Розділ 2. Закономірність і випадковість

Шкали вимірювання

Обробити статистичними методами можна тільки те, що піддається вимірюванню. В зв'язку з цим необхідно розглянути існуючі шкали вимірювання.

Вимірювання – надання числових значень предметам або подіям (процесам), що основані на деякій системі правил. Необхідно, щоб для величин, які є результатами вимірювання властивостей (ознак), що вивчаються, виконувалися такі умови.

Тотожність

1. Або A=B або A≠B
2. Якщо A=B, то B=A.

Транзитивність

Якщо A=B і B=C, то A=C

Ранговий порядок

1. Якщо A>B, то B<A
2. Якщо A>B і B>C, то A>C

Аддитивність

1. Якщо A=B і C > 0, то A+C>B
2. A+B=B+A
3. Якщо A=B і C=D, то A+C=B+D
4. (A+B)+C=A+(D+C).

Залежно від виконання цих умов, а також операцій над вимірюваними величинами (дорівнює, не дорівнює, більше, менше, додавання, віднімання, множення і ділення) існують такі шкали вимірювання:

– шкала класифікацій (найменувань);

– шкала порядку;

– шкала інтервалів;

– шкала відношень.

Шкала класифікацій (найменування, номінальна). Ніякі операції порівняння, окрім дорівнює або не дорівнює, неможливі. Нумерація або найменування служить тільки для ідентифікації об'єкту (номер лікувальної картки, номер методики лікування тощо).

Шкала порядку. Можливі порівняння об'єктів за величиною: більше або менше. Інші операції тут не можливі. Прикладом є шкала твердості матеріалів, що містить еталонні матеріали, які розташовані в порядку їх твердості. У медицині може служити: ступінь тяжкості захворювання; стан здоров'я – ”добрий“, ”задовільний“, ”поганий“; стадії розвитку захворювання тощо. Тут можливі операції порівняння: більше, менше, дорівнює.

Шкала інтервалів – можливе порівняння не лише за величиною, але і визначення на скільки більше (тобто можливі операції додавання і віднімання). Прикладом можуть бути шкали вимірювання температури (за Цельсієм, Кельвіном, Фаренгейтом, Реомюром).

Шкала відношень – тут можливе з'ясування питання ”в скільки разів“ (тобто припустимі всі операції: порівняння, додавання, віднімання, множення і ділення). Наприклад, вага, довжина тощо. В цих випадках існує природна шкала відліку.

В процесі розвитку науки і засобів вимірювання можливий перехід від однієї шкали до іншої, більш досконалої. Інколи говорять про безперервні і дискретні шкали вимірювання. До дискретних відносяться – шкали класифікації і шкали порядку. У цих шкалах не існує проміжних значень, їх називають некількісними. Безперервні шкали характерні для інтервалів і відношень.

Можливі операції в різних шкалах вимірів.

Назва шкали	Вид шкали	Можливі операції
Класифікації	Дискретна	= ≠
Порядку	Дискретна	= ≠ > <
Інтервалів	Безперервна	= ≠ > < + -
Відношень	Безперервна	= ≠ > < + – / *

Статистичні характеристики, які можна обчислити

Назва шкали	Статистичні характеристики, які можна обчислити
Класифікацій	Частоти, модальний клас
Порядку	Частоти, мода, медіана, центилі, рангова корекція
Інтервалів	Частоти, мода, медіана, центилі, рангова корекція, середнє, дисперсія
Відношень	Всі наявні

Зв'язок шкал вимірів і використовуваних методів

Шкала виміру змінних, що впливають	Шкала вимірів залежних змінних	Вживані методи
Інтервалів або відношень	Інтервалів або відношень	Регресійний і кореляційний аналіз
Час	Інтервалів або відношень	Аналіз часових рядів
Найменування або порядку	Інтервалів або відношень	Дисперсійний аналіз
Змішана	Інтервалів або відношень	Коваріаційний і регресійний аналіз
Найменування або порядку	Найменувань або відношень	Аналіз рангових кореляцій і таблиць зв'язаності
Найменування або порядку	Інтервалів або відношень	Дискримінантний аналіз Кластерний аналіз

Випадкові величини

Величини, точне значення яких невідоме, називаються випадковими.

Незважаючи на випадковий характер величин, при їх дослідженні можливе знаходження певних закономірностей, які і використовуються в практичній діяльності.

Для дослідження закономірностей, що проявляють себе через випадковість, визначають закони розподілу випадкових величин і їх числові значення.

Ймовірність (вірогідність) p – це відношення кількості сприятливих можливостей до загальної їх кількості (класичне визначення). Змінюється p від 0 до 1: 0 – подія неможлива. 1 – подія достовірна. $0<p<1$ подія може статися, а може і не статися.

Оскільки в більшості випадків теоретичне визначення ймовірності (вірогідності) неможливе, оскільки не можна розрахувати кількість загальних (спільних) і сприятливих можливостей, вводиться статистичне визначення вірогідності. За цим визначенням p дорівнює відношенню кількості випадків, в яких подія спостерігається, до загальної кількості спостережень. Така вірогідність ще називається відносною частотою. Відносна частота ніколи не збігається з теоретичною вірогідністю.

Але, згідно теоремі Бернуллі, при досить великому числі випробувань ймовірність того, що відхилення відносної частоти від теоретичної вірогідності буде скільки завгодно малим, прагне до одиниці:

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1.$$

Випадковою величиною (ВВ) називається величина, яка в результаті експерименту може набувати невідоме заздалегідь значення.

Дискретною випадковою величиною називається величина, яка набуває окремих значень (наприклад кількість дітей, що народилися).

Безперервною випадковою величиною називається величина, можливі значення якої безперервно заповнюють який-небудь інтервал (наприклад, маса тіла, зріст тощо).

Незалежні випадкові величини – величини, які є результатом незалежних випадкових подій. Тобто таких подій, для яких настання однієї події ніяк не впливає на вірогідність настання іншого.

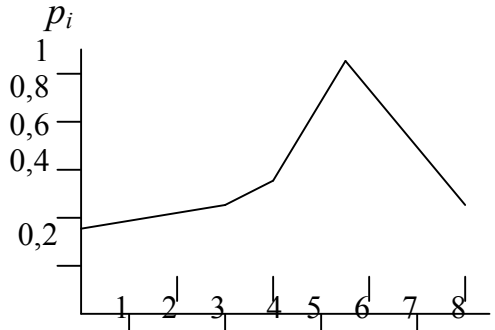
Закони розподілу випадкових величин

Закон розподілу – відповідність між значеннями випадкових величин і вірогідністю їх реалізації. Може бути заданий у вигляді таблиці, формули або графіка.

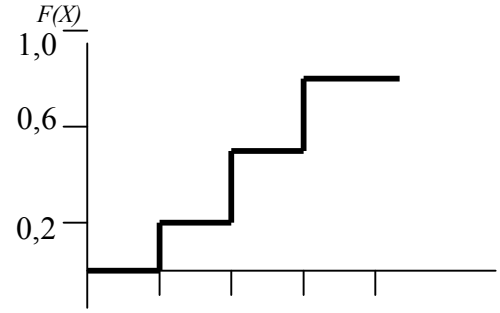
Для дискретної випадкової величини зазвичай закон розподілу задається рядом розподілу.

x_i	1	3	4	8
p_i	0,2	0,3	0,4	0,1

x_i	1	3	4	8
$\sum p_i$	0,2	0,5	0,9	1,0



Многокутник розподілу.



Функція розподілу дискретної ВВ.

Для безперервної випадкової величини табличне її представлення неможливе, тому використовують функцію розподілу(ФР).

ФР – це функція *F(x)*, яка задає ймовірність того, що випадкова величина *X* у випробуванні набуває значення менше *x*:

$$F(x)=p(X<x).$$

Інколи її називають інтегральною функцією *F(x)*, функцією, що не убуває. Таким чином. якщо $a>b$, то $F(a)>F(b)$. При цьому $F(-\infty)=0$, а $F(+\infty)=1$. Для дискретних випадкових величин функція розподілу є ступінчатою функцією.

Вірогідність того, що випадкова величина потрапляє в інтервал, визначається за формулою:

$$P(a<X<b)=F(b)-F(a).$$

Щільність вірогідності (ЩВ) це

$$f(x)=F'(x)$$

При цьому $\int_{-\infty}^{+\infty} f(x)dx = 1$

Ймовірність попадання випадкової величини в інтервал (*a, b*) дорівнює:

$$P(a<X<b)=\int_a^b f\{x\}dx.$$

Математичне очікування (МО)

Для дискретної випадкової величини математичне очікування визначається за формулою:

$$M(X)=\frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{i=1}^{\infty} x_i p_i ,$$

де x_i – значення ВВ;

p_i – ймовірність появи ВВ.

Якщо маємо генеральну сукупність, тобто всі реалізації випадкової величини, то $M(X)=\bar{X}$, де $\bar{X}$ – середнє арифметичне.

Для безперервних випадкових величин

$$M(X)=\int_{-\infty}^{+\infty} xf(x)dx.$$

Дисперсія визначається як МО квадрата центрування випадкової величини

$$D(X)=M((\bar{X}-M(X))^2).$$

Зазвичай дисперсія визначається за формулою:

$$D(X)=M(X^2)-(M(X))^2.$$

Середнє квадратичне відхилення визначається таким чином:

$$\sigma(X)=\sqrt{D(X)}.$$

Квантіль – це розв'язання відносного x рівняння

$$F(x)=p,$$

де p – задана ймовірність.

Виділяють такі окремі випадки квантилів, які мають власну назву.

Квартіль – це три значення ознаки Q_1, Q_2, Q_3, які поділяють варіаційний ряд на чотири рівні частини.

Q_1 – нижній квартіль – це значення, для якого виконується умова, що чверть спостережень менше його.

Q_3 – верхній квартіль – значення, що менше чверті спостережень.

Тобто, медіана і квартилі ділять ранжируваний ряд на чотири рівні частини. Значення $(Q_3–Q_1)$ називається інтерквартильною широтою.

$(Q_3-Q_1)/2$ – називається семиінтерквартильною широтою.

Вона є медіаною абсолютних відхилень від середнього квартиля (Q_3+Q_1) /2.

Децилі – це такі значення, які ділять ранжируваний ряд на 10 рівних за об'ємом частин (центилі – на 100).

Під p-процентним квантилем розуміють такі значення ознаки, які не перевершують p % спостережень.

Розглянемо деякі важливі закони розподілу.

До безперервних законів розподілу відносяться:

– нормальний закон розподілу (Гауса) і пов'язані з ним інші закони розподілу;
– хі-квадрат (Пірсона);
– Стьюдента;
– Фішера;
– показниковий;
– Гамма-розподіл;
– Бета-розподіл;
– Вейбулла;
– Коші;
– Ерланга тощо.

До дискретних законів розподілу відносяться:

– біноміальний;
– Пуассона;

– Паскаля;
– Парето
– рівномірний тощо.

Нормальний закон розподілу (Гауса)

Щільність розподілу

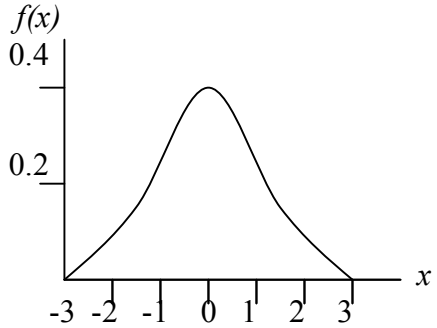
$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\left(\frac{(x-m)^2}{2\sigma^2}\right)}$$

m – МО; σ^2 – дисперсія.

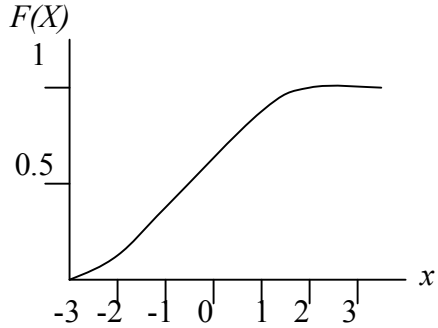
Цей закон широко використовується в ТВ і МС.

Стандартним нормальним розподілом називається розподіл з нульовим МО і одиничною D, щільність якого виглядає таким чином:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x^2}{2}\right)}$$



Щільність вірогідності

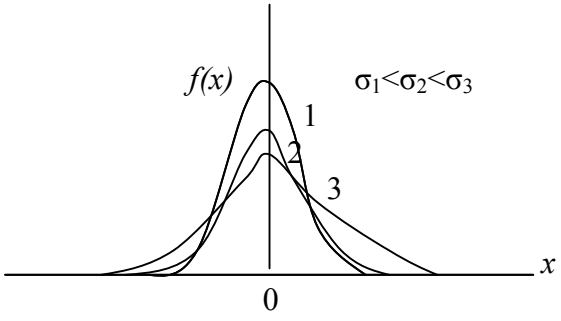


Функція розподілу стандартного нормального закону випадкової величини

Функція розподілу стандартного нормального закону

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{U^2}{2}} dU$$

Зміна МО не змінює форму кривої, а лише переміщає її уздовж осі x. При зміні D форма кривої змінюється:



Чим більше дисперсія, тим більше пологою і розтягнутою стає крива (і навпаки).

На нормальному законі розподілу базується практично вся параметрична статистика. Це пов'язано з тим, що більшість розподілів, використовуваних для перевірки статистичних гіпотез (Фішера, Стьюдента і ін.) є перетворенням НЗР.

Головна особливість НЗР в тому, що він є граничним законом, до якого прагнуть, при виконанні деяких умов, всі інші закони розподілу. Це слідує також і від центральної граничної теореми.

Найбільш важливою є теорема Ляпунова, згідно якої закон розподілу суми незалежних ВВ наближається до нормального закону при необмеженому зростанні числа ВВ і виконанні наступних умов: всі величини мають кінцеві МО і D та ні одна з величин за значенням не відрізняється різко від інших.

Розподіл Стьюдента

За законом Стьюдента розподілена випадкова величина, що визначається як:

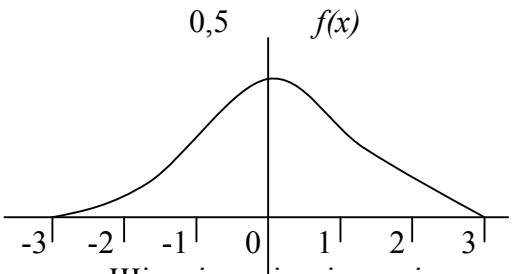
$$t_n = \frac{x_0}{\sqrt{\frac{\sum_{i=1}^{n} x_i^{\,2}}{n}}},$$

де випадкові величини x_i мають стандартний нормальний розподіл. Щільність розподілу:

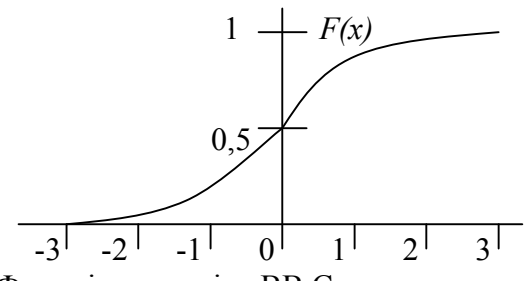
$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi \cdot n}}\left(1+\frac{x^2}{n}\right)^{-\frac{(n+1)}{2}}.$$

МО розподілу Стьюдента (С) дорівнює 0, а дисперсія – $\frac{n}{(n-2)}$

Тут $\Gamma(\,)$ – гамма функція. $\Gamma(z)=\int_0^{\infty} x^{z-1} e^{x}\,dx,\ z>0.$

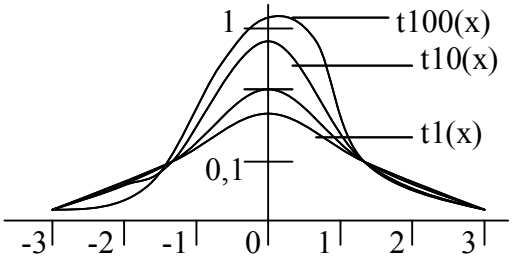


Щільність вірогідності розподілу ВВ з числом степенів свободи 1



Функція розподілу ВВ Стьюдента з одною степеню свободи

Щільність вірогідності розподілу С для різних степенів свободи:



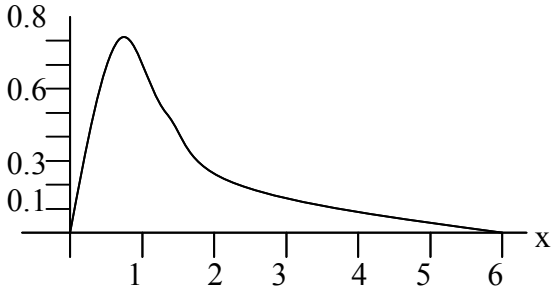
Вона дуже подібна на щільність нормального розподілу.

Розподіл Фішера

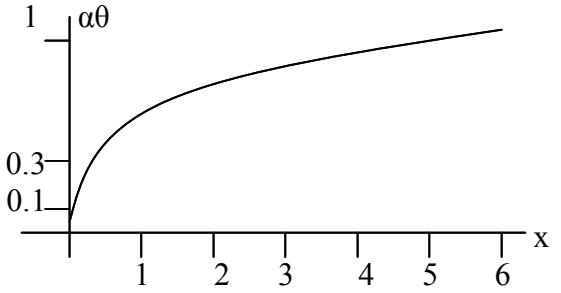
За критерієм Фішера розподілена випадкова величина такого виду:

$$F_{n,m}=\frac{\sum_{i=1}^{m}\frac{x_i^2}{m}}{\sum_{j=1}^{n}\frac{y_j^2}{n}}.$$

Випадкові величини x_i і y_i розподілені за стандартним нормальним закону розподілу.

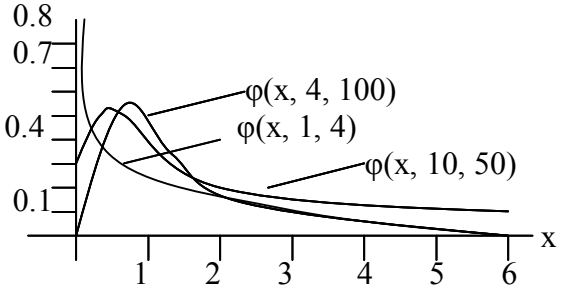


Функція щільності розподілу для степенів свободи (4, 40).



Функція розподілу для степенів свободи (4, 40).

Функції щільності вірогідності розподілу Фішера з різними степенями свободи:



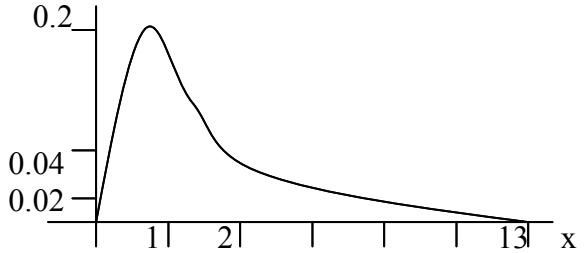
Розподіл χ^2 (*хі-квадрат*) Пірсона

Розподіл χ^2 має ВВ, що є сумою квадратів ВВ, кожна з яких розподілена за нормальним законом. Щільність розподілу χ^2 має такий вигляд:

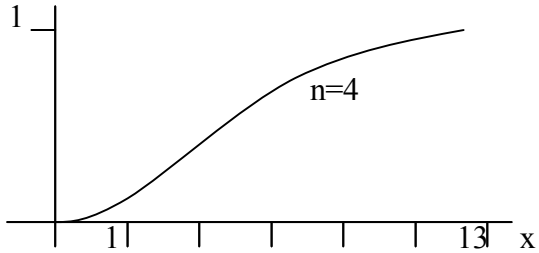
$$f(x)=\frac{1}{2^{\frac{n}{2}}}\frac{1}{\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{\left(\frac{x}{2}\right)}.$$

МО розподілу χ^2 дорівнює n, а дисперсія – $2n$.

Графіки функцій щільності вірогідності і функцій розподілу *хі-квадрат*:



Щільність вірогідності χ^2 при різних степенях свободи.



Функція розподілу χ^2.

Біноміальний розподіл

Це розподіл вірогідності m подій, що настали, в незалежних випробуваннях при постійній вірогідності події в кожному випробуванні p. Вірогідність можливого числа подій визначають за формулою Бернуллі:

$$P_n(X=m)=C_n^m p^m q^{n-m},$$

де p – вірогідність появи події в кожному випробуванні; m – очікуване число подій; n – загальна кількість випробувань; $q=1-p$

$$C_n^m=\frac{n!}{m!(n-m)!}.$$

Біноміальний розподіл може бути заданий у вигляді ряду:

$X=m$	0	1	...	k		...	n
$P_n(m)$	q^n	p^1q^{n-1}	…	$C_n^k p^k q^{n-k}$		…	p^n

МО біноміального розподілу дорівнює np, а дисперсія – npq.

При великій кількості випробувань біноміальний розподіл стає досить близьким до нормального. Форма розподілу із зростанням n наближається до нормальної, але тим повільніше, чим менше p.

Розподіл Пуассона

У разі, коли p або q малі при зростанні числа випробувань біноміальний розподіл прагне до розподілу Пуассона, який задається формулою:

$$P_n(m) = \frac{\lambda^m e^{-\lambda}}{m!},$$

де $\lambda = np$ – найбільш ймовірна частота очікуваної події;
m – частота очікуваної події в n незалежних випробуваннях.

В розподілі Пуассона МО$=np$, $D=np$.

Показниковий розподіл (ПР)

Дуже багато ВВ (наприклад інтервали виклику швидкої допомоги тощо) відповідають експоненціальному закону розподілу.

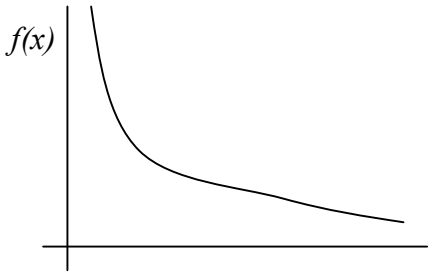
Щільність розподілу для ПР має вигляд

$$p(x,\lambda)=\lambda e^{-\lambda x}.$$

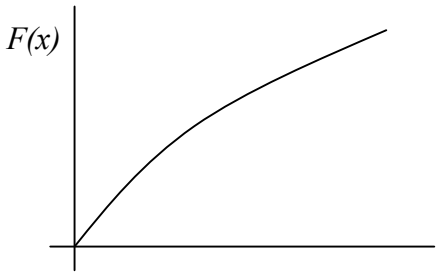
Функція розподілу

$$F(x,\lambda) = \begin{cases} 1 - e^{-\lambda \cdot x}, \forall x \geq 0 \\ 0, \forall x \prec 0 \end{cases},$$

МО дорівнює $1/\lambda$, а дисперсія – $1/\lambda^2$.



Щільність вірогідності.



Функція розподілу вірогідності α.

В більшості випадків при розв'язанні реальних задач закон розподілу і його параметри невідомі. Тому для визначення виду ЗР і його параметрів необхідно виконати ряд дій з аналізу отриманих початкових даних.

У прикладних дослідженнях задача визначення ЗР досить важлива, оскільки різні критерії оцінки вибірок (відповідність тому або іншому ЗР) дають різний результат. Зміна розбиття на інтервали може змінити висновки, емпіричний розподіл може бути „засмічений" (зашумлений) в порівнянні з теоретичним. Для експериментатора необхідність приділяти цьому питанню існує в наступних випадках:

– вживані ним методи, як передумова, вимагає певного закону Р;

– вирішуване задача вимагає значення вигляду і параметрів ЗР.

У першому випадку зазвичай обмежуються несуворими простими перевірками, достатніми для ухвалення рішення (або вибирають параметричні методи, що не вимагають знання закону розподілу).

У другому – апроксимують емпіричний закон Р функціями Пуассона.

Можна вважати, що ВВ розподілена за НЗ, якщо виконуються нижче наведені умови, що випливають з НЗ Р. Для цього визначають середнє абсолютне відхилення:

$$\Delta_{abs} = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)}{N}$$

Потім перевіряють виконання наступних умов:

– кількість додатних і від'ємних відхилень від середнього Δ_{abs} приблизно однакова;

– половина (або трошки більше) відхилень від середніх за абсолютною величиною менше середнього абсолютного відхилення

$$\Delta_i < \Delta_{abs}$$

– жодне з відхилень не перевищує середнє абсолютне відхилення більш ніж в 3–4 рази $\Delta_{imax} < (3–4)\Delta_{abs}$.

Пропонуються і інші можливі перевірки.

Наприклад, достатньо перевірити виконання наступної умови:

$$\left|\frac{\Delta_{abs}}{\sigma} - 0{,}7979\right| < \frac{0{,}4}{\sqrt{N}}.$$

Для тих, хто не хоче обчислювати середнє абсолютне відхилення, можна використовувати перевірку виконання наступного комплексу умов:

– майже всі (99,7%) відхилення від середнього менше 3σ: $\Delta_i < 3\sigma$;

– 2/3 (68,3%) відхилень менше σ;

– половина відхилень менше 0,625 σ.

При виконанні цих умов можна вважати, що дані розподілені за НЗ.

Характеристики ВВ

При аналізі даних виникають такі проблеми:

– об'єм і спосіб відбору даних;

– правомірність поширення висновків, що зроблені на підставі вибіркових даних, на всю генеральну сукупність;

– вибір оптимальних способів оцінювання;

– вибір способів узагальнення, класифікації і представлення даних.

Властивості оцінок параметрів

Оцінки параметрів повинні відповідати наступним вимогам:

– незміщенність – це означає, що при проведенні дуже великої кількості випробувань з вибірками однакового розміру середнє значення кожної вибірки прагне до дійсного значення генеральної сукупності. Зміщенність зазвичай обумовлена наявністю систематичної похибки.

– спроможність. Із зростанням розміру вибірки оцінка повинна прагнути до значення відповідного параметра генеральної сукупності з вірогідністю, що прагне до 1.

– ефективність. Вибрана оцінка для вибірки рівного об'єму повинна мати мінімальну дисперсію.

– достатність – оцінка повинна містити всю необхідну інформацію і вимагати додаткову.

Для оцінювання параметрів використовуються різні методи, особливе місце серед них займає метод максимальної правдоподібності. Він застосовується в тих випадках, коли відомий закон розподілу. Суть його в тому, що оцінки мають дорівнювати значенням, при яких вибірка має максимальну вірогідність появи.

До характеристик одновимірного розподілу відносяться:

– міри розташування (середнє, медіана, мода та ін.);

– міри розсіювання (розмах, коефіцієнт варіації, дисперсія, середньоквадратичне відхилення);

– міри форми (асиметрія, ексцес, моменти третього і четвертого порядку).

Середнє арифметичне (вибіркове)

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Властивості середнього:

– сума відхилень від середнього дорівнює 0;

– якщо всі значення вибірки збільшити або зменшити, помножити або розділити на одне і те ж число, то середнє значення зміниться аналогічно;

– із збільшенням кількості вимірів точність оцінки зростає і середнє наближається до МО, але тільки в тому випадку, якщо немає систематичних похибок і спостереження незалежні;

– середнє суми двох вибірок дорівнює сумі їх середніх, якщо вибірки однакових розмірів (аналогічно для різниці) $\overline{X+Y} = \overline{X} + \overline{Y}$;

– якщо ряд спостережень складається з k групп, то середнє арифметичне всього ряду дорівнює взваженій груповій середній, вагами при цьому є об'єми груп

$$\bar{X} = \frac{\sum_{i=1}^{k} X_i n_i}{\sum_{i=1}^{k} n_i},$$

де n_i – розмір i-ї групи; $\overline{X_i}$ – середнє i-ї групи.

Зауваження:

– середнє зовсім не означає „типове” (середній прибуток зовсім не є типовим);

– середнє не співпадає з МО. За виключенням нормального розподілу.

Арифметичне середнє навіть не є незміщеною оцінкою МО з найменшою дисперсією (за винятком НЗР). Більш того, навіть для НЗР, можна вказати оцінку, яка буде ближча до МО.

Середнє геометричне (вибіркове)

СГ застосовується, якщо:

– змінна змінюється в часі з постійними співвідношеннями між її вимірами (наприклад, збільшення числа бактерій, експлуатаційні витрати, зростання капіталу на рахунку):

$$\frac{X_{i-1}}{X_i} = \frac{X_i}{X_{i+1}} = const$$

– окремі значення у вибірці знаходяться дуже далеко один від одного (наприклад, відрізняються на порядок).

ГС визначається за формулою:

$$\bar{\bar{X}}_0 = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

Середнє гармонійне

У ряді випадків (наприклад, розрахунок середньої тривалості життя, визначення середньої швидкості тощо) використовують гармонійне середнє:

$$\bar{\bar{X}}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}.$$

Мода

Це значення, яке спостерігається найбільшу кількість разів (найбільш вірогідна величина). Для інтервального варіаційного ряду мода розраховується за формулою:

$$M_0 = X_{M0} + \frac{h\left(m_{M0} - m_{M0-1}\right)}{2m_{M0} - m_{M0+1} - m_{M0-1}},$$

де $X_{м0}$ – початок модального інтервалу (такого, якому відповідає найбільша частота);

h – величина модального інтервалу;

m_{M0} – частота модального інтервалу;

m_{Mo-1} – частота модального інтервалу, що передує модальному;

m_{Mo+1} – частота інтервалу, наступного за модальним.

Мода не застосовується в тому випадку, якщо розподіл мультимодальний (багатовершинний).

Медіана (вибіркова)

Це значення, яке ділить ранжируваний варіаційний ряд на дві рівні за об'ємом групи. Варіаційний ряд ранжирується. Якщо кількість членів ряду є непарною, то медіаною є значення ряду, яке розташоване посередині, тобто елемент з номером *(N+1)/2*.

Якщо число членів парне, то медіана дорівнює середньому числу ряду з номерами (*N/2*) і (*N/2+1*).

Для інтервального варіаційного ряду медіана обчислюється за формулою:

$$M_e = X_{Me} + \frac{h\left(\frac{\left(\sum m_x\right)}{2} - m_x^{\max}\right)}{m_m},$$

де X_{Me} – початок медіанного інтервалу;

h – величина медіанного інтервалу;

m_x – частоти за всіма інтервалами;

m_x^{max} – частота накопичена на початок медіанного інтервалу;

m_m – частота медіанного інтервалу.

Медіанним інтервалам називається інтервал, в якому знаходиться значення медіани.

Властивості медіани:

– сума абсолютних величин відхилень варіантів від медіани, помножених на відповідні частоти, менше від будь-якої іншої величини:

$$\sum_{\forall x} |x - M_e| m_x < \sum_{\forall x} |x - a| m_x ;$$

– на значення медіани не впливає зміна крайніх значень варіаційного ряду, якщо тільки менше медіани залишається меншим, а більше – продовжує залишатися більше її.

Показники варіації (розсіювання)

Варіаційний розмах

$$RB = X_{max} - X_{min}$$

Варіаційний розмах є ненадійною оцінкою варіації, оскільки на нього впливають крайні значення. Він не змінюється при будь-яких змінах варіаційного ряду, що не зачіпають крайні значення.

Емпірична дисперсія (дисперсія вибіркова)

$$D = S^2 = \sigma^2 = \frac{\sum_{i=1}^{N}\left(X_i - \bar{\bar{X}}\right)^2}{N-1}.$$

Число *N-1* є кількістю степенів свободи для дисперсії вибіркової.

Властивості *D*:

– *D* постійної величини дорівнює *0*;

– якщо всі результати збільшити або зменшити на одне і те ж число, то *D* не змінюється;

– якщо всі результати змінити в *k* раз, то *D* зміниться в k^2 разів;

– корінь квадратний з *D* є середньоквадратичним відхиленням *S* або σ.

Незміщеною оцінкою *D* є:

$$D = S^2 = \sigma^2 = \frac{\sum_{i=1}^{k}\left(X_i - \overline{\overline{X}}\right)^2}{N}, \text{ або } D = \frac{\sum_{i=1}^{k} f_i (X_i - \overline{X})}{N}$$

якщо МО відомо.

В наведених формулах X_i – значення варіант, а $\overline{X}$ – їх середнє значення, f_i – частоти відхилень в межах від першого до *k*–го класу.

На практиці: якщо МО отримано за іншою вибіркою, використовують формулу з *N* в знаменнику, якщо по тій же вибірці – зміщену оцінку з *N-1* в знаменнику.

Коефіцієнт варіації

$$V = \frac{S}{\overline{\overline{X}}} \cdot 100\%$$

У випадку, якщо полігон частот варіаційного ряду не має значної зкошенності, а всі члени ряду додатні, то $V<30\%$. Якщо коефіцієнт варіації більше 100%, то зазвичай це означає, що дані неоднорідні.

Міри форми

Асиметрія і ексцес. Моменти розподілу.

Серед емпіричних розподілів асиметрія і ексцес зустрічаються досить часто. Замітити їх можна за характером розподілу частот в класах варіаційних рядів. Графічно асиметрія виражається в вигляді скошеної варіаційної кривої, вершина якої може знаходитись зліва або справа від центру розподілу. В першому випадку асиметрія називається правосторонньою або позитивною (додатною), а в другому – лівосторонньою або від'ємною.

Наряду з асиметричними зустрічаються гостро вершинні і плоско вершинні розподіли. Гостровершинність кривої розподілу викликана надмірним накопиченням частот в центральних класах варіаційного ряду. В таких випадках має місце додатній (позитивний) ексцес розподілу. Окрім одновершинних зустрічаються двох - та багато вершинні криві, а також плоско вершинні і

двогорбі криві, що вказує про наявність у такого розподілу негативного (від'ємного) ексцесу.

Величина асиметрії та ексцесу може бути різною. Тому важливо їх не тільки виявити, але й виміряти. Для цього використовують центральні моменти розподілу третього та четвертого порядків.

Моментами розподілу називають суми відхилень варіант x_i від якого-небудь числа *A,* які взведенні в *k*- ту степінь і віднесені до загальної кількості варіант, що складають дану сукупність:

$$M = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^k .$$

Якщо відхилення варіант обчислюють по відношенню до нульової точки, то моменти називають початковими (їх позначають літерою *m*), якщо – від середньої арифметичної, то моменти називають центральними (μ), а при визначенні відхилень від довільно вибраного числа *A* моменти називають умовними і позначають літерою *b*. Показник степені *k* вказує на порядок моменту розподілу.

Для визначення асиметрії і ексцесу використовують центральні моменти розподілу третього і четвертого порядків.

В якості показника асиметрії *As* слугує центральний момент третього порядку μ_3, що віднесений до кубу середньоквадратичного відхилення *S*, тобто:

$$As = \frac{\mu_3}{S^3} = \frac{\sum_{i=1}^{k} f_i (x_i - \overline{X})^3}{S^3 n}.$$

При суворо симетричних розподілах сума третіх степенів відхилень варіант x_i від середньоарифметичної $\overline{X}$ дорівнює нулю і *As=0.* При наявності скошеності розподілу цей показник буде мати доданню (при правосторонній асиметрії) або від'ємну величину (при лівосторонній асиметрії), яка і слугує мірою асиметрії.

Показник ексцесу позначається як *Ex* і виражається формулою:

$$Ex = \frac{\mu_4}{S^4} - 3 = \frac{\sum_{i=1}^{k} f_i (x_i - \overline{X})^4}{S^4 n} - 3 .$$

При відсутності ексцесу *Ex*=0. В випадку позитивного ексцесу цей показник отримує позитивний знак (+) и може мати довільну форму. При плосковершинності або двогорбості варіаційної кривої коефіцієнт *Ex* має від'ємний знак (–), а гранична величина від'ємного ексцесу дорівнює мінус два.

Центральні моменти зручно обчислювати непрямим шляхом – через умовні моменти розподілу, які визначеним чином пов'язані з центральними моментами.

При обчисленні показників *As* і *Ex* способом умовної середньої *A* статистичні моменти визначають за формулами:

$$b_1 = \sum_{i=1}^{k} \frac{f_i a}{n};$$

$$b_2 = \sum_{i=1}^{k} \frac{f_i a^2}{n};$$

$$b_3=\sum_{i=1}^{k}\frac{f_i a^3}{n};$$

$$b_4=\sum_{i=1}^{k}\frac{f_i a^4}{n},$$

де $a=(f_i-A)/\lambda$; n – загальна кількість спостережень; f_i – частоти варіаційного ряду; λ – класовий інтервал.

Малі значення показників асиметрії і ексцесу вказує на близькість розподілу до нормальної кривої.

Довірчий інтервал (ДІ)

ДІ – це такий інтервал, щодо якого з наперед заданою вірогідністю $p=1-\alpha$ можна стверджувати, що він містить невідоме значення параметра Q.

$$p(Q_1<Q<Q_2)=1-\alpha,$$

де $1-\alpha$ – довірча вірогідність; α – рівень значущості.

Властивості:

– при збільшенні кількості вимірів точність підвищується. Це справедливо тільки в тому випадку, якщо немає систематичних похибок і спостереження незалежні;

– збільшення надійності при фіксованій вибірці призводить до збільшення довірчого інтервалу і зниження точності.

Якщо збільшується кількість вимірів, то оцінка параметра стає більш точною і довірчий інтервал зменшується.

ДІ означає не вірогідність попадання значення оцінюваного параметра в межі певних границь, а то, що, якщо ми візьмемо достатнє число вибірок, *100 p* % випадків параметр буде знаходиться в заданому інтервалі.

Рівень значущості вибирається зазвичай в інтервалі від 0,1 до 0,001. При цьому 0,05 – звичайна вимога надійності; 0,01 – підвищена; 0,001 – дуже висока; 0,1 – знижена.

ДІ для середнього

Визначають за формулою:

$$\left[\bar{\bar{X}}-t_{n,p}\frac{S}{\sqrt{n}},\bar{\bar{X}}+t_{n,p}\frac{S}{\sqrt{n}}\right],$$

де S – середнє квадратичне відхилення;

n – число дослідів;

$t_{n,p}$ – табличне значення розподілу Стьюдента з числом степенів свободи n і довірчій вірогідності p.

Цю формулу застосовують, коди D невідома і використовують її оцінку за експериментальними даними.

ДІ для середнього квадратичного відхилення

$$P\left(\frac{\sqrt{nS}}{X_2} < \sigma < \frac{\sqrt{nS}}{X_1}\right) < 1-\alpha,$$

де X_1 і X_2 – квантилі розподілу χ^2 (*хі-квадрат*).

Квантилі X_1 і X_2 мають по (n-*1*) степенів свободи і рівні значущості (*1*-$\alpha/2$) і $\alpha/2$ відповідно.

Поняття параметричної, непараметричної і робастної статистики

Для перевірки будь-якої гіпотези необхідно спиратися на деяку сукупність припущень з яких і виводяться формули, необхідні для цієї перевірки. При цьому серед інших завжди присутні припущення про закон розподілу вибірки. Невиконання цих передумов робить некоректне вживання відповідних методів.

Параметричні методи передбачають конкретний розподіл з певними параметрами. Практично всі традиційні статистичні критерії і методи відносяться до цієї групи. Вони зазвичай строго обґрунтовані і добре вивчені. Але в переважній більшості практичних завдань передумови нормальності не виконуються або перевірити з неможливо.

Робастні методи також передбачають конкретний розподіл, але допускають відхилення від нього. Форма і величина цих відхилень залежать від конкретного методу. Не для всіх видів задач розроблено робастні методи.

Непараметричні методи не роблять конкретних припущень про закон розподілу – тільки найзагальніші. Вони зазвичай досить суворо обґрунтовані. Для багатьох задач, особливо багатовимірних, не існує відповідних непраметричних методів.

Аналіз даних – зазвичай є деякими евристичними процедурами, що розроблені для вирішення конкретного класу задач. Обґрунтованість їх спирається лише на логіку і проведений обчислювальний експеримент із спеціально сформульованими наборами даних.

Вибір методів здійснюється залежно від мети дослідження і особливостей наявних даних.