Розрахункова Робота 2
з математичної статистики
Варіант 133

Воробйов Георгій

23 травня 2021 р.

Зміст

0 Завдання

1. Проведіть первинний аналіз вибірки. Це включає статистичний ряд (для розподілів — інтервальний), емпіричну функцію розподілу (для неперервних розподілів інтервальну), її графік, полігон частот (для дискретних розподілів), гістограму (неперервнихрозподілів), box-and-whisker plot.

2. Знайдіть вибіркове середнє, вибіркову дисперсію, виправлену вибіркову дисперсію, вибіркову медіану, вибіркову моду, вибіркові коефіцієнти асиметрії та ексцесу.

3. Обґрунтуйтета висуньте (нову) гіпотезу про розподіл генеральної сукупності.
4. Методом моментів та методом максимальної вірогідності знайдіть оцінки параметрів розподілу. В деяких випадках це може бути не дуже просто (як, наприклад, для параметра N біноміальної генеральної сукупності). Це чудовий спосіб проявити креативність та/або вміння користуватися Google.
5. Для кожного параметра кращу з цих двох оцінок перевірте на (асимптотичну)незміщеність, консистентність та ефективність.
6. Побудуйте довірчі інтервали надійністю 0.95 для параметрів розподілу.
7. Нарешті, перевірте висунуту гіпотезу про розподіл генеральної сукупності за допомогою критерію χ^2
8. Проявіть всі свої літературні здібності та напишіть висновки

Задана вибірка:
2 1 1 4 4 3 4 3 2 7 6 1 5 3 3 1 4 3 2 3 2 3 2 3 4 5 3 5 5 1 2 3 6 3 5 5 2 5 2 2 0 3 0 2 6 2 3 4 3 2 4 1 4 3 4 2 4 1 4 5 5 3 3 3 2 4 3 2 4 3 3 3 3 4 2 3 6 1 2 3 3 4 0 3 5 1 4 4 3 3 1 3 3 6 2 2 3 2 5 3

1 Розв'язок

1.1 Вступ

Запишемо відсортовану вибірку:
0 0 0 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 7

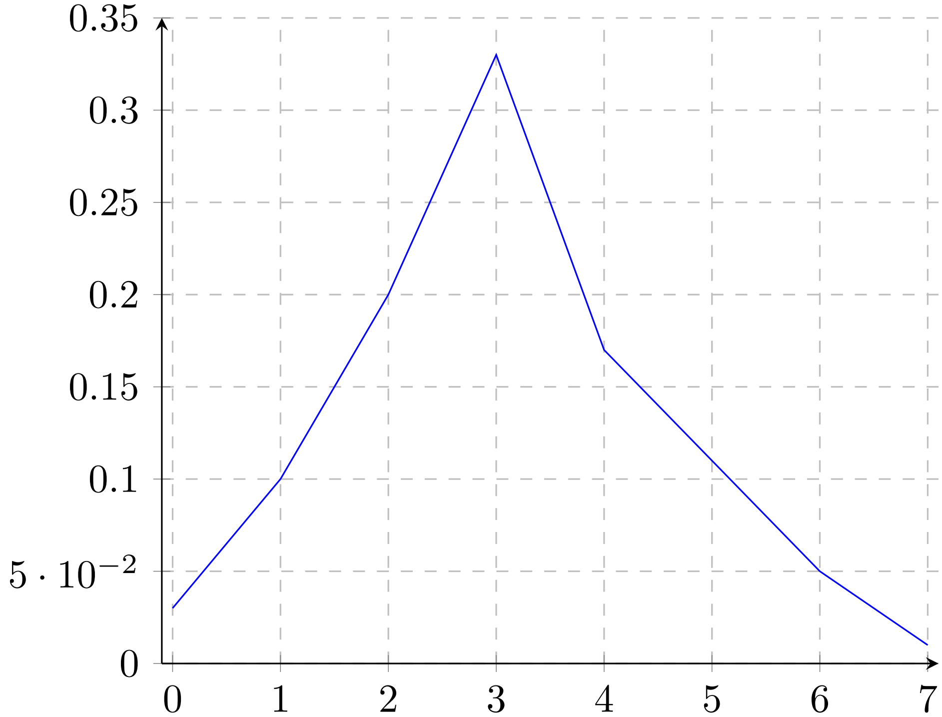
1.2 Аналіз вибірки

Побудуємо статистичний ряд даної вибірки

елементи	Частота n_i	Кумулятивна частота n_i^*	Відносна частота ν_i	Відносна кумулятивна частота ν_i^*
0	3	3	0.03	0.03
1	10	13	0.1	0.13
2	20	33	0.2	0.33
3	33	66	0.33	0.66
4	17	83	0.17	0.83
5	11	94	0.11	0.94
6	5	99	0.05	0.99
7	1	100	1	1

За даними таблиці можемо побудувати полігон відносних частот та емпіричну функцію розподілу

Рис. 1: Полігон частот



Маємо наступну емпіричну функцію розподілу

$$F_n^*(x) = \begin{cases} 0 & x \leq 0 \\ 0.03 & 0 < x \leq 1 \\ 0.13 & 1 < x \leq 2 \\ 0.33 & 2 < x \leq 3 \\ 0.66 & 3 < x \leq 4 \\ 0.83 & 4 < x \leq 5 \\ 0.94 & 5 < x \leq 6 \\ 0.99 & 6 < x \leq 7 \\ 1 & x > 7 \end{cases}$$

Відповідний її графік:

Рис. 2: Епірична функція розподілу

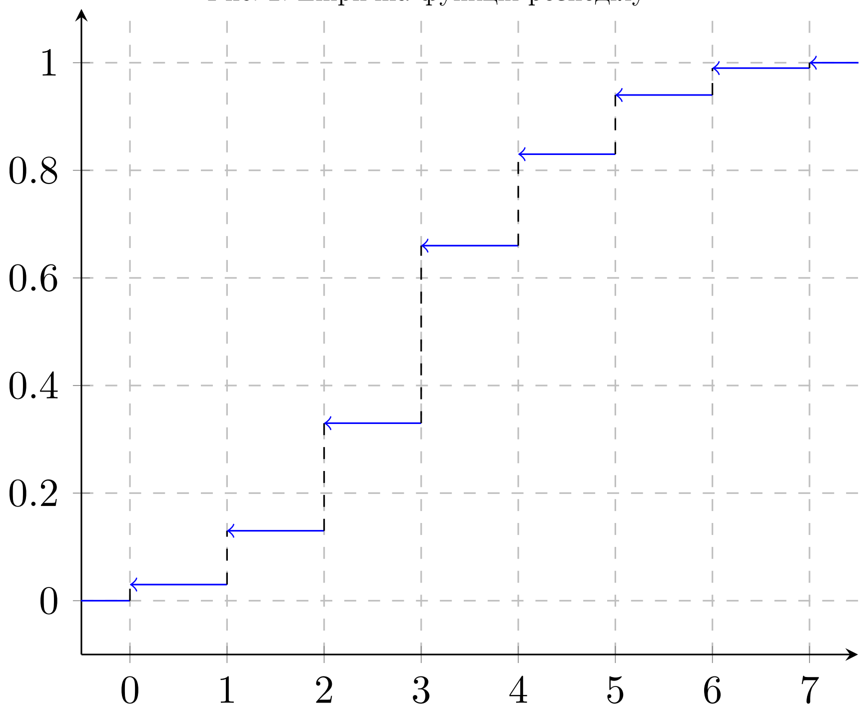
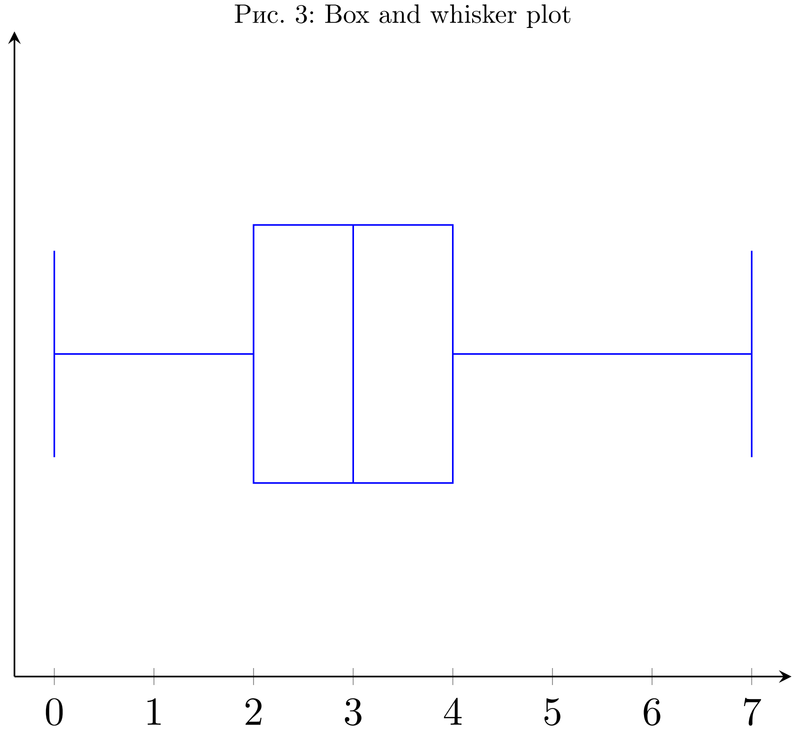


Рис. 3: Box and whisker plot



1.3 Вибіркові статистики

Порахуємо значення вибіркового середнього:

$$\overline{\xi} = \frac{1}{n}\sum_{i=1}^{100}\xi_i$$

$$\overline{\xi}_{\text{Знач}} = 3.09$$

Порахуємо значення вибіркової дисперсії:

$$\mathbb{D}^{**} = \frac{1}{n}\sum_{i=1}^{100}(\xi_i - \overline{\xi})^2$$

$$\mathbb{D}^{**}\xi_{\text{Знач}} = 2.08$$

Порахуємо значення виправленої вибіркової дисперсії:

$$\mathbb{D}^{***} = \frac{n}{n-1}\mathbb{D}^{**}$$

$$\mathbb{D}^{***}\xi_{\text{Знач}} = 2.1$$

Порахуємо вибіркову медіану:

$$Me^{*}\xi_{\text{Знач}} = \langle\text{середина вибірки}\rangle = 3$$

Порахуємо вибіркову моду - значення вибірки, що зустрічається найчастіше:

$$Mo^{*}\xi = 3$$

Для розрахунку вибіркових коефіцієнтів асиметрії та ексцесу розрахуємо потрібні початкові та центральні вибіркові моменти

$$As^{*}\xi = \frac{\mu_3^*}{(\sigma^*)^3}$$

$$Ex^{*}\xi = \frac{\mu_4^*}{(\sigma^*)^4} - 3$$

$$\mu_3^* = \mathbb{E}\left[\xi - \overline{\xi}\right]^3 = 0.667$$

$$(\mu_3^*)_{\text{Знач}} = 0.667$$

$$\mu_4^* = \mathbb{E}\left[\xi - \overline{\xi}\right]^4 = 12.429$$

$$(\mu_4^*)_{\text{Знач}} = 12.429$$

$$\sigma^* = \sqrt{\mathbb{D}^{**}} = 1.443$$

$$\sigma^*_{\text{Знач}} = 1.443$$

$$As^*\xi_{\text{Знач}} = 0.221$$

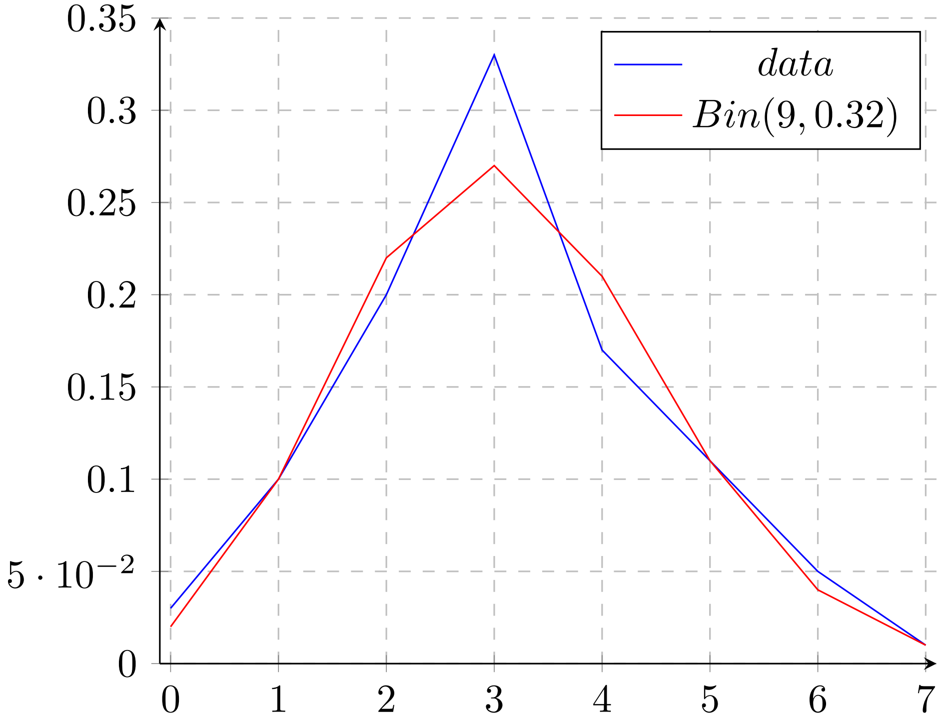
$$Ex^*\xi_{\text{Знач}} = -0.14$$

1.4 Висунення Гіпотези

1. Маємо дискретну ГС
2. Значення дисперсії менше ніж математичного сподівання(Що схоже на npq та np відповідно)
3. Має вибіркову медіану рівну вибірковій моді

Тож можемо припустити, що данна ГС є розподіленою біноміально
$\xi \sim Bin(N, p)$

Рис. 4: Полігон частот



1.5 Оцінки параметра розподілу

Оцінка параметрів біноміального розподілу при невідомих N і p є дуже складною задачею. По-перше зазначимо наступну теорему(без доведення)

Теорема 1 *Нехай $\xi_1, \dots, \xi_n$ є вибіркою з Генеральної сукупності, розподіленої за біноміальним законом, з невідомими значеннями n та p. Тоді*

1. *Якщо $g(n)$ є неконстантною функцією, то не існує незміщенної оцінки $g(n)$*

2. *Якщо $g(p)$ є такою, що в околі точки 0 існує $g'(p)$ таким чином, що $\lim_{p \to 0} g'(p)$ не дорівнює нулю або нескінченності. Тоді не існує незміщенної оцінки $g(p)$.*

Це означає, що не можливо знайти незміщену оцінку параметра n та p. Фішер не вважав значною проблему пошуку оцінки n, вважаючи, що достатньо оцінки $\max \xi_i$, але, як показують досліди, для отримання навіть $P\left\{\max \xi_i \geq \frac{N}{2}\right\} \geq 0.5$ потрібна вибірка розміром приблизно 31500 для реальних значень параметрів $n = 100$, $p = 0.3$

1.5.1 Метод моментів

Маємо відомі нам значення матсподівання та дисперсії:

$$\begin{cases} \mathbb{E}\xi = Np \\ \mathbb{D}\xi = Np(1-p) \end{cases}$$

Звідси, підставивши значення $\mathbb{E}^*\xi = \overline{\xi}$ та D^{***} можемо виразити значення N* та p* :

$$\begin{cases} 1 - p^* = \frac{\mathbb{D}^{***}\xi}{\overline{\xi}} \\ \overline{\xi} = N^* p^* \end{cases}$$

$$\begin{cases} p^* = \frac{\overline{\xi} - \mathbb{D}^{***}\xi}{\overline{\xi}} \\ \overline{\xi} = N^* p^* \end{cases}$$

$$\begin{cases} p^* = \frac{\overline{\xi} - \mathbb{D}^{***}\xi}{\overline{\xi}} \\ N^* = \frac{\overline{\xi}^2}{\overline{\xi} - \mathbb{D}^{***}\xi} \end{cases}$$

Для значення нашої вибірки отримуємо значення p^* та N^*

$$\begin{cases} p^* = \frac{3.09 - 2.1}{3.09} \\ N^* = \frac{3.09^2}{3.09 - 2.1} \end{cases}$$

$$\begin{cases} p^* = 0.32 \\ N^* = 9.64 \end{cases}$$

1.5.2 Метод максимальної правдоподібності

Запишемо функцію Правдоподібності

$$\mathcal{L}(\overrightarrow{x}, N, p) = \prod_{i=1}^{n} \mathbb{P}\{\xi = x_i\} = \prod_{i=1}^{n} C_N^{x_i} p^{x_i} (1-p)^{N-x_i}$$

$$\ln \mathcal{L}(\overrightarrow{x}, N, p) = \sum_{i=1}^{n} lnC_N^{x_k} + \ln p \sum_{i=1}^{n} x_k + \ln(1-p) \left(nN - \sum_{i=1}^{n} x_k \right)$$

Так як ми не можемо за допомогою даного методу оцінити значення параметра N, бо C_N^k не є неперервною функцією, оцінимо лише параметр p.

$$\frac{\partial \ln \mathcal{L}}{\partial p} = \frac{1}{p} \sum_{i=1}^{n} x_k - \frac{1}{1-p} \left(nN - \sum_{i=1}^{n} x_k \right)$$

Прирівнюючи до нуля знайдемо оцінку p.

$$(\frac{1}{p^*} + \frac{1}{1-p^*}) \sum_{i=1}^{n} x_k = \frac{nN}{1-p^*}$$

$$\frac{1}{p^*(1-p^*)} \sum_{i=1}^{n} x_k = \frac{nN}{1-p^*}$$

$$p^* = \frac{\sum_{i=1}^{n} x_k}{nN} = \frac{\overline{x}}{N}$$

Як бачимо, для оцінки p методом максимальної правдоподібності, ми повинні знайти значення N. Для цього розглянемо інші оцінки.

1.5.3 Інші методи оцінки параметру

Інший метод моментів Розглянемо метод моментів, який вигористовує $\max \xi_i, \overline{\xi}, \mathbb{D}^{***}\xi$. Розглянемо наступну рівність:

$$N = \frac{N^{\alpha+1}(Npq)^{\alpha}}{(Np)^{\alpha}(Np)^{\alpha}}$$

Підставивши у якості n будь-яку оцінку, наприклад $\max \xi_i$, а у якості Np - значення матсподівання, $Npq = \mathbb{D}^{***}\xi$

$$N^* = \frac{(\max \xi_i)^{\alpha+1} (\mathbb{D}^{***}\xi)^{\alpha}}{\overline{\xi}^{\alpha} (\max \xi_i - \overline{\xi})^{\alpha}}$$

Параметр α вибирається самостійно експериментатором, у [**1**] показано, що одним із найкращих значень є $\alpha = 1$, для нього маємо оцінку, яку і порахуємо:

$$N^* = \frac{(\max \xi_i)^2 (\mathbb{D}^{***}\xi)}{\overline{\xi}(\max \xi_i - \overline{\xi})}$$

$$N^*_{\text{Знач}} = \frac{7^2(2.1)}{3.09(7-3.09)} = 8.52$$

Оцінка Керолла-Ломбарда Інша оцінка, яку називають оцінкою Керолла-Ломбарда, є наступна оцінка, яка виводиться з методу максимальної правдоподібності:
У цьому методі ми припускаємо, що p розподілений за бета розподілом з деякимии параметрами (a, b), тоді ми можемо записати функцію правдоподібності

$$L(N) = \prod_{i=1}^{k} C_N^{x_i} \left[(nN + a + b + 1) C_{nN+a+b}^{a+\sum_{j=1}^{n} x_j} \right]^{-1}$$

Дана функція може бути оцінена лише на основі даних, що в нас є, тобто мінімізуємо ї пошуком локальних мінімумів за означенням локального мінімуму.

Покращена оцінка за допомогою максимума Останній метод полягає у тому, що ми додаємо деяке значення, яке визначається за допомогою іншої оцінки:

$$N^{**} = \max \xi_i + \sum_{i=0}^{N^*-2} F_{i+1,N^*-i}^{-1} \left(\frac{1}{n} \right)$$

Де $F_{u,v}(x)$ - фукнція розподілу $Beta(u, v)$.
виведення цієї оцінки розглянуто в [**1**].

1.6 Перевірка параметра на незміщенність, консистентність, ефективність.

Незміщенність Перевіримо на незміщенність оцінку N, отриману за стандартним методом моментів.

Хоча ми і маємо теорему про те, що оцінки N та p не можуть бути незміщенними, але вона виконується лише при невідомих обох параметрах. При відомому N, оцінка для p з ММП є незміщенною.

Але це не наша ситуація, тому перевіримо на асимптотичну незміщенність оцінки:

$$N^* = \frac{\overline{\xi}^2}{\overline{\xi} - \mathbb{D}^{***}\xi}$$

$$p^* = \frac{\overline{\xi}}{N^*}$$

Дослідження двох невідомих параметрів гіпотетичного закону розподілу, отримані оцінки яких пов'язані одна з одною, є досить проблематичним. Цьому питанню можна присвятити цілу наукову статтю. Том у в подільшому дослідженні спочатку зробимо припущення, що параметр N відомий, і будемо досліджувати оцінку параметра p. Після цього дослідимо оцінку параметра N, припустивши що параметр відомий.

$$N^* = \frac{\overline{\xi}}{p}$$

$$p^* = \frac{\overline{\xi}}{N}$$

Перевіримо на незміщенність ці параметри:

$$\mathbb{E}N^* = \mathbb{E}\frac{\overline{\xi}}{p} = \frac{\mathbb{E}\overline{\xi}}{p} = \frac{Np}{p} = N$$

$$\mathbb{E}p^* = \mathbb{E}\frac{\overline{\xi}}{N} = \frac{\mathbb{E}\overline{\xi}}{N} = \frac{Np}{N} = p$$

Консистентність

$$N^* = \frac{\overline{\xi}_i}{p} \xrightarrow{n \to \inf} \frac{Np}{p} = N$$

$$p^* = \frac{\overline{\xi}_i}{N} \xrightarrow{n \to \inf} \frac{Np}{N} = p$$

Отже, обидва параметри є консистентними.

Ефективність Використаємо наслідок з критерію Рао-Крамера.

$$\frac{\partial \mathcal{L}}{\partial p} = C(p)(p^* - p)$$

$$\frac{1}{p}\sum x_k - \frac{1}{1-p}\left(nN - \sum x_k\right) = C(p)\left(\frac{\sum x_k}{nN} - p\right)$$

$$\sum x_k \left(\frac{1}{p} + \frac{1}{1-p}\right) - \frac{nN}{1-p} = C(p)\left(\frac{\sum x_k - nNp}{nN}\right)$$

$$\frac{\sum x_k - nNp}{p(1-p)} = C(p)\frac{\sum x_k - nNp}{nN}$$

$$C(p) = \frac{nN}{p(1-p)}$$

Отже бачимо, що $C(p)$ не залежить від значення реалізації вибірки. Отже оцінка для p є ефективною.

Оцінку для N ми перевірити не можемо через те, що не можемо взяти похідну по N.

1.7 Довірчі інтервали

Побудуємо довірчий інтервал для даних оцінок. Знаємо, що

Для p можемо знайти довірчий інтервал як

$$\mathbb{P}\left\{\frac{|p^* - p|\sqrt{n}}{\sqrt{pq}} < t_\gamma\right\} \geq \gamma$$

Для значення $\gamma = 0.95$ $t_\gamma = 1.96$

Розглянемо квадрат значення під ймовірнісною мірою.

$$\frac{(p^* - p)^2 n}{pq} < t_\gamma^2$$

$$\left(1 + \frac{t_\gamma^2}{n}\right) p^2 - \left(2p^* + \frac{t_\gamma^2}{n}\right) p + (p^*)^2 \leq 0$$

$$1.43p^2 - 1.07p + 0.1024 \leq 0$$

Отримали

$$p \in (0.113, 0.636)$$

Для розрахунку довірчого інтервалу для N маємо наступну характеристику:

$$N^* = \frac{\overline{\xi}}{p} \langle \text{Приймемо } p \text{ за відоме нам значення} \rangle$$

тоді ми можемо порахувати дисперсію N^*:

$$\mathbb{D}N^* = \frac{\mathbb{D}\xi}{pn^2} = \frac{Npq}{np} = \frac{Nq}{n}$$

Звідси

$$\mathbb{P}\left\{-t_\gamma < \frac{N^* - N}{\sqrt{\mathbb{D}N^*}} < t_\gamma\right\} \geq \gamma$$

$$\frac{(N^* - N)^2}{\mathbb{D}N^*} < t_\gamma^2$$

$$(N^* - N)^2 \leq \frac{t_\gamma^2 N(1-p)}{n}$$

$$N^2 - 2NN^* + (N^*)^2 \leq \frac{t_\gamma^2 N(1-p)}{n}$$

Для значення $\gamma = 0.95$ $t_\gamma = 1.96$

$$N^2 - N\left(2N^* + \frac{t_\gamma^2(1-p)}{n}\right) + (N^*)^2 < 0$$

$$N^2 - 18.026N + 81 = 0$$

Отримали

$$N \in (8.529, 9.497)$$

1.8 Критерій про розподіл

Перевіримо, наступну гіпотезу:

$$H_0 = \{\xi \sim Bin(9, 0.32)\}$$

Побудуємо таблицю для χ^2 критерію

Елементи	n_i	np_i	$n_i - np_i$	$\frac{(n_i-np_i)^2}{np_i}$
0	3	2	1	0.5
1	10	10	0	0
2	20	22	-2	0.18
3	33	27	5	0.9
4	17	21	-4	0.76
5	11	11	0	0
6	5	4	1	0.25
7	1	1	0	0
$[8, \infty)$	0	2	-2	2

Бачимо що не для всіх значень виконується $np_i > 10$, тому об'єднаємо деякі діапазони:

Елементи	n_i	np_i	$n_i - np_i$	$\frac{(n_i-np_i)^2}{np_i}$
$[0, 2)$	13	12	1	0.08
$[2, 3)$	20	22	-2	0.18
$[3, 4)$	33	27	5	0.9
$[4, 5)$	17	21	-4	0.76
$[5, \infty)$	17	18	-1	0.05

тоді маємо наступне значення:

$$\chi^2(n) = \sum_{0=1}^{8} \frac{(n_i - np_i)^2}{np_i} = 1.97$$

А

$$\chi^2_{critcal,2,\alpha=0.95} = 7.38$$

Так як наша область правостороння, то, ми не відхилюємо гіпотезу H_0.

1.9 Висновки

У даній роботі ми дослідили реалізацію вибірки із 100 чисел. Було перевірено, що наші дані не суперечать тому, що дана ГС є біноміально розподіленою та знайдено точкові та інтервальні параметри даної ГС. Була наведена теорема, яка показує, що при обох параметрах N та p невідомі, то не існує незміщеної оцінки для цих параметрів. Було перевірено значення N та p на незміщеність, консистентність та ефективність (при умові що ми знаємо інший параметр).

Окрім того були показані приклади інших оцінок параметра N, які можуть дати меншу зміщенність, аніж стандартний метод моментів.

Література

[1] A. DasGupta, Herman Rubin *Estimation of binomial parameters when both n, p are unknown*

[2] R. J. Carroll, F. Lombard *A note on N estimators for the binomial distribution*