УДК 519.2:303.5(168.2+004.421.2)

*Ф. В. МОЦНИЙ,*
*доктор фізико-математичних наук, професор,*
*завідувач кафедри прикладної математики,*
*Національна академія статистики, обліку та аудиту*

# Сучасний базовий інструментарій математичної статистики

*Ця стаття – оглядова. У ній узагальнено результати численних досліджень сучасного базового інструментарію математичної статистики. Розглянуто основні поняття і вибіркові характеристики масових випадкових явищ з урахуванням їх особливостей. Розв'язано типові задачі. Проілюстровано теорему Гливенка–Кантеллі. Оцінено параметри статистичних розподілів.*

***Ключові слова:*** *генеральна сукупність, вибірка, варіаційний ряд, розподіл вибірки, полігон частот, гістограма частот, функція розподілу, крива розподілу, теорема Гливенка–Кантеллі, оцінка, дисперсія, середнє квадратичне відхилення, статистичний розподіл.*

**Постановка проблеми.** Математична статистика відображає кількісні дані спостережень масових випадкових явищ та об'єднує різні методи їх дослідження, спрямовані на встановлення закономірностей та отримання наукових і практичних висновків [1–10]. Фундаментом цих методів служить математика, для якої природа досліджуваних об'єктів не має суттєвого значення. Саме тому математичні методи знаходять широке застосування при опрацьовуванні статистичних даних [11–15].

**Аналіз останніх досліджень і публікацій.** Різні аспекти інструментарію математичної статистики розглянуто в численних монографіях, навчальних посібниках і довідниках, зокрема в [5–9; 13; 15; 16]. Усвідомлене використання практиками математичного апарату в статистиці активізувало в останні роки інтерес до її базових знань [17; 18]. Це послужило автору стимулом для виконання цієї роботи з урахуванням особливостей окремих питань.

**Метою** статті є огляд та узагальнення сучасного базового інструментарію математичної статистики, необхідного для коректного збирання та аналізу статистичних даних масових явищ в умовах невизначеності.

## Частина І. Основні поняття математичної статистики

В математичній статистиці використовується низка понять. Розглянемо основні з них.

### 1.1. Генеральна сукупність і вибірка

Нехай фірма виготовляє світлодіоди, обмежуючись дослідженнями випадково відібраних, і на цій основі робить висновок про всю сукупність. Для великої, але кінцевої сукупності досліджуваних об'єктів, наприклад щорічної партії виготовлених світлодіодів, використовується термін "генеральна сукупність".

*Генеральна сукупність* – повна сукупність усіх елементів (предметів, об'єктів) дослідження.

*Вибіркова сукупність (вибірка)* – частина випадково відібраних елементів із генеральної сукупності для вивчення їхньої якісної чи кількісної ознаки.

*Обсяг сукупності (генеральної, вибіркової)* – кількість елементів сукупності. Зокрема, якщо зі 100 тис. світлодіодів відібрати 1 тис., то обсяг генеральної сукупності дорівнюватиме 100 тис., а обсяг вибірки – 1 тис. В залежності від обсягу генеральної сукупності застосовується *повторна* або *безповоротна* випадкова вибірка. Різниця між ними полягає у тому, що при поворотній вибірці кожен відібраний елемент після дослідження повертається знову до генеральної сукупності. Для отримання надійних висновків про досліджувану ознаку генеральної сукупності необхідно, щоб вибірка була репрезентованою. Згідно із законом великих чисел вибірка буде саме такою, якщо кожен окремий елемент вибиратиметься випадково із генеральної сукупності за умови, що усі елементи мають однакову ймовірність потрапляння до цієї вибірки.

*Способи відбору* одиниць елементів із досліджуваної генеральної сукупності можуть бути різні. Відомі такі способи: *простий випадковий*, який не вимагає розбиття генеральної сукупності на складові частини; *типовий (районований), механічний або серійний,* при яких генеральна сукупність розбивається на складові частини; і *комбінований.* При *простому випадковому відборі* вибирається навмання по одному визначена наперед кількість елементів із генеральної сукупності; при *типовому* – по одному з типових частин генеральної сукупності, виділених за якою-небудь суттєвою ознакою; при *механічному* – по одному з кожної механічно відібраної частини, кількість яких відображає генеральну сукупність і збігається з обсягом майбутньої вибірки; при *серійному* – не по одному, а серіями (цілими групами) із генеральної сукупності; при *комбінованому* відборі поєднуються усі названі вище способи, а саме: *простий випадковий*, *типовий*, *механічний* та *серійний*. Якщо генеральна сукупність неоднорідна за досліджуваними ознаками, застосовується типовий спосіб відбору. Якщо ж досліджувана ознака слабо відрізняється у різних серіях, використовується серійний спосіб. Точність серійної вибірки залежатиме від того, наскільки адекватно середні показники серій відображатимуть генеральну сукупність. Зрозуміло, що чим менше вони відхилятимуться від генеральної середньої, тим більш точними будуть результати серійної вибірки.

## 1.2. Статистичні ряди

*Статистичні ряди* є складовою частиною методу статистичних зведень, групувань і належать до найважливіших елементів статистики. Вони поділяються на дискретні й інтервальні.

### *1.2.1. Дискретний варіаційний ряд*

Нехай із генеральної сукупності відібрано вибірку обсягом $n$ для дослідження на яку-небудь кількісну ознаку $X$. Вважатимемо, що ознака $X$ є дискретною і значення $x_1$ приймає $n_1$ разів, значення $x_2 - n_2$ разів, ..., значення $x_k - n_k$ разів. Знайдені експериментально значення $x_1$, $x_2$, ..., $x_k$, називаються *варіантами,* а числа спостережень $n_1$, $n_2$, ..., $n_k$, – *частотами* відповідних варіант. *Обсяг вибірки* $n = \sum_{i=1}^{k} n_i$, де $k$ – кількість варіант, які відрізняються числовим значенням. Відношення частоти $n_i$ варіанти $x_i$ до обсягу вибірки $n$ називається її *відносною частотою* $W_i$, тобто $W_i = n_i / n$. Характерно, що для кожної вибірки справедлива рівність: $\sum_{i=1}^{k} W_i = 1$. Послідовність варіант, розташованих у порядку зростання, називається *дискретним варіаційним рядом.* Перелік варіант варіаційного ряду і відповідних їм частот або відносних частот називається *дискретним статистичним розподілом вибірки.*

### *1.2.2. Інтервальний варіаційний ряд*

Якщо ознака $X$ генеральної сукупності неперервна (безперервна), то варіант буде багато. За наявності досить великої кількості можливих значень ознаки первинний варіаційний ряд стає важко оглядовим і відображає по суті певну кількість рівних або нерівних частинних інтервалів варіант зі своїми частотами. Послідовність частинних інтервалів варіант, розташованих у порядку зростання, називається *інтервальним варіаційним рядом.* Для його побудови необхідно вибрати оптимальну кількість інтервалів і знайти їхні довжини. При цьому довжина кожного інтервалу має залишатися сталою, оскільки при аналізі варіаційного ряду порівнюються частоти з різних груп. Окрім того, оптимальна кількість інтервалів у варіаційному ряді має відображати різноманітність ознак сукупності і враховувати їх закономірний розподіл, унеможливлюючи спотворення сукупності випадковими коливаннями частот. Оптимальна кількість інтервалів у варіаційному ряді визначається за формулою Стерждеса:

$$n^* = 1 + \log_2 N . \qquad (1)$$

Тут $n^*$ – кількість інтервалів, $N$ – чисельність генеральної сукупності. Якщо $n^*$ виявиться дробовим, то кількість інтервалів дорівнюватиме найближчому до нього цілому числу. Величина інтервалу знаходиться за формулою:

$$l = (x_{max} - x_{min})/n^* , \qquad (2)$$

де $x_{max}$ і $x_{min}$ – максимальне і мінімальне значення варіант відповідно. За початок першого інтервалу $x_0$ приймається значення $x_0 = x_{min} - l/2$. Наступні інтервали визначаються послідовним додаванням ширини інтервалу $l$ до попереднього значення кордону: $x_1 = x_0 + l$; $x_2 = x_1 + l$; $\ldots i$, нарешті, за кінець $k$-го інтервалу приймається значення $x_k$, що відповідає верхній границі $k$-го інтервалу і початку $(k-1)$-го: $x_k = x_{k-1} + l$. Побудова шкали інтервалів продовжуватиметься до тих пір, поки виконуватиметься нерівність $x_k < x_{max} + l/2$. Перелік часткових інтервалів і відповідних їм частот (відносних частот) називається *інтервальним статистичним розподілом вибірки.* Послідовне додавання відносної частоти $W_i$ чергового інтервалу називається *накопиченням частоти:* $W_t^a = \sum_{i=1}^{t} W_i$, причому для останнього інтервалу $W_k^a = \sum_{i=1}^{k} W_i = 1$. Розглянемо подачу вибірки у вигляді статистичного ряду частот та відносних частот на такому прикладі.

*Приклад 1.* Із генеральної сукупності навмання відібрано вибірку, значення ознаки $X$ якої наведено в таблиці. Подати вибірку у вигляді інтервального статистичного розподілу.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 9 | 3 | 8 | 11 | 5 | 7 | 3 | 10 | 8 | 5 |
| 10 | 4 | 10 | 6 | 4 | 10 | 9 | 5 | 9 | 6 | 9 |
| 5 | 13 | 11 | ***2*** | ***2*** | 11 | 8 | 4 | 3 | 7 | 6 |
| 3 | 9 | 8 | 10 | ***16*** | 6 | **9** | 9 | 7 | 8 | 8 |
| 11 | 7 | 9 | 7 | 3 | 7 | 11 | 8 | 9 | 9 | 7 |

*Розв'язання.*

Обсяг вибірки $N = 55$, $x_{max} = 16$, $x_{min} = 2$, оптимальна кількість інтервалів у варіаційному ряді $n^* = 7$, довжина інтервалу $l = 2$. Згрупуємо далі варіанти за інтервалами і занесемо дані в табл. 1. Контроль:

$$0{,}0182 + 0{,}0909 + 0{,}1636 + 0{,}2364 + 0{,}2909 + 0{,}1818 + 0{,}0182 = 1.$$

Отже, вибірку як статистичний ряд складено вірно. Інтервальний статистичний розподіл вибірки матиме вигляд, наведений в табл. 1.

*Таблиця 1*

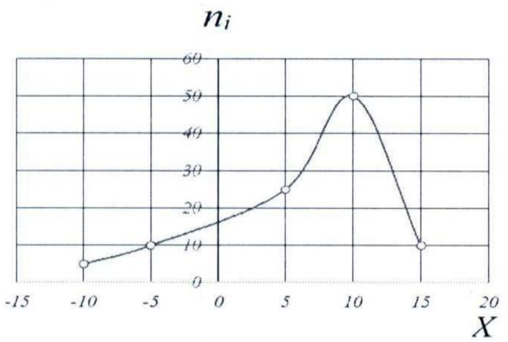
**Інтервальний статистичний розподіл вибірки**

| № інтервалу $i$ | Частинний інтервал $l = 2$ | Середина інтервалу $x_i$ | Частота $n_i$ | Відносна частота $W_i$ |
|---|---|---|---|---|
| 1 | 1-3 | 2 | 1 | 0,0182 |
| 2 | 3-5 | 4 | 5 | 0,0909 |
| 3 | 5-7 | 6 | 9 | 0,1636 |
| 4 | 7-9 | 8 | 13 | 0,2364 |
| 5 | 9-11 | 10 | 16 | 0,2909 |
| 6 | 11-13 | 12 | 10 | 0,1818 |
| 7 | 13-15 | 14 | 1 | 0,0182 |

Найбільш використовуваними графіками для зображення варіаційних рядів є *полігон*, *гістограма* і *емпірична (експериментальна, вибіркова, статистична) функція (кумулята)*. Зупинимося на них детальніше, акцентуючи увагу на їхніх особливостях.

## 1.3. Полігон і гістограма частот і відносних частот

*Полігон частот (відносних частот)* – ламана лінія, яка в прямокутній системі координат з'єднує координати крапок $(x_i;\ n)$ $((x_i;\ W_i))$. Їхні графіки показано на рис. 1 для дискретного статистичного розподілу вибірки, наведеної в табл. 2.



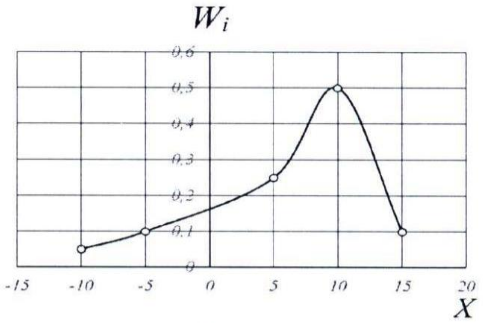




**Рис. 1. Полігон частот (а) і відносних частот (б) для даних табл. 2**

*Таблиця 2*

**Дискретний статистичний розподіл вибірки**

| $X=x_i$ | -10 | -5 | 5 | 10 | 15 |
|---|---|---|---|---|---|
| $n_i$ | 5 | 10 | 25 | 50 | 10 |
| $W_i$ | 0,05 | 0,1 | 0,25 | 0,50 | 0,10 |

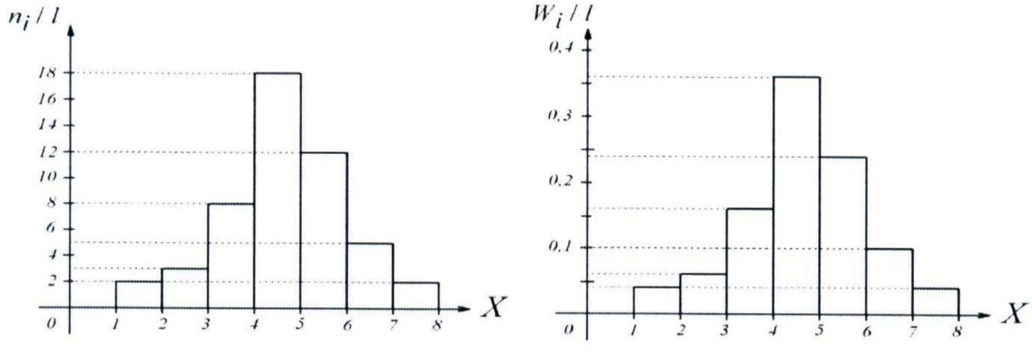
*Інтервальний статистичний розподіл вибірки* можна також зобразити графічно, але у вигляді не ламаної лінії, а фігури зі “сходів” – *гістограми частот (відносних частот).*

*Гістограма частот (відносних частот)* – фігура, яка в прямокутній системі координат складається з прямокутників, кожний з яких має основу $l$ і висоту $h_i = n_i / l$ ( $h_i^* = W / l$). Для побудови гістограми частот (відносних частот) відкладемо на осі абсцис частинні інтервали, а на осі ординат проведемо паралельно осі абсцис відрізки на висоті $h_i = n_i / l$ ( $h_i^* = W / l$). Площа $i$-го частинного прямокутника дорівнює

$$S_i = l \cdot h_i = l \cdot h_i / l = n_i \; (S_i^* = l \cdot h_i^* = l \cdot W_i / l = W_i)$$

і визначається таким чином частотою (відносною частотою) варіант $i$-го інтервалу. Тоді загальна площа усіх гістограм частот (відносних частот):

$S = \sum_{i=1}^{k} S_i = \sum_{i=1}^{k} n_i = n (S^{**} = \sum_{i=1}^{k} S_i^* = \sum_{i=1}^{k} W_i = 1)$, тобто дорівнює обсягу вибірки (одиниці). Гістограму частот (відносних частот) наведено на рис. 2 для даних табл. 3.




а б

**Рис. 2. Гістограма частот (а) і відносних частот (б) для даних табл. 3.**

*Таблиця 3*

**Інтервальний статистичний розподіл вибірки**

| Частинний інтервал довжиною $l$=1 | Сума частот варіант частинного інтервалу $n_i$ | Густина частоти $n_i/l$ | Сума відносних частот частинного інтервалу $W_i$ | Густина відносної частоти $W_i/l$ |
|---|---|---|---|---|
| 1 – 2 | 2 | 2 | 0,04 | 0,04 |
| 2 – 3 | 3 | 3 | 0,06 | 0,06 |
| 3 – 4 | 8 | 8 | 0,16 | 0,16 |
| 4 – 5 | 18 | 18 | 0,36 | 0,36 |
| 5 – 6 | 12 | 12 | 0,24 | 0,24 |
| 6 – 7 | 5 | 5 | 0,10 | 0,10 |
| 7 – 8 | 2 | 2 | 0,04 | 0,04 |

(Контроль: $\sum_{i=1}^{7} W_i = 0{,}04 + 0{,}06 + 0{,}16 + 0{,}36 + 0{,}24 + 0{,}10 + 0{,}04 = 1$ $\Rightarrow$ вибірку як статистичний ряд складено вірно).

## 1.4. Емпірична функція статистичного розподілу вибірки

*Дискретні* та *інтервальні статистичні розподіли вибірок* можна описати аналітично, увівши емпіричну функцію $F_n{}^*(x)$.

*Емпірична функція* $F_n{}^*(x)$ *статистичного розподілу вибірки* визначає відносну частоту події $X < x$:

$$F_n^*(x) = W(X < x) = n_x / n, \qquad (3)$$

де $n_x$ – кількість варіант статистичного розподілу вибірки, значення яких менше фіксованої варіанти $x$; $n$ – обсяг вибірки. Зокрема, щоб знайти $F_n{}^*(x_i)$, потрібно кількість варіант, які менше $x_i$, розділити на обсяг вибірки. Перерахуємо властивості функції $F_n{}^*(x)$, які слідують із її означення: а) $0 \le F_n{}^*(x) \le 1$), б) $F_n{}^*(x_1) \le F_n{}^*(x_2)$ при $x_1 < x_2$, і в) $F_n{}^*(x)=0$ при $x \le x_{\min}$ і $F_n{}^*(x)=1$ при $x > x_{\max}$. Отже, емпірична функція $F_n{}^*(x)$ статистичного розподілу вибірки містить усю інформацію про результати спостережень події $X < x$. Наголосимо також, що при побудові функції $F_n{}^*(x)$ для інтервального варіаційного статистичного ряду передбачається, що досліджувана ознака $X$ на кожному частинному інтервалі має рівномірну густину (щільність) ймовірностей.

*Приклад 2.* Для заданого в таблиці дискретного статистичного розподілу вибірки

| | | | | |
|---|---|---|---|---|
| $X = x_i$ | 12 | 15 | 18 | 21 |
| $n_i$ | 10 | 18 | 14 | 8 |

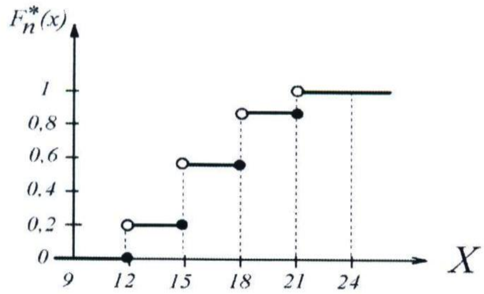
знайти емпіричну функцію $F_n{}^*(x)$ та побудувати її графік .

Розв'язання.

Обсяг вибірки $n = 50$. Знайдемо функцію $F_n^*(x)$:

$$F_n^*(x) = \begin{cases} 0; & x \le 12. \\ W_1 = 0{,}20; & 12 < x \le 15. \\ W_1 + W_2 = 0{,}20 + 0{,}36 = 0{,}56; & 15 < x \le 18. \\ W_1 + W_2 + W_3 = 0{,}20 + 0{,}36 + 0{,}28 = 0{,}84; & 18 < x \le 21. \\ W_1 + W_2 + W_3 + W_4 = 0{,}20 + 0{,}36 + 0{,}28 + 0{,}16 = 1; & x > 21. \end{cases}$$

Графік цієї функції показано на рис. 3(а). Видно, що він має форму ступінчатої лінії, яка на всьому проміжку змінних $x$ характеризується стрибками в крапках вибірки (варіаційного ряду), причому величина стрибка в крапці $x_i$ дорівнює $k / n$, де $k$ – кількість елементів вибірки, які збігаються із $x_i$. Зокрема, величина стрибка в крапці $x_1$ дорівнює 0,20; в крапці $x_2$ – 0,56; в крапці $x_3$ – 0,84 і в крапці $x_4$ – 1.



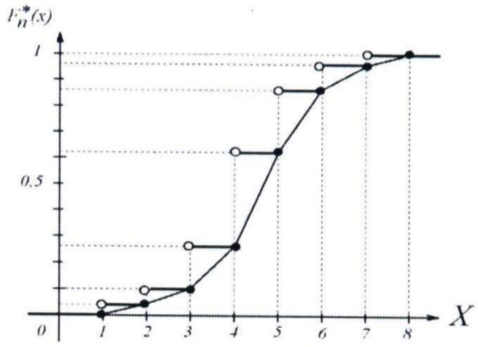




**Рис. 3. Графік емпіричної функції $F_n^*(x)$ для дискретного (а) та інтервального (б) статистичних розподілів вибірок**

*Приклад 3.* Для заданого в табл. 3 інтервального статистичного розподілу вибірки знайти емпіричну функцію $F_n^*(x)$ та побудувати її графік.

*Розв'язання.*
Функція $F_n^*(x)$ має вигляд:

$$F_n^*(x) = \begin{cases} 0; & x \le 1. \\ W_1 = 0{,}04; & 1 < x \le 2. \\ W_1 + W_2 = 0{,}4 + 0{,}06 = 0{,}10; & 2 < x \le 3. \\ W_1 + W_2 + W_3 = 0{,}10 + 0{,}16 = 0{,}26; & 3 < x \le 4. \\ W_1 + W_2 + W_3 + W_4 = 0{,}26 + 0{,}36 = 0{,}62; & 4 < x \le 5. \\ W_1 + W_2 + W_3 + W_4 + W_5 = 0{,}62 + 0{,}24 = 0{,}86; & 5 < x \le 6. \\ W_1 + W_2 + W_3 + W_4 + W_5 + W_6 = 0{,}86 + 0{,}10 = 0{,}96; & 6 < x \le 7. \\ W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7 = 0{,}96 + 0{,}04 = 1; & 7 < x \le 8. \end{cases}$$

Видно, що для інтервального статистичного розподілу вибірки графік функції $F_n^*(x)$ має також форму ступінчатої лінії (рис. 3(б)), яка на всьому проміжку змінних $x$ характеризується стрибками в крапках вибірки (варіаційного ряду), але досягає одиниці на останньому інтервалі, а не в кінці його, як це мало місце у випадку дискретної статистичної вибірки. Окрім того, оскільки досліджувана ознака $X$ на кожному частинному інтервалі має рівномірну густину (щільність) ймовірностей, то графік цієї функції можна подати в іншому вигляді, а саме, у вигляді ламаної лінії, яка зростає на кожному частинному інтервалі і наближається до одиниці.

### 1.5. Теоретична функція розподілу генеральної сукупності

*Теоретична функція $F(x)$ розподілу генеральної сукупності* є аналогом емпіричної функції $F_n{}^*(x)$ статистичного розподілу вибірки. Відмінність між ними полягає у тому, що функція $F(x)$ визначає ймовірність випадкової події $X < x$ $(F(x) = P(X < x))$, тоді як функція $F_n{}^*(x)$ визначає відносну частоту цієї події $(F_n{}^*(x) = W(X < x))$. Із теореми Якоба Бернуллі (закону великих чисел) слідує, що при необмеженому зростанні кількості експериментів $(n_{\exp} \to \infty)$ відносна частота події $X < x$ збігається по ймовірності до ймовірності цієї події. Це означає, що при $n_{\exp} \to \infty$ функція $F_n{}^*(x)$ збігається по ймовірності до функції $F(x)$ $(F_n{}^*(x) \xrightarrow{p} F(x))$ і $\lim\limits_{n_{\exp} \to \infty} P(|F_n{}^*(x) - F(x)| \le \varepsilon) = 1$ для будь-якого малого $\varepsilon > 0$, тобто числа $F_n{}^*(x)$ і $F(x)$ відрізнятимуться мало. З іншого боку, при зростанні обсягу вибірки функція $F_n{}^*(x)$ наближатиметься до функції $F(x)$ рівномірно по $x$, як доведено в теоремі В. І. Гливенка (1897 – 1940) для неперервної функції $F(x)$ [1] та узагальнено в теоремі Ф. П.Кантеллі (1875 – 1966) для довільної функції [2], а саме:

$$\lim_{n \to \infty} \sup_{x \in R} |F_n(x) - F(x)| = 0 \tag{4}$$

майже напевно, де символ sup (supremum) позначає точну верхню грань модуля відхилення $F_n{}^*(x)$ від $F(x)$ (найбільше їх різниці). Слід наголосити також, що теорема Гливенка – Кантеллі є центральною теоремою математичної статистики, яка ґрунтується на випадковості вибірки та ймовірнісних закономірностях і не використовує жодних специфічних властивостей генеральної сукупності. Сутність цієї теореми проілюстровано на рис. 4, де наведено графіки емпіричної і теоретичної функцій при різних обсягах вибірок. Видно, що зі збільшенням обсягу вибірки емпірична крива збігається до теоретичної і співпадає з нею при $n = 2500$. Характерно, що швидкість збігу до нуля у формулі (4), як встановлено А. Н. Колмогоровим (1903 – 1987) у теоремі [4] для неперервної функції $F(x)$, порядку $1/\sqrt{n}$. Більше того, функція $F_n{}^*(x)$ має всі властивості функції $F(x)$. Таким чином, усе разом узяте переконливо свідчить, що при скінченому обсязі вибірки емпірична функція $F_n{}^*(x)$ статистичного розподілу вибірки може служити наближенням для невідомої функції $F(x)$ розподілу генеральної сукупності.

Отже, знаючи розподіл теоретичної функції $F(x)$ генеральної сукупності, можна встановити, наскільки адекватно (правильно) емпірична функція

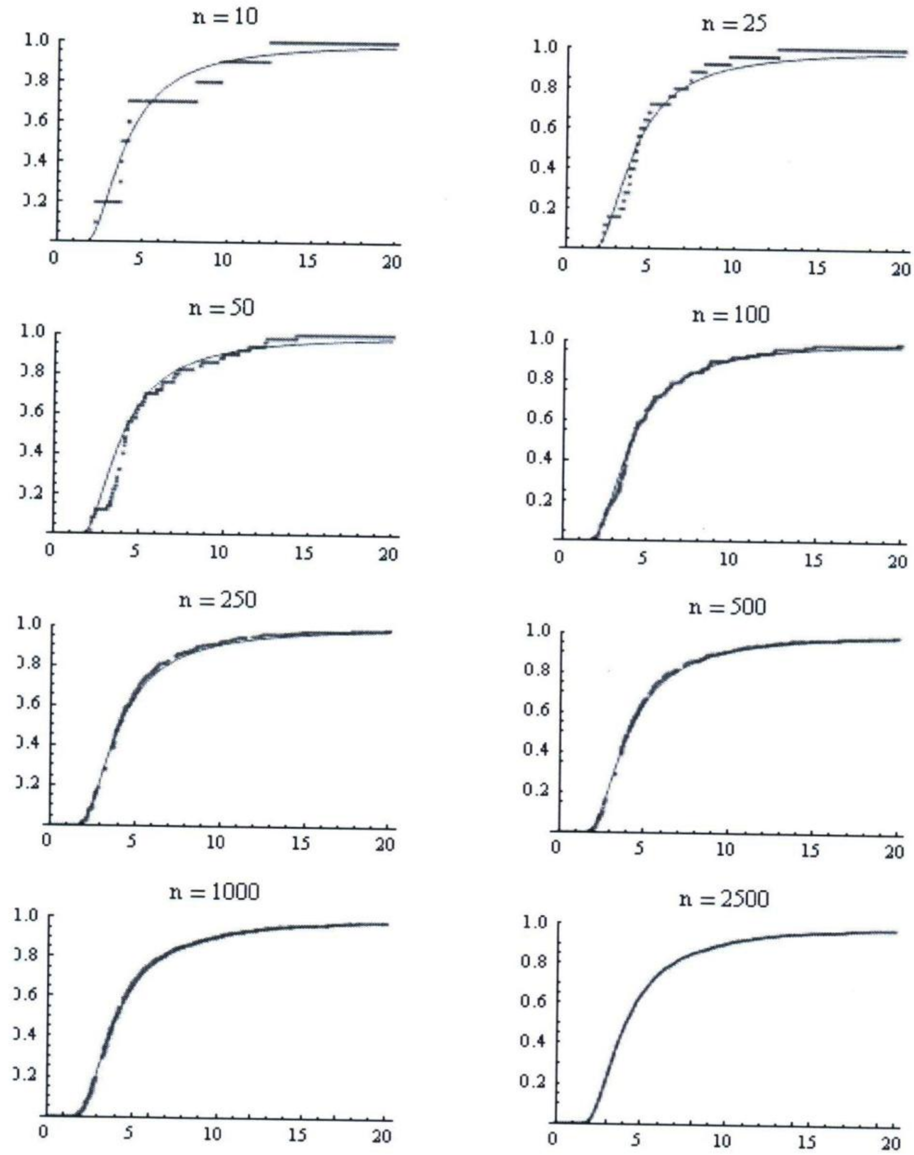



**Рис. 4. Ілюстрація до теореми Гливенка – Кантеллі: *експеримент* – ступінчата лінія, *теорія* – суцільна крива (*n* – обсяг вибірки) [3]**

$F_n^*(x)$ статистичного розподілу вибірки відображає теорію, і навпаки, знаючи розподіл емпіричної функції $F_n^*(x)$, можна уявити вигляд невідомої теоретичної функції $F(x)$.

### 1.6. Згладжена крива розподілу

Однією із основних задач математичної статистики є пошук теоретичної функції $F(x)$ розподілу генеральної сукупності. Знайдемо її на прикладі гістограми відносних частот (рис. 5), побудованої на основі даних табл. 1. Проведемо на цьому графіку плавну криву через середини вертикальних і горизонтальних складових сходинок. Тоді згідно з методом найменших квадратів сума площ гістограми, що залишилася над згладженою кривою, компенсується сумою площ, що увійшли до площі під нею. Отримана у такий спосіб експериментальна крива називаються *згладженою емпіричною кривою розподілу,* а процедура підбору відповідної їй теоретичної функції $f^*(x)$ – *вирівнюванням статистичного ряду*.

Функція $f^*(x)$ належатиме густині (щільності) $f(x)$ розподілу ймовірностей при виконанні таких двох умов:

- $f^*(x) \geq 0$ (5)
- $F^*(x) = \int_{-\infty}^{+\infty} f^*(x)dx = 1.$ (6)

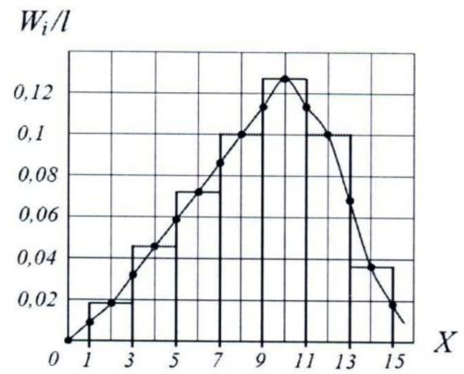



**Рис. 5. Згладжена емпірична крива розподілу статистичного ряду для даних табл. 1.**

Вигляд графіка функції $f^*(x)$ часто дає можливість висунути гіпотезу про закон розподілу ймовірностей. Цей закон містить невідомі параметри, які можна знайти з експериментальних даних. Будь-яке припущення про вигляд теоретичної інтегральної функції $F^*(x)$ розподілу генеральної сукупності називається *статистичною гіпотезою.*

### 1.7. Статистичні оцінювання

Статистичні оцінки ознак генеральної сукупності на основі обробки вибірок завжди містимуть певні похибки, бо обсяги вибірок завжди менші обсягу її сукупності. Для забезпечення максимального ступеня довіри до параметрів генеральної сукупності або закону розподілу її ознаки необхідно, щоб вибірка була репрезентовною і відображала у такий спосіб генеральну сукупність найбільш повно. Нехай $\theta^*$ – статистична оцінка невідомого сталого параметра $\theta$ генеральної сукупності. Виберемо із генеральної сукупності випадкові вибірки обсягом $n$ і на їхній основі знайдемо числа $\theta_1^*, \theta_2^*, \ldots \theta_n^*$, вважаючи їх різними. Тоді оцінка $\theta^*$ є випадковою величиною, а числа $\theta_1^*, \theta_2^*, ... \theta_n^*$ – її можливими значеннями. Характерно, що наближена оцінка $\theta^*$ може бути як більше, так і менше істинного значення $\theta$. Тому окремі числа $\theta_i^* (i = 1,2,..., n)$ можуть також бути як більше, так і менше цього значення. У такому випадку математичне сподівання випадкової величини $\theta^*$ не буде дорівнювати очікуваному параметру $\theta$ $(M(\theta^*) \neq \theta)$, що, в свою чергу, приведе до систематичних похибок вимірювань, які спотворять експериментальні дані. Отже, гарантією від отримання систематичних похибок є збереження вимог $M(\theta^*) = \theta$. Враховуючи це, дамо означення якості оцінювання параметрів статистичних розподілів.

*Незміщена статистична оцінка* $\theta^*$ – це оцінка невідомого параметра $\theta$, математичне сподівання якої співпадає з цим параметром при будь-якому обсязі вибірки $(M(\theta^*) = \theta)$. Точність незміщених оцінок визначається дисперсією – чим дисперсія менша, тим точність вища.

*Асимптотично незміщена оцінка* – статистична оцінка $\theta^*$, зміщення якої від невідомого параметра $\theta$ збігається до нуля, коли обсяг вибірки збігається до нескінченності, тобто $\lim\limits_{n \to \infty}(\theta^* - \theta) = 0$. Практично всі оцінки параметрів, які використовуються у ймовірно-статистичних методах прийняття рішень, є або незміщеними, або асимптотично незміщеними.

*Зміщена оцінка* $M(\theta^* - \theta)$ – це статистична оцінка $\theta^*$, математичне сподівання якої не дорівнює оцінювальному параметру $\theta$ $(M(\theta^*) \neq \theta)$. Точність зміщених оцінок визначається математичним сподіванням квадрата оцінки $M(\theta^* - \theta)^2$. Із основних властивостей математичного сподівання і дисперсії маємо:

$$M(\theta^* - \theta)^2 = D(\theta^*) + (M(\theta^*) - \theta)^2, \tag{7}$$

тобто математичне сподівання квадрата похибки є сумою двох доданків: дисперсії оцінки і квадрата різниці між математичним сподіванням оцінки і оцінювальним параметром. Для численних імовірно-статистичних методів прийняття рішень дисперсія і зміщення є величинами одного порядку $1/n$. Тому при великих обсягах $n$ вибірки можна знехтувати другим доданком у формулі (7). Тоді точність оцінки математичного сподівання квадрата зміщення буде наближено такою:

$$M(\theta^* - \theta)^2 \approx D(\theta^*) \approx \frac{c(\theta^*; \theta)}{n}, \tag{8}$$

де $c(\theta^*; \theta)$ – число, яке визначається методом розрахунку оцінок $\theta^*$ та істинним значенням оцінювального параметра $\theta$. Із формули (8) слідує, що ця точність буде найкращою, коли величина $c(\theta^*; \theta)$ буде мінімальною.

Виникає питання, чи завжди *незміщена оцінка* буде задовільною статистичною оцінкою параметра $\theta$? Відповідь однозначна – ні не завжди, бо дисперсія $D(\theta^*)$ може бути великою. І, як наслідок, оцінка $\theta_i^*$, знайдена на основі однієї вибірки, може виявитися значно віддаленою від середнього значення $\bar{\theta}^*$, а, отже, і від оцінювального параметра $\theta$, що, зрозуміло, приведе до значних похибок. З іншого боку, якщо дисперсія $D(\theta^*)$ буде малою, це унеможливлюватиме допущення великих похибок, і тому, щоб врахувати це, до статистичної оцінки висуваються вимоги *ефективності* – наступної важливої властивості методу оцінювання.

*Ефективна оцінка* – статистична оцінка $\theta_i^*$ яка при заданому обсязі вибірки $n$ має найменшу дисперсію з усіх можливих незміщених оцінок даного параметра.

Якщо вивчаються вибірки великого обсягу, то до статистичних оцінок висуваються вимоги спроможності. *Спроможна оцінка* – статистична оцінка $\theta^*$, яка при $n \to \infty$ наближається по ймовірності до істинного значення параметра $\theta$ генеральної сукупності, тобто $\lim_{n\to\infty} P(|\theta^* - \theta| < \varepsilon) = 1$. Іншими словами, статистика $\theta_n^*$ буде *спроможною оцінкою* параметра $\theta$ тоді і лише тоді, коли для будь-якого наперед заданого малого числа $\varepsilon > 0$ буде справедливим граничне співвідношення $\lim_{n\to\infty} P(|\theta^* - \theta| > \varepsilon) = 0$. Зокрема, якщо звернутися до теореми В. І. Гливенка [1], то функція $F_n^*(x)$ є спроможною оцінкою функції $F(x)$. Зазначимо також, що при розробці нових методів оцінювання параметрів статистичних розподілів необхідно перевіряти спроможність цих оцінок.

Таким чином, знайшовши *незміщену*, *асимптотично незміщену, зміщену*, *ефективну* і *спроможну оцінки* параметра статистичного розподілу, можна порівняти між собою різні методи оцінювання і з'ясувати, наскільки наближена оцінка $\theta^*$є достовірною оцінкою невідомого параметра $\theta$ теоретичного розподілу.

Статистичні оцінювання параметрів генеральної сукупності можна розділити на *крапкову* (*точкову*), *інтервальну* та *довірчий інтервал* (*кінці інтервалу)*.

*Крапкова (точкова) оцінка* $\theta^*$ – статистична оцінка, яка визначається одним числом, крапкою (точкою). До крапкових статистичних оцінок належать незміщена, асимптотично незміщена, зміщена, ефективна і спроможна оцінки параметрів генеральної сукупності. Якщо вибірка мала, то крапкові оцінки можуть значно відрізнятися від оцінювальних параметрів і, як наслідок, стати причиною грубих помилок. Тому у випадках малих вибірок доцільно користуватися інтервальною оцінкою.

*Інтервальна оцінка* $\theta^*$ – статистична оцінка, яка визначається двома числами, що належать кінцям інтервалу. Якщо абсолютна величина зміщення $|\theta^* - \theta| < \delta$,то очевидно, що чим менше $\delta$, тим зміщення $|\theta^* - \theta|$ також менше, тобто точність оцінки визначатиметься $\delta$. З іншого боку, $\theta^*$ може задовольняти нерівності $\left|\theta^{*}-\theta\right|>\delta$ лише з певною ймовірністю $\gamma$ , а саме:

$$P(\left|\theta^{*}-\theta\right|<\delta)=\gamma\ . \qquad (9)$$

Тому параметр $\gamma$ називається параметром *надійності* (*довірчої ймовірності*) і приймається рівним одиниці чи меншому за одиницю числу, наприклад 0,96; 0,99 або 0.999. Якщо формулу (9) переписати у вигляді

$$P[\theta^{*}-\delta<\theta<\theta^{*}+\ \delta]=\gamma\ , \qquad (10)$$

то можна зробити висновок, що ймовірність обмеження невідомого параметра $\theta$ інтервалом $[\theta\ ^{*}-\delta,\theta\ ^{*}+\delta]$ дорівнює $\gamma$ . Цей інтервал називається *довірчим,* бо визначає границі, у межах яких знаходиться $\theta$ із заданою ймовірністю $\gamma$ .

**Висновки**

1. Для забезпечення максимальної довіри до ознаки генеральної сукупності необхідно, щоб вибірка була репрезентованою. Відбір елементів можна здійснювати простим випадковим, типовим, механічним, серійним або комбінованим способами.

2. Статистичні ряди поділяються на дискретні й інтервальні. Перші базуються на дискретних значеннях кількісної ознаки, а другі – на неперервних змінних, заданих інтервалами. Оптимальна кількість інтервалів визначається формулою Стерждеса, а довжина кожного інтервалу – різницею між максимальним і мінімальним значеннями вибірки, поділеної на їх кількість.

3. Дискретний та інтервальний статистичні розподіли вибірок можна подати як у табличній формі, так і графічно у вигляді ламаної лінії (полігону частот (відносних частот)) і фігури зі "сходів" (прямокутників) (гістограм частот (відносних частот)). Ці ряди можна описати також аналітично емпіричною функцією $F_n^{*}(x)$ статистичного розподілу вибірки, що визначає відносну частоту події $X<x$ і містить всю інформацію про результати спостережень.

4. Теоретична функція $F(x)$ розподілу генеральної сукупності визначає ймовірність випадкової події $X<x$ і є аналогом емпіричної функції $F_n^{*}(x)$ статистичного розподілу вибірки. Зі збільшенням обсягу вибірки функція $F_n^{*}(x)$ збігається до функції $F(x)$, причому швидкість їх збігу порядку $1/\sqrt{n}$. Знаючи функцію $F(x)$, можна встановити, наскільки адекватно функція $F_n^{*}(x)$ відображає теорію і навпаки.

5. Згладжені емпіричні криві розподілу можна використати для знаходження теоретичної функції $F(x)$ розподілу генеральної сукупності. Якщо підібрана на основі згладженої емпіричної кривої розподілу функція $f^{*}(x)$ задовольнятиме умовам (5) і (6), то вона відповідатиме густині $f(x)$

розподілу ймовірностей генеральної сукупності, а інтегральна функція $F^*(x)$ – функції $F(x)$ розподілу цієї сукупності.

6. Незміщена, асимптотично незміщена, зміщена, ефективна і спроможна оцінки параметра статистичного розподілу дозволяють порівняти між собою різні методи оцінювання і з'ясувати, наскільки наближена оцінка $\theta^*$ буде достовірною статистичною оцінкою невідомого параметра $\theta$ теоретичного розподілу.

7. При малій вибірці крапкова (точкова) оцінка $\theta^*$ є грубою. Нерівність $|\theta^* - \theta| < \delta$ виконується лише з певною ймовірністю $\gamma$, що дорівнює одиниці або числу, меншому одиниці. Чим менше $\delta$, тим точніше $\theta$.

*Продовження статті знайдете в наступному номері.*