j = \zeta_j/N$ - відносна частота j-го інтервалу, $j \in \overline{1,K}$. Тоді висота стовпчика визначається як $\gamma_j = \nu_j/h$. Множник $1/(Nh)$ обраний для того, щоб можна було γ_j використовувати у якості оцінки щільності розподілу спостережень.

Серед графічних методів, для перевірки узгодженності вхідних даних з наперед заданим розподілом, також використовують **P-P та Q-Q діаграми**. Такі методи хороші тим, що вони є непараметричними, за виключенням того, що потрібно першочергово знати для якої ймовірнісної моделі будується діаграма. "Ймовірність проти ймовірності" порівнює емпіричну та теоретичну функції розподілу, а "Квантиль проти квантиля" емпіричні з теоретичними квантилями розподілу.

Побудова P-P діаграми полягає у виконанні наступних кроків. Нехай Ξ - кратна вибірка та сформульована нульова гіпотеза:

$$H_0 = \{\forall t \in \mathbb{R} : \; F_{\xi_1}(t) = F(t)\}$$

Припускаючи істиність цієї гіпотези, тоді емпірична функція розподілу вибірки $\hat{F}_N$ буде консистентною оцінкою теоретичної функції розподілу F. Тому ордината і абсциса кожної точки $\left(F(\xi_j), \hat{F}_N(\xi_j)\right)$, $j \in \overline{1,N}$ повинні бути близькими одна до одної, отже вишукуються поблизу від бісектриси першого координатного кута. Якщо це не так, то нульову гіпотезу слід відхилити.

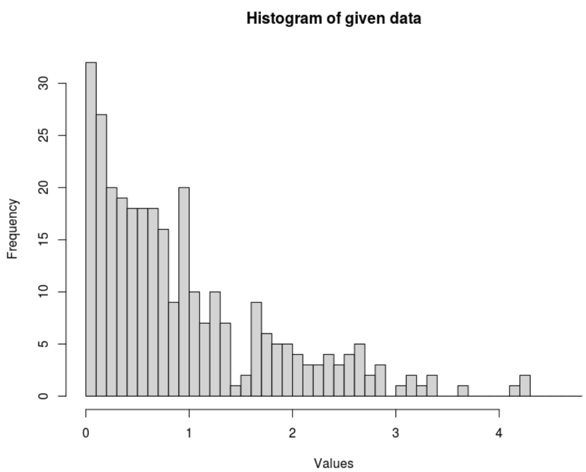
Побудова Q-Q діаграми зводиться до аналогічних кроків, що розглядалися при побудові P-P діаграми, за виключенням визначення точок на площині: $\left(Q^F(j/N - 1/(2N)), \xi_{(j)}\right)$, $j \in \overline{1,N}$.

2 Побудова гістограми спостережень. Підгонка форми та параметрів розподілу.

Маємо вибірку з $N = 300$ спостережень, які зосереджені на інтервалі $\mathbb{R}^+ = [0, \infty)$. Екстремальні значення вибірки лежать в підмножині $\mathcal{P} = [0, 4.8]$, тому розбиття робимо на цій множині. Зафіксуємо $K = 48$ з кроком $h = 0.1$. Тоді отримаємо розбиття:

$$P_1 = [0.0, 0.1], P_2 = (0.1, 0.2], \ldots, P_{K-1} = (4.6, 4.7], P_K = (4.7, 4.8]$$

Гістограми абсолютних та відносних частот наведені нижче.



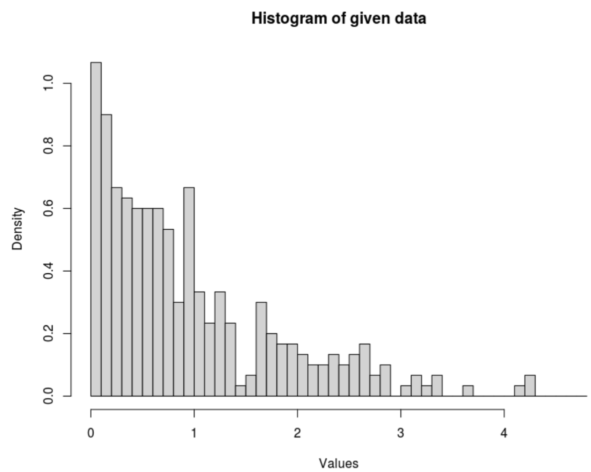


Рис. 1: Гістограми абсолютних та відносних частот для спостережень вхідної вибірки.

З рисунків видно, що форма гістограми може бути доречною для моделі з Г-розподілом, оскільки вона різко спадає при переході у праву частину графіка, є асиметричною, а носій - це підмножина з $\mathbb{R}^+$. Тому припущення про те, що спостереження можуть мати нормальний розподіл, стає сумнівним. Також бачимо провал в околі точки 1.5 - можливо він випадковий, оскільки в загальному картина поблизу нього майже не псується в сенсі абсолютних частот, які спостерігаються в сусідніх інтервалах. Це можна було б додатково перевірити за допомогою тестів узгодженості розподілу, однак цього разу ми не будемо цим користуватися.

Покажемо як добре розміщуються стовпці гістограми під кривою щільності для таких розподілів: нормальний, логнормальний, експоненційний, хі-квадрат. Оскільки параметри підбираємо на око, тому подивимося на цю проблему з математичної точки зору: параметри цих чотирьох розподілів будуть оцінені за допомогою методу моментів.

2.1 Оцінювання параметрів заданих розподілів.

Обчислюємо перші та другі емпіричні моменти вибірки:

$$\hat{\mu}_N = 0.9542,\ \sqrt{\hat{\sigma}_N^2} = 0.8818$$

Покажемо отримані оцінки, отримання деяких детально опишемо:

- **Нормальний розподіл.** $\hat{\mu}_{MM} = \hat{\mu}_N$, $\hat{\sigma^2}_{MM} = \hat{\sigma}_N^2$. Але в роботі для σ ми використали незміщену оцінку: $\hat{\sigma}_N = \sqrt{(n/(n-1))\hat{\sigma^2}_{MM}}$;

- **Експоненційний розподіл.** $\hat{\lambda}_{MM} = 1/\hat{\mu}_N$. В роботі довизначили оцінку для незміщеності: $\hat{\lambda}_N = ((N-1)/N)\hat{\mu}_N^{-1}$. Множник у формулі не зовсім тривіальний, тому покажемо обчислення:

$$\mathbb{M}\left[\hat{\mu}_N^{-1}\right] = \left|\xi_i \sim Exp(\lambda),\ \xi_1 + \ldots + \xi_N \sim \Gamma(N,\lambda)\right| = \frac{N}{(N-1)!}\int\limits_0^\infty \lambda^N t^{N-2}\exp(-\lambda t)dt =$$

$$= \frac{N}{\lambda(N-1)!}\int\limits_0^\infty u^{N-2}\exp(-u)du = \frac{N}{\lambda(N-1)!}\Gamma(N-1) = \frac{N}{(N-1)}\mathbb{M}\left[\xi_1\right]$$

- **Хі-квадрат розподіл.** Припустимо, що розглядається деяка величина $\eta \sim \Gamma(\theta,\lambda)$. Обчислимо математичне сподівання, далі застосуємо метод моментів для θ:

$$\mathbb{M}\left[\eta\right] = \frac{1}{\Gamma(\theta)}\int\limits_0^\infty t\lambda^\theta t^{\theta-1}\exp(-\lambda t)dt = \frac{1}{\lambda\Gamma(\theta)}\int\limits_0^\infty u^\theta\exp(-u)du = \frac{\Gamma(\theta+1)}{\lambda\Gamma(\theta)} = \frac{\theta}{\lambda}$$

  Тепер розглянемо рівняння:

$$\frac{\hat{\theta}_N}{\lambda} = \hat{\mu}_N \Leftrightarrow \hat{\theta}_N = \lambda\hat{\mu}_N \Rightarrow \left|\lambda = 0.5\right| \Rightarrow \frac{\hat{\mu}_N}{2} \Rightarrow \left|\theta = 0.5k\right| \Rightarrow \hat{k}_{MM} = \hat{\mu}_N$$

  Якщо ми розглядаємо хі-квадрат розподіл, то при великих N бажано взяти цілу частину від значення $\hat{\mu}_N$.

- **Логнормальний розподіл.** Припустимо, що розглядається деяка величина $\eta \sim LN(\mu,\sigma^2)$. Для обчислення моментів η, знайдемо твірну функцію моментів величини $\zeta \sim N(0,1)$:

$$M_\zeta(t) = \mathbb{M}\left[e^{t\zeta}\right] = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^\infty e^{tu}e^{-\frac{1}{2}u^2}du = \left| -\frac{1}{2}\left(u^2 - 2tu + t^2 - t^2\right) = -\frac{1}{2}\left(u-t\right)^2 + \frac{t^2}{2}\right| =$$

$$= e^{\frac{t^2}{2}}\int\limits_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(u-t)^2}du = e^{\frac{t^2}{2}} \Rightarrow M_{\mu+\sigma\zeta}(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}};\ t\in\mathbb{R},\ \mu\in\mathbb{R},\ \sigma > 0$$

  При $t = 1$ можна побачити, що $\mathbb{M}\left[\eta\right] = M_{\mu+\sigma\zeta}(1) = e^{\mu+\frac{\sigma^2}{2}}$, а $\mathbb{M}\left[\eta^2\right] = M_{\mu+\sigma\zeta}(2) = e^{2\mu+2\sigma^2}$
  Отже маємо наступні оцінки: $\hat{\sigma^2}_{MM} = \ln\left(\frac{\hat{\sigma}_N^2}{\hat{\mu}_N^2} + 1\right)$, $\hat{\mu}_{MM} = \ln\left(\hat{\mu}_N/\sqrt{\hat{\sigma^2}_{MM}}\right)$.

2.2 Гістограми та щільності. Порівняльна характеристика.

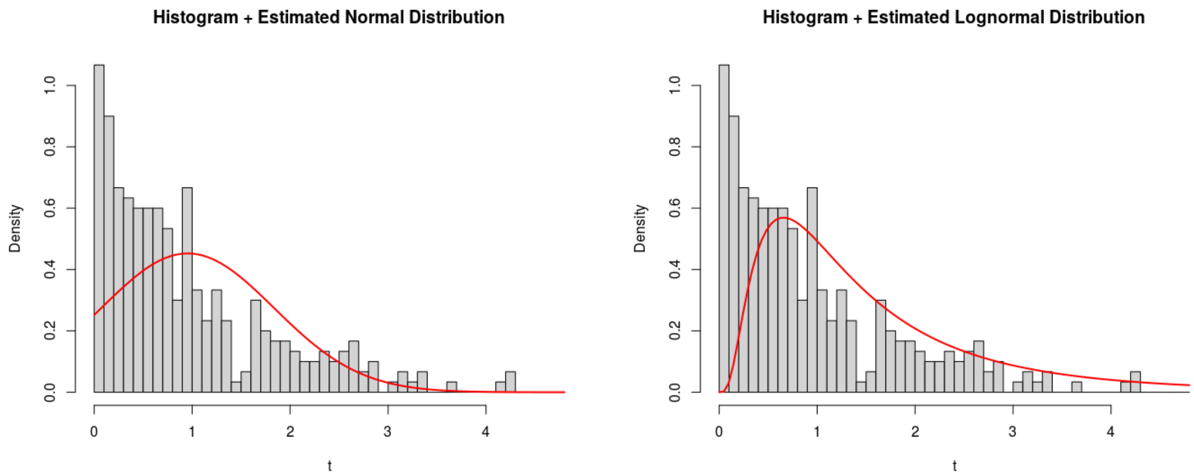


Рис. 2: Гістограма відносних частот з щільностями нормального (зліва) та логнормального (справа) розподілів.

Наведені раніше припущення прийняті: з гістограм відносних частот видно, що розподіл запропонованих даних не співпадає не з гаусовим, так і не з логнормальним. Цікаве попереду, з щільностями гамма-розподілу.

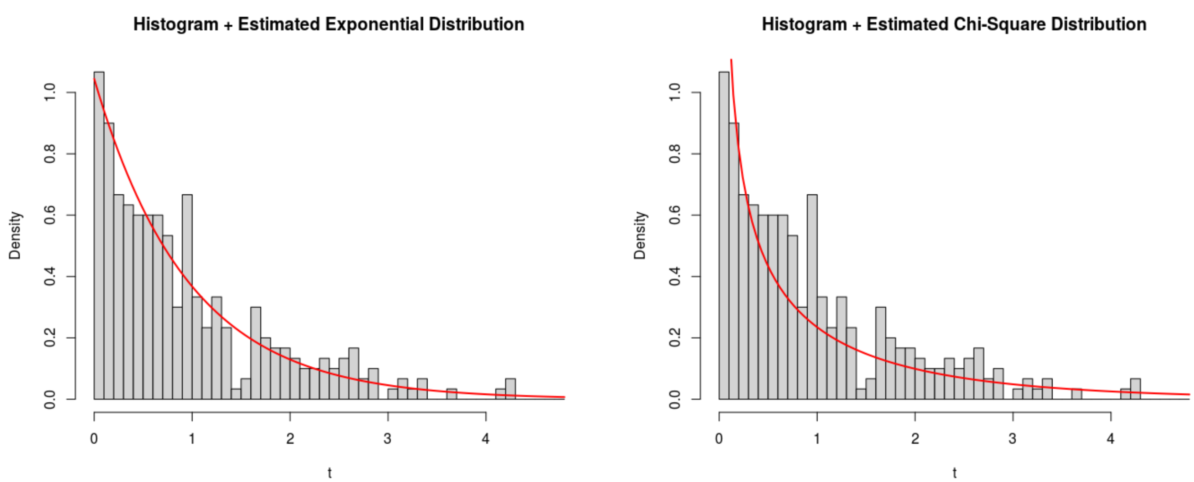


Рис. 3: Гістограма відносних частот з щільностями експоненційного (зліва) та χ^2 (справа) розподілів.

Бачимо, наразі можна брати до уваги два розподіли, а саме експоненційний та хі-квадрат з "підігнаними" параметрами за допомогою методу моментів. Однак видно, що крива щільності першого розподілу більш природньо розмістилася над стовпцями відносних частот.

3 P-P та Q-Q діаграми.

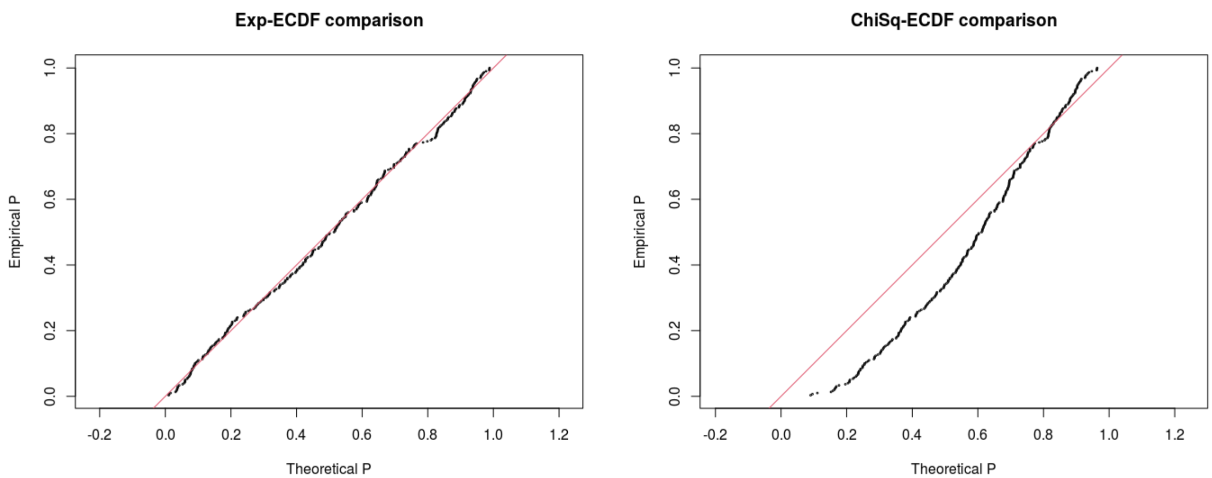


Рис. 4: P-P діаграми для експоненційного та хі-квадрат розподілів.

Вже на P-P діаграмі суттєві відхилення ймовірностей емпіричної функції розподілу від значень функції розподілу хі-квадрат. Але у випадку експоненційного розподілу, точки розташовані вздовж бісектриси першого координатного кута. Тому надалі будемо проводити перевірку узгодженості розподілу для експоненційного.

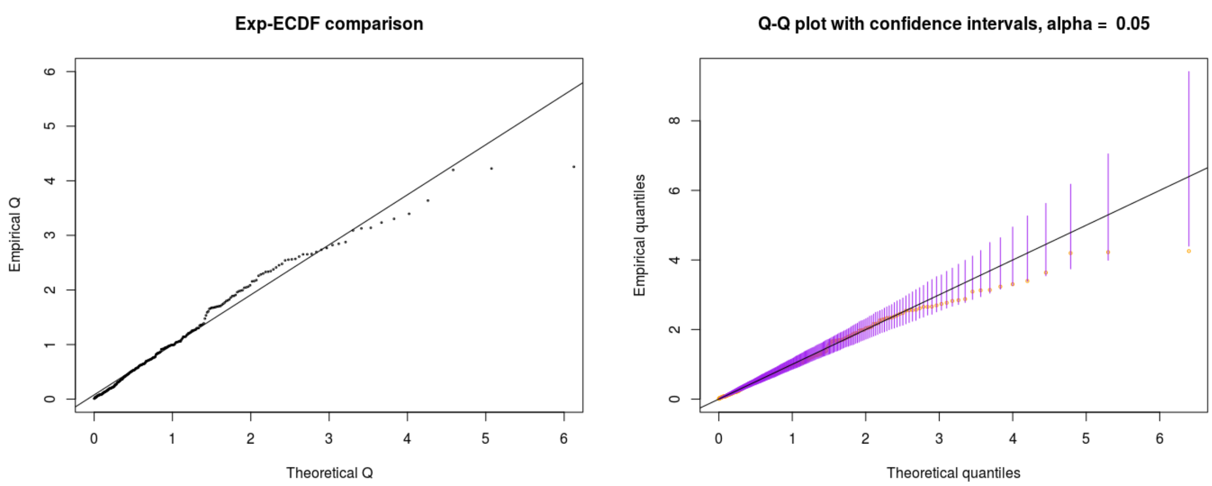


Рис. 5: Q-Q діаграми для експоненційного розподілу. На рисунку справа побудовані прогнозні інтервали рівня $\alpha = 0.05$.

З Q-Q діаграми видно, що точки близько розташовані до бісектриси, однак розкид посилюється, якщо йти вздовж прямої вгору. Також показано, що більшість точок, окрім незначної кількості справа, лежать в межах прогнозних інтервалів. Отже спостереження добре описуються експоненційним розподілом з параметром інтенсивності, який був оцінений за допомогою методу моментів.

4 Гістограма з прогнозними інтервалами для відносних частот.

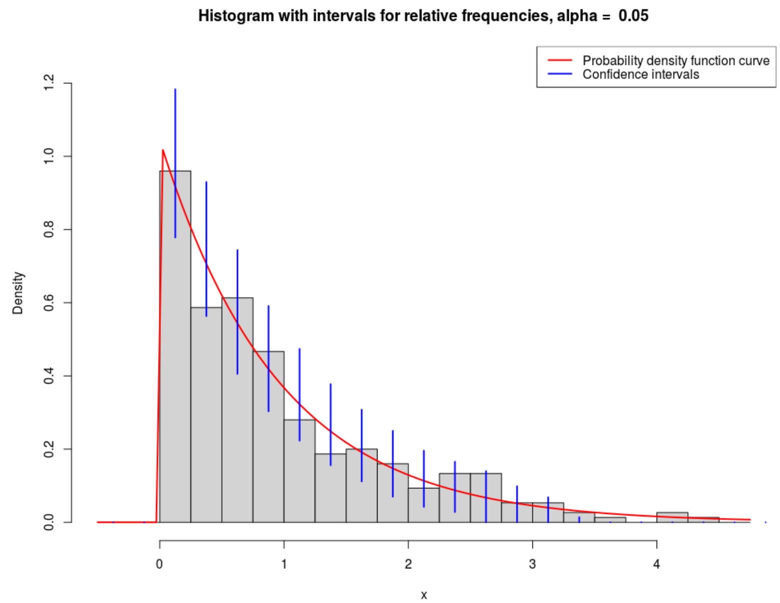


Рис. 6: Гістограма з прогнозними інтервалами для відносних частот. Додатково побудована крива щільності експоненційного розподілу.

У цьому розділі лише покажемо алгоритм побудови гістограми з пронозними інтервалами для відносних частот:

1. Маючи деякий розподіл, згенерувати k копій вибірок обсягу N;
2. Визначити межі носія $\mathcal{P}$, на якому зосереджені спостереження, не враховуючи викиди;
3. Задати розбиття, кількість K та ширину h інтервалів на отриманому носії;
4. Для кожної копії вибірки зберегти дані про відносні частоти;
5. Для кожного інтервалу P_j зібрати відносні частоти усіх копій, обчислити квантили рівня $\alpha, 1-\alpha$;
6. Побудова гістограми відносних частот та кривої щільності заданого розподілу;
7. Маючи масиви квантилів заданих рівнів, побудувати вертикальні відрізки для кожного стовпця гістограми - це прогнозні інтервали відносних частот.

5 Висновки.

Завдяки графічним методам та коректно знайденим оцінкам невідомих параметрів розподілу, для запропонованих даних вдалося підібрати теоретичний розподіл спостережень. Цей розподіл виявився експоненційним. Реалізація алгоритму побудови гістограми з інтервалами відносних частот виявився хорошим, однак при генеруванні досить великої кількості послідовностей обчислення тривають довго (помітно вже при $k > 10^4$). Тому програмну реалізацію ще можна покращити у деяких місцях.