Лабораторна роботи №3
дисципліни "комп'ютерна статистика"
Варіант №4

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

21 листопада 2020 р.

1 Вступ.

У даній роботі вказана інформація про отримані результати під час виконання роботи №3: знайдена оцінка методу моментів для невідомої дисперсії однієї з компонент суміші двох нормальних розподілів. Ця оцінка була реалізована в середовищі мови програмування R. Обчислено граничну дисперсію оцінки за теоретичною формулою та програмними методами. В кінці наведені дані про довірчі інтервали невідомої дисперсії. Додатково наводиться застосування методу медіан для невідомого параметру заданого розподілу.

2 Теоретичні відомості.

Узагальнений метод моментів є одним із способів оцінювання невідомих параметрів розподілу. Розглядається кратна вибірка $\Xi = (\xi_1, \ldots, \xi_n)$, $\xi_j \in \mathbb{R}^p$ з розподілом

$$\mathbb{P}_\nu(A) = \mathbb{P}\left(\xi_j \in A\right),$$

тут $\nu \in \Theta \subset \mathbb{R}^d$ - невідомий параметр, який необіхдно оцінити. Для цього задамо вимірну функцію $h : \mathbb{R}^p \to \mathbb{R}^d$ таким чином, щоб для довльного $t \in \Theta$:

$$H(t) = \mathbb{M}_t\left[h(\xi_1)\right] < \infty$$

При великих обсягах n вибірки Ξ, внаслідок закону великих чисел, $\hat{\mu}_{n,h} = \sum_{j=1}^n h(\xi_j) \approx H(\nu)$. Розглянемо $\hat{\mu}_{n,h} = H(t)$, $t \in \Theta$. Статистика $\hat{\nu}_n$, яка при підстановці в останній вираз дає рівність майже напевно, називається моментною оцінкою параметра ν. Тут $H, \hat{\mu}_{n,h}$ - узагальнені теоретитичний та емпіричний моменти відповідно.

Теорема (Про асимптотичну нормальність оцінки УММ). Нехай ν - невідомий d-вимірний параметр і використовується консистентна моментна оцінка $\hat{\nu}_n$ з моментною функцією $h(\xi) = (h_1(\xi), \ldots, h_d(\xi))^\top$, і вектором теоретичних моментів —

$$H(t) = (H_1(t), \ldots, H_d(t))^\top = \mathbb{M}_t \left[h(\xi_1)\right], \; t = (t_1, \ldots, t_d)^\top \in \Theta.$$

Через $H^{'}(t)$ позначимо матрицю Якобі:

$$H^{'}(t) = \frac{\partial}{\partial t^\top} H(t) = \begin{pmatrix} \frac{\partial H_1(t)}{\partial t_1} & \cdots & \frac{\partial H_1(t)}{\partial t_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial H_d(t)}{\partial t_1} & \cdots & \frac{\partial H_d(t)}{\partial t_d} \end{pmatrix}$$

Якщо виконуються наступні умови:

1. Елементи коваріаційної матриці $D_\nu = \mathrm{Cov}\left(h(\xi_1)\right)$ є скінченними;
2. Існує обернена функція H^{-1};
3. Функція $H^{'}(t)$ є неперервною по t у деякому околі ν.

Тоді $\hat{\nu}_n$ є асимптотично нормальною з матрицею розсіювання

$$V_{\hat{\nu}}(\nu) = (H^{'}(\nu))^{-\top} D_\nu (H^{'}(\nu))^{-1}.$$

Якщо $d = 1$, тоді вищезгадана формула перетворюється на вираз

$$V_{\hat{\nu}}(\nu) = \frac{\mathbb{D}_\nu \left[h(\xi_1)\right]}{(H(\nu)')^2}$$

3 Пошук моментної оцінки параметру.

У нашому випадку, задано гауссову суміш ζ з двох компонент ($\xi_1 \sim N(\mu_1, \theta^2), \xi_2 \sim N(\mu_2, \sigma^2)$), де імовірності змішування є сталими та дорівнюють p, $q = 1 - p$. Дисперсія розподілу першої компоненти суміші є невідомою, тому її необхідно оцінити. Для цього застосуємо узагальнений метод моментів, поклавши в якості моментної функції $h(t) = t^2$. Такий крок пояснюється досить просто: перший теоретичний момент заданої суміші не містить невідомий параметр. Тому обчислимо другий теоретичний момент:

$$\mathbb{M}\left[\zeta^2\right] = \int_{\mathbb{R}} t^2 \left(p f_{\xi_1}(t) + q f_{\xi_2}(t)\right) dt = \left| \text{Лінійність інтеграла Лебега} \right| =$$
$$= p\mathbb{M}\left[\xi_1^2\right] + q\mathbb{M}\left[\xi_2^2\right] = p\left(\mu_1^2 + \theta^2\right) + q\left(\mu_2^2 + \sigma^2\right)$$

Нехай $\hat{\mu}_{2,n} = \dfrac{1}{n}\sum_{j=1}^n \zeta_j^2$, де $(\zeta_j)_{j=1}^n$ є н.о.р., $\zeta_1 \stackrel{\mathrm{d}}{=} \zeta$. Отримавши вираз для другого моменту суміші, тепер розв'яжемо рівняння методу моментів відносно невідомої дисперсії θ^2:

$$p\left(\mu_1^2 + \theta^2\right) + q\left(\mu_2^2 + \sigma^2\right) = \hat{\mu}_{2,n} \Leftrightarrow \hat{\sigma}^2_{MM,n} = \frac{1}{p}\left(\hat{\mu}_{2,n} - q\left(\mu_2^2 + \sigma^2\right)\right) - \mu_1^2$$

У нашому випадку були задані конкретні значення для ймовірностей змішування, математичних сподівань компонент суміші та дисперсію ξ_2, тобто $p = 0.6, \mu_1 = 1, \; \mu_2 = 0, \; \sigma^2 = 0.75^2$. Тоді оцінка методу моментів набуває вигляду:

$$\hat{\sigma}^2_{MM,n} = 0.6^{-1} \cdot \left(\hat{\mu}_{2,n} - 0.4 \cdot 0.75^2\right) - 1 = \frac{5}{3} \cdot \hat{\mu}_{2,n} - \frac{11}{8}$$

Щоб не заплутатися у записах, підставляти числові значення відомих параметрів будемо лише тоді, коли це необхідно.

4 Обчислення теоретичного значення граничної дисперсії моментної оцінки.

Для обчислення граничної дисперсії перевіримо виконання умов теореми про асимптотичну нормальність оцінки методу моментів, яка була сформульована в другому розділі. Спочатку беремо похідну від $H(t) = p\left(\mu_1^2 + t\right) + q\left(\mu_2^2 + \sigma^2\right)$ по t: $\left(H(t)\right)_t' = p$, $t > 0$. Тому похідна H є неперевною на всій області визначення. Знайдемо обернену до H функцію, розв'язавши відповідне рівняння (тут m береться з області значень H):

$$m = p\left(\mu_1^2 + t\right) + q\left(\mu_2^2 + \sigma^2\right) \Leftrightarrow \frac{1}{p}\left(m - q\left(\mu_2^2 + \sigma^2\right)\right) - \mu_1^2 = t$$

Тому обернена функція до H існує і дорівнює $H^{-1}(m) = \dfrac{1}{p}\left(m - q\left(\mu_2^2 + \sigma^2\right)\right) - \mu_1^2$. Залишається показати, що $\mathbb{D}\left[\zeta^2\right] < \infty$. Відомо, що $\mathbb{D}\left[\zeta^2\right] = \mathbb{M}\left[\zeta^4\right] - \left(\mathbb{M}\left[\zeta^2\right]\right)^2$. Другий теоретичний момент було знайдено раніше, а обчислення четвертого буде важчим. Доцільно порахувати математичне сподівання "в лоб" для гаусового розподілу, а потім використати результат для суміші (тут $\eta \sim N(b, s^2), \eta_0 \sim N(0,1)$; $b \in \mathbb{R}$, $s > 0$):

$$\begin{aligned}
\mathbb{M}\left[\eta^4\right] &= \frac{1}{\sqrt{2\pi}s}\int\limits_{\mathbb{R}} t^4 \exp\left(-\frac{(t-b)^2}{2s^2}\right)dt = \left|z = \frac{t-b}{s}\right| = \frac{1}{\sqrt{2\pi}}\int\limits_{\mathbb{R}} (sz+b)^4 \exp\left(-\frac{z^2}{2}\right)dz = \\
&= s^4\mathbb{M}\left[\eta_0^4\right] + 4\cdot s^3 b\cdot 0 + 6\cdot (sb)^2\cdot 1 + 4\cdot sb^3\cdot 0 + b^4 = s^4\mathbb{M}\left[\eta_0^4\right] + 6s^2b^2 + b^4 = \\
&= 3s^4 + 6s^2b^2 + b^4 < \infty,
\end{aligned}$$

де четвертий момент стандартного нормального розподілу нескладно обчислити:

$$\mathbb{M}\left[\eta_0^4\right] = \sqrt{\frac{2}{\pi}}\int\limits_0^\infty z^4 \exp\left(-\frac{z^2}{2}\right)dz = \left|u = \frac{z^2}{2}\right| = \frac{4}{\sqrt{\pi}}\int\limits_0^\infty u^{\frac{3}{2}}\exp(-u)du = \frac{4}{\sqrt{\pi}}\Gamma\left(\frac{5}{2}\right) = 3$$

Звідси маємо, що $\mathbb{D}\left[\zeta^2\right] < \infty$, оскільки

$$\mathbb{M}\left[\zeta^4\right] = p\mathbb{M}\left[\xi_1^4\right] + q\mathbb{M}\left[\xi_2^4\right] < \infty$$

Отже ми показали, що умови теореми виконані, тому оцінка $\hat{\sigma}^2_{MM,n}$ є асимптотично нормальною з граничною дисперсією $\mathbb{D}\left[\zeta^2\right]/(H(\theta^2)')^2$. Якщо підставити отримані вирази у запис граничної дисперсії, то отримаємо:

$$\begin{aligned}
V(\theta^2) = \frac{\mathbb{D}\left[\zeta^2\right]}{(H(\theta^2)')^2} &= \frac{p\mathbb{M}\left[\xi_1^4\right] + q\mathbb{M}\left[\xi_2^4\right] - \left(\mathbb{M}\left[\zeta^2\right]\right)^2}{p^2} = \\
&= \frac{p(3\theta^4 + 6\theta^2\mu_1^2 + \mu_1^4) + q(3\sigma^4 + 6\sigma^2\mu_2^2 + \mu_2^4) - \left(p\left(\mu_1^2 + \theta^2\right) + q\left(\mu_2^2 + \sigma^2\right)\right)^2}{p^2}
\end{aligned}$$

Якщо підставити задані числа відомих параметрів розподілу, то коефіцієнт розсіювання за теоретичною формулою, припускаючи що справжнє значення дисперсії $\theta^2 = 0.05^2$, приблизно дорівнює 0.8488792. Значення не зовсім хороше, але не є поганим. Розглянемо в наступному розділі поведінку отриманої оцінки на змодельованих даних.

5 Оцінка якості статистики за допомогою імітаційного експерименту.

Далі будемо вважати, що $\theta^2 = 0.05^2$ - справжнє значення невідомої дисперсії у першій компоненті суміші, $N = \{100, 250, 500, 1000, 2000, 5000, 10000\}$. Наступним завданням було дослідити якість оцінки на змодельованих даних. Для кожного $n \in N$ було згенеровано по $B = 1000$ кратних вибірок обсягу n. Для кожної з цих вибірок були пораховані моментні оцінки $\hat{\sigma}^2_{MM,n,B}$ та внесені у масив $\hat{\Sigma}^2_{MM,n} = \{\hat{\sigma}^2_{MM,1,B}, \ldots, \hat{\sigma}^2_{MM,n,B}\}$. На основі отриманих числових значень у масиві $\hat{\Sigma}^2_{MM,n}$, мали наближені значення для зміщення та коефіцієнту розсіювання оцінки. Для цього використали вибіркове середнє та нормовану вибіркову дисперсію, тобто $\sqrt{n}\,(\hat{\mu}_n - \theta^2)$ та $n\hat{\sigma}^2_{0,n}$. Нижче покажемо вибіркові значення дисперсії та зміщення при деяких обсягах вибірки.

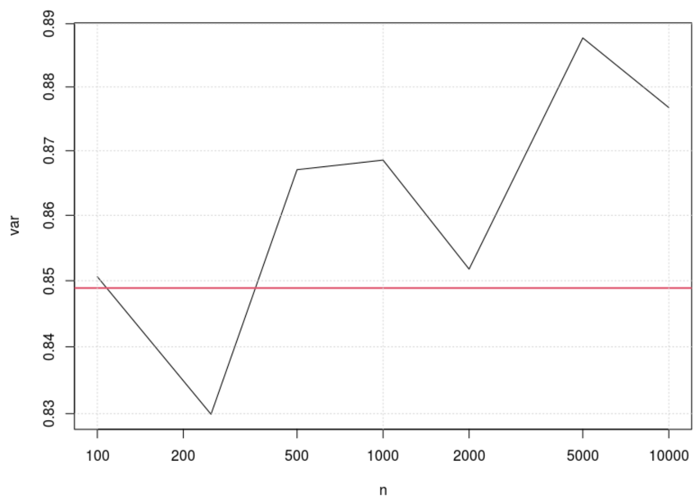

Рис. 1: Графік зміни коефіцієнту розсіювання за вибіркою в залежності від її обсягу. Червона лінія розміщена на рівні теоретичного значення коефіцієнту розсіювання.

```
"theoretical: 0.8489"
"#####"
"obtained: 0.8299"
"bias: 0.0129"
"n: 500"
"obtained: 0.8671"
"bias: -0.0008"
"n: 1000"
"obtained: 0.8685"
"bias: 0.0023"
"n: 2000"
"obtained: 0.8518"
"bias: 0.0148"
```

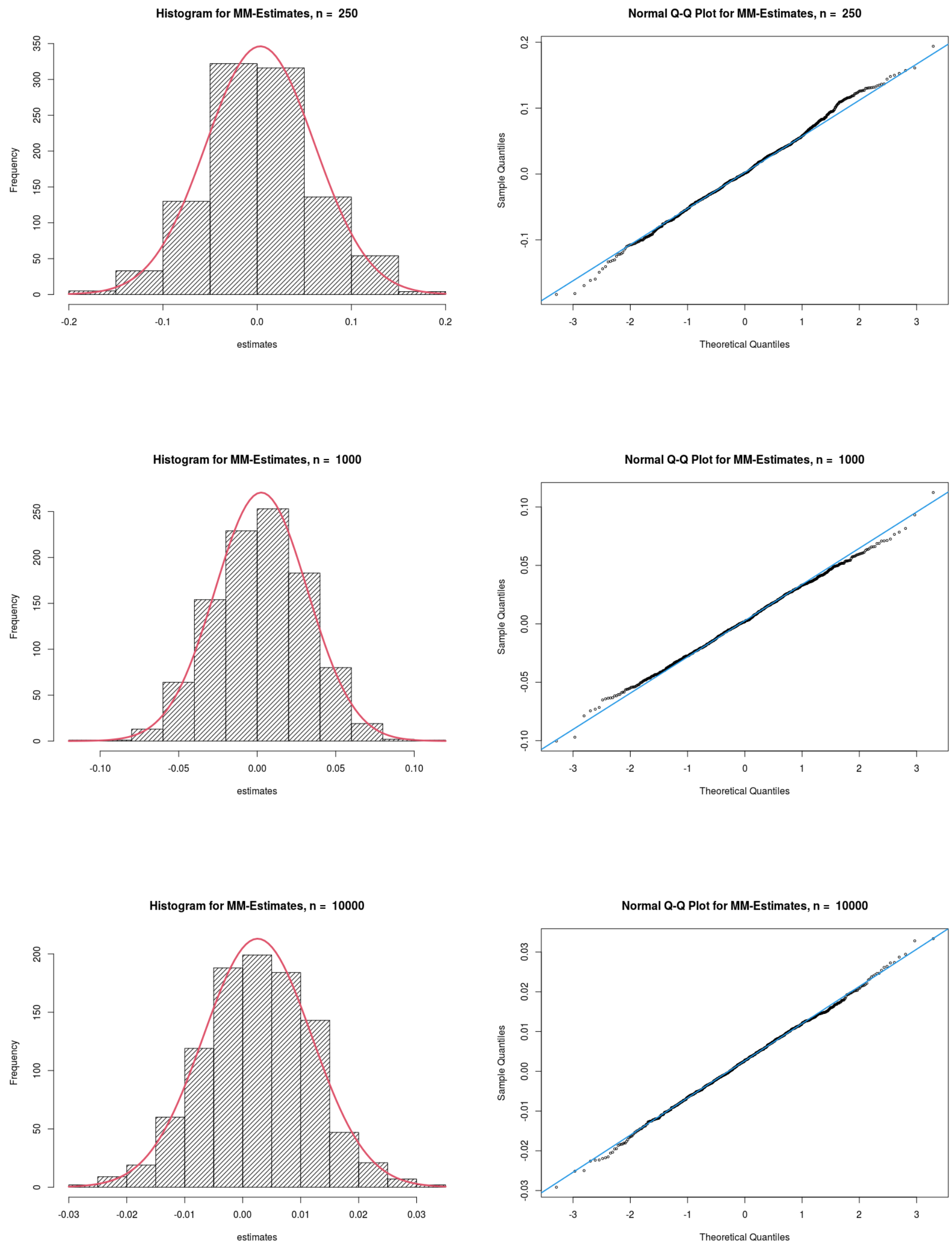

Рис. 2: Гістограми та QQ-діаграми для значень оцінок з B вибірок різного обсягу. Показані при $n = 250, 1000, 10000$.

З рисунків видно, що непогане наближення до нормального розподілу вже маємо при $n \geqslant 250$.

6 Побудова асимптотичних довірчих інтервалів. Рівень значущості інтервалу.

Маючи теоретичне значення асимптотичної дисперсії оцінки методу моментів для суміші з двох компонент, можемо побудувати асимптотичний довірчий інтервал для дисперсії першої компоненти, тобто такий проміжок $[\hat{\theta}_{1,n}, \hat{\theta}_{2,n}]$, що

$$\lim_{n\to\infty} \mathbb{P}\left(\theta^2 \in [\theta_{1,n}, \theta_{2,n}]\right) = 1 - \alpha,$$

де $\alpha \in (0,1)$ - фіксований рівень значущості. У якості кінців такого проміжку, покладемо

$$\hat{\theta}_{j,n} = \hat{\sigma}^2_{MM,n} + (-1)^j q_{\frac{\alpha}{2}} \sqrt{\frac{V(\hat{\sigma}^2_{MM,n})}{n}},\ j = 1, 2$$

де $V(\cdot)$ - асимптотична дисперсія отриманої оцінки, $q_{\frac{\alpha}{2}}$ визначається з рівності:

$$\mathbb{P}\left(|\xi| < q_{\frac{\alpha}{2}}\right) = 1 - \alpha,\ \xi \sim N(0, 1)$$

Програмна реалізація має вигляд:

```
conf.interval.mm <- function(x, mu.1, mu.2, s.2, p, alpha = 0.05)
{
  mm.estimate <- est.mm(x, mu.1, mu.2, s.2, p)
  quant.norm <- qnorm(1 - alpha/2)
  asympt.v <- asympt.var.bicycle(mm.estimate, mu.1, mu.2, s.2, p)
  c.interval <- mm.estimate + c(-1,1) * quant.norm * sqrt(asympt.v/length(x))
  c.interval
}
```

Визначення точності побудованого інтервалу виконана за допомогою імітаційного експерименту, зробивши 2000 копій вибірок з гауссової суміші обсягу 1000 одиниць кожна. Для кожного інтервалу перевіряємо, чи потрапляє справжнє значення параметру до нього чи ні. В кінці обчислюємо частоту похибки.

```
N <- 2000
B <- 1000
counts <- c()
for(i in 1:B)
{ # Далі генеруємо величини з гауссової суміші, будуємо відповідний інтервал
  u.mxt <- rgaussmixt(N, given.mu.1, given.mu.2, true.theta, given.sigma.2, given.p)
  intervals <- conf.interval.mm(
    u.mxt, given.mu.1, given.mu.2, given.sigma.2, given.p, alpha = 0.05
  ) # В кінці перевіряємо чи належить значення параметру до інтервалу
  counts <- c(counts, intervals[1] < true.theta^2 && true.theta^2 < intervals[2])
}
print(1 - mean(counts))
```

В результаті маємо, що при менших обсягах вибірки (наприклад, до 1500 елементів), побудовані інтервали виявляються нестрогими. При $n > 1500$, частота досить коливається близько теоретичного значення. Наприклад, при $\alpha = 0.05$, $N = 2000$ та B отримали частоту похибок рівній 0.048.

7 Висновки.

Побудована оцінка методу моментів та перевірена її якість за допомогою методу імітаційного моделювання. Виявилося, що отримана статистика є асимптотично нормальною оцінкою невідомого параметра дисперсії другої компоненти суміші. Вибіркова дисперсія при більших обсягах вибірки не збігається до теоретичної граничної дисперсії, але коливається близько цього значення (різниця лише в сотих). Асимптотичні довірчі інтервали, отримані на основі властивостей оцінки методу моментів, хороші, коли вибірка складається з великої кількості елементів.