Лекція 1.

Предмет і задачі математичної статистики

Математична статистика виникла в XVII ст. і розвивалася паралельно з теорією ймовірностей. Великий внесок у розвиток математичної статистики внесли російські вчені в XIX в. – початку XX ст.: в першу чергу П.Л. Чебишев, А.А. Марков, А.М. Ляпунов, а також вчені інших країн - К. Гаусс, К. Пірсон, Ф. Гальтон і т.д.

У XX ст. істотний внесок у розвиток математичної статистики був зроблений радянськими математиками, зокрема, О.М. Колмогоровим, В.І. Романовським, Е.Е. Слуцьким, Н. В. Смирновим, а також англійськими – Стьюдентом, Р. Фішером, Е. Пірсоном і американськими вченими – Ю. Нейманом, А. Вальдом і ін.

До початку XX ст. теорія ймовірностей та математична статистика відносились до фізичних дисциплін, а після закладення аксіоматичного фундаменту О.М. Колмогоровим стала повноправною математичною дисципліно.

Сьогодні математична статистика є основою таких напрямків наукових досліджень як науковий експеримент, теорія надійності і т.д. та знайшла своє застосування у фізиці, медицині, соціології та ін..

Теорія ймовірностей вивчає математичні моделі випадкових явищ, при цьому сама математична модель вважається заданою. В задачах теорії ймовірностей виходять з того, що задано ймовірнісний простір, множину елементарних фіналів і ймовірність будь-якої події.

Так, наприклад, якщо вивчається деяка випадкова подія А, то відомо Р(А). Якщо ж мова йде про випадкову величину Х, то відомий закон розподілу ймовірностей в будь-якій формі і, як наслідок, числові характеристики досліджуваної випадкової величини.

У практичних завданнях ці характеристики, як правило, невідомі, але є деякі експериментальні дані про подію або випадкову величину. Потрібно на підставі цих даних побудувати відповідну вірогідну модель досліджуваного явища, тобто приблизно оцінити невідомий закон розподілу і числові характеристики досліджуваної випадкової величини на основі експериментальних даних. Це і є завданням математичної статистики. У математичній статистиці єдиний об'єкт це дані експерименту. Результати експерименту виражаються значеннями деякої випадкової величини.

У теорії ймовірностей імовірнісний простір задано і потрібно передбачити можливу поведінку випадкової величини. У математичній статистиці навпаки, відомі лише результати (значення випадкової величини), за якими відновлюється імовірнісний простір. За експериментальними даними будується імовірнісна модель явища, що відповідає цим даним, тобто інтерпретація даних.

Математичну статистику визначають як науку про методи отримання та обробки результатів спостережень (вимірювань) для виявлення закономірностей у масових випадкових явищах.

Перше завдання математичної статистики: вказати способи збору і угруповання статистичних даних, отриманих в результаті експериментів.

Друге завдання математичної статистики: розробити методи аналізу статистичних даних.

До другої задачі відносяться:

- Оцінка невідомих параметрів (ймовірності події, функції розподілу і її параметрів і т.д.) з побудовою довірчих інтервалів (методи оцінювання).
- Перевірка статистичних гіпотез про вид невідомого розподілу і параметрів розподілу (методи перевірки гіпотез).

При цьому вирішуються такі в порядку складності і важливості завдання:

Опис явищ, тобто, впорядкування отриманого статистичного матеріалу, подання його в найбільш зручному для огляду і аналізу вигляді (таблиці, графіки).

Аналіз і прогноз, тобто наближена оцінка характеристик на підставі статистичних даних. Наприклад, наближена оцінка математичного очікування і дисперсії випадкової величини і визначення похибок цих оцінок.

Вироблення оптимальних рішень. Наприклад, визначення числа дослідів n, достатнього для того, щоб помилка від заміни теоретичних числових характеристик їх експериментальними оцінками не перевищувала заданого значення. У зв'язку з цим виникає завдання перевірки правдоподібності гіпотез про параметри розподілу і про закони розподілу випадкової величини, рішенням якої є можливість зробити один з висновків:

- відкинути гіпотезу, як таку, що суперечить дослідним даним;
- прийняти гіпотезу, вважати її прийнятною.

Теоретичним базисом математичної статистики э теорія ймовірностей, зокрема, особливу роль відіграють граничні теореми.

Основні поняття математичної статистики

Генеральна сукупність –сукупність однорідних елементів, що підлягає вивченню і характеризується деякою ознакою. Наприклад, нас цікавить поширеність даного захворювання в певному регіоні, тоді генеральна сукупність, це все населення регіону. Якщо необхідно виділити чоловіків і жінок окремо по цьому захворюванню, то отримуємо 2 генеральні сукупності.

Кількість об'єктів, що входять в генеральну сукупність називається об'ємом генеральної сукупності (N)

Генеральну сукупність можна вивчати за деякою її частини.

Вибіркова сукупність (вибірка) – частина генеральної сукупності, що обирається, для статистичної обробки. Обсяг вибірки n. Властивості об'єктів вибірки повинні відповідати властивостям генеральної сукупності.

Результати дослідження деякого ознаки генеральної сукупності, будуть більш надійні, якщо вибірку утворювати випадковим чином. Елементи вибірки беруться навмання. Кожен об'єкт може потрапити до вибірки з однаковою ймовірністю. Головним питанням є: як визначити обсяг вибірки, необхідної для отримання достовірного результату.

Основними властивостями вибірки є репрезентативність і достовірність.

Репрезентативність вибірки – властивість вибірки, яке характеризує її представництво: здатність вибірки представляти досліджувані явища досить повно з точки зору їх мінливості в генеральної сукупності. Іншими словами,

репрезентативна вибірка представляє собою меншу за розміром, але точну модель тієї генеральної сукупності, яку вона повинна відображати. Досягненню репрезентативності може сприяти така організація експерименту, при якій елементи вибірки витягуються з генеральної сукупності випадковим чином. Правда, на практиці, ця вимога не завжди може дотримуватися в силу певних причин. Наприклад, при з'ясуванні думки з якогось питання у студентів вузу, вибірка, складена тільки зі студентів 1 курсу не буде репрезентативною.

Статистична достовірність вибірки (статистична значимість) результатів дослідження визначається за допомогою методів математичної статистики.

Помилка вибірки (довірчий інтервал) – відхилення результатів, отриманих за допомогою вибіркового спостереження від справжніх даних генеральної сукупності. Помилка вибірки буває двох видів – статистична і систематична. Статистична похибка залежить від розміру вибірки. Чим більше розмір вибірки, тим вона нижче.

Приклад:

Для простої випадкової вибірки розміром 400 одиниць максимальна статистична помилка (з 95% довірчою ймовірністю) становить 5%, для вибірки в 600 одиниць - 4%, для вибірки в 1100 одиниць - 3% Зазвичай, коли говорять про помилку вибірки, мають на увазі саме статистичну помилку .

Систематична помилка залежить від різних факторів, що постійно впливають на дослідження і зміщують результати дослідження в певну сторону.

Приклад:

Використання будь-яких імовірнісних вибірок занижує частку людей з високим доходом, що ведуть активний спосіб життя. Відбувається це в силу того, що таких людей набагато складніше застати в якомусь певному місці (наприклад, будинки).

Проблема респондентів, які відмовляються відповідати на питання анкети (частка «відмовників», для різних опитувань, коливається від 50% до 80%)

У деяких випадках, коли відомй істинний розподіл, систематичну помилку можна нівелювати введенням квот або перезважуванню даних, але в більшості реальних досліджень навіть оцінити її буває досить проблематично.

Способи отримання вибірки

Вибірки діляться на два типи: імовірнісні, неімовірнісні

1. Імовірнісні вибірки

1.1 *Випадкова вибірка* (простий випадковий відбір). Така вибірка передбачає однорідність генеральної сукупності, однакову ймовірність доступності всіх елементів, наявність повного списку всіх елементів. При відборі елементів, як правило, використовується таблиця випадкових чисел.

1.2 *Механічна (систематична) вибірка*. Різновид випадкової вибірки, впорядкована за будь-якою ознакою (алфавітний порядок, номер телефону, дата народження і т.д.). Перший елемент відбирається випадково, потім, з кроком 'n' відбирається кожен 'k'-ий елемент. Розмір генеральної сукупності, при цьому - N = n * k

1.3 *Стратифікована (районована)*.Застосовується в разі неоднорідності генеральної сукупності. Генеральна сукупність розбивається на групи (страти). У кожній страті відбір здійснюється випадковим або механічним чином.

1.4 *Серійна (гніздова або кластерниа)* вибірка. При серійної вибірці одиницями відбору виступають не самі об'єкти, а групи (кластери або гнізда). Групи відбираються випадковим чином. Об'єкти всередині груп обстежуються суціль.

2. Неімовірнісні вибірки

Відбір в такій вибірці здійснюється не за принципами випадковості, а за суб'єктивними критеріями – доступності, типовості, рівного представництва і т.д.

2.1. *Квотна вибірка*. Спочатку виділяється певна кількість груп об'єктів (наприклад, чоловіки у віці 20-30 років, 31-45 років та 46-60 років) Для кожної групи задається кількість об'єктів, які повинні бути обстежені. Кількість об'єктів, які повинні потрапити в кожну з групу, задається, найчастіше, або пропорційно заздалегідь відомої частці групи в генеральній сукупності, або однаковим для кожної групи. Усередині груп об'єкти відбираються довільно. Квотні вибірки використовуються в маркетингових дослідженнях досить часто.

2.2. *Метод сніжної грудки*. Вибірка будується наступним чином. У кожного респондента, починаючи з першого, просяться контакти його друзів, колег, знайомих, які підходили б під умови відбору і могли б взяти участь в дослідженні. Таким чином, за винятком першого кроку, вибірка формується за участю самих об'єктів дослідження. Метод часто застосовується, коли необхідно знайти і опитати важкодоступні групи респондентів (наприклад, респондентів, які мають високий дохід, респондентів, які належать до однієї професійної групи, респондентів, що мають будь-які схожі хобі/захоплення і т.д.)

2.3 *Стихійна вибірка*. Опитуються найбільш доступні респонденти. Типові приклади стихійних вибірок – опитування в газетах/журналах, анкети, віддані респондентам на самозаповнення, більшість інтернет-опитувань. Розмір і склад стихійних вибірок заздалегідь не відомий, і визначається тільки одним параметром - активністю респондентів.

2.4 *Вибірка типових випадків.* Відбираються одиниці генеральної сукупності, що володіють середнім (типовим) значенням ознаки. При цьому виникає проблема вибору ознаки і визначення його типового значення.

Повторна вибірка (з поверненням), коли випадково відібраний і вже обстежений об'єкт, повертається в загальну сукупність і теоретично може бути повторно відібраний;

Бесповторна вибірка, коли відібраний елемент не повертається в загальну сукупність.

На практиці користуються поєднанням вищевказаних способів і видів відбору.

Якщо для кожного об'єкта в вибірці виміряється значення однієї змінної (параметра), вибірка називаються одновимірними. Якщо ж для кожного об'єкта реєструються значення двох або кількох змінних (параметрів), такі вибірки називаються багатовимірними.

Нехай досліджується деякий об’єкт, що характеризується певною ознакою. Ознаки можуть бути якісними і кількісними.

Якісні ознаки описуються словами, наприклад масть чорна, червона, руда, чала і тд. Якщо є два взаємовиключних варіанти, то такі якісні ознаки називають альтернативними, наприклад стать-чоловіча і жіноча.

Кількісні ознаки змінюються, підраховуються і виражаються цифрами, Вони діляться на неперервні (маса тіла) і дискретні (кількість волосся)

Ознака, що змінює числове значення від одного об'єкта до іншого називається варіюємою. Якщо кількісна ознака лежить в певному інтервалі, її називають інтервальною.

Зрозуміло, що кількісні ознаки є більш інформативними, їх можна порівнювати на відміну від якісних.

Частота – кількість об'єктів з конкретним числовим значенням ознаки: n_i.

Відносна частота – частка з даними значенням ознаки ($w_i=n_i/N$), де N – об’єм сукупності. Відносна частота є статистичною оцінкою теоретичної ймовірності.

Варіаційний ряд – це пара, яка складається із значення випадкової величини (варіанти) та відповідної частоти або відносної частоти, причому значення випадкової величини впорядковані за зростанням:

Варіаційний ряд частот							
x_i	35	36	37	38	39	40	41
n_i (p_i)	2	4	5	6	7	5	1

Об'єм вибірки $N=\sum_{i=1}^{7} n_i = 30$

Варіаційний ряд відносних частот							
x_i	35	36	37	38	39	40	41
w	0,0(6)	0,1(3)	0,1(6)	0,2	0,2(3)	0,2	0,0(3)

Варіаційний ряд може задавати значення випадкової величини за інтервальною шкалою. Для переходу до абсолютної шкали в якості значення береться середина інтервалу.

Навпаки для переходу до інтервальної шкали увесь діапазон значень розбивається на інтервали і в якості частоти береться загальна кількість значень, що потрапили у відповідний інтервал.

Приклад. Нехай задано варіаційний ряд за інтервальною шкалою

x_i	2-5	5-8	8-11	11-14	14-17
n_i (p_i)	9	10	25	6	4

Перейдемо до абсолютної шкали

x_i	3,5	6,5	9,5	12,5	15,5
n_i (p_i)	9	10	25	6	4

Для полегшення обчислень, часто, переходять до умовних варіант:

$$u_i = (x_i - C)/h,$$

де С – хибний нуль (новий початок відліку, найчастіше вибирається як середина інтервалу значень варіант), h – крок або різниця між сусідніми варіантами.

Якщо крок постійний (варіанти рівновіддалені), такий перехід дозволяє отримати цілочисельний симетричний варіаційний ряд.

Приклад. Нехай задана вибіркова сукупність об'ємом N=100

x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i
1,00	1	1,10	4	1,20	4	1,30	6	1,39	1	1,49	4	1,25	8	1,33	5
1,03	3	1,12	3	1,23	4	1,32	4	1,40	2	1,50	2	1,26	4		
1,05	6	1,15	6	1,25	8	1,33	5	1,44	3	1,19	2	1,29	4		
1,06	4	1,16	5	1,26	4	1,37	6	1,45	3	1,20	4	1,30	6		
1,08	2	1,19	2	1,29	4	138	2	1,46	2	1,23	4	1,32	4		

$x_{\min}=1{,}00,\ \ x_{\max}=1{,}50$

Розібємо весь відрізок значень на 5 часткових інтервалів і в якості нових варіант візьмемо середини часткових відрізків. Підрахуємо відповідні частоти:

y_i	1,05	1,15	1,25	1,35	1,45
n_i	18	20	25	22	15

В якості умовного нуля візьмемо середнє значення C=1,25, h=0,1. Перейдемо до умовних варіант:

u_i	-2	-1	0	1	2
n_i	18	20	25	22	15

Графічне представлення варіаційного ряду

Для наочного представлення варіаційного ряду застосовують спеціальні графіки. Якщо варіаційний ряд заданий за абсолютною шкалою будується полігон, якщо за інтервальною – будують гістограму.

Полігоном частот називають ламану, відрізки якої з'єднують точки з координатами (x_i, n_i);

Полігоном умовних частот називають ламану, відрізки якої з'єднують точки з координатами $\left(x_i, {n_i}/{N}\right)$

Гістограмою частот (умовних частот) називають ступінчасту фігуру, що складається з прямокутників, основи яких розташовані на осі x і довжини їх дорівнюють довжинам часткових інтервалів h, а висоти дорівнюють відношенню:

${n_i}/{h}$ – для частот ${n_i}/{Nh}$ – для умовних частот.

Площа гістограми частот дорівнює n, а умовних частот – 1.

Можна побудувати полігон для інтервального ряду, якщо його перетворити в дискретний ряд.

Приклад. Дана вибірка значень випадкової величини об'ємом 20:

12, 14, 19, 15, 14, 18, 13, 16, 17, 12, 18, 17, 15, 13, 17, 14, 14, 13, 14, 16

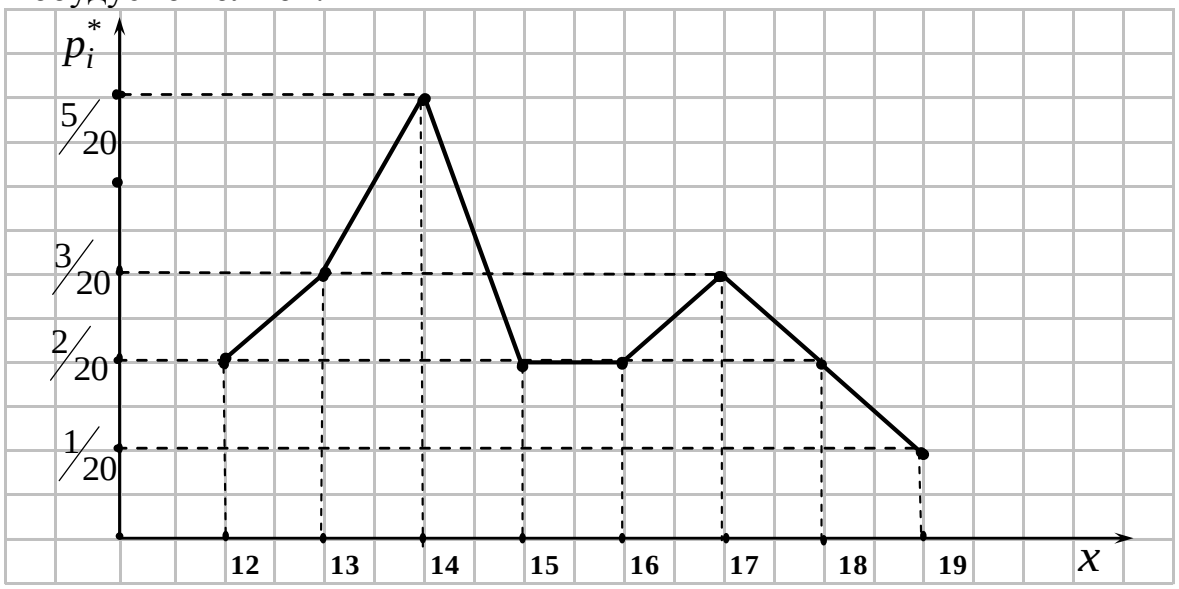
Побудувати полігон.

Спочатку побудуємо варіаційний ряд та обчислимо умовні частоти

Значення варіант x_i	12	13	14	15	16	17	18	19

Частоти n_i	2	3	5	2	2	3	2	1	$\sum_{i=1}^{8} n_i = 20$
Умовні частоти $p_i^* = \frac{n_i}{n}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{5}{20}$	$\frac{2}{20}$	$\frac{2}{20}$	$\frac{3}{20}$	$\frac{2}{20}$	$\frac{1}{20}$	$\sum_{i=1}^{8} p_i = 1$

Побудуємо полігон:



Приклад. Результати вимірювань відхилень від норми діаметрів 50 підшипників дали чисельні значення (в мкм), наведені в табл.

-1,760	-0,291	-0,110	-0,450	0,512
-0,158	1,701	0,634	0,720	0,490
1,531	-0,433	1,409	1,740	-0,266
-0,058	0,248	-0,095	-1,488	-0,361
0,415	-1,382	0,129	-0,361	-0,087
-0,329	0,086	0,130	-0,244	-0,882
0,318	-1,087	0,899	1,028	-1,304
0,349	-0,293	0,105	-0,056	0,757
-0,059	-0,539	-0,078	0,229	0,194
0,123	0,318	0,367	-0,992	0,529

Для даної вибірки:

- побудувати інтервальний варіаційний ряд;
- побудувати гістограму та полігон умовних частот.

$x_{\min} = -1{,}76$; $x_{\max} = 1{,}74$,

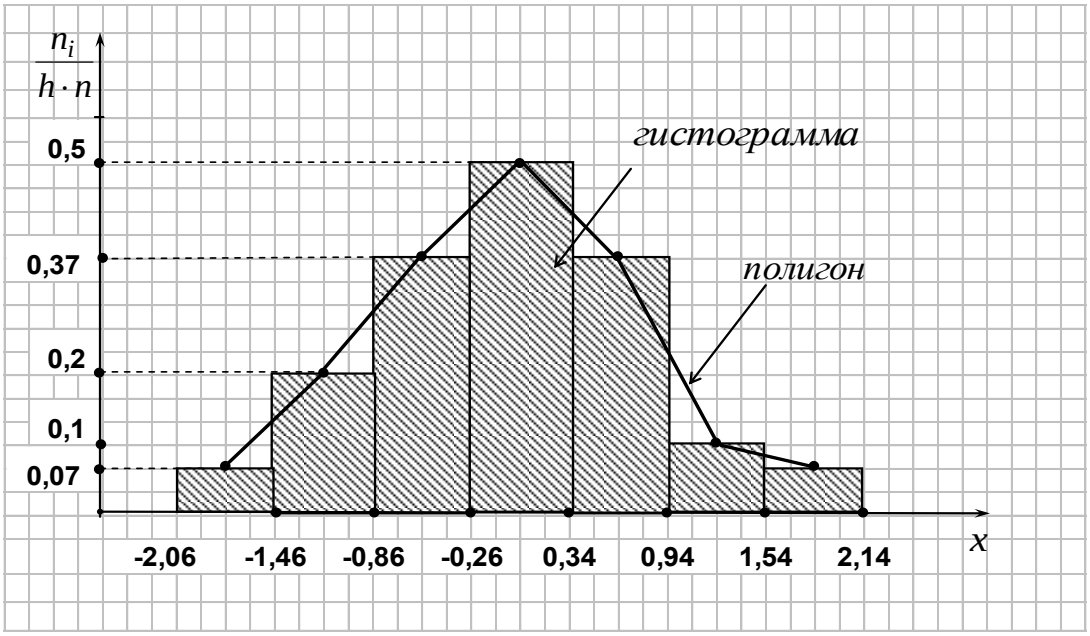
Для визначення довжини інтервалу використовуємо формулу Стерджеса:

$$h = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \lg 50} \approx 0{,}526$$

Приймемо h=0,6, тоді кількість інтервалів m=7

Інтервали	$[-2{,}06; -1{,}46)$	$[-1{,}46; -0{,}86)$	$[-0{,}86; -0{,}26)$	$[-0{,}26; 0{,}34)$

Частоти n_i	2	6	11	15
Умовні частоти p_i	$\frac{2}{50}$	$\frac{6}{50}$	$\frac{11}{50}$	$\frac{15}{50}$
Інтервали	$[0{,}34;\ 0{,}94)$	$[0{,}94;1{,}54)$	$[1{,}54;\ 2{,}14)$	
Частоти n_i	11	3	2	$\sum_{i=1}^{7} n_i = 50;$
Умовні частоти p_i	$\frac{11}{50}$	$\frac{3}{50}$	$\frac{2}{50}$	$\sum_{i=1}^{7} p_i = 1.$



Кумулятою називається крива накопичених частот, якає ламаною лінією, що сполучає точки ($x_i, n_{накопичене}$)

Емпірична функція розподілу

Будь яка випадкова величина описується своїм законом розподілу. У статистиці, так як значення випадкової величини визначається дослідним шляхом, для опису закону розподілу випадкової величини застосовують емпіричну функцію розподілу.

Емпіричної функцією розподілу, що відповідає вибірці $(x_1, x_2, ..., x_n)$, називається функція $F^*(x)$, яка для кожного значення *x* визначає відносну частоту події $X < x$: $F^*(x) = \frac{n_x}{N}$, де n_x – кількість спостережень $X < x$, N – об'єм вибірки

Принципова відмінність емпіричної функції розподілу $F^*(x)$ від звичайної функції розподілу $F(x)$ полягає в тому, що вона може змінюватися від вибірки до вибірки і притому випадковим чином. Найважливішою властивістю емпіричної функції розподілу як випадкової функції є те, що вона при збільшенні обсягу виборкі наближається (в сенсі збіжності за ймовірністю) до істинної функції розподілу. Тому кажуть, що емпірична функція розподілу є статистичним

аналогом (оцінкою) невідомої функції розподілу, яку називають при цьому теоретичною, а отже має всі її властивості:

1. $F^*(x) \in [0,1]$;
2. $F^*(x)$ – неспадна функція;
3. $F^*(x) = 0$ дляв всіх значень менших від найменшого значення випадкової величини і $F^*(x) = 1$ для всіх значень більших за максимальне значення випадкової величини.

Графіком емпіричної функції розподілу ознаки являється розривна ступінчаста фігура, неперервна зліва, рівна нулю лівіше найменшого значення ознаки, що спостерігається, дорівнює одиниці правіше найбільшого значення. У точках можливих значень графік емпіричної функції розподілу має розриви першого роду, величина стрибка в цих точках дорівнює відносній частоті відповідного значення.

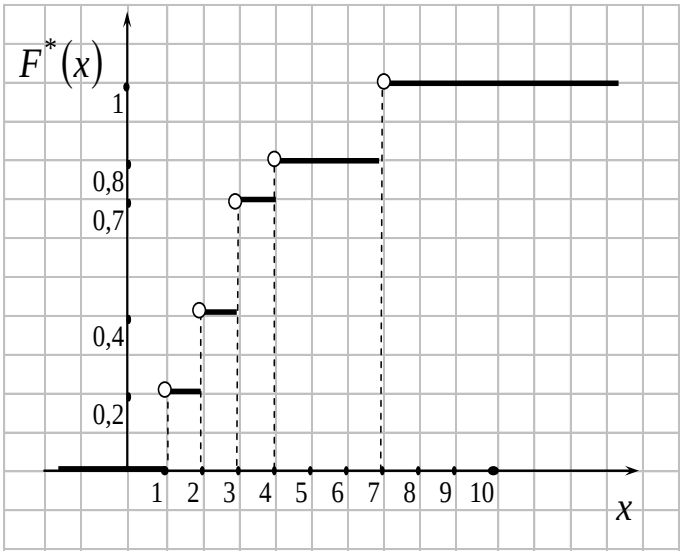
Приклад. За заданим варіаційним рядом побудувати графік функції розподілу

x_i	0	2	3	4	7
n_i	2	2	3	1	2
$\frac{n_i}{n}$	$\frac{2}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$

$$\sum_{i=1}^{5} n_i = 10;$$

$$\sum_{i=1}^{5} p_i^* = 1.$$

$$F^*(x) = \begin{cases} 0 & x \le 1; \\ 0{,}2 & 1 < x \le 2; \\ 0{,}4 & 2 < x \le 3; \\ 0{,}7 & 3 < x \le 4; \\ 0{,}8 & 4 < x \le 7; \\ 1 & x > 7; \end{cases}$$



Лінія, огинаюча графік функції розподілу буде кумулятою.

Емпірична щільність розподілу

Для інтегральної функції розподілу справедлива наближене рівність: $F(x+\Delta) - F(x) \approx f(x) \cdot \Delta x$, де $f(x)$ – диференціальна функція розподілу (функція щільності ймовірності).

Тому природно вибірковим аналогом функції $f(x)$ вважати функцію:

$$f^*(x)=\frac{F^*(x+\Delta x)-F^*(x)}{\Delta x}$$

де $F^*(x+\Delta x)-F^*(x)$ – умовні частоти попадання спостережуваних значень випадкової величини *X* в інтервал $[x;\, x+\Delta x)$. Таким чином, значення $f^*(x)$ характеризує щільність частості на цьому інтервалі.

$$f^*(x)=\begin{cases}0\ ,при\ \ x<a_1,\\ \dfrac{p_i^*}{h}\ ,при\ \ a_i\le x\le a_{i+1},\ \ i=1,2,\ldots m,\\ 0\ ,при\ \ x>a_{m+1},\end{cases}$$

де $[a_i\,;\, a_i+h)$– частинні інтервали