# ЗМІСТОВИЙ МОДУЛЬ 2
# МАТЕМАТИЧНА СТАТИСТИКА

## Первинне опрацювання статистичних даних

### Генеральна та вибіркова сукупності

Нехай потрібно вивчити сукупність однорідних об’єктів відносно деякої якісної або кількісної ознаки, яка характеризує ці об’єкти. Наприклад, якщо досліджується партія товарів, то якісною ознакою може слугувати стандартність товару, я кількісною – наприклад, розмір товару.

Іноді проводять повне дослідження, тобто аналізується кожний об’єкт сукупності відносно ознаки, якою цікавляться. Проте на практиці, таке повне дослідження проводиться відносно рідко, і найчастіше із сукупності обирається та досліджується лише якась контрольна частина об’єктів у обмеженій кількості. Введемо відповідні означення.

**Означення.** Сукупність всіх можливих об’єктів даного виду, над якими проводяться спостереження, або сукупність всіх можливих спостережень, які проводяться в незмінних умовах над деякою випадковою величиною, називається *генеральною сукупністю.*

Генеральна сукупність може містити скінченну або нескінченну кількість елементів.

**Означення.** Відібрані з генеральної сукупності об’єкти (або результати спостережень) називаються *вибірковою сукупністю* або просто *вибіркою.*

**Означення.** Число $N$ елементів *генеральної сукупності* та число $n$ елементів *вибіркової сукупності* будемо називати відповідно *об’ємами генеральної* та *вибіркової сукупностей*.

**Повторна та безповторна вибірки. Репрезентативна вибірка**

При складанні вибірки можна діяти двома способами: після того, як об’єкт був відібраний та над ним було проведене спостереження, він може бути повернутий або не повернутий у генеральну сукупність. Відповідно до цього вибірки розділяють на *повторні* та *безповторні.*

**Означення.** *Повторною* називають *вибірку*, при якій відібраний об’єкт (перед відібранням наступного) повертають до генеральної сукупності.

**Означення.** *Безповторною* називають *вибірку*, при якій відібраний об’єкт не повертають до генеральної сукупності.

На практиці зазвичай користуються безповторним випадковим відбором.

Для того щоб за даними вибірки можна було достатньо впевнено судити про ознаку генеральної сукупності, яка нас цікавить, необхідно, щоб об’єкти вибірки правильно її представляли. Цю вимогу формулюють так: вибірка повинна бути *репрезентативною* (або *представницькою*). Вважається, що якщо кожний об’єкт вибірки відібраний із генеральної сукупності випадково, тобто всі об’єкти мають однакову ймовірність потрапити у вибірку, і кількість об’єктів, відібраних для спостереження, є досить великою, то вибірка буде репрезентативною. Різниця між показниками вибіркової та генеральної сукупностей становить *помилку репрезентативності.* Ці помилки виникають тому, що вибіркова сукупність неточно відображає генеральну сукупність.

**Означення.** Вибірка називається *репрезентативною*, якщо вона достатньо добре відтворює генеральну сукупність.

Це означення не дозоляє робити конкретні виводи, бо не вказана загальна міра відповідності між репрезентативною вибіркою та генеральною сукупністю, тому питання про репрезентативність необхідно вирішувати у конкретних задачах, спираючись на конкретні критерії відповідності.

**Дискретний варіаційний ряд розподілу**

Нехай із генеральної сукупності проведена вибірка об'єму $n$, і досліджувана випадкова величина в цій вибірці прийняла різні $k$ значень $(x_1, x_2, \ldots, x_k)$ – ці елементи називають *варіантами*:

$$\underbrace{(\underbrace{x_1,\ldots,x_1}_{n_1\ разів},\underbrace{x_2,\ldots,x_2}_{n_2\ разів},\ldots,\underbrace{x_k,\ldots,x_k}_{n_k\ разів})}_{n\ разів}$$

Нехай у вибірці $(x_1, x_2, \ldots, x_k)$ варіанта $x_1$ спостерігалась $n_1$ разів, $x_2$ спостерігалась $n_2$ разів, і так далі відповідно до $x_k$, яка спостерігалась $n_k$ разів, тобто з того, що об'єм вибірки дорівнює $n$, випливає:

$$n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i = n.$$

**Означення.** Значення $n_1, n_2, \ldots, n_k$, тобто чисельність окремої групи згрупованого ряду вибірки, називають *частотами* або *вагами варіант*. Відношення $n_i$ до загального об'єму вибірки $n$, називається *відносною частотою варіанти* $x_i$ та позначається:

$$p_i^* = \frac{n_i}{n}.$$

З означення відносної частоти випливає:

$$p_1^* + p_2^* + \ldots + p_k^* = \sum_{i=1}^{k} p_i^* = 1.$$

**Означення.** Якщо значення випадкової величини, яке відповідає окремій групі згрупованого ряду даних, називається *варіантою*, то змінення цього значення – *варіюванням*.

**Означення.** Розташування вибіркових спостережень значень випадкової

величини в порядку неспадання називається *ранжуванням*.

**Означення.** Різниця між максимальним та мінімальним значеннями варіант, тобто інтервал варіювання, називається *розмахом вибірки* та позначається:

$$R = x_{\max} - x_{\min}.$$

**Означення.** *Дискретним варіаційним рядом розподілу* (або *дискретним статистичним розподілом,* або *розподілом частот*) називається ранжована сукупність варіант $x_i$ з відповідними частотами або відносними частотами.

У *табличній формі* він має такий вигляд:

| $x_i$ | $x_1$ | $x_2$ | $x_3$ | … | $x_k$ |
|---|---|---|---|---|---|
| $n_i$ | $n_1$ | $n_2$ | $n_3$ | … | $n_k$ |
| $p_i^*$ | $p_1^*$ | $p_2^*$ | $p_3^*$ | … | $p_k^*$ |

---

**Приклад.** Проводяться спостереження над значеннями грошових виграшів у миттєвій лотереї. У результаті отримані наступні значення (у тис. грн.):

0, 1, 0, 0, 5, 0, 10, 0, 1, 0, 0, 1, 5, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 5, 0, 5, 0, 0, 1, 1, 1, 5, 10, 0, 1, 1, 0, 5, 0, 0, 0, 0, 1, 0, 1, 0, 5, 0, 0, 0, 0, 1, 0.

Складемо варіаційний ряд розподілу у табличній формі, для цього визначимо які різні значення прийняла досліджувана випадкова величина в цій вибірці, тобто визначимо варіанти, обрахуємо загальний об’єм вибірки $n$, відповідні частоти та відносні частоти варіант:

| $x_i$ | 0 | 1 | 5 | 10 |
|---|---|---|---|---|
| $n_i$ | 31 | 14 | 7 | 2 |
| $p_i^*$ | $\frac{31}{54}$ | $\frac{14}{54}$ | $\frac{7}{54}$ | $\frac{2}{54}$ |

$$n = 31 + 14 + 7 + 2 = 54,$$

$$\sum_{i=1}^{k} p_i^* = 1.$$

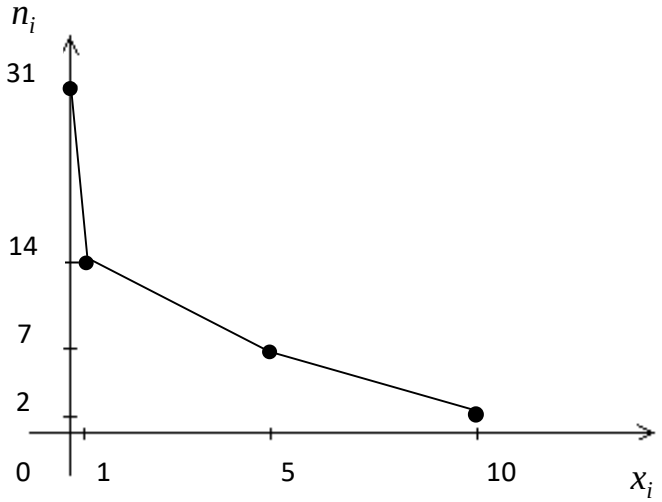
---

Для графічного представлення дискретного варіаційного ряду розподілу будують *полігон (або багатокутник) частот* та *полігон (або багатокутник) відносних частот*.

**Означення.** Ламана лінія, відрізки якої послідовно з'єднують точки з координатами $(x_i, n_i)$, $i = \overline{1,k}$, називається *полігоном частот*, а ламана лінія, відрізки якої послідовно з'єднують точки з координатами $(x_i, p_i^*)$, $i = \overline{1,k}$, називається *полігоном відносних частот*.

---

**Приклад.** Побудуємо полігон частот для попереднього прикладу. Для цього відмітимо на декартовій площині точки з координатами $(x_i, n_i)$, $i = \overline{1,k}$, які у нашому випадку будуть дорівнювати: (0,31), (1,14), (5,7), (10,2), та з'єднаємо їх послідовно відрізками прямих.




---

Зауважимо, що полігон, побудований за дискретним варіаційним рядом, є вибірковим аналогом багатокутника розподілу дискретної випадкової

величини.

---

**Емпірична функція розподілу**

**Означення.** *Вибірковою (емпіричною) функцією розподілу* називається функція $F^*(x)$, яка задає для кожного значення *x* відносну частоту події $X < x$. Отже, за означенням $F^*(x) = \frac{n_x}{n}$, де $n_x$ – число вибіркових значень величини *X*, менших за *x*, а *n* – об'єм вибірки.

Вибіркову функцію можна задати у табличному та у графічному виді. Розберемо це на прикладі.

---

**Приклад.** Побудуємо вибіркову функцію розподілу для розглянутого вище прикладу.

Об'єм вибірки $n = 54$, найменша варіанта дорівнює 0, отже $n_x = 0$ при $x \le 0$, а $F^*(x) = \frac{n_x}{n} = \frac{0}{54}$, при $x \le 0$.

При $0 < x \le 1$ нерівність $X < x$ виконується для варіанти $x_1 = 0$, і відповідно $n_x = n_1 = 31$, а $F^*(x) = \frac{n_x}{n} = \frac{31}{54} = p_1^*$.

При $1 < x \le 5$ нерівність $X < x$ виконується для варіант $x_1 = 0$ та $x_2 = 1$, і відповідно $n_x = n_1 + n_2 = 31 + 14 = 45$, а $F^*(x) = \frac{n_x}{n} = \frac{45}{54} = p_1^* + p_2^*$.

При $5 < x \le 10$ нерівність $X < x$ виконується для варіант $x_1 = 0$, $x_2 = 1$ та $x_3 = 5$, і відповідно $n_x = n_1 + n_2 + n_3 = 31 + 14 + 7 = 52$, а $F^*(x) = \frac{n_x}{n} = \frac{52}{54} = p_1^* + p_2^* + p_3^*$.

При $10 < x$ нерівність $X < x$ виконується для варіант $x_1 = 0$, $x_2 = 1$, $x_3 = 5$

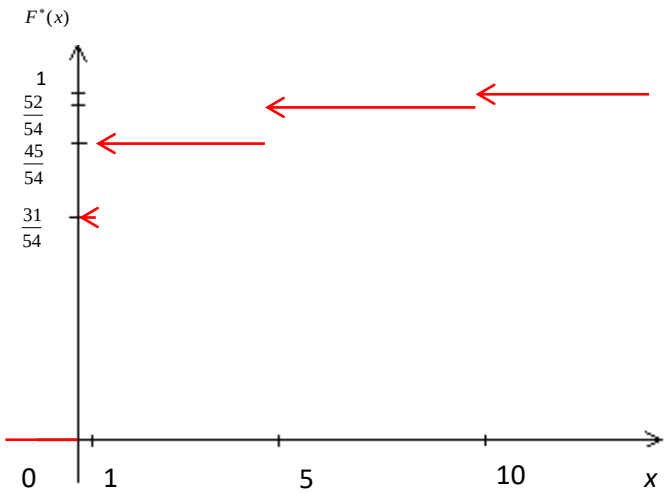
та $x_4 = 10$, і відповідно $n_x = n_1 + n_2 + n_3 + n_4 = 31 + 14 + 7 + 2 = 54$, а $F^*(x) = \frac{n_x}{n} = \frac{54}{54} = 1 = p_1^* + p_2^* + p_3^* + p_4^*$.

Результати обчислень занесемо у таблицю (для зручності).

| $x$ | $F^*(x)$ |
|---|---|
| $x \le 0$ | $0$ |
| $0 < x \le 1$ | $p_1^* = \frac{31}{54}$ |
| $1 < x \le 5$ | $p_1^* + p_2^* = \frac{45}{54}$ |
| $5 < x \le 10$ | $p_1^* + p_2^* + p_3^* = \frac{52}{54}$ |
| $10 < x$ | $p_1^* + p_2^* + p_3^* + p_4^* = 1$ |

Можна замість таблиці записувати вибіркову функцію розподілу так само, як записували функцію розподілу дискретної випадкової величини.

Побудуємо графік вибіркової функції розподілу.



**Інтервальний варіаційний ряд**

Якщо досліджувана випадкова величина неперервна або дискретна величина така, що число її можливих значень достатньо велике, то для побудови варіаційного ряду використовують *інтервальний ряд розподілу*. У цьому випадку весь інтервал варіювання розбивають на декілька частинних інтервалів однакової довжини (яку називають *кроком*) та підраховують частоту попадання значень випадкової величини у кожний частинний інтервал.

**Означення.** *Інтервальним варіаційним рядом* (або *інтервальним розподілом частот*) називається упорядкована послідовність інтервалів варіювання випадкової величини із відповідними частотами $n_i$ або відносними частотами $p_i^*$ попадання у кожен з них значень випадкової величини.

У *табличній формі* він має такий вигляд:

| $(x_i, x_{i+1}]$ | $[x_1, x_2]$ | $(x_2, x_3]$ | $(x_3, x_4]$ | … | $(x_k, x_{k+1}]$ |
|---|---|---|---|---|---|
| $n_i$ | $n_1$ | $n_2$ | $n_3$ | … | $n_k$ |
| $p_i^*$ | $p_1^*$ | $p_2^*$ | $p_3^*$ | … | $p_k^*$ |

**Зауваження.** Часто замість інтервалів $(x_i, x_{i+1}]$ записують $[x_i, x_{i+1})$.

Для кожної випадкової величини та для кожного об'єму вибірки необхідно визначати оптимальне число частинних інтервалів. Існують спеціальні формули для визначення цього числа. В задачах це може бути вказано у формулюванні, але часто оптимальну кількість інтервалів визначають за формулою Стерджесса:

$$k = 1 + 3{,}322 \cdot \lg n,$$

де $n$ – обсяг вибірки, $k$ – кількість частинних інтервалів.

Кількість інтервалів, що обчислюється за формулою Стерджесса, округляється до цілого числа. Залежно від кількості вимірювань це число знаходиться в межах від 8 до 12. Для визначення довжини інтервалу, тобто

кроку $h$, необхідно розмах $R$ поділити на кількість інтервалів $k$:

$$h = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg n}.$$

Значення, яке отримують за цією формулою, округлюють до найбільш зручного, якщо це необхідно (часто округлюють до меншого цілого). У результаті округлення шагу нижньою границею першого інтервалу, тобто $x_1$, можна зсунути відносно $x_{\min}$ у бік менших значень, а верхню границю останнього інтервалу відповідно у бік більших значень, але цей зсув не повинен перевищувати $\frac{h}{2}$.

---

Для графічного представлення інтервального варіаційного ряду розподілу будують *гістограму частот* або *відносних частот*. Для її побудови у прямокутній системі координат на осі $Ox$ відкладають відрізки частинних інтервалів варіювання та на цих відрізках як на основах будують прямокутники з висотами, які дорівнюють частотам або відносним частотам відповідних інтервалів.

Для інтервального варіаційного ряду розподілу, як і для дискретного, можна побудувати *полігон частот*, але для цього таблицю треба доповнити ще одним рядком, в якому записати відповідні для кожного інтервалу $(x_i, x_{i+1}]$ точки $y_i$, які є їх серединами:

$$y_i = \frac{x_{i+1} + x_i}{2}, \quad i = \overline{1,k}.$$

| $(x_i, x_{i+1}]$ | $[x_1, x_2]$ | $(x_2, x_3]$ | $(x_3, x_4]$ | … | $(x_k, x_{k+1}]$ |
|---|---|---|---|---|---|
| $y_i$ | $y_1$ | $y_2$ | $y_3$ | | $y_k$ |
| $n_i$ | $n_1$ | $n_2$ | $n_3$ | … | $n_k$ |
| $p_i^*$ | $p_1^*$ | $p_2^*$ | $p_3^*$ | … | $p_k^*$ |

**Означення.** Ламана лінія, відрізки якої послідовно з'єднують точки з координатами $(y_i, n_i)$, $i = \overline{1,k}$, називається *полігоном частот*, а ламана лінія, відрізки якої послідовно з'єднують точки з координатами $(y_i, p_i^*)$, $i = \overline{1,k}$, називається *полігоном відносних частот.*

**Означення.** Якщо відносну частоту розділити на довжину інтервалу, то отримана величина буде *вибірковою оцінкою щільності ймовірності:*

$$f^*(x_i) = \frac{p_i^*}{h}.$$

*Емпірична функція розподілу* неперервної випадкової величини, тобто у випадку інтервального розподілу частот, будується дещо інакше, ніж для дискретної випадкової величини – використовуючи інтервальний варіаційний ряд з відносними частотами $p_i^*$, значення функції $F^*(x)$ обчислюються у кінцевих точках інтервалів у вигляді «накопичувальної відносної частоти». Після чого точки $(x_i, F^*(x_i))$ відмічаємо на прямокутній системі координат та з'єднуємо ці точки відрізками прямих.

---

**Приклад.** При вимірюванні зросту 105 студентів (хлопців) отримана вибірка результатів:

155, 170, 185, 180, 188, 152, 173, 178, 178, 168, 185,
173, 170, 183, 175, 173, 170, 183, 175, 180, 175, 193,
178, 183, 180, 197, 178, 181, 187, 168, 174, 179, 184,
183, 178, 180, 178, 163, 166, 178, 175, 182, 190, 167,
170, 178, 183, 170, 178, 181, 173, 168, 185, 175, 170,
155, 169, 186, 179, 189, 155, 174, 179, 179, 169, 186,
174, 171, 184, 175, 193, 178, 184, 180, 196, 175, 181,
188, 168, 179, 178, 183, 184, 178, 181, 177, 163, 166,
178, 175, 183, 190, 167, 170, 178, 183, 170, 178, 182,

173, 168, 186, 176, 171, 188.

Складемо інтервальний варіаційний ряд. Знайдемо мінімальне та максимальне значення випадкової величини: $x_{\min}=152$, $x_{\max}=197$. Тобто розмах – інтервал варіювання дорівнює $R=x_{\max}-x_{\min}=197-152=45$. Нехай за умовами задане число частинних інтервалів, яке дорівнює 10. Зсунувши нижню границю з 152 до 150, а верхню з 197 до 200, отримаємо довжину частинного інтервалу, яка дорівнює $\frac{200-150}{10}=5$. Складемо інтервальний варіаційний ряд (запишемо його у зручному для подальших обчислень в *MS Excel* вигляді).

| Номер інтервалу $i$ | Інтервал $(x_i, x_{i+1}]$ | Середина інтервалу $y_i$ | Частота $n_i$ | Відносна частота $p_i^*$ |
|---|---|---|---|---|
| 1 | (150,155] | 152,5 | 4 | 0,0380952 |
| 2 | (155,160] | 157,5 | 0 | 0 |
| 3 | (160,165] | 162,5 | 1 | 0,0095238 |
| 4 | (165,170] | 167,5 | 19 | 0,1809524 |
| 5 | (170,175] | 172,5 | 18 | 0,1714286 |
| 6 | (175,180] | 177,5 | 27 | 0,2571429 |
| 7 | (180,185] | 182,5 | 22 | 0,2095238 |
| 8 | (185,190] | 187,5 | 10 | 0,0952381 |
| 9 | (190,195] | 192,5 | 2 | 0,0190476 |
| 10 | (195,200] | 197,5 | 2 | 0,0190476 |

Для графічного представлення інтервального варіаційного ряду розподілу побудуємо *гістограму частот* (проілюструємо цей приклад у подальшому на лабораторному занятті).