ЛЕКЦІЯ 4. Точкове та інтервальне оцінювання параметрів.

ПЛАН:

4.1.Оцінка параметрів генеральної сукупності по вибірці

Точкові та інтервальні оцінки. Властивості оцінок

За даними генеральної сукупності та відомим законом розподілу можна визначити параметри цього розподілу. На практиці є тільки дані вибірки і деякі припущення відносно закону розподілу. В таких умовах необхідно оцінити параметри розподілу.

Статистичною оцінкою невідомого параметра теоретичного розподілу називають його наближене значення, яке залежить від даних вибірки, тобто деяку функцію цих величин. Цю функцію називають ***статистикою***.

Статистична оцінка є випадковою величиною і відхиляється від параметру генеральної сукупності на деяку невідому величину δ . Позначимо через θ параметр генеральної сукупності що оцінюється, а через θ^* – його вибіркову оцінку. Величину $\left|\theta - \theta^*\right|$ називають ***точністю оцінки***.

Чим вона менша, тим краще, точніше визначено невідомий параметр.

Незміщеною називають статистичну оцінку, математичне сподівання якої дорівнює оцінюваному параметру за довільного об’єму вибірки тобто

$$M(\theta^*) = \theta,$$

і ***зміщеною,*** якщо

$$\boldsymbol{M}(\theta^*) \neq \theta.$$

Однак, випадково може статися, що $\boldsymbol{M}(\theta^*) = \theta$ при сильному розсіюванні даних, коли вибірка сильно відрізняється від генеральної сукупності. Тому вводиться поняття ***ефективної*** оцінки.

Ефективною називають статистичну оцінку, яка при заданому об’ємі вибірки має найменшу можливу дисперсію.

Спроможною називають оцінку, яка наближається до істинного параметру при $\boldsymbol{n} \to \infty$.

Тобто можна з деякою ймовірністю $\boldsymbol{P}$ стверджувати, що $\left|\boldsymbol{\theta} - \boldsymbol{\theta}^*\right| < \boldsymbol{\delta}$, де $\delta > \mathbf{0}$.

Наприклад, вибіркова середня $\overline{\boldsymbol{X}}_{\boldsymbol{в}}$ є незміщеною та спроможною оцінкою математичного сподівання.

Вибіркове дослідження не дає точної характеристики всієї генеральної сукупності. Одержана похибка називається ***похибкою репрезентативності*** (представництва). Чим більший об'єм вибірки – тим менша похибка.

Оцінка, що визначається одним числом, називається ***точковою (***наприклад $\overline{\boldsymbol{X}}_{\boldsymbol{в}}; \boldsymbol{D}_{\boldsymbol{в}}; \boldsymbol{S}^2; \boldsymbol{S}$***).***

Оцінка, що визначається двома числами – межами інтервалу, називається ***інтервальною.***

Випадковий інтервал $\left(\theta^* - \delta ; \theta^* + \delta\right)$, у межах якого з ймовірністю γ знаходиться невідомий параметр, що оцінюється, називається ***довірчим інтервалом I***, тобто

$$I = \left(\theta^* - \delta ; \theta^* + \delta\right).$$

Ймовірність γ виконання нерівності $\left|\theta - \theta^*\right| < \delta$ називається ***довірчою ймовірністю*** або ***надійністю*** оцінки θ^*, тобто

$$P\left(\left|\theta - \theta^*\right| < \delta\right) = \gamma .$$

Величина $p = 1 - \gamma$ називається ***рівнем значимості*** і показує, з якою ймовірністю припущення про надійність оцінки помилкове. Рівень значимості задається як правило рівним 0,05; 0,01; 0,001.

Повернутись до плану.

4.2.Точність оцінок. Оцінка параметрів нормальної сукупності

Довірчий інтервал для генерального середнього

Позначимо через $M(X)$ середнє значення генеральної сукупності X. Припустимо, що генеральна сукупність X має нормальний закон розподілу ймовірностей. Розглянемо два випадки.

1) Дисперсія $D(X)$ генеральної сукупності відома.

Довірчий інтервал для генерального середнього з заданою ймовірністю γ знаходять за формулою

$$\overline{X}_в - u\cdot\frac{\sigma}{\sqrt{n}} < M(X) < \overline{X}_в + u\cdot\frac{\sigma}{\sqrt{n}},$$

де $\sigma = \sqrt{D(X)}$,

u визначається з рівності:

$$\Phi(u) = 1 - p/2 \Rightarrow u = \Phi^{-1}(1 - p/2).$$

Функція $\Phi(u)$ є функцією Лапласа. Вона табульована (див.табл.2 додатку).

По заданій довірчій ймовірності γ знаходиться рівень значимості p по якому з таблиці функції Лапласа знаходиться відповідний квантиль u, якому відповідає значення функції Лапласа ***1-p/2.***

Дисперсія генеральної сукупності невідома і знаходиться за даними вибірки.

Англійський статистик Госсет (псевдонім – Стьюдент) довів, що у випадку нормального розподілу ознаки X у генеральній сукупності нормування, випадкова величина

$$T = \frac{\overline{X}_в - M(X)}{\frac{S}{\sqrt{n}}}$$

залежить тільки від об'єму вибірки.

Була знайдена функція розподілу випадкової величини T та ймовірність $P(T < t_\gamma)$, де γ – точність оцінки.

Функція

$$S(n, t_\gamma) = P(|T| < t_\gamma) = \gamma,$$

називається ***t-розподілом Стьюдента*** з $(n-1)$ ступенями свободи.

Ця функція зв'язує випадкову величину ***T,*** довірчий інтервал ***I*** та довірчу ймовірність γ. Знаючи дві з них, можна знайти третю.

Оскільки функція щільності такого закону парна, то

$$P\left(\frac{\left|\overline{X}_в - M(X)\right|}{S/\sqrt{n}} < t_{\gamma}\right) = \gamma .$$

Довірчий інтервал для знаходження математичного сподівання генеральної сукупності будується за формулою

$$\overline{X}_в - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{p}{2}} < M(X) < \overline{X}_в + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{p}{2}},$$

де $t_{1-\frac{p}{2};n-1}$ – квантиль розподілу Стьюдента з $(n-1)$ ступенями свободи на рівні значимості $p = 1-\gamma$,

де γ – ймовірність твердження (див.табл.3 Додатка).

Практичний приклад 4.1.

Отримано наступний варіаційний ряд довжин колосків обсягом 30 спостережень (см) :

8;8;9;9;9;9;10;10;10;10;10;11;11;11;11;11;11;12;12;12;12;13;13;13;14;14;15;15;16;16.

Побудувати таблицю інтервального розподілу та зробити оцінки математичного сподівання, дисперсії, коефіцієнтів асиметрії та ексцесу.

Розв'язання.

За формулою Стерджеса знайдемо крок гістограми

$(x_{\min} = 8см; \qquad x_{max} = 16см; \qquad n = 30)$:

$h = (16 - 8)/(1 + 3{,}322\lg 30) \approx 1{,}35$.

Ціла частина від цього числа дорівнює 1. Тобто крок дорівнює 2см. Початкова точка розбивки дорівнює 7см. Знайдемо інтервали та відповідні частоти (табл.)

Номер інтервалу	1	2	3	4	5
Межі інтервалу	(7;9]	(9;11]	(11;13]	(13;15]	(15:17]
Середина інтервалу	8	10	12	14	16
Частота	6	11	7	4	2
Відносна частота	0,2	0,37	0,23	0,13	0,07
Кумулята	0,2	0,57	0,8	0,93	1,0

За згрупованими даними зробимо оцінку статистичних характеристик:

$\overline{X}_в = 0{,}2\cdot 8 + 0{,}37\cdot 10 + 0{,}23\cdot 12 + 0{,}13\cdot 14 + 0{,}07\cdot 16 = 11см$

$\sigma_в^2 = 0{,}2\cdot 8^2 + 0{,}37\cdot 10^2 + 0{,}23\cdot 12^2 + 0{,}13\cdot 14^2 + 0{,}07\cdot 16^2 - 11^2 = 5{,}32; \qquad \Rightarrow \qquad \sigma_в = 2{,}31$

$\gamma = (0{,}2(8-11)^3 + 0{,}37(10-11)^3 + 0{,}23(12-11)^3 + 0{,}13(14-11)^3 + 0{,}07(16-11)^3)/2{,}31^3 \approx 0{,}55$

$\mu = (0{,}2(8-11)^4 + 0{,}37(10-11)^4 + 0{,}23(12-11)^4 + 0{,}13(14-11)^4 + 0{,}07(16-11)^4)/2{,}31^4 - 3 \approx -0{,}50$

За даними попередньої задачі знайти 90% довірчий інтервал для математичного сподівання довжин колосків.

<u>Розв’язання.</u>

Оцінки середнього та значення середньоквадратичного відхилення за даними попередньої задачі:

$\overline{X}_в = 11см; \qquad S = 2{,}31см.$ Обсяг вибірки $n=30$. Значення квантілю розподілу Стьюдента $t_{0,05;29} \approx 2{,}04$. Звідси 90% довірчий інтервал для математичного сподівання:

$$11-\frac{2{,}31}{\sqrt{30}}\cdot 2{,}04\le M(X)\le 11+\frac{2{,}31}{\sqrt{30}}\cdot 2{,}04\Rightarrow 10{,}14\le M(X)\le 11{,}86$$

Повернутись до плану.

4.3.Довірчий інтервал для генеральної дисперсії

Нехай σ^2 – невідоме значення дисперсії генеральної вибірки, S^2 – оцінка дисперсії за вибіркою об'єму n. Тоді випадкова величина

$$\chi^2=\frac{S^2}{\sigma^2}\cdot(n-1),$$

має закон розподілу Пірсона (χ^2) з $(n-1)$ ступенями свободи.

З ймовірністю $\gamma=1-p$ шукане значення дисперсії генеральної сукупності знаходиться в інтервалі

$$\left(\frac{S^2}{\chi^2_{\frac{p}{2}}}\cdot(n-1);\frac{S^2}{\chi^2_{1-\frac{p}{2}}}\cdot(n-1)\right).$$

Значення $\chi^2_{1-\frac{p}{2}}$ та $\chi^2_{\frac{p}{2}}$ беруть з таблиці квантилей χ^2 розподілу (табл.4, додаток) за ступенями свободи $(n-1)$ на рівні значимості p.

Практичний приклад 3.3.1.

За даними попередньої задачі знайти 90% довірчі інтервали для дисперсії генеральної сукупності.

Розв’язання.

З табл. 4 додатку знайдемо відповідні квантилі розподілу χ^2 , кількість ступеней свободи дорівнює $\gamma = 29$ (в табл.4 додатку відсутнє, тому обираємо середнє значення для $\gamma = 28$ та $\gamma = 30$. $\chi^2_{0,05;29} = 42,0; \chi^2_{0,95;29} = 17,0$.

Оскільки, вибіркове значення дисперсії дорівнювало 5,32, то отримаємо наступний 90% довірчий інтервал для дисперсії генеральної вибірки:

$$5,32 \cdot 29 / 42 \le \sigma^2 \le 5,32 \cdot 29 / 17$$

$$3,67 \le \sigma^2 \le 9,08$$

Повернутись до плану.

4.4.Тестування різниці між оцінками середнього, що отримано за різними вибірками

Існує ряд практичних задач, коли необхідно зробити висновок про існування значимої різниці між середніми значеннями, що отримані за різними вибірками.

Якщо ця різниця незначна , то цілком можливо вона обумовлена похибкою в оцінці середнього. Наприклад, існує задача порівняння рівня оплати праці на двох фірмах, що випускають аналогічну продукцію, або для аграрного сектору потрібно порівняти дві технології вирощування певної культури .

Висувається нульова гіпотеза, яка полягає в тому, що математичні сподівання генеральних сукупностей не відрізняються одне від одного: $H_0 : M(X_1) = M(X_2)$.

Будемо вважати, що кількість спостережень за якою робиться оцінка математичного сподівання ($\overline{x}_1$) для першого випадку дорівнює n_1, дисперсія σ_1^2, для другого випадку аналогічні величини: $\overline{x}_2;\ \ n_2;\ \ \ \sigma_2^2.$

Нехай виконується умова: $n_1 \geq 30; n_2 \geq 30.$

Тоді величина z має нормальний розподіл не залежно від того відомі дисперсії генеральних сукупностей чи шукаються за вибірками. Ця величина дозволяє прийняти або відхилити гіпотезу про рівність середніх значень:

$$z = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

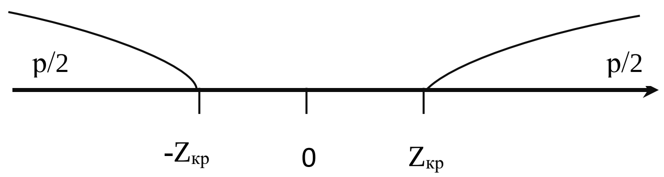
Надалі будемо називати її мірою або статистикою і позначатимемо як $z_{сп}$, тобто спостереження.

Перевірка гіпотези робиться в залежності від виду конкуруючої гіпотези.

Випадок 1.

Конкуруюча гіпотеза має вид: $H_1 : M(\overline{X}_1) \neq M(\overline{X}_2)$.

Будується двохстороння критична область:



Якщо $\left|Z_{сп}\right| < Z_{кр}$, немає підстав відкидати нульову гіпотезу;

Якщо $|Z_{сп}| > Z_{кр}$, нульова гіпотеза відкидається. Судження робляться на рівні значимості ***p***.

Величина $Z_{кр}$ береться з таблиць інтегральних функцій Лапласа.

$$\phi(z_{кр}) = (1 - p)/2.$$

Практичний приклад 3.4.1

За вибірками об'єму $n_x = 60$ та $n_y = 55$ знайдені точкові оцінки генеральних середніх $\bar{x} = 1200$ і $\bar{y} = 1250$. Дисперсії генеральних сукупностей відомі $D(X) = 120$ та $D(Y) = 110$. Перевірити гіпотезу про рівність генеральних середніх за конкуруючої гіпотези $H_1 : M(\bar{X}) \neq M(\bar{Y})$.

Судження провести на рівні значимості ***0,01***.

<u>Розв'язання:</u>

Знаходимо критичне значення величини ***z*** з таблиць значень інтегральної функції Лапласа за значенням:

$$\phi(z_{кр}) = (1 - p)/2 = 0{,}495$$

$$z_{кр} = 2{,}58$$

Розрахуємо величину, що є мірою

$$Z_{сп} = \frac{|\bar{x} - \bar{y}|}{\sqrt{D(X)/n_x - D(Y)/n_y}} =$$

$$= \frac{|1200 - 1250|}{\sqrt{120/60 + 110/55}} = |-25| = 25$$

Як бачимо $|z_{сп}| > z_{кр}$, тому з ймовірністю 0,99 H_0 – відкидається.

Практичний приклад 3.4.2.

Середня місячна зарплата на першій фірмі 10000 грн, кількість співробітників – 40, дисперсія зарплати 1 млн. У другій фірмі аналогічні показники: 11000 грн., 50 співробітників, 2 млн. – дисперсія.

<u>Розв’язання.</u>

Знайдемо величину z:

$$\boldsymbol{z = \frac{11000 - 10000}{\sqrt{\frac{10^6}{40} + \frac{2 \cdot 10^6}{50}}} = 3{,}92}$$

Рівень значущості, що відповідає цьому значенню (табл.2 додатку) дорівнює 0,0005, тобто з ймовірністю 0,9995 можна вважати, що середній рівень оплати праці у другій фірмі вище ніж у першій.

Випадок 2

Конкуруюча гіпотеза

$$\boldsymbol{H_1 : M(\overline{X}_1) > M(\overline{X}_2)}$$

Маємо правосторонню критичну область

$$\boldsymbol{\phi(z_{кр}) = (1 - 2p) / 2}$$

Якщо $Z_{сп} < Z_{кр}$, $\boldsymbol{H_0}$ приймається, якщо $Z_{сп} > Z_{кр}$, $\boldsymbol{H_0}$ відкидається.

Випадок 3

$H_1 : M(\overline{X}_1) < M(\overline{X}_2)$

Маємо лівосторонню критичну область

$z'_{кр} = -z_{кр}$

Якщо $Z_{сп} < Z'_{кр}$, - H_0 приймається;

$Z_{сп} > Z'_{кр}$ - H_0 відкидається.

Практичний приклад 3.4.3

За вибірками об'єму $n_x = 6$ та $n_y = 5$ знайдені точкові оцінки генеральних середніх $\bar{x} = 12,2$ і $\bar{y} = 14,1$. Дисперсії генеральних сукупностей відомі $D(X) = 12$ та $D(Y) = 11$. Перевірити гіпотезу про рівність генеральних середніх за конкуруючої гіпотези $H_1 : M(\overline{X}) < M(\overline{Y})$. Судження провести на рівні значимості $0,01$.

Розв'язання:

$$\phi(z_{кр}) = (1 - 2\alpha)/2 = (1 - 2 \cdot 0,01)/2 = 0,49$$

З таблиць значень інтегральної функції Лапласа знаходимо $z_{кр} = 2,33$. Тоді $z'_{кр} = -2,33$.

Розрахуємо статистичну величину

$$Z_{сп} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n_x - D(Y)/n_y}} =$$

$$= \frac{12,2 - 14,1}{\sqrt{12/6 + 11/5}} = -0,95$$

Можна зробити висновок, що $Z_{сп} < Z'_{кр}$, тому з ймовірністю 0,99 H_0 – відкидається.

Якщо дисперсії генеральних сукупностей невідомі і шукаються за вибірками і об'єм вибірок невеликий (менше 30), то для перевірки нульової гіпотези використовується t статистика (розподіл Стьюдента):

$$T=\frac{\overline{X}_1-\overline{X}_2}{\sqrt{(n_1-1)S_1^2+(n_2-1)S_2^2}}\sqrt{\frac{n_1 n_2(n_1+n_2-2)}{n_1+n_2}}$$

Значення знаходиться з таблиць значень розподілу Стьюдента. Вхід по ступенях свободи $k=n_x+n_y-2$.

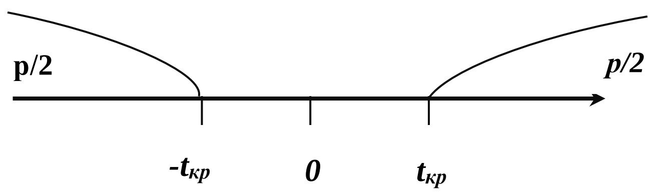
Перевірка нульової гіпотези гіпотези відбувається в залежності від виду суперечливої гіпотези.

Випадок 1.

Суперечлива гіпотеза має вигляд:

$$H_1 : M(\overline{X}) \neq M(\overline{Y})$$

Будується двохстороння критична область



Вечину $t_{кр}$ знаходимо з таблиць. Вхід в таблиці за величиною p та ступенями свободи k.

Якщо $|T_{сп}|<t_{кр}(p;k)$, то H_0 приймається;

Якщо $|T_{сп}|>t_{кр}(p;k)$, то H_0 відкидається.

<u>Примітка 1</u>

Якщо статистика розраховується у вигляді $|\overline{X}_1-\overline{X}_2|$, то в якості перевірки використовують величину:

$$T=\frac{\frac{S_1^2}{n_1}t_{1-\frac{p}{2}}(n_1-1)+\frac{S_2^2}{n_2}t_{1-\frac{p}{2}}(n_2-1)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}},$$

де $t_{1-\frac{p}{2}}$ - квантилі розподілу Стьюдента з ($n_1 - 1$) та ($n_2 - 1$) ступенями свободи, на рівні значимості $p = 1 - P$, де P надана ймовірність майбутнього висновку.

Якщо величина $\left|\bar{x}_1 - \bar{x}_2\right|$ більша за T, то гіпотеза H_0 відкидається з наданою ймовірністю P. Якщо величина $\left|\bar{x}_1 - \bar{x}_2\right|$ менша T, нульова гіпотеза не відкидається, а розбіжність значень $\bar{x}_1$ *та* $\bar{x}_2$ можна пояснити лише випадковими причинами.

У випадку, коли попередня гіпотеза H_0 про рівність дисперсій справедлива, для оцінки значимості різниці $\left|\bar{x}_1 - \bar{x}_2\right|$ величина T обчислюється за формулою:

$$T = t_{1-\frac{p}{2}} \cdot S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

де $t_{1-\frac{p}{2}}$ – квантиль розподілу Стьюдента з ($n_1 + n_2 - 2$) ступенями свободи.

Об’єднана дисперсія розраховується за формулою:

$$S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

Практичний приклад 3.4.4.

Потрібно оцінити ефективність двох технологій вирощування жита. Для оцінки першої використовувалося 6 ділянок, та отримано наступні показники врожайності: $\bar{x}_1 = 35ц/га; \quad S_1^2 = 20$, другої 9 ділянок: $\bar{x}_2 = 40ц/га; \quad S_2^2 = 25$.

Розв’язання.

Оскільки перевірка гіпотези про рівність дисперсій показує, що вони статистично не значимо відрізняються одна від одної ($F_{кр}$=1,25, а табличне його значення, взяте з таблиць Фішера за рівнем значущості 0,05 та за 8 і 5 ступенями свободи відповідно у чисельнику і знаменнику дорівнює 4,9), то обчислимо загальну вибіркову дисперсію:

$$S^2 = \frac{5 \cdot 20 + 8 \cdot 25}{6 + 9 - 2} = 23{,}1;$$

Для розрахунку міри користаємось наступною формулою:

$$t = \frac{40 - 35}{\sqrt{23{,}1(1/6 + 1/9)}} = 1{,}97.$$

Відповідне критичне значення розподілу Стьюдента $t_{0,05;13} = 1{,}77$, тобто на рівні значущості 0,05 можна зробити висновок, що генеральні середні статистично різні. Перевагу, очевидно, слід віддати другій технології вирощування жита.

Повернутись до плану.

4.5.Оцінка обсягу вибірки

При проведенні експерименту надзвичайно важливим питанням є визначення обсягу вибірки, який потрібен для отримання надійних оцінок.

Спочатку зробимо спробу оцінити обсяг вибірки, коли потрібно із із заданою точністю визначити, наприклад, частку населення що підтримує деякого кандидата. Однак проведення кожного дослідження коштує деяку суму, тому краще заздалегідь визначиться з мінімальним обсягом спостережень, що забезпечить необхідну точність.

Відповідно до центральної граничної теореми, вибіркова оцінка частки, що підтримує одного з кандидатів $\hat{p}$ при достатньому обсязі спостережень розподілена нормально відносно частки p в генеральній сукупності.

Щоб побудувати довірчі інтервали для цієї частки нам потрібно оцінити середньоквадратичне відхилення цієї частки. Оскільки кількість виборців X, що підтримує кандидатуру при загальній кількості виборців n розподілена відповідно біноміального розподілу то дисперсія X дорівнює:

$$\sigma^2(X) = np(1-p)$$

Прорахуємо дисперсію частки:

$$\sigma^2(\widehat{p}) = \sigma^2\left(\frac{X}{n}\right) = \frac{1}{n^2}\sigma^2(X) = \frac{p(1-p)}{n}$$

Змінюємо p на його оцінку $\widehat{p}$ отримаємо для середньоквадратичного відхилення частки оцінку:

$$\sigma_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Незміщена оцінка:

$$\widehat{\sigma}_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}$$

Тоді $100(1-\alpha)$ довірчі інтервали для частки виборців в генеральній сукупності:

$$\widehat{p} - z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \le p \le \widehat{p} + z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \qquad (3.5.1)$$

$z_{\alpha/2}$ – табличне значення хвостів стандартного нормального розподілу, яке відповідає ймовірності $\alpha/2$.

Наведемо Практичний приклад.

У випадкової вибірки з 1000 потенційних виборців 21 0 підтримали діючого президента. Це дає точкову оцінку для частки $\hat{p} = 0{,}21$.

Тоді незміщена оцінка середньоквадратичного відхилення:

$\hat{\sigma}_{\hat{p}} = \sqrt{0{,}21 \cdot 0{,}79 / 999} \approx 0{,}013$

Розрахуємо 95% $\alpha / 2 = 0{,}025$довірчі інтервали для загальної сукупності (всіх виборців).

З таблиці хвостів нормального розподілу $z_{0,025} = 1{,}96$.

Звідси 95% довірчі інтервали для частки:

$0{,}21 - 1{,}96 \cdot 0{,}013 \le p \le 0{,}21 + 1{,}96 \cdot 0{,}013 \Rightarrow 0{,}185 \le \hat{p} \le 0{,}235$

Це означає, що підтримка колишнього президента на час опитування з 95% ймовірністю складає від 18,5% до 23,5% населення.

Відхилення від очікуваного значення до верхньої або нижньої межі має назву похибки вибірки. З виразу (3.5.1) слідує, що похибка частки вибірки дорівнює:

$$\Delta = z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Розрахуємо, яким потрібен бути обсяг вибірки, щоб похибка частки не перевищувала деякий фіксований відсоток d.

Отримаємо оцінку $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$.

Ця оцінка здійснюється до визначення обсягу вибірки n, тому обсяг вибірки не є відомим. Визначмо p при якому його середньоквадратичне відхилення має максимальне значення. При фіксованому n задача зводиться до максимізації виразу:

$y(p) = p - p^2$

Знайдемо критичні точки:

$y' = 1 - 2p = 0 \Rightarrow p = 1/2$

Перевірімо, що це є максимум: $y'' = -2 < 0$

Тобто для знаходження обсягу вибікиі ми використаємо припущення повної невизначеності (максимуму ентропії). Звідси:

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{\sqrt{0,5 \cdot 0,5}}{\sqrt{n}} = \frac{0,5}{\sqrt{n}}$$

Оскільки похибка оцінки частки обмежена величиною d, то цю умову можна подати:

$$z_{\alpha/2} \cdot \frac{0,5}{\sqrt{n}} \le d$$

Звідси достатній обсяг вибірки для забезпечення точності d на рівні значимості α:

$$n \ge \frac{0,25 \cdot z_{\alpha/2}^2}{d^2}$$

Наприклад, для забезпечення 2% максимальної похибки з 5% рівнем значимості $(d = 0,02; z_{0,025} = 1,96)$ потрібна вибірка не менш ніж:

$$n \ge \frac{0,25 \cdot 1,96^2}{0,02^2} = 2401$$

Якщо задати максимальну похибку в 1% то обсяг вибірки зросте в 4 рази і настільки ж зросте вартість проведення спостережень.

Розглянемо питання планування обсягу спостережень, який відповідає заданим вимогам точності. Мова йде про оцінку математичного очікування.

Існують дві оцінки математичного очікування: для достатньо великих вибірок (більш100 спостережень) і для малих вибірок.

Для великих вибірок 100 (1-α) довірчі інтервали для математичного очікування:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le E(x) \le \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

Похибка оцінки h дорівнює;

$$h = \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

Звідси обсяг вибірки необхідний для отримання оцінки математичного очікування з точністю h і з довірчою ймовірністю 100(1-α):

$$n = \frac{z^2_{\alpha/2}}{h^2}\sigma^2$$

Розглянемо це на прикладі оцінки очікуваної урожайності пшениці в деякому регіоні. Нехай нам потрібна ця оцінка з 99% ймовірністю та з похибкою що не перевищує 3га (h= 3 ц/га).

Значення середньоквадратичного відхилення для урожайності в даному регіоні дорівнює 15 ц/га ($\sigma = 15 ц/га$).

З таблиці хвостів нормального розподілу знайдемо значення нормованої змінної $z_{0,005} = 2{,}57$

Звідси обсяг вибірки можна оцінити:

$$n = \frac{2{,}57^2 \cdot 15^2}{3^2} = 165$$

Тобто для того, щоб оцінити очікувану регіональну урожайність з похибкою, що не перевищує 3 ц/га на рівні значимості 1% потрібно використати спостереження не менш ніж на 165 фермах.

У випадку незначної кількості спостережень (n<100) замість нормального розподілу використовується розподіл Стьюдента:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} t_{n-2;\alpha/2} \le E(x) \le \bar{x} + \frac{\sigma}{\sqrt{n}} t_{n-2;\alpha/2}$$

де $t_{n-2;\alpha/2}$ – критичне значення розподілу Стьюдента з n-2 ступенями свободи та рівнем значимості $\alpha/2.$.

Повернутись до плану.