Перевірка статистичних гіпотез

Означення 1. **Статистичною гіпотезою** H, називають будь-яке припущення відносно виду або параметрів розподілу випадкової величини ξ, яке може бути перевірене за результатами вибірки.
Означення 2. **Простою** називають гіпотезу, що містить лише одне припущення. Якщо гіпотеза складається з кількох простих гіпотез, то вона називається **складною**. При цьому одне з припущень обирається за основне і називається **нульовою** (основною) гіпотезою H_0. Інші гіпотези (припущення або можливості), що суперечать нульовій, називають **альтернативними** гіпотезами і позначаються $H_1, H_2, \ldots$.
Означення 3. **Статистичним критерієм** будемо називати правило, згідно з яким гіпотеза, що перевіряється, приймається або відхиляється.
Означення 4. **Статистикою критерію** називають функцію від вибірки $g(x_1, x_2, \ldots, x_n)$, яка характеризує міру розбіжності емпіричних даних від гіпотетичних законів розподілу.
Означення 5. Ймовірність відхилення гіпотези H_0, тоді коли вона є вірною, називається **ймовірністю помилки першого роду** або **рівнем значущості**. Ця ймовірність позначається α і є досить малою $\alpha \le 0,2$.

Перевірка гіпотези про вигляд розподілу

Нехай задано вибірку $x_1, x_2, \ldots, x_n$, яка представляє собою спостереження над випадковою величиною ξ з невідомою функцією розподілу $F_\xi(x)$. Тоді нам потрібно перевірити гіпотезу $H_0 : F_\xi(x) = F(x)$, де $F(x)$ - відома функція розподілу.

1) Критерій Колмогорова.

Зауваження. Цей критерій можна застосовувати лише для розподілів абсолютно неперервного типу.

$$H_0 : F_\xi(x) = F(x), \quad \alpha.$$

Статистика критерію:

$$D_n = \sqrt{n} \max_x \left| F_n^*(x) - F(x) \right|,$$ де

$F_n^*(x)$ - емпірична функція розподілу.

Відомо, що статистика D_n має розподіл Колмогорова. Таблиця критичних величин цього розподілу приведена нижче.

α	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
λ_α	0,828	0,895	0,974	1,073	1,224	1,358	1,52	1,627	1,95

Якщо $D_n < \lambda_\alpha$, тоді H_0 приймається. У протилежному випадку гіпотеза H_0 - відхиляється.

Приклад 1.

Нехай вибірка записана у вигляді групованого статистичного ряду

Інтервал	[0; 1)	[1; 2)	[2; 3)	[3; 4)	[4; 5)	[5; 6)	[6; 7)	[7; 8)	[8; 9)	[9;10)
Частота	35	16	15	17	17	19	11	16	30	24

Перевірити гіпотезу $H_0 : F(x) = \begin{cases} 1, \ x > 10 \\ \dfrac{x}{10}, \ x \in [0,10]. \\ 0, \ x < 0 \end{cases}$

Рівень значущості вважати рівним 0,05.

Розв'язання.

Ми перевіряємо гіпотезу, що наша вибірка відповідає рівномірному розподілу на відрізку $[0;10]$. Для цього ми заповнюємо таблицю:

Інтервал	[0; 1)	[1; 2)	[2; 3)	[3; 4)	[4; 5)	[5; 6)	[6; 7)	[7; 8)	[8; 9)	[9;10)
m_i	35	16	15	17	17	19	11	16	30	24
x_i^*	0,5	1,5	2,5	3,5	4,5	5,5	6,5	7,5	8,5	9,5
$F_n^*(x_i^*)$	0,087	0,215	0,292	0,372	0,457	0,547	0,622	0,69	0,805	0,94
$F(x_i^*)$	0,05	0,15	0,25	0,35	0,45	0,55	0,65	0,75	0,85	0,95
$\left\lvert F_n^*(x_i^*) - F(x_i^*) \right\rvert$	0,037	0,065	0,042	0,022	0,007	0,002	0,027	0,06	0,045	0,01

У цій таблиці використано такі позначення:

x_i^* - середина і-го інтервалу;

$F_n^*(x_i^*)$ - значення емпіричної функції розподілу на середині і-го інтервалу, яке шукається за наступною формулою

$$F_n^*(x_i^*) = \frac{1}{n}\left(\sum_{k=1}^{i-1} m_k + 0,5 m_i \right);$$

$F(x_i^*)$ - значення гіпотетичної функції розподілу на середині і-го інтервалу;

$\left\lvert F_n^*(x_i^*) - F(x_i^*) \right\rvert$ - різниця за модулем емпіричної та гіпотетичної функцій розподілів на середині і-го інтервалу.

Розглянемо детальніше, як шукаються значення емпіричної функції розподілу.

Спочатку знайдемо об'єм вибірки

$m_1 + m_2 + \ldots + m_s = n = 200$.

Далі

$$F_n^*(x_1^*) = \frac{1}{200}(0 + 0,5 \cdot 35) \approx 0,087;$$

$$F_n^*(x_2^*) = \frac{1}{200}(35 + 0,5 \cdot 16) \approx 0,215;$$

$$F_n^*(x_3^*) = \frac{1}{200}(35 + 16 + 0,5 \cdot 15) \approx 0,292;$$

$$F_n^*(x_4^*) = \frac{1}{200}(35 + 16 + 15 + 0,5 \cdot 17) \approx 0,372$$

і так далі.

Останній рядок заповнюється, як різниця двох передостанніх рядків за модулем. І з останнього рядка обирається найбільше значення.
Тепер можемо знайти статистику критерію

$D_{200} = \sqrt{200} \cdot 0,065 = 0,919$. Порівнюємо цей результат з відповідною критичною величиною для розподілу Колмогорова $\lambda_{0,05} = 1,358$. Бачимо, що $D_{200} < \lambda_{0,05}$. Отже, гіпотеза H_0 приймається і ми можемо працювати з нашими даними, як з рівномірним розподілом на відрізку $[0;10]$.

2) Критерій згоди χ^2.

Зауваження. Цей критерій можна застосовувати для будь-яких розподілів.

Розіб'ємо множину всіх можливих значень спостерігаємої випадкової величини ξ на r інтервалів, що не перетинаються $\Delta_1, \Delta_2, ..., \Delta_r$. Якщо спостерігається дискретна випадкова величина, то Δ_i, $i = 1, 2, ..., r$ - це різні значення цієї величини. Нехай m_i - частота елемента Δ_i. Введемо позначення

$$p_i = P\left(\xi \in \Delta_i \middle| H_0\right),\ i = 1, 2, ..., r.$$

p_i - це ймовірність того, що значення спостерігаємої випадкової величини ξ потрапить в інтервал Δ_i, при умові виконання гіпотези H_0. У дискретному випадку, ця ймовірність визначається наступним чином

$$p_i = P\left(\xi = \Delta_i \middle| H_0\right),\ i = 1, 2, ..., r.$$

Статистика критерію:

$$\chi_n^2 = \sum_{i=1}^{r} \frac{\left(m_i - np_i\right)^2}{np_i}.$$

Ця статистика наближено має χ^2 розподіл з $(r-1)$ ступенем свободи. Таке наближення можна використовувати з великою точністю, якщо $n \geq 50,\ m_i \geq 5,\ i = 1, 2, ..., r$.

Критерій перевірки гіпотези H_0 будується наступним чином. Спочатку обчислюється значення статистики χ_n^2 і обирається рівень значущості α. Далі, за таблицею критичних величин χ^2 розподілу шукаємо величину $\chi_{r-1,\alpha}^2$. Для цього заходимо в таблицю з такими параметрами $P\left(\chi^2(r-1) \geq \chi_{r-1,\alpha}^2\right) = \alpha$. Якщо виконується умова $\chi_n^2 < \chi_{r-1,\alpha}^2$, то гіпотеза H_0 про вигляд розподілу приймається, у протилежному випадку гіпотеза H_0 відхиляється.

Приклад 2.

Монету підкинули 50 разів. Герб з'явився 20 разів. Чи можна вважати монету симетричною, якщо рівень значущості $\alpha = 0{,}1$.

Розв'язання.

Появу решки будемо кодувати «0», а появу герба – «1». Тоді гіпотезу про симетричність монети можна записати наступним чином

$$H_0 : p_1 = P(\xi = 0) = p_2 = P(\xi = 1) = \frac{1}{2} = 0{,}5.$$

Об'єм вибірки $n = 50$. Кількість можливих значень нашої випадкової величини $r = 2$. Щоб знайти значення статистики, заповнимо таблицю

Δ_i	0	1
m_i	30	20
p_i	0,5	0,5
$\frac{(m_i - np_i)^2}{np_i}$	1	1

Тоді статистика буде дорівнювати $\chi^2_{50} = 1 + 1 = 2$.

Далі шукаємо критичну величину $\chi^2_{r-1;\alpha} = \chi^2_{1;0,1} = 2{,}7$.

Бачимо, що $\chi^2_{50} < \chi^2_{1;0,1}$. Отже гіпотеза H_0 про симетричність монети приймається.

Критерій χ^2 можна використовувати для перевірки складних гіпотез. Нехай по вибірці $\xi_1, \xi_2, ..., \xi_n$ потрібно перевірити гіпотезу $H_0 : F_\xi(x) \in m$, де $m = \{F(x,\theta),\ \theta \in \Theta\}$ - задане сімейство функцій розподілу. Тоді для статистики критерію $\chi^2_n = \sum_{i=1}^{r} \frac{(m_i - np_i)^2}{np_i}$ граничним буде χ^2- розподіл з $(r - l - 1)$ ступенями свободи, де l - розмірність вектора θ. Далі критерій перевірки гіпотези будується аналогічно наведеному вище критерію. Якщо виконується умова $\chi^2_n < \chi^2_{r-l-1,\alpha}$, то гіпотеза H_0 про вигляд розподілу приймається, у протилежному випадку гіпотеза H_0 відхиляється.

Приклад 3.

По спостереженням наведеним у таблиці, за допомогою критерію χ^2 перевірити гіпотезу, що випадкова величина має нормальний розподіл

Інтервал	[-4; 0)	[0; 2)	[2; 4)	[4; 6)
m_i	20	40	30	10

$\alpha = 0{,}05$.

Розв'язання.

Згадаємо операцію стандартизації нормального розподілу.

Якщо випадкова величина ξ має нормальний розподіл з параметрами a і σ^2 $\left(\xi \sim N(a,\sigma^2)\right)$, тоді випадкова величина $\eta = \dfrac{\xi - a}{\sigma}$ буде мати стандартний нормальний розподіл $\left(\eta \sim N(0,1)\right)$.

Обчислимо оцінки параметрів a і σ^2 за нашою вибіркою (нагадаю, що для групованих даних, ми будемо працювати з серединами інтервалів)

$$\overline{x} = \frac{1}{100}\left(-2\times 20 + 1\times 40 + 3\times 30 + 5\times 10\right) = 1,4;$$

$$\hat{S}^2 = \frac{1}{99}\left((-2-1,4)^2 20 + (1-1,4)^2 40 + (3-1,4)^2 30 + (5-1,4)^2 10\right) = 4,48;$$

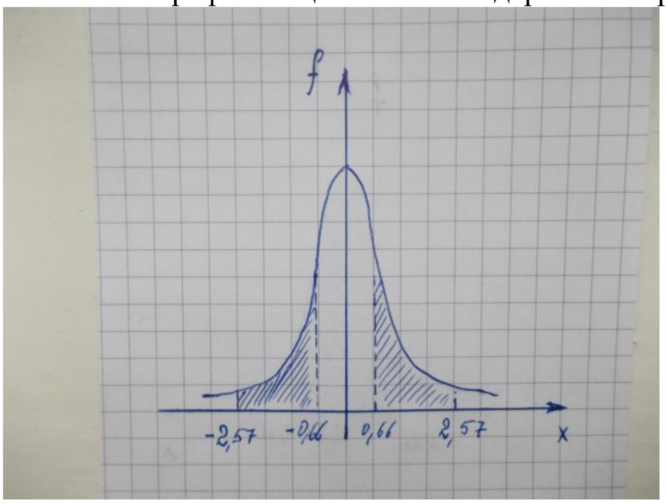
$$\hat{S} = \sqrt{\hat{S}^2} = 2,1.$$

Тепер перейдемо до підрахунку ймовірностей $p_i,\ i = \overline{1,4}$.

$$p_1 = P\left(-4 \le \xi < 0\right) = P\left(\frac{-4-1,4}{2,1} \le \frac{\xi - a}{\sigma} < \frac{0-1,4}{2,1}\right) =$$

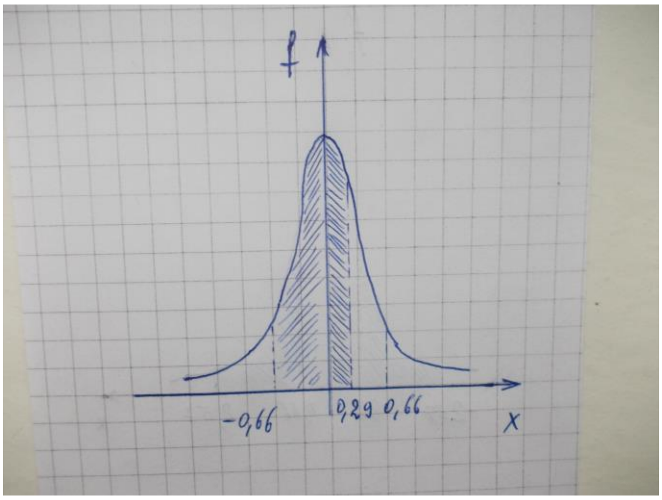
$$= \Phi_1(2,57) - \Phi_1(0,66) = 0,4949 - 0,2455 = 0,2494.$$

Щоб одержати цей результат, потрібно скористатись таблицею значень для функції Лапласа та графіком щільності стандартного нормального закону.



$$p_2 = P\left(0 \le \xi < 2\right) = P\left(\frac{0-1,4}{2,1} \le \frac{\xi - a}{\sigma} < \frac{2-1,4}{2,1}\right) =$$

$$= \Phi_1(0,286) + \Phi_1(0,66) = 0,1140 + 0,2455 = 0,3595.$$



$$p_3 = P\left(2 \le \xi < 4\right) = P\left(\frac{2-1,4}{2,1} \le \frac{\xi - a}{\sigma} < \frac{4-1,4}{2,1}\right) =$$

$$= \Phi_1(1,23) - \Phi_1(0,286) = 0,3905 - 0,1140 = 0,2765.$$

$$p_4 = P\left(4 \le \xi < 6\right) = P\left(\frac{4-1,4}{2,1} \le \frac{\xi - a}{\sigma} < \frac{6-1,4}{2,1}\right) =$$

$$= \Phi_1(2,19) - \Phi_1(1,23) = 0,4855 - 0,3905 = 0,095.$$

Далі результати наведено у таблиці.

Інтервал	[-4; 0)	[0; 2)	[2; 4)	[4; 6)
m_i	20	40	30	10
p_i	0,2494	0,3595	0,2765	0,095
np_i	24,94	35,95	27,65	9,5
$\frac{\left(m_i - np_i\right)^2}{np_i}$	0,98	0,46	0,2	0,02

$\chi_n^2 = 1,66$. Кількість інтервалів $r = 4$, а кількість невідомих параметрів $l = 2$. Тоді кількість ступенів свободи буде $k = r - l - 1 = 1$.

$\chi_{1;0,05}^2 = 3,8$. Бачимо, що $1,66 < 3,8$. Отже гіпотеза приймається.

Перевірка гіпотези однорідності вибірок

Нехай задано дві вибірки $x_1, x_2, \ldots, x_n$ та $y_1, y_2, \ldots, y_m$ з невідомими функціями розподілу $F_1(x)$ і $F_2(x)$. Тоді для заданого рівня значущості α сформулюємо гіпотезу

$$H_0 : F_1(x) = F_2(x)$$

1) Критерій однорідності Смірнова-Колмогорова.

Зауваження. Цей критерій можна застосовувати лише для розподілів абсолютно неперервного типу.

$H_0 : F_1(x) = F_2(x), \alpha$.

Статистика критерію:

$$\lambda_{nm} = \sqrt{\frac{nm}{n+m}} \max_x \left| F^*_{n1}(x) - F^*_{m2}(x) \right|, \text{ де}$$

$F^*_{n1}(x)$ та $F^*_{m2}(x)$ - емпіричні функції розподілів, побудовані за першою та другою вибірками відповідно.

Відомо, що статистика λ_{nm} має розподіл Колмогорова. Таблиця критичних величин цього розподілу приведена нижче.

α	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
λ_α	0,828	0,895	0,974	1,073	1,224	1,358	1,52	1,627	1,95

Якщо $\lambda_{nm} < \lambda_\alpha$, тоді H_0 приймається. У протилежному випадку гіпотеза H_0 - відхиляється.

Приклад для цього критерію наводити не будемо, оскільки з ним можна працювати, майже так само, як з критерієм Колмогорова .

2) Критерій однорідності χ^2.

Зауваження. Цей критерій можна використовувати для перевірки дискретних даних. Окрім того, за його допомогою можна перевіряти однорідність будь-якого скінченого числа вибірок (критерій Смирнова-Колмогорова може аналізувати лише дві вибірки).

Припустимо, що проведено k послідовних серій незалежних спостережень (тобто є k вибірок), які включають $n_1, n_2, \dots, n_k$ спостережень відповідно. При цьому в кожному експерименті може з'явитися один з l наслідків.

Позначимо через v_{ij} - число появ i-го наслідку в j-тій серії.

Тоді $n_j = \sum_{i=1}^{l} v_{ij}$ - об'єм j-тої вибірки, $n = \sum_{j=1}^{k} n_j$ - загальна кількість спостережень.

Гіпотезу H_0 запишемо словами.

H_0: всі спостереження проводились над однією і тією ж випадковою величиною (або всі k вибірок є однорідними).

Статистика критерію

$$\chi^2_n = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{\left(v_{ij} - \frac{v_{i\circ} n_j}{n} \right)^2}{\frac{v_{i\circ} n_j}{n}}, \text{ де}$$

$$v_{i\circ}=\sum_{j=1}^{k}v_{ij}\ .$$

Статистика критерію χ_n^2 має наближено χ^2-розподіл з $(l-1)(k-1)$ ступенями свободи. За таблицею критичних величин χ^2 розподілу (див. заняття №11 таблиця 2) шукаємо величину $\chi^2_{(l-1)(k-1),\alpha}$. Для цього заходимо в таблицю з такими параметрами $P\left(\chi^2(l-1)(k-1)\ge\chi^2_{(l-1)(k-1),\alpha}\right)=\alpha$. Якщо виконується умова $\chi_n^2<\chi^2_{(l-1)(k-1),\alpha}$, то гіпотеза однорідності H_0 приймається, у протилежному випадку гіпотеза H_0 відхиляється.

Приклад 1.

За допомогою критерію χ^2 перевірити гіпотезу однорідності двох вибірок при рівні значущості $\alpha=0,05$.

x_i	1	2	3	4
v_{i1}	40	26	24	10
v_{i2}	30	20	30	20
$v_{i\circ}$	70	46	54	30

Знаходимо

$n_1=40+26+24+10=100$;

$n_2=30+20+30+20=100$;

$n=100+100=200$.

Шукаємо статистику критерію

$$\chi_n^2=\sum_{i=1}^{l}\sum_{j=1}^{k}\frac{\left(v_{ij}-\dfrac{v_{i\circ}n_j}{n}\right)^2}{\dfrac{v_{i\circ}n_j}{n}}$$

$$\chi_n^2=\frac{(40-35)^2}{35}+\frac{(26-23)^2}{23}+\frac{(24-27)^2}{27}+\frac{(10-15)^2}{15}+$$
$$+\frac{(30-35)^2}{35}+\frac{(20-23)^2}{23}+\frac{(30-27)^2}{27}+\frac{(20-15)^2}{15}=6,2$$

Визначаємо $k=2,\quad l=4$. Далі шукаємо критичну величину

$\chi^2_{(l-1)(k-1);\alpha}=\chi^2_{3;0,05}=7,8$.

Бачимо, що статистика критерію менша за критичну величину. Отже гіпотеза однорідності вибірок приймається.

Перевірка гіпотези незалежності вибірок

Критерій незалежності χ^2.

H_0 : випадкові величини ξ і η є незалежними.

Статистика критерію:

$$\chi_n^2 = \sum_{i=1}^{l}\sum_{j=1}^{k}\frac{\left(v_{ij} - m_{ij}\right)^2}{m_{ij}}, \text{ де}$$

v_{ij} - число випадків, коли одночасно спостерігались $\xi = x_i, \eta = y_j$ (для неперервних випадкових величин i та j - номери відповідних інтервалів);

$m_{ij} = \frac{v_{i\circ}v_{\circ j}}{n}$, $v_{\circ j} = \sum_{i=1}^{l} v_{ij}$, $v_{i\circ} = \sum_{j=1}^{k} v_{ij}$;

l, k - число значень, що приймають випадкові величини ξ і η відповідно;

$n = \sum_{i=1}^{l}\sum_{j=1}^{k} v_{ij}$ - об’єм вибірки.

Далі, за таблицею критичних величин χ^2 розподілу (див. заняття №11 таблиця 2) шукаємо величину $\chi^2_{(l-1)(k-1),\alpha}$. Для цього заходимо в таблицю з такими параметрами $P\left(\chi^2(l-1)(k-1) \geq \chi^2_{(l-1)(k-1),\alpha}\right) = \alpha$. Якщо виконується умова $\chi_n^2 < \chi^2_{(l-1)(k-1),\alpha}$, то гіпотеза H_0 про вигляд розподілу приймається, у протилежному випадку гіпотеза H_0 відхиляється.

Приклад 2.

Проведено одночасно 300 спостережень над випадковими величинами ξ і η, які приймають значення 1, 2 та 1, 2, 3 відповідно. Кількості спостережень пар v_{ij} наведено у таблиці:

η / ξ	1	2	3	$v_{i\circ}$
1	32	68	50	150
2	40	70	40	150
$v_{\circ j}$	72	138	90	300

Перевірити за допомогою критерію χ^2 чи є незалежними випадкові величини ξ і η при рівні значущості $0,01$.

Розв'язання.

У нашій задачі зручно ввести такі матриці:

1) матриця, яка складається з елементів m_{ij}

$$\left(m_{ij}\right)=\begin{pmatrix}36 & 69 & 45\\ 36 & 69 & 45\end{pmatrix};$$

2) матриця, яка складається з елементів $(v_{ij}-m_{ij})^2$

$$\left((v_{ij}-m_{ij})^2\right)=\begin{pmatrix}16 & 1 & 25\\ 16 & 1 & 25\end{pmatrix};$$

3) матриця, яка складається з елементів $\dfrac{(v_{ij}-m_{ij})^2}{m_{ij}}$

$$\left(\frac{(v_{ij}-m_{ij})^2}{m_{ij}}\right)=\begin{pmatrix}0,44 & 0,014 & 0,55\\ 0,44 & 0,014 & 0,55\end{pmatrix}.$$

Звідси знаходимо, що статистика критерію $\chi^2_{300}=2,008$.

З таблиці критичних величин знаходимо $\chi^2_{(l-1)(k-1),\alpha}=\chi^2_{(2-1)(3-1);0,01}=\chi^2_{2;0,01}=9,2$.

Бачимо, що $\chi^2_{300}<\chi^2_{2;0,01}$. Отже гіпотеза H_0 про незалежність випадкових величин ξ і η приймається.

Задача. У таблиці наведено результати опитування 100 студентів перших трьох курсів, яким ставилося одне запитання: «Чи вважаєте ви, що куріння заважає навчанню?»

З'ясувати, чи підтверджують ці дані припущення про те, що ставлення до куріння студентів на різних курсах різне? Прийняти $\alpha=0,05$.

Відповідь	Курс		
	Перший	Другий	Третій
Так	-	30	25
Не знаю	8	5	7
Ні	15	10	-

Застосуємо критерій однорідності χ^2. Для цього запишемо таблицю у наступному вигляді

x_i	Так	Не знаю	Ні
v_{i1}	0	8	15
v_{i2}	30	5	10
v_{i3}	25	7	0
$v_{i\circ}$	55	20	25

$$k=3;\ n_1=23;\ n_2=45;\ n_3=32;\ n=100;\ l=3.$$

$$\chi_n^2=\sum_{i=1}^{l}\sum_{j=1}^{k}\frac{\left(v_{ij}-\frac{v_{i\circ}n_j}{n}\right)^2}{\frac{v_{i\circ}n_j}{n}}=\frac{\left(0-\frac{55\cdot 23}{100}\right)^2}{\frac{55\cdot 23}{100}}+\frac{\left(30-\frac{55\cdot 45}{100}\right)^2}{\frac{55\cdot 45}{100}}+$$

$$+\frac{\left(25-\frac{55\cdot 32}{100}\right)^2}{\frac{55\cdot 32}{100}}+\ldots=44,2.$$

$$\chi^2_{(l-1)(k-1);\alpha}=\chi^2_{4;0,05}=9,488.$$

Бачимо, що $\chi_n^2>\chi^2_{(l-1)(k-1);\alpha}$. Отже, гіпотеза однорідності вибірок відхиляється. Тоді відповідь на питання задачі буде позитивною.