Критерій хі-квадрат.
Перевірка простої основної гіпотези.

1 Теоретичні відомості

1.1 Узагальнення схеми випробувань Бернуллі. Поліноміальний розподіл

Припустимо, що проводиться n незалежних випробувань Бернуллі, у кожному з яких може трапитися один із r результатів. Імовірність того, що в одному випробуванні матимемо k-ий результат дорівнює p_k, $1 \le k \le r$, де $p_k > 0$, $\sum_{k=1}^{r} p_k = 1$.

Нехай вектор $(X_1, \ldots, X_n)$ результати проведених n випробувань. Через ν_k позначимо кількість появ k-го результату у проведених випробуваннях, тобто

$$\nu_k = \sum_{j=1}^{n} \mathbf{1}\{X_j = k\},\ 1 \le k \le r.$$

Неважко побачити, що $\sum_{k=1}^{r} \nu_k = n$.

Знайдемо розподіл вектора $(\nu_1, \ldots, \nu_r)$. Припустимо, що k-ий результат трапився n_k разів, $1 \le k \le r$, $\sum_{j=1}^{r} n_k = n$. Тоді, ймовірність однієї такої реалізації вектора становить $p_1^{n_1} \cdot \ldots \cdot p_r^{n_r}$. Кількість усіх можливих реалізацій, коли $\{\nu_k = n_k\}$, дорівнює

$$C_{n_1,\ldots,n_r;n} = C_n^{n_1} \cdot C_{n-n_1}^{n_2} \cdot C_{n-n_1-n_2}^{n_3} \cdot \ldots \cdot C_{n-n_1-\ldots-n_{r-1}}^{n_r} =$$

$$= \frac{n!}{n_1!\cancel{(n-n_1)!}} \cdot \frac{\cancel{(n-n_1)!}}{n_2!\cancel{(n-n_1-n_2)!}} \cdot \frac{\cancel{(n-n_1-n_2)!}}{n_3!\cancel{(n-n_1-n_2-n_3)!}} \cdot \ldots \cdot \frac{\cancel{(n-n_1-\ldots-n_{r-1})!}}{n_r!(n-n_1-\ldots-n_r)!} =$$

$$= \frac{n!}{n_1! \cdot \ldots \cdot n_r!(n-n_1-\ldots-n_r)!} = |n-n_1-\ldots-n_r = n-n = 0| = \frac{n!}{n_1! \cdot \ldots \cdot n_r!}$$

Коефіцієнт $C_{n_1,\ldots,n_r;n}$ в обчисленнях вище називають мультиноміальним коефіцієнтом. Отже,

$$\mathbf{P}\left((\nu_1, \ldots, \nu_r) = (n_1, \ldots, n_r)\right) = C_{n_1,\ldots,n_r;n} \cdot p_1^{n_1} \cdot \ldots \cdot p_r^{n_r}. \tag{1}$$

Можна переконатися в тому, що розподіл зосереджено лише на тих n_j, що $\sum_{k=1}^{r} n_j = n$.

Розподіл в (1) називають поліноміальним розподілом (в англ. – multinomial distribution).

Вищеописана модель є узагальненням схеми випробувань Бернуллі з біноміальним розподілом відповідно у разі, коли в одному випробуванні допускається більше, ніж два результати.

Приклад. Нехай $r = 2$. Тоді $\nu_2 = n - \nu_1$, $p_2 = 1 - p_1$, $n_2 = n - n_1$ та

$$\mathbf{P}(\nu_1 = n_1) = \mathbf{P}\left((\nu_1, \nu_2) = (n_1, n_2)\right) = C_{n_1,n_2;n} \cdot p_1^{n_1} \cdot p_2^{n_2} = \frac{n!}{n_1! n_2!} p_1^{n_1} p_2^{n_2} = \frac{n!}{n_1!(n-n_1)!} p_1^{n_1} (1-p_1)^{n-n_1}.$$

1.2 Проста основна гіпотеза про розподіл результатів

Нехай дослідник проводить n незалежних випробувань, у кожному з яких може трапитися один із r результатів. Тобто, він має вибірку $X = (X_1, \ldots, X_n)$ з результатів кожного з випробувань. На основі цієї вибірки він може порахувати абсолютні (вибіркові) частоти $\hat{\nu}_k = \sum_{j=1}^n \mathbf{1}\{X_j = k\}$, $1 \le k \le r$.

Вектор абсолютних частот $(\hat{\nu}_1, \ldots, \hat{\nu}_r)$ має поліноміальний розподіл (співставте з попереднім пунктом). Припускається, що імовірності результатів випробування $\{p_k\}$ вважаються невідомими. Тобто невідомим параметром є вектор імовірностей $\theta = (p_1, \ldots, p_r)$, а параметричним простором є $\Theta = \{(x_1, \ldots, x_r) \mid x_j > 0,\ 1 \le j \le r,\ \sum_{j=1}^r x_k = 1\}$.

Дослідник хоче перевірити гіпотези на розподіл результатів випробування. Припустимо, що він хоче перевірити, чи узгоджений розподіл $(p_1, \ldots, p_r)$ із повністю відомим розподілом $(q_1, \ldots, q_r)$. Тобто, перевіряється основна гіпотеза

$$\mathrm{H}_0 :\ p_k = q_k,\ 1 \le k \le r$$

проти альтернативи

$$\mathrm{H}_1 :\ \text{Гіпотеза } \mathrm{H}_0 \text{ не виконується}$$

Альтернативу можна інтерпретувати наступним чином: існує принаймні один $k_0 = 1, \ldots, r$, для якого $p_{k_0} \ne q_{k_0}$.

Як перевірити ці гіпотези?

1.3 Хі-квадрат тест для перевірки простої основної гіпотези

Побудуємо тест для перевірки статистичних гіпотез, описаних у попередньому пункті.

Розглянемо вибіркову частоту $\hat{p}_k = \nu_k/n$ появи k-го успіху, $1 \le k \le r$. Відомо, що вибіркова частота є строго консистентною оцікою імовірності p_k: $\hat{p}_k \to^{P_1} p_k$ при $n \to \infty$.

Якби мала місце основна гіпотеза, то $\hat{p}_k \to^{P_1} q_k$ при $n \to \infty$ для всіх $k = 1, \ldots, r$. Можна було б очікувати, що для досить великих n, відхилення $(\hat{p}_k - q_k)^2$ були б невеликими. Отже, можна ввести таку 'відстань' між розподілами:

$$\sum_{k=1}^r c_k (\hat{p}_k - q_k)^2, \tag{2}$$

де c_k – деякі коефіцієнти. Якщо розглянути $c_k = n/q_k$, отримаємо

$$\hat{\chi}_n^2 = \sum_{k=1}^r \frac{n}{q_k} (\hat{p}_k - q_k)^2 = \sum_{k=1}^r \frac{(\hat{\nu}_k - nq_k)^2}{nq_k}.$$

Величина $\hat{\chi}_n^2$ називається статистикою χ^2 Пірсона. Величини $\hat{\nu}_k$ та nq_k можна інтерпретувати як спостережувані та очікувані частоти відповідно.

Пояснимо чим зумовлено саме такий вибір c_k. Якщо основна гіпотеза вірна, тоді статистика $\hat{\chi}_n^2$ має слабку границю до невиродженого розподілу:

$$\hat{\chi}_n^2 \to^W \chi_{r-1}^2,\ n \to \infty$$

це дозволяє обирати поріг тесту, тобто базуючись на граничному розподілі статистики. Але з яких міркувань?

Коли вірна альтернатива, тоді принаймні для одного k: $\hat{p}_k - q_k \to^{P_1} p_k - q_k \neq 0$. Значить, принаймні один доданок в сумі

$$\sum_{k=1}^{r} \frac{n}{q_k}(\hat{p}_k - q_k)^2$$

буде необмежено зростати при $n \to \infty$. Отже, підозрілими для дослідника будуть великі значення статистики тесту.

Визначимо **тест хі-квадрат** наступним чином:

$$\pi(X) = \begin{cases} 0, & \hat{\chi}_n^2 < h_\alpha, \\ 1, & \hat{\chi}_n^2 \geq h_\alpha, \end{cases}$$

де $h_\alpha = Q^{\chi_{r-1}^2}(1-\alpha)$. Який рівень значущості такого тесту?

Якщо основна гіпотеза вірна, тоді імовірність помилки першого роду наближено дорівнює

$$\mathbf{P}_{H_0}(\pi(X) = 1) = \mathbf{P}_{H_0}\left(\chi_n^2 \geq h_\alpha\right) \approx \mathbf{P}\left(\chi_{r-1}^2 \geq h_\alpha\right) = \alpha,$$

тобто $\alpha \in (0,1)$ є рівнем значущості тесту π.

Досягнутий рівень значущості. Розглянемо умову, коли тест π приймає основну гіпотезу. Тобто,

$$\chi_n^2 < h_\alpha$$

Якщо остання нерівність має місце, то в силу монотонності $f(t) = \mathbf{P}\left(\chi_{r-1}^2 < t\right)$ матимемо

$$f(\chi_n^2) < f(h_\alpha) = 1 - \alpha$$

Розв'яжемо нерівність відносно α:

$$1 - f(\chi_n^2) > \alpha.$$

Величина p-value $= 1 - f(\chi_n^2)$ є досягнутим рівнем значущості тесту хі-квадрат. У термінах p-value, тест π можна переподати так:

$$\pi(X) = \begin{cases} 0, & \text{p-value} > \alpha, \\ 1, & \text{p-value} \leq \alpha. \end{cases}$$

Коментарі щодо використання хі-квадрат тесту:

1. Хі-квадрат тест є асимптотичним, тобто базується на асимптотичних властивостях $\hat{\chi}_n^2$ при $n \to \infty$. Тому бажано використовувати тест на 'великих' вибірках.
2. Коли спостережень небагато, групи з малою кількістю частот (наприклад, до 5) бажано об'єднувати.

2 Задачі

2.1 Задача 1

З 1871 по 1900 рр. у Швейцарії народилося 1359671 хлопчиків та 1285086 дівчаток.

Чи узгоджується з цими даними гіпотеза

$$\mathrm{H}_0: \text{ імовірність народження хлопчика становить } q = 0.5?$$

Вважати, що рівень значущості становить $\alpha = 0.05$.

Розв'язання. Потрібно перевірити гіпотези на розподіл хлопчиків та дівчаток, використовуючи хі-квадрат тест. Зведемо вхідні дані до потрібної моделі.

Маємо вибірку $X = (X_1, \ldots, X_n)$ з $n = 1359671 + 1285086 = 2644757$ спостережень. Кожне спостереження X_j приймає лише два значення: 1 (хлопчик) або 2 (дівчинка). Всього $r = 2$ результатів. Тобто, маємо вектор невідомих імовірностей $\theta = (p_1, p_2) = (p, 1-p)$, де p вважається ймовірністю народити хлопчика.

В термінах зведеної моделі, відповісти на питання задачі – це перевірити гіпотези вигляду:

$$\mathrm{H}_0: \ p = q = 0.5, \ \mathrm{H}_1: \ p \neq q.$$

Обчислимо значення статистики хі-квадрат тесту. Наведемо всі необхідні величини у табличному вигляді:

k	ν_k	nq_k	$\nu_k - nq_k$	$(\nu_k - nq_k)^2/nq_k$
1	1359671	13223781	37292.5	1051.689
2	1285086	1322378	-37292.5	1051.689
-	-	-	Σ	2103.377

У таблиці вище $q_1 = q$, $q_2 = 1 - q$. Тобто, $\hat{\chi}_n^2 \approx 2103.377$.

Визначимо поріг тесту. Оскільки вважається, що рівень значущості тесту є $\alpha = 0.01$, то

$$h_\alpha = Q^{\chi^2_{r-1}}(1-\alpha) = Q^{\chi^2_1}(0.95) \approx 6.6349.$$

Перевіряємо, чи потрапляє значення статистики тесту $\hat{\chi}_n^2$ у критичну область $[h_\alpha, \infty)$. Легко бачити, що це так, бо $\chi_n^2 \approx 2103.377 > 6.6349 \approx h_\alpha$.

Значить, ми маємо підстави на користь альтернативи. Основна гіпотеза відхиляється на рівні $\alpha = 0.01$.

2.2 Задача 2

У експериментах з селекцією гороха Мендель спостерігав частоти різних видів насіння, що отримані при схрещуванні рослин з круглим жовтим насінням і рослим зі зморшкуватим зеленим насіннями. Ці дані наведені у таблиці: Перевірити гіпотезу за критерієм хі-квадрат про

Насіння	Частота	Ймовірність
кругле жовте	315	9/16
зморшкувате жовте	101	3/16
кругле зелене	108	3/16
зморшкувате зелене	32	1/16
сума	556	1

відповідність спостережень теоретичним частотам. Вважати, що рівень значущості становить $\alpha = 0.05$.

Розв'язання. Потрібно перевірити гіпотези на розподіл різновидів насіння, використовучи хі-квадрат тест. Зведемо вхідні дані до потрібної моделі.

Маємо вибірку $X = (X_1, \ldots, X_n)$ з $n = 556$ спостережень. Кожне спостереження X_j приймає лише чотири значення: 1 (кр. ж.) або 2 (зм. ж) або 3 (кр. з.) або 4 (зм. з.). Всього $r = 4$ результатів. Тобто, маємо вектор невідомих імовірностей $\theta = (p_1, p_2, p_3, p_4)$.

В термінах зведеної моделі, відповісти на питання задачі – це перевірити гіпотези вигляду:

$$\mathrm{H}_0 : (p_1, p_2, p_3, p_4) = (q_1, q_2, q_3, q_4), \ \mathrm{H}_1 : (p_1, p_2, p_3, p_4) \neq (q_1, q_2, q_3, q_4),$$

де $(q_1, q_2, q_3, q_4) = (9/16, 3/16, 3/16, 1/16)$.

Обчислимо значення статистики хі-квадрат тесту. Наведемо всі необхідні величини у табличному вигляді:

k	ν_k	nq_k	$\nu_k - nq_k$	$(\nu_k - nq_k)^2/nq_k$
1	315	312.75	2.25	0.01618705
2	101	104.25	-3.25	0.10131894
3	108	104.25	3.75	0.13489209
4	32	34.75	-2.75	0.21762590
-	-	-	Σ	0.47

Тобто, $\hat{\chi}_n^2 \approx 0.47$.

Визначимо поріг тесту. Оскільки вважається, що рівень значущості тесту є $\alpha = 0.05$, то

$$h_\alpha = Q^{\chi^2_{r-1}}(1-\alpha) = Q^{\chi^2_3}(0.95) \approx 7.8147.$$

Перевіряємо, чи потрапляє значення статистики тесту $\hat{\chi}_n^2$ у критичну область $[h_\alpha, \infty)$. Легко бачити, що це не так, бо $\chi_n^2 \approx 0.47 < 7.81479 \approx h_\alpha$.

Значить, ми маємо підстави на користь основної гіпотези. Основна гіпотеза приймається на рівні $\alpha = 0.05$.

2.3 Задача 3

Далі наведено дані про кількість студентів, які не склали успішно сесію протягом 200 навчальних семестрів: де k – кількість студентів, не склавших сесію в одному семестрі, ν_k –

k	0	1	2	3	4	5 і більше	разом
ν_k	36	68	48	27	14	7	200

кількість семестрів (спостережень), під час якого k студентів пішли на перескладання.

Перевірити гіпотезу про Пуассонів семестровий розподіл кількості двійочників з параметром $\lambda = 3/2$. Вважати, що рівень значущості становить $\alpha = 0.05$.

Розв'язання. Потрібно перевірити гіпотези на розподіл кількості двійочників, використовуючи хі-квадрат тест. Зведемо вхідні дані до потрібної моделі.

Маємо вибірку $X = (X_1, \ldots, X_n)$ з $n = 200$ спостережень (семестрів). Кожне спостереження X_j приймає лише шість значень:

$$X_j = \begin{cases} 1, & 0 \text{ студентів на перескладанні в } j\text{-му семестрі}, \\ 2, & 1 \text{ студент на перескладанні в } j\text{-му семестрі}, \\ 3, & 2 \text{ студенти на перескладанні в } j\text{-му семестрі}, \\ 4, & 3 \text{ студенти на перескладанні в } j\text{-му семестрі}, \\ 5, & 4 \text{ студенти на перескладанні в } j\text{-му семестрі}, \\ 6, & \geq 5 \text{ студентів на перескладанні в } j\text{-му семестрі}. \end{cases}$$

Всього $r = 6$ результатів. Тобто, маємо вектор невідомих імовірностей $\theta = (p_1, \ldots, p_6)$.

В термінах зведеної моделі, відповісти на питання задачі – це перевірити гіпотези вигляду:

$$\mathrm{H}_0 : \ p_k = q_k, 1 \leq k \leq r, \ \mathrm{H}_1 : \text{ Гіпотеза } \mathrm{H}_0 \text{ не справджується},$$

де q_k визначено таким чином:

$$q_k = \begin{cases} \mathbf{P}\left(\xi = k - 1\right), & k = 1, \ldots, 5 \\ \mathbf{P}\left(\xi \geq 5\right), & k = 6 \end{cases}$$

Деякі підрахунки дають наближені значення для q_k:

k	1	2	3	4	5	6
q_k	0.2231	0.3347	0.2510	0.1255	0.0471	0.0186

Обчислимо значення статистики хі-квадрат тесту.

Наведемо всі необхідні величини у табличному вигляді:

k	ν_k	nq_k	$\nu_k - nq_k$	$(\nu_k - nq_k)^2/nq_k$
1	36	44.6260	-8.6260	1.6674
2	68	66.9390	1.0610	0.0168
3	48	50.2043	-2.2043	0.0968
4	27	25.1021	1.8979	0.1435
5	14	9.4133	4.5867	2.2349
6	7	3.7152	3.2848	2.9043
-	-	-	Σ	7.0637

Тобто, $\hat{\chi}^2_n \approx 7.0637$.

Визначимо поріг тесту. Оскільки вважається, що рівень значущості тесту є $\alpha = 0.05$, то

$$h_\alpha = Q^{\chi^2_{r-1}}(1-\alpha) = Q^{\chi^2_5}(0.95) \approx 11.0705.$$

Перевіряємо, чи потрапляє значення статистики тесту $\hat{\chi}^2_n$ у критичну область $[h_\alpha, \infty)$. Легко бачити, що це не так, бо $\chi^2_n \approx 7.0637 < 11.0705 \approx h_\alpha$.

Значить, ми маємо підстави на користь основної гіпотези. Основна гіпотеза приймається на рівні $\alpha = 0.05$.

Питання. Поекспериментуйте з α. Згрупуйте останні дві групи: як зміниться результат?

2.4 Задача 4

Нижче наведені інтервали в експлуатаційних годинах між послідовними відмовами апаратури кондиціювання повітря на літаку:

1, 1, 3, 3, 3, 4, 4, 5, 5, 5, 5, 6, 7, 7, 11, 11, 11,
1112, 13, 14, 14, 14, 15, 16, 16, 16, 18, 18, 18, 18, 20, 21, 22,
22, 23, 23, 24, 30, 31, 36, 39, 39, 42, 43, 44, 46, 46, 47, 50, 51,
52, 54, 62, 63, 68, 71, 71, 72, 77, 79, 80, 82, 85, 87, 88, 90, 91,
95, 97, 97, 98, 100, 102, 106, 111, 120, 120, 130, 139, 141, 142, 163, 188, 191,
197, 206, 210, 216, 225, 230, 246, 261, 487.

Перевірте гіпотезу про показниковий розподіл з параметром 0.01 часу безвідмовної роботи апаратури кондиціювання повітря. Вважати, що рівень значущості становить $\alpha = 0.01$.

Розв'язання. Вхідні дані взяті з неперервного розподілу. Потрібно зробити дискретизацію вибірки, тобто подати у вигляді скінченної кількості результатів випробувань.

Це зробимо за допомогою групування даних.

Розіб'ємо носій гіпотетичного розподілу, $(0, \infty) = \cup_{k=1}^{r} I_k$ на такі інтервали

$$I_k = (x_k, x_{k+1}],\ k = 1, \dots, r-1,\ I_r = (x_r, \infty),$$

щоб $\mathbf{P}(\xi \in I_k) = 1/r$ для всіх $k = 1, \dots, r$. Як знайти x_k?

Очевидно, що $x_1 = 0$. Далі, знайдемо x_2:

$$\mathbf{P}(\xi \in I_1) = \mathbf{P}(\xi < x_2) - \mathbf{P}(\xi < x_1) = \mathbf{P}(\xi < x_2) - 0 = \mathbf{P}(\xi < x_2) = 1/r$$

Тобто $x_2 = Q^{\mathrm{Exp}(\lambda)}(1/r)$. Знайдемо у явному вигляді:

$$\mathbf{P}(\xi < x_2) = 1 - e^{-\lambda x_2} = 1/r \Rightarrow x_2 = -\frac{\ln(1 - 1/r)}{\lambda}.$$

Далі, знайдемо x_k, $k = 2, 3, \dots, r-2, r-1$:

$$\mathbf{P}(\xi \in I_k) = \mathbf{P}(\xi < x_{k+1}) - \mathbf{P}(\xi < x_k) = \mathbf{P}(\xi < x_{k+1}) - \sum_{l=1}^{k-1} \mathbf{P}(\xi \in I_l) = \mathbf{P}(\xi < x_{k+1}) - (k-1)/r = 1/r$$

Тобто

$$\mathbf{P}(\xi < x_{k+1}) = k/r$$

Звідси

$$x_{k+1} = -\frac{\ln(1 - k/r)}{\lambda}$$

Отже, I_k мають вигляд:

$$I_k = (x_k, x_{k+1}],\ k = 1, \dots, r-1,\ I_r = (x_r, \infty),\ x_k = -\ln(1 - k/r)/\lambda.$$

Тепер перейдемо до використання цього результату в нашій задачі.

Розіб'ємо $(0, \infty)$ на $r = 4$ інтервали з рівними імовірностями ($1/r = 1/4 = 0.25$). З попередніх викладок маємо такі кінці інтервалів I_k:

$$I_1 = (0, 0.2877],\ I_2 = (0.2877, 0.6931],\ I_3 = (0.6931, 1.3863],\ I_4 = (1.3863, +\infty).$$

Власне, робимо дискретизацію X: для всіх $j = 1, \dots, n$

$$Y_j = \sum_{k=1}^{4} k\mathbf{1}\{X_j \in I_k\} = \begin{cases} 1, & X_j \in I_1, \\ 2, & X_j \in I_2, \\ 3, & X_j \in I_3, \\ 4, & X_j \in I_4. \end{cases}$$

Перейшовши від неперервної вибірки X до дискретної альтернативи $Y = (Y_1, \dots, Y_n)$. Для останньої можемо застосувати хі-квадрат тест.

Інтерпретація вибірки Y: маємо n незалежних випробувань з $r = 4$ результатів, імовірність кожного результату становить p_k, $1 \leq k \leq r$. Потрібно перевірити основну гіпотезу

$$\mathrm{H}_0 : \ p_k = q_k,\ k = 1, \dots, r$$

проти альтернативи

$$\mathrm{H}_1 : \ \text{Гіпотеза } \mathrm{H}_0 \text{ не виконується,}$$

де $q_k = \mathbf{P}\left(\xi \in I_k\right) = 1/4$.

Обчислимо значення статистики хі-квадрат тесту. Наведемо всі необхідні величини у табличному вигляді:

k	ν_k	nq_k	$\nu_k - nq_k$	$(\nu_k - nq_k)^2/nq_k$
1	36	23.25	12.75	6.9919
2	18	23.25	-5.25	1.1855
3	23	23.25	-0.25	0.0027
4	16	23.25	-7.25	2.2608
-	-	-	Σ	10.4409

У таблиці $\hat{\nu}_k = \sum_{j=1}^{n} \mathbf{1}\{Y_j = k\} = \sum_{j=1}^{n} \mathbf{1}\{X_j \in I_k\}$. З таблиці видно, що $\hat{\chi}_n^2 \approx 10.4409$.

Визначимо поріг тесту. Оскільки вважається, що рівень значущості тесту є $\alpha = 0.05$, то

$$h_\alpha = Q^{\chi^2_{r-1}}(1-\alpha) = Q^{\chi^2_3}(0.95) \approx 7.8147.$$

Перевіряємо, чи потрапляє значення статистики тесту $\hat{\chi}_n^2$ у критичну область $[h_\alpha, \infty)$. Легко бачити, що це так, бо $\chi_n^2 \approx 10.4409 > 7.81479 \approx h_\alpha$.

Значить, ми маємо підстави на користь основної гіпотези. Основна гіпотеза відхиляється на рівні $\alpha = 0.05$.

Зауваження. Зауважимо, що гіпотези ми *підмінили*. По суті в дискретизованій версії ми перевіряємо, чи характерно неперервному розподілу потрапляти у задані інтервали із конкретними імовірностями. Тобто припущення дещо відштовхуються від припущення експоненційної розподіленості спостережень, оскільки по факту можна підібрати інший розподіл, для якого ймовірності потрапляння в інтервали будуть такими ж, як в експоненційному випадку. Таким чином пояснюючи чому тут отриманий тест може бути менш чутливим до випадків, що не узгоджуються з H_0.

Додаток. Програмна реалізація розв'язків в R

Задача 1

```
# Кількість хлопчиків та дівчаток, сумарна кількість
O <- c(1359671, 1285086)
r <- length(O)
n <- sum(O)

# Припущення на імовірність народження хлопчика
q <- 0.5

# Очікувані частоти згідно H0
E <- n * c(q, 1-q)
print(E)

# Різниця між спостережуваними та очікуваними частотами
D <- O - E
print(D)

# Доданки в статистиці хі-квадрат
X2 <- D^2 / E
print(X2)

# Значення статистики хі-квадрат
Chisq2.emp <- sum(X2)
print(round(Chisq2.emp, 4))

# Поріг тесту
alpha <- 0.01
h.alpha <- qchisq(1 - alpha, df = r - 1)
print(round(h.alpha, 4))
```

Задача 2

```
# Спостережувані частоти
O <- c(315, 101, 108, 32)
r <- length(O)
n <- sum(O)

# Гіпотеза на розподіл результатів
q <- c(9, 3, 3, 1) / 16

# Очікувані частоти
E <- n * q
print(E)

# Різниця між спостережуваними та очікуваними частотами
D <- O - E
print(D)

# Доданки в статистиці хі-квадрат
X2 <- D^2 / E
print(X2)

# Значення статистики хі-квадрат
Chisq2.emp <- sum(X2)
print(round(Chisq2.emp, 4))

# Поріг тесту
alpha <- 0.05
h.alpha <- qchisq(1 - alpha, df = r - 1)
print(round(h.alpha, 4))
```

Задача 3.

```
# Спостережувані частоти
O <- c(36, 68, 48, 27, 14, 7)
r <- length(O)
n <- sum(O)

# Гіпотеза на розподіл результатів
l.h0 <- 3/2
q <- c(dpois(0:4, lambda = l.h0), 1-ppois(5-1, lambda = l.h0))
print(round(q, 4))

# Очікувані частоти
E <- n * q
print(E)

# Різниця між спостережуваними та очікуваними частотами
D <- O - E
print(D)

# Доданки в статистиці xi-квадрат
X2 <- D^2 / E
print(X2)

# Значення статистики xi-квадрат
Chisq2.emp <- sum(X2)
print(round(Chisq2.emp, 4))

# Поріг тесту
alpha <- 0.05
h.alpha <- qchisq(1 - alpha, df = r - 1)
print(round(h.alpha, 4))
```

Задача 4.

```
x <- c(
  1, 1, 3, 3, 3, 4, 4, 5, 5, 5, 5, 6, 7, 7, 11, 11, 11,
  1112, 13, 14, 14, 14, 15, 16, 16, 16, 18, 18, 18, 18, 20, 21, 22,
  22, 23, 23, 24, 30, 31, 36, 39, 39, 42, 43, 44, 46, 46, 47, 50, 51,
  52, 54, 62, 63, 68, 71, 71, 72, 77, 79, 80, 82, 85, 87, 88, 90, 91,
  95, 97, 97, 98, 100, 102, 106, 111, 120, 120, 130, 139, 141, 142, 163, 188, 191,
  197, 206, 210, 216, 225, 230, 246, 261, 487
)

# Кінці інтервалів
r <- 4
l.given <- 0.01
xk <- c(-log(1-((1:r)-1)/r)/l.given, +Inf)

# Дискретизована вибірка
y <- findInterval(x, xk)

# Спостережувані частоти
O <- table(y)
n <- sum(O)

# Гіпотеза на розподіл результатів
q <- diff(pexp(xk, rate = l.given))

# Очікувані частоти
E <- n * q
print(E)

# Різниця між спостережуваними та очікуваними частотами
D <- O - E
print(D)

# Доданки в статистиці хі-квадрат
X2 <- D^2 / E
print(X2)

# Значення статистики хі-квадрат
Chisq2.emp <- sum(X2)
print(round(Chisq2.emp, 4))

# Поріг тесту
alpha <- 0.05
h.alpha <- qchisq(1 - alpha, df = r - 1)
print(round(h.alpha, 4))
```