МІНІСТЕРСТВО ОСВІТИ І НАУКИ
УЖГОРОДСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МАТЕМАТИЧНИЙ ФАКУЛЬТЕТ

ПОЛЯК І.Й.

МЕТОДИЧНІ ВКАЗІВКИ ДО ПРАКТИЧНИХ ЗАНЯТЬ З МАТЕМАТИЧНОЇ СТАТИСТИКИ ДЛЯ СТУДЕНТІВ МАТЕМАТИЧНОГО ФАКУЛЬТЕТУ

Ужгород -2007

Заняття 1.
Предмет і основні задачі математичної статистики. Генеральна сукупність і вибірка.

Теоретичні відомості.

Нехай для вивчення кількісної ознаки X із генеральної сукупності вилучена вибірка $x_1, x_2, ..., x_k$ об'єму n. Спостерігальні значення x_i ознаки X називають варіантами, а послідовність варіант записаних в зростаючому порядку,- варіаційним рядом.

Статистичним розподілом вибірки називають список варіант x_i варіаційного ряду і відповідних їх частот n_i.

Емпіричною функцією розподілу називають функцію $F^*(x)$, що визначає для кожного значення x відносну частоту події $X < x$:

$F^*(x) = \frac{n_x}{n}$, n_x – число варіант, менших за x; n – об'єм вибірки.

Розв'язування типових задач.

1.Вибірка задана у вигляді розподілу частот.

x_i	4	7	8	12
n_i	5	2	3	10

Знайти розподіл відносних частот. Знайти емпіричну функцію розподілу і зобразити графік. Побудувати полігон відносних частот.

Розв'язання:

Запишемо шуканий розподіл відносних частот

x_i	4	7	8	12
ω_i	0,25	0,1	3/20	0,5

де $\omega_i = \frac{n_i}{n}$, та запишемо емпіричну функцію розподілу:

$$F^*(x) = \begin{cases} 0 & при \quad x \le 4 \\ 0{,}25 & при \quad 4 < x \le 7 \\ 0{,}35 & при \quad 7 < x \le 8. \\ 0{,}5 & при \quad 8 < x \le 12 \\ 1 & при \quad x > 12 \end{cases}$$

Задачі і вправи для самостійного розв'язання.

1. В результаті обстеження одержано дані про число пташенят в гніздах лісової ластівки (Iridoprocne bicolor): 4 5 4 5 5 4 5 4 3 5 4 5 6 1 6 4 4 4 5 5 3 5 5 4 6 4 6 2 3 4 5 5 5 5 5 5 4 5 5 6 4 6 2 5 5 3 5 5 4 5 5 6 4 6 2 5 5 3 5 5 5 7 5 5 5 5 4 3 7 6 4 4 5 5 6 6 4 4 6. Подати дані у вигляді варіаційного ряду, емпіричного розподілу, побудувати полігон частот. Знайти емпіричну функцію розподілу.

2. Довжина 100 листків садової суниці (в см) характеризується такими даними: 8.2 9.7 5.6 7.4 8.0 6.4 6.6 6.8 8.4 7.1 9.0 6.9 7.6 8.1 11.8 5.8 9.3 7.3 8.2 7.2 7.2 6.4 7.7 9.0 8.1 7.1 7.1 8.8 7.5 9.2 7.5 6.8 7.0 6.4 7.4 8.2 6.3 7.0 8.1 10.0 7.0 7.1 8.7 6.3 8.6 7.7 7.3 8.0 8.4 9.3 7.3 6.0 7.7 6.1 9.6 7.4 7.2 7.2 8.7 7.5 9.1 6.4 8.3 6.5 8.2 7.2 6.9 6.9 8.2 9.0 7.4 8.0 8.4 7.0 7.1 7.4 6.6 6.4 8.3 7.9 8.3 7.2 7.2 6.6 6.6 7.7 8.7 5.6 7.5 5.7 6.9 7.4 7.2 6.2 6.9 6.8 9.2 9.2 7.1 6.5. Побудувати інтервальний розподіл, розбивши статистичні дані на 6 – 7 інтервалів однакової довжини, побудувати гістограму частот.

3.Вибірка задана у вигляді розподілу частот :

а).

x_i	2	5	7
n_i	1	3	6

б).

x_i	3	6	8
n_i	6	2	2

Знайти розподіл частот.

4.Знайти емпіричну функцію розподілу та побудувати графік по заданому розподілу вибірки.

x_i	1	4	5
n_i	10	15	25

б).

x_i	3	5	6
n_i	15	20	15

5.Побудувати полігон частот по заданому розподілу вибірки.

а).

x_i	1	4	5	7
n_i	20	10	14	6

x_i	2	3	5	6
n_i	10	15	5	20

6.Побудувати гістограму частот по заданому розподілу вибірки.

а).

$x_i - x_{i+1}$	1 - 5	5 - 9	9 – 13	13 - 17	17 – 21
n_i	10	20	50	12	8

б).

$x_i - x_{i+1}$	3 – 5	5 - 7	7 - 9	9 -11	11 - 13	13 – 15	15 – 17
n_i	4	6	20	40	20	4	6

7. Обчислити моду , медіану слідуючи вибірок:

а). 7;3;3;6;5;1;2;1;3. б).3,1;3,0;1,5;1,8;2,5;3,1;2,4;2,8;1,3.

Заняття 2. Основні характеристики вибірки.

Теоретичні відомості.

Будь-які характеристики, які знаходяться на основі статистичний даних, називаються *емпіричними* або статистичними характеристиками, а характеристики, які знаходяться на основі розподілу досліджуваної величини, називаються *теоретичними* характеристиками.

Використання всіх статистичних даних для аналізу ознаки не завжди доцільно. Для аналізу властивостей досліджуваної ознаки на основі статистичних даних використовують числові характеристики вибірки (статистичні характеристики). До основних характеристик вибірки відносять величини, які характеризують середнє значення та розсіювання можливих значень досліджуваної величини.

Однією із основних характеристик середнього значення є *вибіркова середня* $\bar{x}$ - *середнє арифметичне* результатів спостережень:

$$\bar{x} = \frac{1}{n}\sum_i x_i \,. \qquad (1)$$

Попередню уяву про розсіювання статистичних даних дає *розмах варіювання* $R = x_{\max} - x_{\min}$, але ця величина є досить грубою характеристикою розсіювання.

До основних характеристик розсіювання статистичних даних відносять *вибіркову дисперсію* (середнє арифметичне квадратів відхилень результатів спостережень від вибіркової середньої)

$$D_B = \frac{1}{n}\sum_i \left(x_i - \bar{x}\right)^2 \qquad (2)$$

і вибіркове середнє квадратичне відхилення

$$\sigma_B = \sqrt{D_B}\,.$$

Для незгрупованих даних вибіркова середня і вибіркова дисперсія знаходяться за формулами (1) і (2). Якщо ж статистичні дані згруповані, то

$$\bar{x} = \frac{1}{n}\sum_i n_i x_i\,, \quad D_B = \frac{1}{n}\sum_i \left(x_i - \bar{x}\right)^2 n_i\,, \qquad (3)$$

де x_i - різні результати спостережень випадкової величини ξ, а n_i - відповідні частоти.

Для вибіркової середньої і вибіркової дисперсії для довільних $c \in R$ і $h \neq 0$ справедливі формули

$$\bar{x} = h\bar{u} + c\,, \quad D_B = h^2\left(\overline{u^2} - \left(\bar{u}\right)^2\right), \qquad (4)$$

де $\bar{u} = \frac{1}{n}\sum_i n_i u_i$, $\overline{u^2} = \frac{1}{n}\sum_i n_i u_i^2$, а $u_i = \frac{x_i - c}{h}$.

Для більш детального вивчення властивостей розподілу досліджуваної величини використовують емпіричні моменти і центральні емпіричні моменти. Емпіричним моментом r – го порядку (r>0) називають величину

$$\hat{\alpha}_r = \frac{1}{n}\sum_i n_i x_i^r\,,$$

а центральним емпіричним моментом -

$$\hat{\mu}_r \quad \text{або} \quad \mu_r = \frac{1}{n}\sum_i \left(x_i - \bar{x}\right)^r n_i .$$

Розвязування типових вправ.

Приклад. Знайти вибіркову середню і вибіркову дисперсію за заданим емпіричним розподілом

x_i	18,4	18,9	19,3	19,6
n_i	5	10	20	15

Розв'язування. Використаємо формули $\bar{x} = h\bar{u} + c$, $D_B = h^2\left(\overline{u^2} - \left(\bar{u}\right)^2\right)$, де $\bar{u} = \frac{1}{n}\sum_i n_i u_i$, $\overline{u^2} = \frac{1}{n}\sum_i n_i u_i^2$, а $u_i = \frac{x_i - c}{h}$. Покладемо c = 19,3, h = 1. У відповідності до вибраних формул складемо розрахункову таблицю

x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$
18,4	5	-0,9	-4,5	4,05
18,9	10	-0,4	-4	1,6
19,3	20	0	0	0
19,6	15	0,3	4,5	1,35
$\sum$	n=50		$\sum n_i u_i$ =4	$\sum n_i u_i^2$ =7

При заповненні третього стовпчика ($u_i = x_i - c$) знаходимо різниці чисел першого і другого стовпчиків, при заповненні четвертого стовпчика знаходимо добуток чисел другого і третього стовпчиків, при заповненні п'ятого стовпчика знаходимо добуток чисел третього і четвертого стовпчиків, що розміщені в одному рядку. В самому нижньому рядку записуємо суми чисел відповідних стовпчиків. Підставляємо одержані значення у формули (4). Тоді $\bar{u} = \frac{1}{50}\cdot(-4)$ = -0,08; $\overline{u^2} = \frac{1}{50}\cdot 7 = 0{,}14$. Отже

$\bar{x}$ = -0,08+19,3 = 19,22; $D_B = 0{,}14 - (-0{,}08)^2 \approx 0{,}13$.

Задачі і вправи для самостійного розв'язання.

8.За статистичними даними вправ 1 і 2 (після їх виконання) знайти вибіркову середню і вибіркову дисперсію.

9.Знайти вибіркову середню і вибіркову дисперсію за заданим інтервальним розподілом розміру денного виробітку тканин 100 ткалями:

Денний виробіток, м	40-44	44-48	48-52	52-56	56-60
Кількість ткаль	12	28	36	16	8

10.Знайти вибіркову середню і вибіркову дисперсію за даними інтервальним розподілом росту(в см) 100 випадково відібраних студентів.

Ріст	154-158	158-162	162-166	166-170	170-174	174-178	178-182
К-ть студентів	10	14	26	28	12	8	2

Заняття 3. Статистичні оцінки параметрів розподілу.

Теоретичні відомості.

Будь-яка борелева функція $f(X_1,...,X_n)$ від результатів спостережень називається *статистикою*. Тому статистика є випадковою величиною, що визначена на вибірковому просторі R^n. Наприклад, $X^{(1)} = \min\{X_1,...,X_n\}$ і $X^{(n)} = \max\{X_1,...,X_n\}$ - крайні члени варіаційного ряду, є статистиками. Прикладами статистик є також вибіркова середня, вибіркова дисперсія, емпіричні моменти, емпірична функція розподілу.

Статистика, яка приймається за наближене значення невідомого параметру, називається *статистичною оцінкою*. Отже статистична оцінка є функцією, що визначена на вибірковому просторі R^n і приймає значення в множині Θ. Оцінка $\hat{\theta}$ називається *незміщеною* оцінкою невідомого параметру θ, якщо $M\hat{\theta} = \theta$. Якщо ж $M\hat{\theta} \neq \theta$, то оцінка називається зміщеною і величина $M\hat{\theta} - \theta$ називається зміщенням оцінки.

Оцінка $\hat{\theta}_n$ називається *спроможною* (*конзистентною*), якщо $\hat{\theta}_n \xrightarrow[n\to\infty]{p} \theta$.

Умова $D\hat{\theta}_n \xrightarrow[n\to\infty]{} 0$ є достатньою умовою спроможності.

Якщо $M\hat{\theta} \neq \theta$ але $M\hat{\theta}_n \xrightarrow[n\to\infty]{} 0$, то оцінку $\hat{\theta}_n$ називають *асимптотично незміщеною* оцінкою параметру θ.

Якщо $\hat{\theta}_n \xrightarrow[n\to\infty]{} \theta$ з ймовірністю 1, то $\hat{\theta}_n$ називають *сильно спроможною* оцінкою.

Розв'язування типових вправ.

Приклад. Нехай досліджувана випадкова величина ξ має рівномірний розподіл на $[a,b]$. Розглянемо оцінку $\hat{\theta} = \min\{X_1,...,X_n\}$. Чи буде $\hat{\theta}$ незміщеною, спроможною оцінкою для деякого із параметрів a або b досліджуваної випадкової величини ξ. Знайти розподіл цієї оцінки.

Розв'язування. Щільність рівномірного розподілу на $[a,b]$ має вигляд

$$p(x) = \begin{cases} \dfrac{1}{b-a}, & x \in (a,b), \\ 0, & x \notin (a,b), \end{cases}$$ а функція розподілу $$F(x) = \begin{cases} 0, & x \le a, \\ \dfrac{x-a}{b-a}, & a < x \le b, \\ 1, & x > b. \end{cases}$$

Знайдемо функцію розподілу $\hat{\theta}$

$$F_{\hat{\theta}}(x)=P\{\min(X_1,...,X_n)<x\}=1-P\{\min(X_1,...,X_n)\geq x\}=1-P\left\{\bigcap_{k=1}^{n}(X_k\geq x)\right\}=$$

$$=1-\prod_{k=1}^{n}P\{X_k\geq x\}=1-\prod_{k=1}^{n}\left(1-P\{X_k<x\}\right)=1-\left(1-F(x)\right)^n.$$

Тобто,

$$F_{\hat{\theta}}(x)=1-\left(1-F(x)\right)^n.$$

Тому $p_{\hat{\theta}}(x)=F'_{\hat{\theta}}(x)=1-\left(1-F(x)\right)^{n-1}F'_{\hat{\theta}}(x)=\begin{cases}0, & x\notin(a,b),\\ n\dfrac{(b-x)^{n-1}}{(b-a)^n}, & x\in(a,b).\end{cases}$ Знайдемо тепер

$$M\hat{\theta}=\int_{-\infty}^{+\infty}xp_{\hat{\theta}}(x)\,dx=\int_a^b nx\frac{(b-x)^{n-1}}{(b-a)^n}dx=\frac{n}{(b-a)^n}\int_a^b x(b-x)^{n-1}dx=$$

$$=\frac{1}{(b-a)^n}\left(-x(b-x)^n\Big|_a^b+\int_a^b(b-x)^n\,dx\right)=a+\frac{b-a}{n+1}.$$

Отже $M\hat{\theta}\neq a$, тому $\hat{\theta}$ є зміщеною оцінкою параметра a. Але $M\hat{\theta}\to a$ при $n\to\infty$, тому $\hat{\theta}$ є асимптотично незміщеною оцінкою параметра a.

Оцінка $\hat{\theta}$ буде спроможною оцінкою параметра $\theta=a$, якщо при $n\to\infty$ $\hat{\theta}\xrightarrow[n\to\infty]{p}\theta$ або $\forall\varepsilon>0\ \lim\limits_{n\to\infty}P\left\{\left|\hat{\theta}-\theta\right|<\varepsilon\right\}=1$. Нехай $\varepsilon<b-a$, тоді

$$P\left\{\left|\hat{\theta}-a\right|<\varepsilon\right\}=P\left\{a-\varepsilon<\hat{\theta}<a+\varepsilon\right\}=F_{\hat{\theta}}(a+\varepsilon)-F_{\hat{\theta}}(a-\varepsilon).$$

Оскільки

$$F_{\hat{\theta}}(a+\varepsilon)=1-\left(1-F(a+\varepsilon)\right)^n=1-\left(1-\frac{b-a-\varepsilon}{b-a}\right)^n=1-\left(\frac{\varepsilon}{b-a}\right)^n,$$

а

$$F_{\hat{\theta}}(a-\varepsilon)=\left(1-\left(1-F(a-\varepsilon)\right)^n\right)=0,$$

бо $F(a-\varepsilon)=0$, то

$$P\left\{\left|\hat{\theta}-a\right|<\varepsilon\right\}=1-\left(\frac{\varepsilon}{b-a}\right)^n\xrightarrow[n\to\infty]{}1.$$

Отже, $\hat{\theta}$ є спроможною оцінкою параметра a рівномірного розподілу.

Задачі і вправи для самостійного розв'язання .

11.Нехай $X_1,...,X_n$ - вибірка із розподілу зі щільністю $p(x,\alpha)=\begin{cases}0, & x\leq 0,\\ \alpha^{-1}e^{-x/\alpha}, & x>0,\end{cases}$ $(\alpha>0)$. Довести, що $\bar{X}=\dfrac{1}{n}\sum\limits_{i=1}^{n}X_i$ є незміщеною і спроможною оцінкою параметра α.

12.Нехай досліджувана випадкова величина ξ має рівномірний розподіл на $[a,b]$. Показати, що оцінка $\hat{\theta} = \max\{X_1,...,X_n\}$ буде асимптотично незміщеною та спроможною оцінкою для параметра b. Знайти розподіл цієї оцінки.

13. Показати, що центральні емпіричні моменти $\hat{\mu}_r = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^r$ є спроможними оцінками відповідних теоретичних моментів.

14. Нехай $\zeta = (\xi_1, \xi_2, ..., \xi_n)$- вибірка з генеральної сукупності з пуасонівським розподілом $P\{\xi = k\} = \frac{\theta^k}{k!}, k = 0,1,..., \theta > 0.$ Показати, що статистика $\hat{\theta}_n = \bar{\xi}$ є незміщеною та ефективною оцінкою параметра θ.

15. Нехай $\zeta = (\xi_1, \xi_2, ..., \xi_n)$- вибірка з генеральної сукупності з розподілом Паскаля $P\{\xi = k\} = \frac{\theta^k}{(1+\theta)^{k+1}}, k = 0,1,..., \theta > 0.$ Показати, що статистика $\hat{\theta}_n = \bar{\xi}$ є незміщеною та ефективною оцінкою параметра θ.

16. Одновимірний вектор ζ набуває скінченого числа значень 0,1,...,n з імовірністю $P\{\zeta = x\} = C_n^x \theta^x (1-\theta)^{n-x}, x = 0,1,...,n, \theta \in (0,1)$. Довести, що статистика $\hat{\theta}_n = \frac{\zeta}{n}$ є незміщеною та ефективною оцінкою параметра q (біномний розподіл).

17.Випадкові величини $\xi_1, \xi_2, ..., \xi_n$ – незалежні й однаково розподілені зі щільністю $p(x,\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, x \geq 0, \theta > 0.$ Знайти достатню статистику і записати щільність її розподілу.

Заняття 4. Ефективні оцінки. Нерівність Крамера-Рао.

Теоретичні відомості.

Розглянемо вибірку (вибірковий вектор) $X = (X_1,..., X_n)$ фіксованого обсягу *n*. Нехай θ - одновимірний параметр, а Θ - інтервал в множині дійсних чисел. Позначимо через $f(x,\theta) = f(x_1,..., x_n, \theta)$- щільність розподілу X, якщо досліджувана випадкова величина ξ неперервна. Якщо розподіл ξ - дискретний, то $f(x,\theta) = P_\theta\{X = x\}$, при цьому, $f(x,\theta) > 0$ тільки для скінченної або зліченної множини точок $x^{(k)} = \left(x_1^{(k)},..., x_n^{(k)}\right)$. Нехай $p(x,\theta)$- щільність розподілу досліджуваної випадкової величини ξ, якщо ξ - неперервна, і $p(x,\theta) = P_\theta\{\xi = x\}$, якщо розподіл ξ - дискретний. Оскільки результати спостережень незалежні, то $f(x,\theta) = \prod_{k=1}^{n} p(x_k,\theta)$. Надалі ми будемо розглядати випадкову функцію $f(X,\theta)$, що залежить від аргумента θ. Цю функцію позначають також через $L(X,\theta)$ і називають функцією правдоподібності. Для

одержання нерівності Крамера-Рао на функцію $f(X,\theta)$ будемо накладати деякі обмеження.

Лема. Нехай для будь-якого $\theta \in \Theta$ існують похідні $\frac{\partial}{\partial\theta} f(x,\theta)$, $\frac{\partial^2}{\partial\theta^2} f(x,\theta)$ і виконуються умови

$$M\left|\frac{\partial \ln f(X,\theta)}{\partial\theta}\right| < \infty, \quad M\left|\frac{\partial^2 \ln f(X,\theta)}{\partial\theta^2}\right| < \infty, \quad M\left|\frac{\partial \ln f(X,\theta)}{\partial\theta}\right|^2 < \infty.$$

Тоді для всіх $\theta \in \Theta$

$$M\frac{\partial \ln f(X,\theta)}{\partial\theta} = 0; \quad M\left|\frac{\partial \ln f(X,\theta)}{\partial\theta}\right|^2 = -M\frac{\partial^2 \ln f(X,\theta)}{\partial\theta^2}. \tag{1}$$

Функцію

$$I_n(\theta) = M\left(\frac{\partial \ln f(X,\theta)}{\partial\theta}\right)^2$$

називають інформацією Фішера, що міститься у вибірці обсягу n, про значення параметра θ. Із леми випливає, що інформацію Фішера можна подати у вигляді

$$I_n(\theta) = -M\left(\frac{\partial^2 \ln f(X,\theta)}{\partial\theta^2}\right).$$

Якщо $f(X,\theta) = \prod_{i=1}^{n} p(X_i,\theta)$, то $\frac{\partial}{\partial\theta}\ln f(X,\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln p(X_i,\theta)$, $\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta) = \sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln p(X_i,\theta)$. Тому

$$I_n(\theta) = -M\left(\frac{\partial^2 \ln f(X,\theta)}{\partial\theta^2}\right) = -\sum_{i=1}^{n} M\frac{\partial^2}{\partial\theta^2}\ln p(X_i,\theta) = nI(\theta),$$

тобто,

$$I_n(\theta) = nI(\theta),$$

де

$$I(\theta) = M\left(\frac{\partial \ln p(X_i,\theta)}{\partial\theta}\right)^2 = -M\left(\frac{\partial^2 \ln p(X_i,\theta)}{\partial\theta^2}\right).$$

Теорема. (Крамера-Рао). Нехай виконуються умови леми і $h(\theta)$ - диференційовна функція на Θ, для якої існує незміщена оцінка $h(X)$ із скінченною дисперсією, що задовольняє умові

$$\int_{R^n}\left|h(x)\frac{\partial}{\partial\theta} f(x,\theta)\right| dx < \infty, \ \forall\theta \in \Theta.$$

Тоді

$$M\left(h(X) - h(\theta)\right)^2 \geq \frac{\left(h'(\theta)\right)^2}{I_n(\theta)}, \tag{2}$$

при цьому знак рівності має місце тоді і тільки тоді, коли

$$\frac{\partial \ln f(x,\theta)}{\partial \theta} = c(\theta)\Big(h(x) - h(\theta)\Big). \qquad (3)$$

Нерівність (2) називають *нерівністю Крамера-Рао*. Якщо $h(\theta) = \theta$, то нерівність Крамера-Рао набуває вигляду

$$M\left(h(X) - \theta\right)^2 \geq \frac{1}{I_n(\theta)}.$$

Нерівність Крамера-Рао встановлює нижню межу для дисперсій всіх незміщених оцінок. Якщо для деякої оцінки $\hat{h}(X)$ нерівність Крамера-Рао перетворюється в рівність

$$D\Big(h(X)\Big) = \frac{\left(h'(\theta)\right)^2}{I_n(\theta)},$$

то оцінка $h(X)$ буде ефективною оцінкою $h(\theta)$.

Рівність (3) є необхідною і достатньою умовою того, що оцінка $h(X)$ є ефективною оцінкою $h(\theta)$. Якщо $M\Big(h(X) - h(\theta)\Big) = 0$ то оцінка $h(X)$ є незміщеною оцінкою $h(\theta)$.

Якщо $h(X)$ є ефективною оцінкою параметра θ, то $D\Big(h(X)\Big) = \frac{1}{nI(\theta)} = O\left(\frac{1}{n}\right)$.

Тобто ефективна оцінка завжди спроможна і незміщена.

Послідовність незміщених оцінок θ_n параметру θ, що побудовані за вибіркою обсягу n, називається *асимптотично ефективною*, якщо відношення дисперсії оцінки до нижньої межі Крамера-Рао дисперсій прямує до 1:

$$I_n(\theta) D\Big(\theta_n\Big) \xrightarrow[n\to\infty]{} 1.$$

Розв'язування типових вправ.

Приклад 1. Задано вибірку $x_1, ..., x_n$ із нормального розподілу. Довести, що вибіркова середня є ефективною оцінкою математичного сподівання нормально розподіленої випадкової величини. Знайти ефективну оцінку дисперсії нормального закону при відомому математичному сподіванні.

Розв'язування. Невідомим параметром є $M\xi = a = \theta$ $(h(\theta) = \theta)$. Розглянемо статистичну оцінку $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Ми розглянемо два підходи. Перший оснований на нерівності Крамера-Рао $D\Big(h(X)\Big) \geq \frac{1}{I_n(\theta)}$. Для цього знайдемо спочатку дисперсію оцінки $D\Big(\hat{\theta}\Big) = D\Big(\overline{X}\Big) = D\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{D\xi}{n} = \frac{\sigma^2}{n}$. А тепер знайдемо інформацію Фішера, для цього вигідніше використовувати таку формулу $I_n(\theta) = -M\frac{\partial^2 \ln f(X,\theta)}{\partial \theta^2}$, бо друга похідна частіше має простіший вигляд і легше знаходити від неї математичне сподівання. Щільність нормального розподілу має вигляд

$$p(x,\theta)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right).$$

Тоді

$$f(x,\theta)=\prod_{i=1}^{n}p(x_i,\theta)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right),$$

а

$$\ln f(X,\theta)=-n\ln\left(\sqrt{2\pi}\sigma\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\theta)^2.$$

Продиференціюємо двічі цю рівність по θ: $\frac{\partial\ln f(X,\theta)}{\partial\theta}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\theta)$, $\frac{\partial\ln^2 f(X,\theta)}{\partial\theta^2}=-\frac{n}{\sigma^2}$. Звідки, $I_n(\theta)=-M\frac{\partial^2\ln f(X,\theta)}{\partial\theta^2}=\frac{n}{\sigma^2}$. Отже,

$$\frac{\sigma^2}{n}=D\left(\overline{X}\right)=\frac{1}{I_n(\theta)}=\frac{\sigma^2}{n}.$$

Тобто, нерівність Крамера-Рао перетворилась у рівність, а це означає, що вибіркова середня є ефективною оцінкою математичного сподівання нормального розподілу.

Другий підхід базується на рівності (10). Першій похідній надамо вигляду

$$\frac{\partial\ln f(X,\theta)}{\partial\theta}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\theta)=\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}X_i-n\theta\right)=\frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\theta\right).$$

Із рівності (10) випливає, що $\hat{\theta}=\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ є ефективною оцінкою параметра $\theta=a=M\xi$ нормального розподілу.

Нехай невідомим є $\theta=\sigma^2$, а a - відоме. Тоді

$$f(X,a,\theta)=\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi\theta}}\exp\left(-\frac{(X_k-a)^2}{2\theta}\right)=\left(\frac{1}{\sqrt{2\pi\theta}}\right)^n\exp\left(-\sum_{k=1}^{n}\frac{(X_k-a)^2}{2\theta}\right),$$

$$\ln f(X,a,\theta)=-n\ln\sqrt{2\pi}-\frac{n}{2}\ln\theta-\frac{1}{2\theta}\sum_{k=1}^{n}(X_k-a)^2,$$

$$\frac{\partial\ln f(X,a,\theta)}{\partial\theta}=-\frac{n}{2\theta}+\frac{1}{2\theta^2}\sum_{i=1}^{n}(X_k-a)^2=\frac{n}{2\theta^2}\left(\frac{1}{n}\sum_{k=1}^{n}(X_k-a)^2-\theta\right).$$

Звідки, $c(\theta)=\frac{n}{2\theta^2}$, $\hat{h}(X)=\frac{1}{n}\sum_{i=1}^{n}(X_i-a)^2$. Умова (10) існування ефективної оцінки виконується і оцінка $\hat{\sigma}^2=\frac{1}{n}\sum_{k=1}^{n}(X_k-a)^2$ є ефективною оцінкою дисперсії σ^2 нормального закону.

Приклад 2. Розглянемо біномний розподіл. За вибіркою $x_1,...,x_n$, де $x_i=0$ або 1, оцінити ймовірність $p=\theta$ настання події в кожному експерименті.

Розв’язування. Оскільки,

$$p(x,\theta)=P\{\xi=x\}=\theta^x(1-\theta)^{1-x},$$

то

$$f(X,\theta)=\prod_{k=1}^{n}\theta^{X_k}(1-\theta)^{1-X_k}=\theta^{\sum_{k=1}^{n}X_k}(1-\theta)^{n-\sum_{k=1}^{n}X_k},$$

$$\ln f(X,\theta)=\sum_{k=1}^{n}X_k\ln\theta+\left(n-\sum_{k=1}^{n}X_k\right)\ln(1-\theta),$$

$$\frac{\partial\ln f(X,\theta)}{\partial\theta}=\sum_{k=1}^{n}X_k\frac{1}{\theta}-\left(n-\sum_{k=1}^{n}X_k\right)\frac{1}{1-\theta}=\frac{n}{\theta(1-\theta)}\left(\frac{1}{n}\sum_{k=1}^{n}X_k-\theta\right).$$

Звідки $\theta^{*}=\overline{X}=\frac{1}{n}\sum_{k=1}^{n}X_k$. Оскільки $\sum_{k=1}^{n}X_k=\mu_n$ є число появ події в n експериментах, то $\hat{\theta}=\frac{\mu_n}{n}$ є відносною частотою події в n експериментах. Отже, з теореми Крамера-Рао випливає, що відносна частота є ефективною оцінкою ймовірності появи події в схемі Бернуллі.

Задачі і вправи для самостійного розв'язання.

18.Нехай $x_1,...,x_n$ вибірка із розподілу Пуассона з параметром λ. Показати, що вибіркова середня є ефективною оцінкою параметра λ.

19.Нехай $x_1,...,x_n$ вибірка із показникового розподілу із щільністю $p(x,\theta)=\begin{cases}\frac{1}{\theta}e^{-\frac{x-m}{\theta}}, & x\ge m,\\ 0, & x<m,\end{cases}$ $\theta>0$, m - відоме. Чи є $\hat{\theta}=\overline{X}-m$ ефективною оцінкою параметра θ.

20.Нехай $x_1,...,x_n$ вибірка із показникового розподілу із щільністю $p(x,\theta)=\theta e^{-\theta x}, x\ge 0.$ Знайти оцінку параметра показникового розподілу. Чи є вона ефективною.

Заняття 5. Методи одержання статистичних оцінок.

Теоретичні відомості.

Метод моментів одержання статистичних оцінок базується на тому факті, що емпіричні моменти є незміщеними і спроможними оцінками відповідних теоретичних моментів.

Нехай розподіл $F(x,\theta)$ залежить від векторного параметра $\theta=(\theta_1,...,\theta_s)$. Тоді на основі розподілу $F(x,\theta)$ можна знайти, за умови їх існування, теоретичні моменти $M\xi,...,M\xi^s$. Ці теоретичні моменти будуть залежати від невідомих параметрів, тобто $M\xi^k=\alpha_k(\theta_1,...,\theta_s)$.*Оцінками за методом моментів* для невідомих параметрів $\theta_1,...,\theta_s$ називають розв'язок системи рівнянь

$$\begin{cases} \alpha_1(\theta_1,...,\theta_s) = \overline{X}, \\ \alpha_2(\theta_1,...,\theta_s) = \overline{X^2}, \\ ..... \\ \alpha_s(\theta_1,...,\theta_s) = \overline{X^s}. \end{cases}$$

Якщо теоретичні моменти є неперервними функціями параметрів, то оцінки за методом моментів будуть спроможними.

Переходимо до розгляду методу *максимальної правдоподібності*. Нехай $p(x,\theta)$ - щільність розподілу випадкової величини ξ, якщо величина неперервна і $p(x,\theta) = P_\theta\{\xi = x\}$, якщо випадкова величина ξ - дискретна. Параметр θ приймає значення із деякої множини $\Theta \subset R^s$, $\theta = (\theta_1,...,\theta_s)$. Нехай маємо вибірку об'єму n $X = (X_1,...,X_n)$. Позначимо через $f(x,\theta)$ - щільність розподілу вибірки X, тоді $L(x,\theta) = f(x,\theta) = p(x_1,\theta)\cdot...\cdot p(x_n,\theta)$. Таку функцію називають *функцією правдоподібності*.

Метод оцінки параметра θ, що одержав назву методу максимальної правдоподібності, грунтується на тому, що в більшій частині випадків у експерименті спостерігається таке значення вектора X, при якому щільність $f(x,\theta)$ близька до максимального значення. Отже, за оцінку θ невідомого параметру θ природно прийняти розв'язок рівняння $f(X,\theta) = \max\limits_{\theta} f(X,\theta)$, якщо це рівняння має зміст і воно має розв'язки.

Розв'язки рівняння

$$f(X,\theta) = \max\limits_{\theta} f(X,\theta)$$

називають *оцінками за методом максимальної правдоподібності*.

$$\begin{cases} \dfrac{\partial \ln L(x,\theta_1,...,\theta_s)}{\partial \theta_1} = 0, \\ ..... \\ \dfrac{\partial \ln L(x,\theta_1,...,\theta_s)}{\partial \theta_s} = 0. \end{cases}$$

Ці рівняння називають рівняннями максимальної правдоподібності. У більшості випадків, розв'язки цієї системи і є оцінками максимальної правдоподібності невідомих параметрів $\theta_1,...,\theta_s$. При розв'язуванні рівнянь правдоподібності необхідно відкидати розв'язки вигляду $\theta = const$ і розглядати лише ті розв'язки, які залежать від $x_1,...,x_n$ і попадають в область допустимих значень параметра Θ.

Рівняння максимальної правдоподібності має вигляд:

$$\frac{\partial \ln f(X,\theta)}{\partial \theta} = c(\theta)\big(h(X) - \theta\big) = 0.$$

Отже, $h(X)$ є оцінкою за методом максимальної правдоподібності параметра θ. За теоремою Крамера-Рао $h(X)$ є незміщеною і ефективною оцінкою параметра θ.

Розв’язування типових вправ.

Приклад. Нехай $x_1,...,x_n$ - вибірка із рівномірного розподілу із щільністю $p(x)=\begin{cases}\dfrac{1}{b-a}, x\in(a,b),\\ 0, \quad x\notin(a,b).\end{cases}$ За методом моментів знайти оцінки параметрів a і b.

Розв’язування. На основі щільності рівномірного розподілу ми раніше уже знайшли $M\xi=\dfrac{a+b}{2}$, $M\left(\xi^2\right)=\dfrac{b^2+ab+a^2}{3}$. Тому для знаходження оцінок одержуємо систему рівнянь: $\dfrac{a+b}{2}=\overline{X}$, $\dfrac{b^2+ab+a^2}{3}=\overline{X^2}$. Піднесемо перше рівняння до квадрату і віднімемо від другого. Тоді система набуде вигляду

$$\begin{cases}\dfrac{a+b}{2}=\bar{X},\\ \dfrac{(b-a)^2}{12}=D_B\end{cases} \text{ або } \begin{cases}a+b=2\bar{X},\\ b-a=2\sqrt{3}\sigma_B.\end{cases} \text{ Звідки, } \hat{a}=\bar{X}-\sqrt{3}\sigma_B,\ \hat{b}=\bar{X}+\sqrt{3}\sigma_B.$$

Приклад. За вибіркою $x_1,...,x_n$ із нормального закону знайти оцінки максимальної правдоподібності параметрів $a=\theta_1$ і $\sigma^2=\theta_2$.

Розв’язування. На основі щільності нормального закону

$$p(x,\theta_1,\theta_2)=\frac{1}{\sqrt{2\pi\theta_2}}\exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right)$$

запишемо функцію правдоподібності

$$L(X,\theta_1,\theta_2)=f(X,\theta_1,\theta_2)=\prod_{i=1}^{n}p(X_i,\theta_1,\theta_2)=\left(\frac{1}{\sqrt{2\pi\theta_2}}\right)^n\exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(X_i-\theta_1)^2\right).$$

Тоді

$$\ln L(X,\theta_1,\theta_2)=-n\ln\left(\sqrt{2\pi}\right)-\frac{n}{2}\ln\theta_2-\frac{1}{2\theta_2}\sum_{i=1}^{n}(X_i-\theta_1)^2.$$

Знаходимо частинні похідні по θ_1 і θ_2:

$$\frac{\partial\ln L(X,\theta_1,\theta_2)}{\partial\theta_1}=\frac{1}{\theta_2}\sum_{i=1}^{n}(X_i-\theta_1),\quad \frac{\partial\ln L(X,\theta_1,\theta_2)}{\partial\theta_2}=-\frac{n}{2\theta_2}+\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(X_i-\theta_1)^2.$$

Прирівнявши похідні до нуля, записуємо систему рівнянь правдоподібності:

$$\begin{cases}\dfrac{1}{\theta_2}\displaystyle\sum_{i=1}^{n}(X_i-\theta_1)=0,\\ -\dfrac{n}{2\theta_2}+\dfrac{1}{2\theta_2^2}\displaystyle\sum_{i=1}^{n}(X_i-\theta_1)^2=0.\end{cases}$$

Із першого рівняння знаходимо

$$\hat{\theta}_1=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}.$$

Підставивши це значення в друге рівняння, одержимо

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = D_B.$$

Тобто оцінками максимальної правдоподібності математичного сподівання і дисперсії нормального закону є відповідно вибіркова середня і вибіркова дисперсія.

Задачі і вправи для самостійного розвязання

21.Нехай $x_1,...,x_n$ - вибірка із геометричного розподілу $p(k,p)=(1-p)^k p$, $k=0,1,...$ Оцінити параметр p за методом моментів.

22.Нехай $x_1,...,x_n$ - вибірка із гамма-розподілу $p(x,\alpha,\beta)=\frac{1}{\beta^{\alpha+1}\Gamma(\alpha+1)}x^{\alpha}e^{-x/\beta}$, ($x>0$, $\alpha>-1$, $\beta>0$). Оцінити параметри α і β за методом моментів.

23.Нехай $x_1,...,x_n$ - вибірка із гамма-розподілу $p(x,\alpha,\beta)=\frac{1}{\beta^{\alpha+1}\Gamma(\alpha+1)}x^{\alpha}e^{-x/\beta}$, ($x>0$, $\alpha>-1$, $\beta>0$). Знайти оцінку параметра β за методом максимальної правдоподібності.

24.Методом максимальної правдоподібності знайти оцінку параметра λ показникового розподілу $p(x,\lambda)=\lambda e^{-\lambda x}$ ($\lambda>0$, $x>0$), якщо в результаті n експериментів випадкова величина прийняла значення $x_1,...,x_n$.

25.За вибіркою $x_1,...,x_n$ знайти методом максимальної правдоподібності оцінку ймовірності успіху в схемі Бернуллі ($\theta=p$).

26.Використовуючи метод моментів, знайти за вибіркою $\xi_1,\xi_2,...,\xi_n$, де $P\{\xi_k=m\}=e^{-\theta}\frac{\theta^m}{m!}, m=0,1,...,$ оцінку θ_n параметра θ. Чи буде оцінка незміщеною, спроможною? Знайти також оцінку максимальної правдоподібності параметра θ.

27.Нехай $x_1,...,x_n$ - вибірка з генеральної сукупності зі щільністю $p(x,\theta)=k(\theta)x^2e^{-\frac{x^2}{\theta^2}}, x\geq 0, \theta>0.$ Знайти функцію $k(\theta)$, оцінку параметра θ методом моментів. Чи буде оцінка незміщеною, спроможною? Знайти також оцінку максимальної правдоподібності параметра θ.

28.Методом максимальної правдоподібності знайти оцінку параметра θ розподілу $P\{\xi_k=m\}=\frac{(\theta-1)^m}{\theta^{m+1}}, m=0,1,...,\theta>1$. Чи буде ця оцінка незміщеною та ефективною?

29.Задано емперічний розподіл

x_i	1,2	1,6	2,0	2,4	2,8
n_i	18	18	16	22	26

Випадкової величини, що має рівномірний розподіл. Знайти методом моментів оцінки параметрів a і b $(a<b)$.

30. Випадкова величина ξ задовольняє біноміальному закону з невідомим параметром p. Знайти методом максимальної правдоподібності точкову оцінку невідомого параметра p біноміального розподілу якщо

x_i	0	1	2	3	4	5	6	7
n_i	2	3	10	22	26	20	12	5

Заняття 6. Інтервальні оцінки невідомих параметрів розподілу.

Теоретичні відомості.

Розглянемо дві оцінки θ_H і θ_B невідомого параметра θ і розглянемо інтервал (θ_H,θ_B), де $\theta_H<\theta_B$. Нехай $P\{\theta\in(\theta_H,\theta_B)\}=\gamma$, тобто $\gamma=P\{\theta_H<\theta<\theta_B\}$, тоді γ називається *надійністю* оцінки, а інтервал (θ_H,θ_B) - *надійним* або *довірковим інтервалом*, який з надійністю γ оцінює невідомий параметр θ (γ-довірковим інтервалом).

Для забезпечення однозначності визначення довіркового інтервалу (θ_H,θ_B) межі інтервалу θ_H і θ_B вибирають так, щоб виконувались умови $P\{\theta\le\theta_H\}=P\{\theta\ge\theta_B\}=\frac{\gamma}{2}$.

Якщо $|\theta-\theta|<\delta$, то в цьому випадку δ називається точністю оцінки. Виберемо δ так, щоб $P\{|\theta-\theta|<\delta\}=\gamma$. Тоді інтервал

$$\theta-\delta<\theta<\theta+\delta$$

буде довірковим інтервалом, який із надійністю γ оцінює невідомий параметр θ. Точність оцінки δ залежить від надійності γ і обсягу вибірки n. Знайдемо тепер довірковий інтервал для невідомої ймовірності $P(A)=p$ у схемі Бернуллі. Задамо надійність γ близьку до 1 і знайдемо t із рівняння $2\Phi(t)=\gamma$. Тоді при великих n із надійністю, що близька до γ, виконується нерівність $|Z|<t$

$$\frac{\tilde{p}n+\frac{1}{2}t^2-t\sqrt{\tilde{p}(1-\tilde{p})n+\frac{1}{4}t^2}}{n+t^2}<p<\frac{\tilde{p}n+\frac{1}{2}t^2+t\sqrt{\tilde{p}(1-\tilde{p})n+\frac{1}{4}t^2}}{n+t^2}.$$

Якщо, аналогічно переходу від (12) до (13), у лівій і правій частинах нерівності (14) p замінити на $\hat{p}$, то одержимо більш простий вигляд довіркового інтервалу для невідомої ймовірності:

$$\hat{p}-t\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}<p<\hat{p}+t\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \qquad (15)$$

Якщо врахувати, що $p(1-p)\le\frac{1}{4}$, то із ймовірністю, що близька до γ, при великих n виконується нерівність

$$\hat{p}-\frac{t}{2\sqrt{n}}<p<\hat{p}+\frac{t}{2\sqrt{n}}.$$

Але остання оцінка є досить грубою. Частіше використовують інтервал у вигляді (15). У цьому інтервалі величина $t\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=\delta$ є точність оцінки невідомої ймовірності із надійністю γ.

Довіркові інтервали для параметрів нормального розподілу. Щільність нормального розподілу з параметрами a і σ^2 має вигляд: $f\left(x,a,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$, де $a=M(\xi)$, $\sigma^2=D(\xi)$.

Розглянемо окремі випадки знаходження довіркових інтервалів для параметрів a і σ^2.

а) Знайдемо довіркові інтервали для невідомого метематичного сподівання $a=M(\xi)$, коли σ^2 - відоме. За оцінку невідомого математичного сподівання виберемо $\overline{X}$. Оскільки $\overline{X}=\frac{X_1+\cdots+X_n}{n}$, де X_i - нормально розподілені випадкові величини, то $\overline{X}$ є також номально розподіленою величиною із параметрами $M\overline{X}=a$ і $D\overline{X}=\frac{\sigma^2}{n}$. Тоді величина $Z=\left(\overline{X}-a\right)\frac{\sqrt{n}}{\sigma}$ буде мати стандартний нормальний розподіл і $P\{|Z|<t\}=2\Phi(t)$. За надійністю γ можна знайти t_γ як розв'язок рівняння $2\Phi(t)=\gamma$. Тоді із надійністю γ буде виконуватись нерівність

$$\overline{X}-t_\gamma\frac{\sigma}{\sqrt{n}}<a<\overline{X}+t_\gamma\frac{\sigma}{\sqrt{n}},$$

де величина

$$t_\gamma\frac{\sigma}{\sqrt{n}}=\delta$$

є точністю оцінки.

Тобто, довірковий інтервал, який із надійністю γ оцінює невідоме метематичне сподівання $a=M(\xi)$, коли σ^2 - відоме, має вигляд

$$\overline{X}-t_\gamma\frac{\sigma}{\sqrt{n}}<a<\overline{X}+t_\gamma\frac{\sigma}{\sqrt{n}},$$

де число t_γ знаходиться із рівняння $2\Phi\left(t_\gamma\right)=\gamma$.

б) Нехай σ^2 - невідоме. Величина $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ є незміщеною і спроможною оцінкою для σ^2. Якщо надійність γ задана, то за розподілом Стьюдента можна знайти таке число $t_{\gamma,n}$, яке задовольняє умову $P\left\{|T| < t_{\gamma,n}\right\} = \gamma$. Тоді із надійністю γ буде виконуватись нерівність

$$\overline{X} - t_{\gamma,n}\frac{s}{\sqrt{n}} < a < \overline{X} + t_{\gamma,n}\frac{s}{\sqrt{n}},$$

де

$$t_{\gamma,n}\frac{s}{\sqrt{n}} = \delta$$

- точність оцінки.

Тобто, довірковий інтервал, який із надійністю γ оцінює невідоме метематичне сподівання $a = M(\xi)$, коли σ^2 - невідоме, має вигляд

$$\overline{X} - t_{\gamma,n}\frac{s}{\sqrt{n}} < a < \overline{X} + t_{\gamma,n}\frac{s}{\sqrt{n}},$$

де число $t_{\gamma,n}$ знаходиться за надійністю γ і числом ступенів вільності $n-1$ за розподілом Стьюдента із умови $P\left\{|T| < t_{\gamma,n}\right\} = \gamma$.

Довірковий інтервал, який із надійністю γ оцінює невідому дисперсію σ^2, коли a - невідоме, має вигляд

$$\frac{(n-1)s^2}{h''} < \sigma^2 < \frac{(n-1)s^2}{h'},$$

де числа h' і h'' вибираються так, щоб виконувались умови $P\left\{\chi^2(n-1) < h'\right\} = \frac{1-\gamma}{2}$ і $P\left\{\chi^2(n-1) < h''\right\} = \frac{1+\gamma}{2}$.

Задачі і вправи для самостійного розв'язання.

31.При проведенні обстеження стану очей у 265 учнів 131 були хворі (49,4%). Встановити межі довіркового інтервалу для долі хворих дітей при надійності $\gamma = 0,95$.

32.За вибіркою об'єму $n = 25$ із нормального розподілу знайдені $\bar{x} = 18,2$ і $s^2 = 96,04$. Знайти довіркові інтервали для параметрів a і σ^2 при надійності $\gamma = 0,95$.

33.Знайти довірковий інтервал для оцінки з надійністю γ невідомого математичного сподівання нормально розподіленої ознаки X генеральної сукупності, якщо відомі

а) $\sigma = 4; \overline{x_b} = 10,2; n = 16; \gamma = 0,99;$ б) $\sigma = 5; \overline{x_b} = 16,8; n = 25; \gamma = 0,99;$

в) $\sigma = 40; \overline{x_b} = 2000; n = 5; \gamma = 0,95;$ г) $\sigma = 40; \overline{x_b} = 1000; n = 100; \gamma = 0,95;$

34.Знайти мінімальний об'єм вибірки при якому з надійністю 0,975 точність оцінки математичного сподівання a генеральної сукупності по вибірковій середній $\delta = 0,3$, якщо $\sigma = 1,2$ нормально розподіленої генеральної сукупності.

35.Із генеральної сукупності вилучена вибірка об'єму $n = 12$

x_i	-0,5	-0,4	-0,2	0	0,2	0,4	0,6	0,8	1	1,2	1,4
n_i	1	2	1	1	1	1	1	1	1	2	1

Оцінити з надійністю 0,95 математичне сподівання a нормально розподіленої ознаки генеральної сукупності за допомогою довіркового інтервалу.

36.По даним 16 незалежних рівно можливих вимірювань деякої фізичної величини знайдені $\bar{x}_b = 42,8; s = 8.$ Оцінити істинне значення величини з надійністю 0,95.

37. По даним вибірки об'єму із генеральної сукупності нормально розподіленої кількісної ознаки знайдено „виправлене" середнє квадратичне відхилення S. Знайти довірковий інтервал, що покриває генеральне середнє квадратичне відхилення σ з надійністю 0,99, якщо а) $n = 10; s = 5,1;$

б) $n = 50; s = 14;$

38. Із 580 інкубованих яєць вилупилось 535 курчат. Знайти долю вилуплених яєць і знайти межі довіркового інтервала для цієї долі при надійності 0,95.

39.Проводяться незалежні випробовування з однаковою, але невідомою ймовірністю p появи події A в кожному випробовуванні. Знайти довірковий інтервал для оцінки ймовірності p з надійністю 0,99, якщо в 100 випробовуваннях подія A зявилась 60 раз.

Заняття 7. **Перевірка статистичних гіпотез**

Теоретичні відомості

Статистичною гіпотезою (або просто *гіпотезою*) називають будь-яке твердження про вигляд або властивості розподілу досліджуваних в експерименті випадкових величин.

Висунуту гіпотезу, яка має піддаватися перевірці, позначають H_0 і називають *основною* або *нульовою* гіпотезою. Поряд із основною розглядають гіпотезу, що протирічить основній, її називають *альтернативною* (альтернативою до гіпотези H_0) або *конкурентною*. Статистики, які вибирають для перевірки гіпотез, називають статистиками критерію або *критеріями*.

Множину значень критерію розбивають на дві області: *область прийняття гіпотези* – множина значень критерію, при яких гіпотеза приймається; *критичну область S* - множина значень критерію, при яких гіпотеза відкидається. Ймовірність того, що критерій прийме значення із критичної області, називається *рівнем значущості* критерію і позначається α.

Застосування процедури перевірки гіпотези пов'язано із такими помилками: відкинути гіпотезу H_0, якщо вона правильна (*помилка першого роду*); прийняти гіпотезу H_0, якщо вона неправильна (*помилка другого роду*).

Теорема (Неймана – Пірсона). Серед всіх критеріїв із заданим рівнем значущості α, що перевіряють дві прості гіпотези H_0 і H_1, критерій відношення правдоподібності є найбільш потужним.

Перевірка гіпотез про ймовірності. Нехай задано деяку сукупність однорідних об'єктів. Необхідно перевірити гіпотезу – доля об'єктів із заданою властивістю дорівнює заданому числу p_0. У зв'язку з цим розглянемо просту гіпотезу $H_0 : p = p_0$ при альтернативній гіпотезі $H_1 : p \neq p_0$. Нехай проведено n експериментів, у яких подія настала μ_n разів. Ймовірність p є невідомою, але відносна частота $\frac{\mu_n}{n}$ є незміщеною і спроможною оцінкою невідомої ймовірності p, тому ми можемо порівняти відносну частоту $\frac{\mu_n}{n}$ з p_0, де $\frac{\mu_n}{n} = \frac{X_1 + \cdots + X_n}{n}$, X_i - число появ події в i-у експерименті. При справедливості висунутої гіпотези $MX_i = p_0$, $DX_i = p_0(1-p_0)$, $D\frac{\mu_n}{n} = \frac{p_0(1-p_0)}{n}$. Для перевірки гіпотези використаємо критерій - випадкову величину $Z = \frac{\frac{\mu_n}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\mu_n - np_0}{\sqrt{np_0(1-p_0)}}$. При справедливості гіпотези H_0 $MZ = 0$, $DZ = 1$, тому при великих n велична Z є асимптотично нормальною (0,1).

Нехай задано рівень значущості α. Виберемо критичну точку z_α так, щоб виконувалась умова $P\{|Z| < z_\alpha\} \approx 2\Phi(z_\alpha) = 1-\alpha$. Знайшовши z_α із рівняння $2\Phi(z_\alpha) = 1-\alpha$, ми поділяємо множину всіх значень Z на область прийняття гіпотези і критичну область. Критична область – це множина тих значень Z, для яких $|Z| \geq z_\alpha$. За результатами спостережень знаходимо спостережуване значення критерію $z_{сп}$. Якщо $|z_{сп}| < z_\alpha$, то гіпотезу приймають, у протилежному випадку - відкидають.

Розглянемо просту гіпотезу $H_0 : p = p_0$ при альтернативній гіпотезі $H_1 : p > p_0$. В цьому випадку критична область буде односторонньою: це множина тих значень Z, для яких $Z \geq z_\alpha$, а критична точка z_α знаходиться із умови $P\{Z \geq z_\alpha\} = \alpha$. критичну точку z_α знаходимо із умови $\Phi(z_\alpha) = 0{,}5 - \alpha$. За результатами спостережень знаходимо спостережуване значення критерію $z_{сп}$. Якщо $z_{сп} < z_\alpha$, то гіпотезу приймають, у протилежному випадку - відкидають.

Нехай розглядаються дві сукупності. Ймовірність настання події A в першій сукупності дорівнює p_1, а в другій - p_2. Часто виникає необхідність перевіряти гіпотезу $H_0 : p_1 = p_2$ при альтернативі $H_1 : p_1 \neq p_2$. Нехай в першій сукупності проведено n_1 експериментів, в яких подія настала m_1 разів, в другій сукупності проведено n_2 експерименти, в яких подія настала m_2 разів.

Тоді оцінками невідомих ймовірностей p_1 і p_2 будуть відповідні відносні частоти $\frac{m_1}{n_1}$ і $\frac{m_2}{n_2}$. Використаємо критерій $Z=\frac{\frac{m_1}{n_1}-\frac{m_2}{n_2}}{\sqrt{\frac{m_1+m_2}{n_1+n_2}\left(1-\frac{m_1+m_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ (*).

Величина Z є асимптотично нормальною з параметрами 0 і 1. Тому за рівнем значущості α можна знайти z_α із умови $P\{|Z|<z_\alpha\}\approx 2\Phi(z_\alpha)=1-\alpha$. За результатами спостережень знаходимо спостережуване значення критерію z_{cn}. Якщо $|z_{cn}|<z_\alpha$, то гіпотезу приймають, у протилежному випадку - відкидають.

Розв'язування типових вправ.

1.По 100 незалежним випробовуванням знайдена відносна частота $\frac{m}{n}=0{,}14$. При рівні значущості 0,05 необхідно перевірити нульову гіпотезу $H_0: p=p_0=0{,}20$ при конкуруючій гіпотезі $H_1: p\neq 0{,}20$.

Розв'язання. Знайдемо спостерігальне значення критерія , враховуючи, що $q_0=1-p_0=0{,}80: Z=\frac{\frac{\mu_n}{n}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}=\frac{\mu_n-np_0}{\sqrt{np_0(1-p_0)}}$=-1,5. По умові, конкуруюча гіпотеза має вигляд $p\neq p_0$, тому критична область двохстороння. Знайдемо критичну точку z_{kp} за формулою $\Phi(z_{kp})=(1-\alpha)/2=(1-0{,}05)/2=0{,}475$(**). По таблиці функції Лапласа знаходимо z_{kp}=1,96. Оскільки $|z_{cn}|<z_{kp}$ - немає підстав відкидати нульову гіпотезу. Іншими словами, спостерігальна відносна частота не суттєво відрізняється від гіпотетичної ймовірності 0,20.

2.За зміну відмовили 15 елементів приладу 1, що складається із 800 елементів і 25 елементів приладу 2, що складається із 1000 елементів. При рівні значущості 0,05 перевірити нульову гіпотезу $H_0: p_1=p_2=p$ про рівність ймовірностей відмови елементів двох приладів при конкуруючій гіпотезі $H_1: p_1\neq p_2$.

Розв'язання. По умові конкуруюча гіпотеза має вид $H_1: p_1\neq p_2$, тому критична область двостороння. Знайдемо спостерігальне значення критерія по формулі (*) і підставив відповідні значення одержимо $z_{cn}=-0{,}89$. Знайдемо критичну точку по формулі (**). По таблиці функції Лапласа знаходимо z_{kp}=1,96. Оскільки $|z_{cn}|<z_{kp}$ - немає підстав відкидати нульову гіпотезу. Іншими словами, ймовірності відмови елемента двох приладів відрізняються не суттєво.

Завдання для самостійного розв'язання.

40. Партія виробів приймається, якщо ймовірність того, що виріб виявиться бракованим, не перевищує 0,02. Серед випадково відібраних 480 виробів виявилось, що 12 дефектні. Чи можна прийняти партію?

Вказівка: $H_0: p=p_0=0{,}02$ при конкуруючій гіпотезі $H_1: p>0{,}20$.

41.Партія виробів приймається, якщо ймовірність того, що виріб виявиться бракованим не перевищує 0,03. Випадково відібрано 400 виробів і виявилось, що 18 бракованих. Чи можна прийняти партію?

42.В результаті тривалих спостережень встановлено, що ймовірність повного одуження хворого, що приймав ліки А дорівнює 0,8. Нові ліки В призначені 800 хворим, причому 660 із них повністю одужали. Чи можна вважати що нові ліки ефективніші ліків А на п'ятипроцентному рівні значущості?

43.В партії із 500 деталей, виготовлених першим станком-автоматом, виявилось 60 не стандартних; із 600 деталей другого станка 42 не стандартні. При рівні значущості 0,01 перевірити нульову гіпотезу $H_0 : p_1 = p_2 = p$ про рівність ймовірностей виготовлення нестандартних деталей станками при конкуруючій гіпотезі $H_1 : p_1 \neq p_2$.

44.Для оцінки якості виробів виготовлених двома заводами, взяті вибірки $n_1 = 200; n_2 = 300$ виробів. В цих вибірках виявилось відповідно $m_1 = 20; m_2 = 15$ бракованих виробів. При рівні значущості 0,05 перевірити нульову гіпотезу $H_0 : p_1 = p_2 = p$ про рівність ймовірностей виготовлення бракованих виробів заводами при конкуруючій гіпотезі $H_1 : p_1 > p_2$.

45. Із 100 пострілів по цілі кожним із двох рушниць зареєстровано відповідно $m_1 = 12; m_2 = 8$ промахів. При рівні значущості 0,05 перевірити нульову гіпотезу $H_0 : p_1 = p_2 = p$ про рівність ймовірностей промаху обох рушниць при конкуруючій гіпотезі $H_1 : p_1 > p_2$

Заняття 8. Перевірка статистичних гіпотез про рівність середніх двох нормально розподілених випадкових величин

Теоретичні відомості.

Нехай $X_1,...,X_n$ і $Y_1,...,Y_m$ дві незалежні нормально розподілені вибірки із параметрами (a_1, σ_1^2) і (a_2, σ_2^2) відповідно. Нехай параметри $MX = a_1$ і $MY = a_2$ невідомі. Часто на практиці виникає потреба встановити, чи суттєво відрізняються середні в цих вибірках. Тобто необхідно перевірити гіпотезу $H_0 : M(X) = M(Y)$. Нехай альтернативна гіпотеза має вигляд $H_1 : M(X) \neq M(Y)$. У цьому випадку критична область буде симетричною. Нехай дисперсії σ_1^2 і σ_2^2 відомі. Оскільки величини $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ і $\overline{Y} = \frac{Y_1 + \cdots + Y_m}{m}$ незалежні і нормальні $(a_1, \sigma_1^2 / n)$, $(a_2, \sigma_2^2 / m)$, то при правильності основної гіпотези $M(\overline{X} - \overline{Y}) = MX - MY = 0$, а $D(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$, тому величина

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \qquad (1)$$

буде мати стандартний нормальний розподіл. За рівнем значущості α із рівняння $1 - \alpha = P\{|Z| < |z_\alpha|\} = 2\Phi(z_\alpha)$ можна знайти z_α (із рівняння

$1-\alpha = 2\Phi(z_\alpha)$), а за вибірками $x_1,...,x_n$ і $y_1,...,y_m$ знаходимо спостережуване значення критерію z_{cn}. Якщо $|z_{cn}| < z_\alpha$, то гіпотеза приймається, інакше - відкидається.

Розглянемо тепер випадок, коли дисперсії σ_1^2 і σ_2^2 невідомі, але рівні між собою, тобто $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Тоді величина $\frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}} = \frac{\overline{X}-\overline{Y}}{\sigma}\sqrt{\frac{mn}{m+n}} = Z$ має стандартний нормальний розподіл. Величини $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ і $s_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2$ є незміщеними та спроможними оцінками дисперсій $D(X)$ і $D(Y)$. Для перевірки гіпотези використаємо критерій

$$T = \frac{\overline{X}-\overline{Y}}{\sqrt{(n-1)s_X^2 + (m-1)s_Y^2}}\sqrt{\frac{mn(m+n-2)}{m+n}}, \qquad (2)$$

що має розподіл Стьюдента з $n+m-2$ ступенями вільності.

За рівнем значущості α і числом ступенів вільності $(m+n-2)$ можна знайти таку точку T_α, щоб $P\{|T| > T_\alpha\} = \alpha$. А за вибірками $x_1,...,x_n$ і $y_1,...,y_m$ знаходимо $\overline{X}$, $\overline{Y}$, s_X^2, s_Y^2 і спостережуване значення критерію T_{cn}. Якщо $|T_{cn}| < T_\alpha$, то гіпотеза приймається, в протилежному випадку - відкидається.

Розв'язування типових вправ.

1.По двох незалежних вибірках, об'єми яких $n = 40; m = 50$ вилучених із нормальних генеральних сукупностей, знайдені вибіркові середні: $\overline{x} = 130; \overline{y} = 140$. Генеральні дисперсії відомі $DX = 80; DY = 100$.При рівні значущості 0,01 перевірити нульову гіпотезу $H_0 : MX = MY$ при конкуруючій гіпотезі $H_1 : MX \neq MY$.

Розв'язання. За формулою (1) знаходимо $z_{cn} = -5$. За умовою конкуруюча гіпотеза має вигляд $H_1 : MX \neq MY$, тому критична область двостороння. Знайдемо праву критичну току за формулою $\Phi(z_{kp}) = (1-\alpha)/2 = (1-0{,}01)/2 = 0{,}495$. По таблиці функції Лапласа знаходимо $z_{kp} = 2{,}58$. Оскільки $|Z_{cn}| > z_{kp}$ то нульову гіпотезу відхиляємо. Іншими словами, вибіркові середні відрізняються суттєво.

2.По двох незалежних малих вибірках об'єми яких $n = 12; m = 18$ вилучених із нормальних генеральних сукупностей, знайдені вибіркові середні $\overline{x} = 31{,}2; \overline{y} = 29{,}2$ і виправлені дисперсії $s_x^2 = 0{,}84$; $s_y^2 = 0{,}40$. При рівні значущості 0,05 перевірити нульову гіпотезу $H_0 : MX = MY$ при конкуруючій гіпотезі $H_1 : MX \neq MY$.

Розв'язання. Виправлені дисперсії різні , тому перевіримо попередньо гіпотезу про рівність генеральних дисперсій, використовуючи критерій Фішера-Снекедора. Знайдемо відношення більшої дисперсії до меншої:

F $F_{сп} = \frac{0{,}84}{0{,}40} = 2{,}1$. Дисперсія $s_x^2 = 0{,}84$ значно більша дисперсії $s_y^2 = 0{,}40$, тому в якості конкуруючої гіпотези приймемо гіпотезу $H_1 : DX > DY$. В цьому випадку критична область – правостороння. По таблиці точок розподілу Фішера-Снекедора, при рівні значущості 0,05 і числам степенем вільності $k_1 = n - 1 = 12 - 1 = 11; k_2 = m - 1 = 18 - 1 = 17$ знаходимо критичну точку $F_{kp}(0{,}05;11;17) = 2{,}41$. Оскільки $F_{сп} < F_{kp}$ - немає підстав відхиляти нульову гіпотезу про рівність генеральних дисперсій. Припущення про рівність генеральних дисперсій виконується, тому порівняємо середні.

Обчислимо значення критерія Стьюдента по формулі (2) $T_{сп} = 7{,}1$. За умовою, конкуруюча гіпотеза має вигляд $H_1 : MX \neq MY$, тому критична область двостороння. При рівні значущості 0,05 і числу степенем вільності $k = n + m - 2 = 12 + 18 - 2 = 28$ знаходимо по таблиці критичних точок розподілу Стьюдента $t(0{,}05;28) = 2{,}05$. Оскільки $T_{сп} > t_{kp}$ - нульову гіпотезу про рівність середніх відхиляємо. Іншими словами, вибіркові середні відрізняються суттєво.

Завдання для самостійного розв'язання.

46. По вибірці об'єму $n = 30$ знайдена середня вага $\bar{x} = 130$ г виробів, що виготовлені на першому станку; по вибірці об'єму $m = 40$ знайдена середня вага $\bar{y} = 125$ г виробів, що виготовлені на другому станку. Генеральні дисперсії відомі $DX = 60; DY = 80$.При рівні значущості 0,05 перевірити нульову гіпотезу $H_0 : MX = MY$ при конкуруючій гіпотезі $H_1 : MX \neq MY$.Припускається, що випадкові величини розподілені нормально і вибірки незалежні.

47. По вибірці об'єму $n = 50$ знайдено середній розмір $\bar{x} = 20{,}1$мм діаметрів валиків, що виготовлені на автоматі № 1; по вибірці об'єму $m = 50$ знайдено середній розмір $\bar{y} = 19{,}8$мм діаметрів валиків, що виготовлені на автоматі № 2.Генеральні дисперсії відомі $DX = 1{,}750; DY = 1{,}375$.При рівні значущості 0,05 перевірити нульову гіпотезу $H_0 : MX = MY$ при конкуруючій гіпотезі $H_1 : MX \neq MY$.Припускається, що випадкові величини розподілені нормально і вибірки незалежні.

48. Із двох партій виготовлених на двох однакових станках вилучені малі вибірки, об'єми яких $n = 10; m = 12$ одержані результати $\bar{x} = 1{,}45; \bar{y} = 0{,}85$. При рівні значущості 0,02 перевірити нульову гіпотезу $H_0 : MX = MY$ при конкуруючій гіпотезі $H_1 : MX \neq MY$. Припускається, що випадкові величини розподілені нормально і вибірки незалежні.

49.При рівні значущості 0,05 перевірити нульову гіпотезу $H_0 : MX = MY$ про рівність генеральних середніх при конкуруючій гіпотезі $H_1 : MX > MY$. Якщо $n = 10; m = 16$; $\bar{x} = 6{,}41; \bar{y} = 2{,}34$. Припускається, що випадкові величини розподілені нормально і вибірки незалежні.

Заняття 9. Перевірка гіпотез про рівність дисперсій двох нормально розподілених випадкових величин

Теоретичні відомості.

Нехай перевіряється гіпотеза про те, що $x_1,...,x_n$ є вибіркою із розподілу $F(x)$. За вибіркою $x_1,...,x_n$ знаходимо емпіричну функцію розподілу $F_n(x)$. При великих n граничну функцію $K(t)$ можна використовувати для практичних розрахунків, тобто $K(t)$ досить добре наближує ймовірність $P\left\{\sqrt{n}D_n < t\right\}$, тому для перевірки гіпотези можна використати статистику $K_n = \sqrt{n}D_n$. Граничний розподіл $K(t)$ не залежить від вигляду розподілу $F(x)$. Важливим є і те, що розподіл величини K_n при великих n (уже при $n \geq 20$) практично не залежить від n. Оскільки емпірична функція розподілу збігається до відповідної теоретичної функції розподілу, то критична точка t_α за рівнем значущості α визначається із умови $P\left\{\sqrt{n}D_n \geq t_\alpha\right\} = \alpha$. Враховуючи наближену рівність $P\left\{\sqrt{n}D_n < t\right\} \approx K(t)$, критичну точку t_α можна знайти із рівняння $K(t_\alpha) = 1-\alpha$ (наприклад, $\alpha = 0{,}1$ $t_\alpha = 1{,}22$, $\alpha = 0{,}05$ $t_\alpha = 1{,}36$, $\alpha = 0{,}01$ $t_\alpha = 1{,}63$). Для знаходження спостережуваного значення критерію K_n за вибіркою $x_1,...,x_n$, для кожної із точок знайдемо спочатку різниці $\left|F_n(x_i) - F(x_i)\right|$, $\left|F_n(x_i+0) - F(x_i)\right| = \left|F_n(x_{i+1}) - F(x_i)\right|$. Значення D_n дорівнює найбільшій із цих $2n$ різниць. Тоді спостережуване значення критерію $K_{cn} = \sqrt{n}D_n$. Якщо $K_{cn} < t_\alpha$ то гіпотезу приймають, інакше – відкидають.

Досить широке використання в статистиці знайшов *критерій* χ^2 (*критерій Пірсона*), що оснований на порівнянні емпіричних та теоретичних частот.

Нехай перевіряється гіпотеза про те, що $x_1,...,x_n$ є вибіркою із розподілу $F(x)$. Нехай множина значень досліджуваної величини X розбита на s підмножин: $\Delta_1,...,\Delta_s$. Позначимо через n_i число результатів спостережень x_j, які попадають до Δ_i, $n_1 + ... + n_s = n$, тоді n_i називають *емпіричними частотами*. Оскільки гіпотеза стверджує, що досліджувана величина X має розподіл $F(x)$, то за розподілом $F(x)$ можна знайти ймовірності $p_i = P\{X \in \Delta_i\}$. Тоді $n_i' = np_i$ називають *теоретичними частотами*. Із закону великих чисел, теорема Бореля, випливає, що при справедливості основної гіпотези $\frac{n_i}{n} \to p_i$ з ймовірністю одиниця. Тобто при великих n різниці $n_i - np_i = n_i - n_i'$ будуть невеликими.

Розглянемо випадкову величину $\chi^2 = \sum_{i=1}^{s} \frac{(n_i - np_i)^2}{np_i}$. Можна довести, що при $n \to \infty$ розподіл величини χ^2 збігається до розподілу χ^2 з $s-1$ ступенями вільності. На практиці граничний розподіл можна використовувати із гарним наближенням уже при $n \geq 50$ і $n_j \geq 5$.

Величина χ^2 є невід‘ємною. За рівнем значущості α і числом ступенів вільності $s-1$ із рівняння $P\{\chi^2 > \chi^2_\alpha\} = \alpha$ знаходимо χ^2_α. За результатами спостережень знаходимо спостережуване значення критерію $\chi^2_{сп} = \sum_{i=1}^{s} \frac{\left(n_i - n_i'\right)^2}{n_i'}$.

Якщо $\chi^2_{сп} < \chi^2_\alpha$, то гіпотеза приймається, інакше – відкидається.

Розв'язування типових вправ.

Приклад. В експериментах з селекцією гороху Мендель спостерігав частоти різного вигляду насіння, що одержані при схрещуванні рослин з круглим жовтим і зморшкуватим зеленим насінням. Ці дані та значення теоретичних ймовірностей, що визначаються за теорією спадковості Менделя, наведені в таблиці:

Насіння	Частота n_i	Ймовірність p_i
Кругле жовте	315	9/16
Зморшкувате жовте	101	3/16
Кругле зелене	108	3/16
Зморшкувате зелене	32	1/16
Всього	$n = 556$	1

При рівні значущості $\alpha = 0,05$ перевірити гіпотезу про узгодження експериментальних даних із теоретичними ймовірностями.

За результатами спостережень знаходимо спостережуване значення критерію

$$\chi^2_{сп} = \sum_{i=1}^{s} \frac{\left(n_i - n_i'\right)^2}{n_i'} =$$

$$= \frac{(315 - 556\cdot 9/16)^2}{556\cdot 9/16} + \frac{(101 - 556\cdot 3/16)^2}{556\cdot 3/16} + \frac{(108 - 556\cdot 3/16)^2}{556\cdot 3/16} + \frac{(32 - 556\cdot 1/16)^2}{556\cdot 1/16} = 0,47.$$

За рівнем значущості $\alpha = 0,05$ і числом ступенів вільності 4-1=3 за таблицею розподілу χ^2 знаходимо $\chi^2_\alpha = 7,8$. Оскільки $\chi^2_{сп} = 0,47 < 7,8 = \chi^2_\alpha$, то можна зробити висновок, що експериментальні дані добре узгоджуються із теоретичними ймовірностями.

При великих значеннях s розподіл χ^2 можна наближено замінити нормальним розподілом із середнім $s-1$ і дисперсією $2(s-1)$.

Нехай розподіл $F(x, \theta_1, ..., \theta_r)$ залежить від r параметрів. Тоді на основі результатів спостережень ми можемо замінити невідомі параметри $\theta_1, ..., \theta_r$ їх відповідними оцінками $\hat{\theta}_1, ..., \hat{\theta}_r$ і на основі функції $F\left(x, \hat{\theta}_1, ..., \hat{\theta}_r\right)$ знаходимо теоретичні частоти n_i'. Тоді при $n \to \infty$ розподіл величини χ^2 збігається до

розподілу χ^2 з $s-1-r$ ступенями вільності, де r - число оцінюваних параметрів розподілу. А далі процедура перевірки гіпотези така ж сама.

Вправи.

50.За вибірками обсягів $n=12$ і $m=15$, вибраних із нормально розподілених генеральних сукупностей *X* і *Y*, знайдені виправлені дисперсії $s_x^2=11,41$ і $s_y^2=6,52$. Перевірити гіпотезу про рівність дисперсій при рівні значущості $\alpha=0,05$.

51При 4040 киданнях монети Бюффон одержав 2048 випадань герба. При рівні значущості $\alpha=0,01$ перевірити гіпотезу H_0: монета симетрична.

52.За вибіркою обсягом $n=1000$ підраховано число обривів ниток. При рівні значущості $\alpha=0,05$ перевірити гіпотезу: число обривів ниток *X* має розподіл Пуассона.

Число обривів x_i	0	1	2	3
Частота n_i	600	320	70	10

53.Використовуючи критерій Персона при рівні значущості 0,05 встановити, чи випадкові або значні розбіжності між емпіричними частотами n_i і теоретичними частотами n_i', які обчислені за гіпотезою про нормальний розподіл генеральної сукупності.

А)

n_i	5	10	20	8	7
n_i'	6	14	18	7	5

Б)

n_i	6	8	13	15	20	16	10	7	5
n_i'	5	9	14	16	18	16	9	6	7

Заняття 11. Вибіркова кореляція і регресія

Теоретичні відомості

Вибірковим коефіцієнтом кореляції називається величина

$$r_B=\frac{\frac{1}{n}\sum_i\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sigma_X\sigma_Y},$$

де $\bar{x}$ і $\bar{y}$ - вибіркові середні, а σ_X і σ_Y - вибіркові середні квадратичні відхилення величин X і Y. Нескладними перетвореннями можна одержати формулу

$$r_B=\frac{\frac{1}{n}\sum_i x_i y_i-\bar{x}\,\bar{y}}{\sigma_X\sigma_Y}.$$

Для обчислення коефіцієнта кореляції можна перейти до нових величин $U=\frac{X-c_1}{h_1}$ і $V=\frac{Y-c_2}{h_2}$, при цьому для довільних c_1, c_2, $h_1\neq 0$, $h_2\neq 0$ справедлива

формула

$$r_B = \frac{\sum n_{uv} uv - n\bar{u}\bar{v}}{n\sigma_u \sigma_v},$$

де $\bar{u} = \frac{1}{n}\sum n_u u$, $\bar{v} = \frac{1}{n}\sum n_v v$, $\sigma_u^2 = \left(\overline{u^2} - \left(\bar{u}\right)^2\right)$, $\sigma_v^2 = \left(\overline{v^2} - \left(\bar{v}\right)^2\right)$, $\overline{u^2} = \frac{1}{n}\sum n_u u^2$, $\overline{v^2} = \frac{1}{n}\sum n_v v^2$, n_{uv} є частота відповідної пари (x, y).

Вибіркове кореляційне відношення Y на X є статистичною оцінкою теоретичного кореляційного відношення (позначимо його $\eta_{y/x}$) і визначається за формулою

$$\eta_{y/x} = \frac{\sigma_{\overline{Y_x}}}{\sigma_Y},$$

де $\sigma_Y^2 = \frac{1}{n}\sum_i n_{y_i}\left(y_i - \bar{y}\right)^2$, $\sigma_{\overline{Y_x}}^2 = \frac{1}{n}\sum_i n_{x_i}\left(\overline{Y_{x_i}} - \bar{y}\right)^2$.

Вибіркове кореляційне відношення використовується для характеристики довільного зв'язку між величинами і має властивості, що аналогічні до властивостей теоретичного кореляційного відношення. Зокрема, $0 \le \eta_{y/x} \le 1$, $|r_B| \le \eta_{y/x}$. Якщо $\eta_{y/x} = 1$, то величина Y пов'язана із X точною функціональною залежністю, якщо ж при цьому $r_B^2 = 1$, то ця залежність буде лінійною. У випадку $\eta_{y/x} < 1$ функціональна залежність відсутня. Якщо $\eta_{y/x} = 0$, то $\overline{Y_x} = \bar{Y}$.

Заняття12.

Знаходження рівняння прямої лінії регресії методом найменших квадратів

Теоретичні відомості.

У найпростішому випадку лінійної регресії емпіричне рівняння регресії Y на X знаходять у вигляді $\overline{Y_x} = \rho x + \beta$. Тобто необхідно знайти пряму, яка б була найближче розміщена до точок (x_i, y_i) у розумінні методу найменших квадратів. У цьому випадку параметри ρ і β вибираються так, щоб сума

$$\sum_{i=1}^{n}\left(y_i - \rho x_i - \beta\right)^2 = F\left(\rho, \beta\right)$$

була мінімальною. Для цього утворимо систему нормальних рівнянь

$$\begin{cases} \dfrac{\partial F}{\partial \rho} = 0, \\ \dfrac{\partial F}{\partial \beta} = 0. \end{cases}$$

Після знаходження частинних похідних одержуємо систему

$$\begin{cases} \dfrac{\partial F}{\partial \rho} = \sum_{i=1}^{n} 2\left(y_i - \rho x_i - \beta\right)\left(-x_i\right) = 0, \\ \dfrac{\partial F}{\partial \beta} = \sum_{i=1}^{n} 2\left(y_i - \rho x_i - \beta\right)\left(-1\right) = 0, \end{cases}$$

яку можна записати у вигляді $\begin{cases}\sum_{i=1}^{n}\left(x_i y_i-\rho x_i^2-\beta x_i\right)=0,\\ \sum_{i=1}^{n}\left(y_i-\rho x_i-\beta\right)=0\end{cases}$ або

$$\begin{cases}\rho\sum_{i=1}^{n}x_i^2+\beta\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}x_i y_i,\\ \rho\sum_{i=1}^{n}x_i+n\beta=\sum_{i=1}^{n}y_i.\end{cases}$$

Введемо позначення $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, $\overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2$, $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$ тоді система набуває вигляду

$$\begin{cases}\rho\overline{x^2}+\beta\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i y_i,\\ \rho\bar{x}+\beta=\bar{y},\end{cases}$$

звідки

$$\beta=\bar{y}-\rho\bar{x},\ \rho=\frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\bar{x}\bar{y}}{\overline{x^2}-\left(\bar{x}\right)^2}=\frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i-\bar{x}\bar{y}}{\sigma_X^2}.$$

Коефіцієнт ρ називається *коефіцієнтом регресії* Y на X. Враховуючи вигляд вибіркового коефіцієнта кореляції, одержимо $\rho=r_B\frac{\sigma_Y}{\sigma_X}$. Підставимо знайдені значення параметрів у рівняння прямої лінії регресії, тоді емпіричне рівняння прямої регресії Y на X буде мати вигляд

$$\overline{Y_x}-\bar{y}=r_B\frac{\sigma_Y}{\sigma_X}\left(x-\bar{x}\right).$$

Вправа.

54.Знайти вибіркове рівняння прямої лінії регресії за такими статистичними даними для випадкового вектора (X,Y):

x_i	18	19	25	20	25	21	23	22	23
	24								
y_i	20	20	35	20	30	25	25	25	30
	30								

55. Знайти вибіркові рівняння прямих ліній У на Х і Х на У регресії за такими статистичними даними для випадкового вектора (X,Y):

У/Х	5	10	15	20	25	30	35	40	n_y
100	2	1							3
120	3	4	3						10
140			5	10	8				23
160				1		6	1	1	9
180							4	1	5
n_x	5	5	8	11	8	6	5	2	N=50

Б).

У/Х	18	23	28	33	38	43	48	n_y
125		1						1
150	1	2	5					8
175		3	2	12				17
200			1	8	7			16
225					3	3		6
250						1	1	2
n_x	1	6	8	20	10	4	1	N=50

В).

У/Х	5	10	15	20	25	30	35	n_y
100						6	1	7
120						4	2	6
140			8	10	5			23
160	3	4	3					10
180	2	1		1				4
n_x	5	5	11	11	5	10	3	N=50