Лекція 14.
Поняття про статистичні оцінки випадкових величин

План лекції

Питання, що розглядаються:

Емпірична функція розподілу, статистична оцінка невідомого параметра теоретичного розподілу, систематичні помилки, незміщеність оцінки, зміщеність оцінки, спроможність оцінки, ефективність оцінки, вибіркове середнє, вибіркова дисперсія, генеральне середнє, генеральна дисперсія, виправлена вибіркова дисперсія, інтервальна оцінка, точність оцінки, надійність оцінки, довірчий інтервал, рівень значущості.

1. Емпірична функція розподілу

Нехай відомий статистичний розподіл частот кількісної ознаки *X*. Позначимо через n_x число спостережень, при яких спостерігалося значення ознаки, менше *x* і через *n* - загальне число спостережень. Очевидно, відносна частота події $X < x$ рівна n_x / n і є функцією *x*. Оскільки ця функція знаходиться емпіричним (дослідним) шляхом, то її називають емпіричною.

Емпіричною функцією розподілу (функцією розподілу вибірки) називають функцію $F^*(x)$, що визначає для кожного значення *x* відносну частоту події $X < x$. Таким чином, за визначенням $F^*(x) = n_x / n$, де n_x - число варіант, менших *x*, *n* - об'єм вибірки.

На відміну від емпіричної функції розподілу вибірки, функцію розподілу $F(x)$ генеральної сукупності називають ***теоретичною функцією розподілу***. Відмінність між цими функціями полягає в тому, що теоретична функція $F(x)$ визначає ***ймовірність*** події $X < x$, тоді як емпірична - ***відносну частоту*** цієї ж події.

При зростанні *n* відносна частота події $X < x$, тобто $F^*(x)$ прямує за ймовірностю до вірогідності $F(x)$ цієї події. Іншими словами $\lim\limits_{n\to\infty} P[\left|F(x) - F^*(x)\right| < \varepsilon] = 1 \qquad (\varepsilon > 0)$

Властивості емпіричної функції розподілу :

1) Значення емпіричної функції належать відрізку [0,1]
2) $F^*(x)$ - неспадна функція
3) Якщо x_1 - найменша варіанту, то $F^*(x) = 0$ при $x \le x_1$, якщо x_k - найбільша варіанту, то $F^*(x) = 1$ при $x > x_k$.

Емпірична функція розподілу вибірки служить для оцінки теоретичної функції розподілу генеральної сукупності.

Приклад. Побудуємо емпіричну функцію по розподілу вибірки :

Варіанти	2	6	10
Частоти	12	18	30

Знайдемо об'єм вибірки : 12+18+30=60. Найменша варіанта рівна 2, тому $F^*(x)$=0 при $x \le 2$. Значення $x<6$, тобто $x_1 = 2$, спостерігалося 12 разів, отже, $F^*(x)$=12/60=0,2 при $2< x \le 6$. Аналогічно, значення $X < 10$, тобто $x_1 = 2$ і $x_2 = 6$ спостерігалися 12+18=30 разів, тому $F^*(x)$=30/60 =0,5 при $6< x \le 10$. Оскільки x=10 - найбільша варіанту, то $F^*(x)$=1 при $x> 10$. таким чином, шукана емпірична функція має вигляд:

$$F^*(x) = \begin{cases} 0 & x \le 2 \\ 0{,}2 & 2 < x \le 6 \\ 0{,}5 & 6 < x \le 10 \\ 1 & x > 10 \end{cases}$$

2.Найважливіші властивості статистичних оцінок

Нехай вимагається вивчити деяку кількісну ознаку генеральної сукупності. Допустимо, що з теоретичних міркувань вдалося встановити, який саме розподіл має ознаку і необхідно оцінити параметри, якими воно визначається. Наприклад, якщо ознака, що вивчається, розподілена в генеральній сукупності нормально, то треба оцінити математичне очікування і середнє квадратичне відхилення; якщо ознака має розподіл Пуассона - то необхідно оцінити параметр λ.

Зазвичай є лише ці вибірки, наприклад значення кількісної ознаки $x_1, x_2, \ldots, x_n$, отримані в результаті n незалежних спостережень. Розглядаючи $x_1, x_2, \ldots, x_n$ як незалежні випадкові величини $X_1, X_2, \ldots, X_n$ можна сказати, що ***знайти статистичну оцінку невідомого параметра теоретичного розподілу - означає знайти функцію від спостережуваних випадкових величин, яка дає наближене значення оцінюваного параметра.*** Наприклад, для оцінки математичного сподівання нормального розподілу роль функції виконує середнє арифметичне $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$

Для того, щоб статистичні оцінки давали коректні наближення оцінюваних параметрів, вони повинні задовольняти деяким вимогам, серед яких найважливішими є вимоги незміщеності і спроможності оцінки.

Нехай Θ^*- статистична оцінка невідомого параметра Θ теоретичного розподілу. Нехай за вибіркою об'єму n знайдена оцінка Θ_1^*. Повторимо дослід, тобто витягнемо з генеральної сукупності іншу вибірку того ж об'єму і за її даними отримаємо іншу оцінку Θ_2^*. Повторюючи дослід багаторазово, отримаємо різні числа $\Theta_1^*, \Theta_2^*, \ldots \Theta_k^*$. Оцінку Θ^* можна розглядати як випадкову величину, а числа $\Theta_1^*, \Theta_2^*, \ldots \Theta_k^*$ - як її можливі значення.

Якщо оцінка Θ^* дає наближене значення Θ **з надлишком**, тобто кожне число Θ_i^* $(i = 1, \ldots, k)$ більше істинного значення Θ то, як наслідок, математичне сподівання (середнє значення) випадкової величини Θ^* більше, ніж $\Theta : M(\Theta^*) > \Theta$. Аналогічно, якщо Θ^* дає оцінку **з недостачею**, то $M(\Theta^*) < \Theta$.

Таким чином, використання статистичної оцінки, математичне сподівання якої не дорівнює оцінюваному параметру, привело б до систематичних (одного знаку) помилок. Якщо, навпаки, $M(\Theta^*) = \Theta$, то це гарантує від систематичних помилок.

Незміщеною називають статистичну оцінку Θ^*, математичне сподівання якої дорівнює оцінюваному параметру Θ при будь-якому об'ємі вибірки $M(\Theta^*) = \Theta$.

Зміщеною називають оцінку, що не задовольняє цій умові**.**

Незміщенність оцінки ще не гарантує отримання хорошого наближення для оцінюваного параметра, оскільки можливі значення Θ_i^* можуть бути ***сильно розсіяні*** навколо свого середнього значення, тобто дисперсія $D\ \Theta$ може бути значною. В цьому випадку знайдена за даними однієї вибірки оцінка, наприклад Θ_1^*, може виявитися значно віддаленою від середнього значення $\overline{\Theta^*}$, а значить, і від самого оцінюваного параметра.

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

При розгляді вибірок великого об'єму до статистичних оцінок пред'являється вимога спроможності.

Спроможною називається статистична оцінка, яка при $n \to \infty$ прямує за ймовірністю до оцінюваного параметра. Наприклад, якщо дисперсія незміщеної оцінки при $n \to \infty$ прямує до нуля, то така оцінка виявляється і спроможною.

3.Вибіркові середнє і дисперсія

Нехай для вивчення генеральної сукупності відносно кількісної ознаки X витягнута вибірка об'єму n.

Вибірковим середнім $\overline{x_B}$ називають середнє арифметичне значення ознаки вибіркової сукупності. Якщо усі значення $x_1, x_2, \ldots, x_n$ ознаки вибірки об'єму n різні, то $\overline{x_B} = (x_1 + x_2 + \ldots + x_n)/n$.

Якщо значення ознаки $x_1, x_2, \ldots, x_k$ мають частоти $n_1, n_2, \ldots, n_k$ відповідно, причому $n_1 + n_2 + \ldots + n_k = n$,

то $\overline{x_B} = (n_1 x_1 + n_2 x_2 + \ldots + n_k x_k)/n = (1/n)\sum_{i=1}^{k} n_i x_i$.

Вибіркове середнє, знайдене за даними однієї вибірки, дорівнює певному числу. При витяганні інших вибірок того ж об'єму вибіркове середнє мінятиметься від вибірки до вибірки. Тобто вибіркове середнє можна розглядати як випадкову величину і говорити про її розподіли (теоретичний і емпіричний) і про числові характеристики цього розподілу (наприклад, про математичне сподіваня і дисперсію).

Для того, щоб охарактеризувати розсіяння спостережуваних значень кількісної ознаки вибірки навколо середнього значення $\overline{x_B}$ вводиться ***вибіркова дисперсія***. ***Вибірковою дисперсією*** D_B називають середнє арифметичне квадратів відхилення спостережуваних значень ознаки від їх середнього значення $\overline{x_B}$. Якщо усі значення $x_1, x_2, \ldots, x_n$ ознаки вибірки об'єму n різні, то

$$D_B = (1/n)\sum_{i=1}^{n}(x_i - \overline{x_B})^2 .$$

Якщо значення ознаки $x_1, x_2, \ldots, x_k$ мають частоти $n_1, n_2, \ldots, n_k$ відповідно, причому $n_1 + n_2 + \ldots + n_k = n$, то

$$D_B = (1/n)\sum_{i=1}^{k} n_i\,(x_i - \overline{x_B})^2 .$$

Аналогічно вибірковим середньому і дисперсії визначаються ***генеральні середнє і дисперсія***, характеризуючі генеральну сукупність в цілому. Для розрахунку цих характеристик досить в наведених вище співвідношеннях замінити об'єм вибірки *n* на об'єм генеральної сукупності *N*.

Фундаментальне значення для практики має знаходження середнього і дисперсії ознаки генеральної сукупності за відповідними відомими вибірковими параметрами. Можна показати, що вибіркове середнє є незміщеною спроможною оцінкою генерального середнього. В той же час, незміщеною спроможною оцінкою генеральної дисперсії виявляється не вибіркова дисперсія D_B, а так звана "***виправлена" вибіркова дисперсія***, що дорівнює $s^2 = \dfrac{n}{n-1}\,D_B$.

Таким чином, в якості оцінок генерального середнього і дисперсії в математичній статистиці приймають вибіркове середнє і виправлену вибіркову дисперсію.

4. Надійність і довірчий інтервал

Досі ми розглядали точкові оцінки, тобто такі оцінки, які визначаються одним числом. При вибірці малого об'єму точкова оцінка може значно відрізнятися від оцінюваного параметра, що призводить до грубих помилок. У зв'язку з цим при невеликому об'ємі вибірки користуються інтервальними оцінками.

Інтервальною називають оцінку, що визначається двома числами - кінцями інтервалу**.** Нехай знайдена за даними вибірки статистична характеристика Θ^* служить оцінкою невідомого параметра Θ. Очевидно, Θ^* тим точніше визначає параметр Θ, чим менше абсолютна величина різниці $|\Theta - \Theta^*|$. Іншими словами, якщо $\delta > 0$ і $|\Theta - \Theta^*| < \delta$, то чим менше δ, тим точніша оцінка. Таким чином, додатне число δ характеризує ***точність оцінки.***

Статистичні методи ***не дозволяють*** стверджувати, що оцінка Θ^* задовольняє нерівності $|\Theta - \Theta^*| < \delta$; можна говорити лише про ймовірність, з якою ця нерівність здійснюється.

Надійністю (довірчою вірогідністю) оцінки Θ за Θ^* називають ймовірність γ, з якою здійснюється нерівність $|\Theta - \Theta^*| < \delta$. Звичайно надійність оцінки задається заздалегідь, причому в якості γ беруть число, близьке до одиниці, - як правило 0,95; 0,99 або 0,999.

Нехай ймовірність того, що $\left|\Theta-\Theta^*\right| < \delta$ дорівнює γ:

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

Замінимо нерівність $\left|\Theta-\Theta^*\right| < \delta$ рівносильною йому подвійною нерівністю

$$P\left(\Theta^* - \delta < \Theta < \Theta^* + \delta\right) = \gamma \ .$$

Це співвідношення слід розуміти так: ймовірність того, що інтервал $(\Theta^* - \delta, \Theta^* + \delta)$ містить в собі (покриває) невідомий параметр Θ, дорівнює γ .

Таким чином, ***довірчим*** називають інтервал $(\Theta^* - \delta, \Theta^* + \delta)$, який покриває невідомий параметр із заданою надійністю γ .

Величину 1 - γ = α називають ***рівнем значущості*** або ***ймовірністю помилки.***

Для побудови інтервальної оцінки параметра необхідно знати закон його розподілу як випадкової величини.

Питання для самоперевірки

1. Дати означення емпіричної функції розподілу.
2. Записати властивості емпіричної функції розподілу.
3. Що означає знайти статистичну оцінку невідомого параметра теоретичного розподілу?
4. Яким вимогам повині задовольняти статистичні оцінки?
5. Що таке систематичні помилки?
6. Яка оцінка називається незміщеною? Зміщеною?
7. Яка оцінка називається спроможною?
8. Яка оцінка називається ефективною?
9. Дати означення вибіркового середнього і вибіркової дисперсії.
10. Дати означення генерального середнього і генеральної дисперсії.
11. Записати формулу для виправленої вибіркової дисперсії.
12. Що таке інтервальна оцінка?
13. Що таке точність оцінки?
14. Що таке надійність оцінки?
15. Дати означення довірчого інтервалу.
16. Що таке рівень значущості?