Лекція 35.

Фіксація й обробка статистичних результатів

У лекції 21 ми докладно познайомилися зі схемою статистичного комп'ютерного експерименту. У лекціях 21—26 ми розглянули практичну реалізацію всіх основних блоків (див. мал. 21.3) цієї схеми. Зараз важливо навчитися організовувати роботу останніх двох блоків - блок обчислення статистичних характеристик (БОСХ) і блок оцінки вірогідності статистичних результатів (БОВ).

Отже, розглянемо, як варто фіксувати статистичні величини в результаті експерименту, щоб одержати надійну інформацію про властивості модельованого об'єкта. Нагадаємо, що узагальненими характеристиками випадкового процесу або явища є середні величини.

Обчислення середніх

Обчислення середніх величин під час експерименту, що багаторазово повторюється, а результат його усереднюється, може бути організовано декількома способами:

- вся статистика обчислюється наприкінці;
- вся статистика обчислюється в процесі обчислення (по рекурсивних співвідношеннях);
- вся статистика обчислюється в класових інтервалах (цей метод сполучає універсальність першого методу й економічність другого).

Спосіб 1. Обчислення всієї статистики наприкінці. Для цього в процесі експерименту значення X_i вихідний (досліджуваної) випадкової величини X накопичується в масиві даних. Після закінчення експерименту підраховується математичне очікування (середнє) $\boldsymbol{X}$ і дисперсія D (характерний розкид величин щодо цього математичного очікування).

$$\boldsymbol{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \qquad (1)$$

$$D = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \boldsymbol{X})^2 \qquad (2)$$

Часто використають середньоквадратичне відхилення σ = sqrt(D).

Помітимо, що недоліком методу є неефективне використання пам'яті, тому що доводиться накопичувати й зберігати велика кількість значень вихідної величини протягом усього експерименту, що може бути досить тривалим.

Другий мінус полягає в тім, що доводиться двічі зчитувати масив X_i, тому що скористатися формулою (2) у тім виді, як вона тут записана, ми можемо, тільки прорахувавши формулу (1) (від 1 до n), а потім ще раз прогнавши для формули (2) масив X_i.

Позитивним моментом є збереження всього масиву даних, що дає можливість більше докладного його вивчення надалі при необхідності розслідування тих або інших ефектів і результатів.

Спосіб 2. Обчислення всієї статистики в процесі обчислення (по рекурсивних співвідношеннях). Цей спосіб передбачає можливість зберігати тільки поточне значення математичного очікування $\boldsymbol{X_i}$ і дисперсії D_i, що підправляє на кожній ітерації. Це рятує нас від необхідності постійного зберігання всього масиву експериментальних даних. Кожне нове дане X_i ураховується в сумі з ваговим коефіцієнтом — чим більше слагаемых i накопичено в сумі $\boldsymbol{X_i}$, тим більше її значення важливо стосовно чергового виправлення X_i, тому співвідношення вагових коефіцієнтів $i/(i + 1) : 1/(i + 1)$.

$$\boldsymbol{X}_{i+1} = \frac{\boldsymbol{X}_i \cdot i + X_{i+1}}{i+1} = \boldsymbol{X}_i \cdot \frac{i}{i+1} + X_{i+1} \cdot \frac{1}{i+1}$$

$$D_{i+1} = \frac{(X_{i+1} - \boldsymbol{X}_i) + i \cdot D_i}{i+1} = \frac{(X_{i+1} - \boldsymbol{X}_i)}{i+1} + \frac{i}{i+1} \cdot D_i$$

де $\boldsymbol{X_i}$ — чергове значення експериментальної вихідної величини.

Спосіб 3. Обчислення всієї статистики в класових інтервалах. Цей спосіб припускає, що в масив будуть накопичувати не всі значення X_i, а тільки по значимих інтервалах, у яких розподілена випадкова вихідна величина X_i. Загальний інтервал зміни X_i розбивається на m підінтервалів, у кожному з яких фіксується кількість n_i, що показує, скільки разів X_i прийняло значення з *i-го* інтервалу. При невеликій кількості інтервалів ($m \approx 1$) ми одержуємо спосіб 1, при кількості інтервалів $m = n$ ми одержуємо спосіб 2. У випадку $1 < m < n$ одержуємо середнє рішення — компроміс між займаною пам'яттю й інформативністю масиву вихідних даних.

$$\boldsymbol{X} = \frac{n_1 \cdot X_1 + n_2 \cdot X_2 + \ldots + n_m \cdot X_m}{n_1 + n_2 + \ldots + n_m} = \frac{1}{N} \cdot \sum_{i=1}^{n} n_i X_i$$

$$D = \frac{1}{n} \cdot \sum_{i=1}^{n} (n_i \cdot X_i - \boldsymbol{X})^2$$

Обчислення геометрії розподілу

Ще більш інформативним є обчислення геометрії розподілу випадкової величини. Воно необхідно для того, щоб уявити собі більш точно характер розподілу. Відомо, що за значенням статистичного моменту можна приблизно судити про геометричний вид розподілу.

Перший момент (або середн арифметичне) обчислюється так:

$$m_1 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - A)^1, \text{ где } A = \begin{cases} 0, (\text{начальный момент}) \\ X, (\text{центральный момент}) \end{cases}$$

Якщо A приймає значення 0, то перший момент називається початковим моментом, якщо A приймає значення X, те перший момент називається центральним. (У принципі A може бути будь-яким числом, заданим дослідником.)

На практиці прийнято використати не сам перший момент, а нормовану його величину $R_1 = m_1/\sigma^1$.

Перший момент указує на центр ваги в геометрії розподілу, див. мал. 34.1.

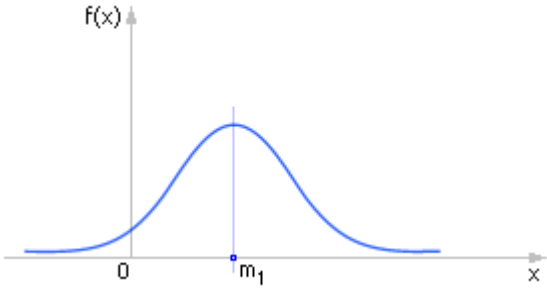

Рис. 34.1. Характерне положення першого моменту
на графіку розподілу статистичної величини

Другий момент (або дисперсія, розкид) обчислюється так:

$$m_2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \boldsymbol{X})^2$$

Ви знайомі з поняттям середньоквадратичного відхилення, зв'язаним із другим моментом:

$$\sigma = \sqrt{m_2}$$

На практиці прийнято використати не сам другий момент, а нормовану його величину $R_2 = m_2/\sigma^2$.

Дисперсія характеризує величину розкиду експериментальних даних щодо центра ваги m_1. Таким чином, по величині m_2 можна судити про другий параметр геометрії розподілу (див. мал. 34.2).

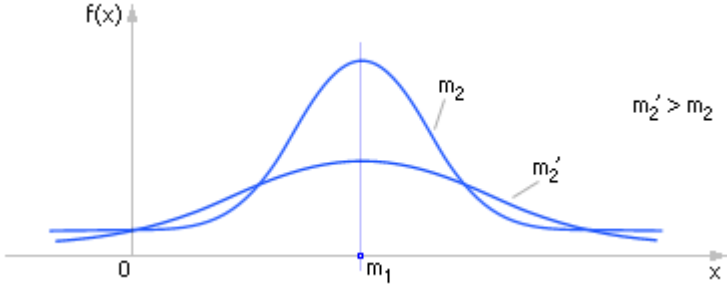

Рис. 34.2. Характерна зміна виду розподілу статистичної величини залежно від величини другого моменту

Третій момент характеризує асиметрію (або скошеність) (див. мал. 34.3.

На практиці прийнято використати не сам другий момент, а нормовану його величину $R_3 = m_3/\sigma^3$.

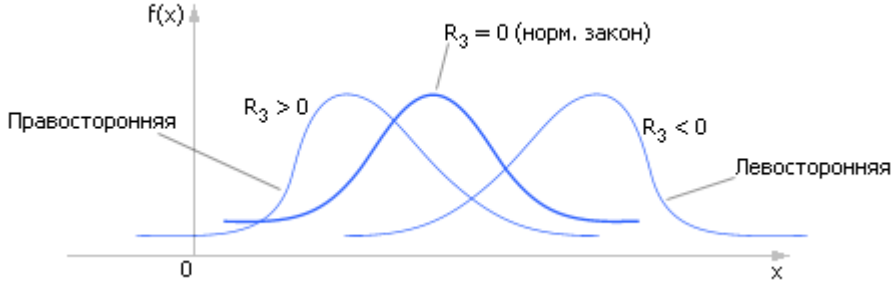

Рис. 34.3. Характерна зміна виду розподілу статистичної величини залежно від величини третього моменту

Визначаючи знак R_3, можна визначити, є чи асиметрія в розподілу (див. мал. 34.3), а якщо є ($R_3 \neq 0$), то в яку сторону.

Четвертий момент (див. мал. 34.4) характеризує ексцес (або островершинность) і обчислюється так:

$$m_4 = \frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \boldsymbol{X})^4$$

Нормований момент дорівнює: $R_4 = m_4/\sigma^4$.

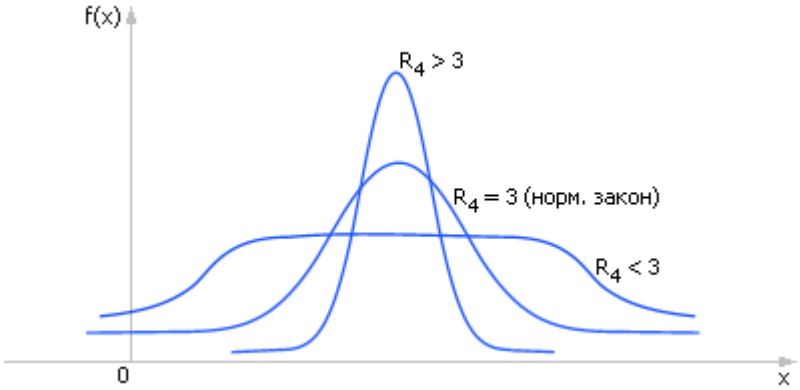

Рис. 34.4. Характерна зміна виду розподілу статистичної величини залежно від величини четвертого моменту

Дуже важливим є з'ясування того, на який розподіл найбільше походить отриманий експериментальний розподіл випадкової величини. Оцінка ступеня збігу емпіричного закону розподілу з теоретичним проводиться у два етапи: визначають параметри експериментального розподілу й далі роблять оцінку по Колмогорову відповідності експериментального розподілу обраному теоретичному.

Оцінка (по Колмогорову) збігу емпіричного закону розподілу з теоретичним

1. Обчислюємо моменти m_1, m_2, m_3, … Число моментів дорівнює числу невідомих у теоретичному законі розподілу.
2. Насамперед, тому що оцінка стосується безперервного розподілу, а ми маємо справу з дискретним розподілом, знятим експериментально, те треба вирішити, на скільки інтервалів треба розбити при дискретизації й те, і інший розподіл.

   Для цього рекомендується користуватися правилом Стерджеса, що добре зарекомендовали себе на практиці: $K = 1 + \log_2 n = 1 + 3.322 \cdot \log_{10} n$, де n — кількість випадкових значень (досвідів), k — кількість інтервалів розподілу.

3. Будується інтегральний (див. мал. 34.5) закон для емпіричного розподілу $F(x) = P(x \le x_i)$.

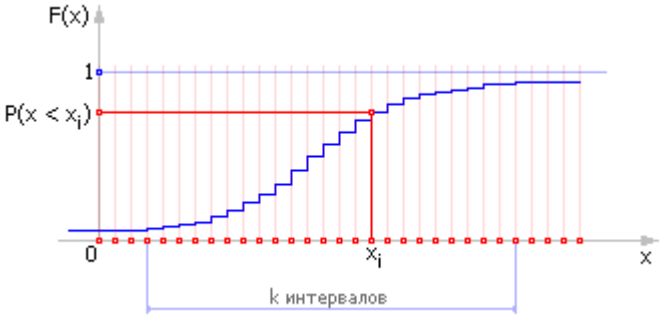

Рис. 34.5. Інтегральний закон емпіричного розподілу, дискретний варіант (приклад)

4. Залежно від числа експериментів n і кількості інтервалів $1 \le i \le k$ можна порахувати число результатів у кожному з інтервалів: $N_i = P_i \cdot n$.

5. Далее варто розрахувати теоретичний розподіл частоти: $N_i^{\text{TEOP.}} = P_i \cdot n$. Якщо в якості теоретичного прийняти нормальний закон розподілу, то можна зробити так:

$$P_i(x < x_i) = F\left(\frac{x - a}{\sigma}\right) + \frac{1}{2}$$

де F — функція Лапласа, а параметри a й σ закону обчислені в п. 1.

6. Зрівняємо отримані частоти: $N_i^{\text{ТЕОР.}}$ і N_i у всіх k інтервалах (див. мал. 34.6) і виберемо найбільше відхилення експериментального розподілу від перевіря теоретичного:

$$D = \max_i \left| N_i - N_i^{\text{теор.}} \right|$$

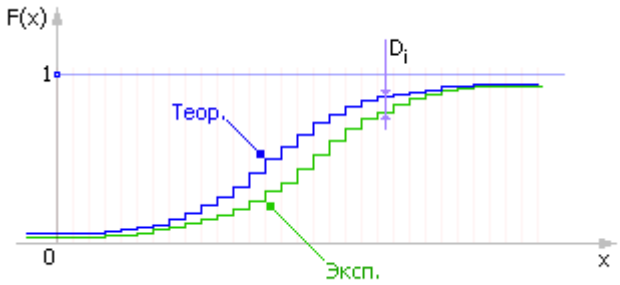

Рис. 34.6. Порівняння теоретичних й емпіричного інтегральних розподілів випадкової величини (дискретний варіант)

7. Параметр Колмогорова λ характеризує відхилення теоретичного розподілу

від експериментального:

$$\lambda = (\sqrt{k} + 0.12 + \frac{0.11}{\sqrt{k}}) \cdot D$$

Далі, використовуючи табл. 34.1 Колмогорова, варто прийняти або відкинути гіпотезу про те, чи є емпіричний розподіл із заданої нами ймовірністю *Q* теоретичним чи ні. Для прийняття гіпотези повинне бути: $\lambda < \lambda_{табл.}$.

Таблиця 34.1.
Таблиця критерію Колмогорова

Q	0.85	0.90	0.95	0.99
λ	1.14	1.22	1.36	1.63

Примітка. Критерій Колмогорова не єдиний можливий до застосування при оцінюванні; можна використати критерій Хі-квадрат, критерій Андерсона-Дарлінга й інших.

Оцінка точності статичних характеристик

Украй важливим є питання, скільки експериментів варто зробити, щоб можна було довіряти знятим характеристикам. Якщо експериментів не досить, то характеристика недостовірна. Звичайно дослідник задає довірчу ймовірність, тобто ймовірність, з якої він готовий довіряти знятим характеристикам. Чим більше буде задана довірча ймовірність, тим більше експериментів буде потрібно зробити. Раніше ми користувалися й іншими способами оцінки необхідної кількості експериментів (див. лекцію 21, приклад з монетою).

Отже, зараз наша оцінка буде ґрунтуватися на центральній граничній теоремі (див. лекцію 25, що затверджує, що сума (або середнє) випадкових величин є величина невипадкова. ЦПТ затверджує, що значення обчисленої нами статистичної характеристики будуть розподілені за нормальним законом, n_i — число *i-их* результатів значення статистичної характеристики в n експериментах, $p_i = n_i/n$ — частота *i-го* результату.

Якщо $n \to \infty$, те $p \to P$ (частота p прагне до теоретичної ймовірності P) і емпіричні характеристики будуть прагнути до теоретичного (див. мал. 34.7). Отже, згідно ЦПТ p буде розподілена за нормальним законом с математичним очікуванням m і середньоквадратичним відхиленням σ.

При цьому $m = P$, $\sigma = \text{sqrt}(p \cdot (1 - p)/n)$.

Позначимо як Q **довірчу ймовірність**, тобто ймовірність того, що частота p відрізняється від імовірності P не більш, ніж на ε. Тоді по теоремі Бернуллі:

$$Q(|p - P| \le \varepsilon) = F(\frac{\varepsilon}{\sigma}) = F(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}})$$

Величина ε називається довірчим інтервалом. Зміст ε полягає в тому, що в серії (кожна вибіркою n) у середньому $\varepsilon \cdot 100\%$ довірчих інтервалів містять щире значення статистичної характеристики p. Як і раніше (див. лекцию 25), F — інтеграл від функції нормального закону розподілу, інтегральна функція Лапласа.

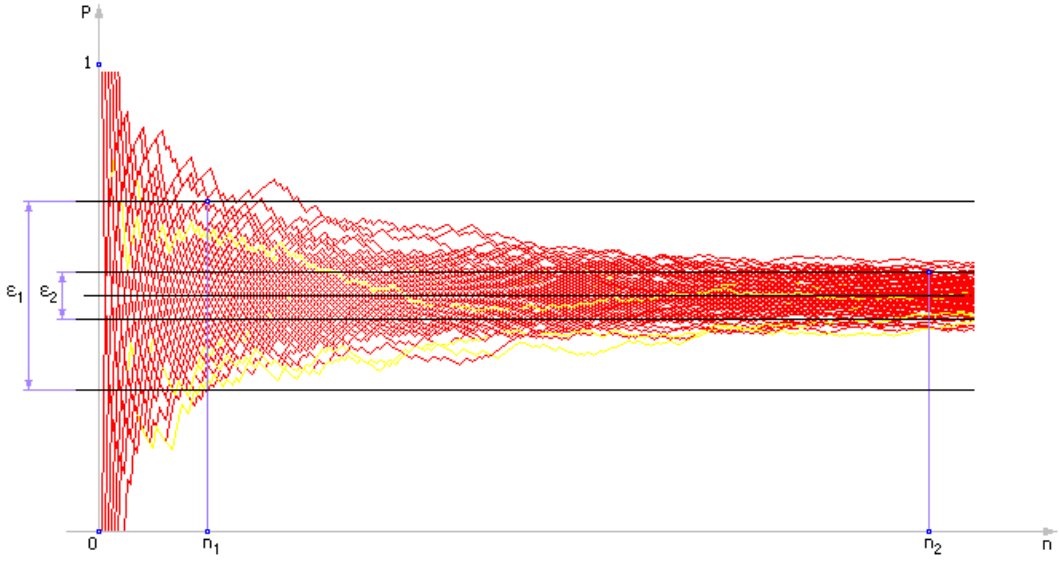

Рис. 34.7. Ілюстрація до обчислення кількості експериментів по величині довірчого інтервалу відповідно до центральної граничної теореми

Звідси можна виразити необхідне для довірчої ймовірності кількість експериментів (F^{-1} — зворотна функція Лапласа):

$$n = \frac{p \cdot (1-p)}{\varepsilon^2} \cdot \left(F^{-1}\left(Q \right) \right)^2$$

Приклад. При моделюванні випускає продукции, що, підприємством у результаті імітації його роботи протягом 50 днів були отримані наступні вихідні дані (див. табл. 34.2).

Таблица 34.2.

Експериментальні статистичні дані моделювання

Якість виробу в балах (випадкова подія i)	1	2	3	4
Кількість результатів (n_i)	15	10	5	20
Частость результату ($p_i = n_i/n$)	0.3	0.2	0.1	0.4

Тобто всього було проведено: 15 + 10 + 5 + 20 = 50 експериментів ($n = 50$). З таблиці експериментів треба відповідь задачі, що частость (імовірність) випуску виробів 1 сорту дорівнює 15/50, частість (імовірність) випуску виробів 2 сорти дорівнює 10/50, частость (імовірність) випуску виробів 3 сорти дорівнює 5/50, частость (імовірність) випуску виробів 4 сорти дорівнює 20/50.

Нехай буде довірча ймовірність до відповідей моделі $Q = 0.9$ і довірчим інтервалом $\varepsilon = 0.05$.

Тепер треба відповістити на запитання: чи можна довіряти з імовірністю Q обчисленій відповіді?

Будемо оцінювати результат статистичних експериментів по найгіршій імовірності, такий у нашій задачі є $p = 0.4$, тому що ймовірність, наприклад, 0.1 визначена набагато краще.

Дуже важлива примітка. Взагалі ймовірності (частости) близькі до 0 або 1 досить привабливі як відповідь, тому що цілком визначають рішення. Імовірності близькі до 0.5 говорять про те, що відповідь досить невизначена, подія трапиться «50 на 50». Така відповідь задовільним назвати складно, вона мало інформативний.

Формула

$$n = \frac{p \cdot (1-p)}{\varepsilon^2} \cdot \left(F^{-1}\left(Q \right) \right)^2$$

після підстановки значень $F^{-1}(0.9) = 1.65$ (див. таблицю Лапласа), далі $(F^{-1}(0.9))^2 = 2.7$, $p = 0.4$, $\varepsilon = 0.05$ дає $N = 0.4 \cdot 0.6 \cdot 2.7/0.05^2$ або остаточно $N = 250$.

Тобто наш експеримент і його відповідь недостовірна щодо заданих Q й ε: 50 експериментів недостатньо для відповіді, потрібно 250. Тобто треба продовжувати експерименти й ще провести 200 експериментів, щоб досягти необхідної точності.

Дуже важлива примітка. Формула використає себе рекурентно. Відразу обчислити з її допомогою кількість експериментів n не вдається. Щоб обчислити n, треба провести пробну серію експериментів, оцінити значення шуканої статистичної характеристики p, підставити це значення у формулу й визначити необхідної число експериментів.

Для впевненості дану процедуру варто провести кілька разів при різних одержуваних послідовно значеннях n.

Отже, у блоці оцінки вірогідності (БОВ) (див. лекцію 21) аналізують ступінь вірогідності статистичних експериментальних даних, знятих з моделі (беручи до уваги точність результату Q й ε, задані користувачем) і визначають необхідне для цього кількість статистичних випробувань n.

При великій кількості досвідів n частота появи події p, отримана експериментальним шляхом, прагне до значення теоретичної ймовірності появи події P. Якщо коливання значень частоти появи подій щодо теоретичної ймовірності менше заданої точності, то експериментальну частоту приймають як відповідь, інакше генерацію випадкових вхідних впливів продовжують, і процес моделювання повторюється. При малому числі випробувань результат може виявитися недостовірним. Але чим більше випробувань, тим точніше відповідь, відповідно до центральної граничної теореми. Кількість необхідних експериментів n дані для порівняння в табл. 34.3 і табл. 34.4 при різних комбінаціях p й ε.

Таблиця 34.3.
Кількість експериментів n, необхідних для обчислення достовірної відповіді з довірчою ймовірністю $Q = 0.95$, $(F^{-1}(0.95))^2 = 3.84$, $p = 0.1$

ε	0. 001	0. 005	0. 010	0. 050	0. 100
Критична кількість експериментів n	345600	13824	3456	138	35

Таблиця 34.4.
Кількість експериментів n, необхідних для обчислення достовірної відповіді з довірчою ймовірністю $Q = 0.95$, $(F^{-1}(0.95))^2 = 3.84$, $p = 0.5$

ε	0. 001	0. 005	0. 010	0. 050	0. 100
Критична кількість експериментів n	960000	38400	9600	384	96

На мал. 34.8 відображений графік залежності $n(\varepsilon)$ при $Q = 0.95$ й $p = 0.5$.

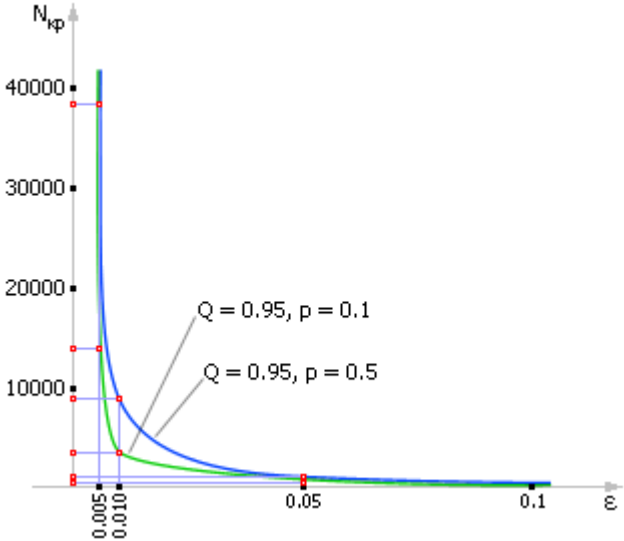

Рис. 34.8. Залежність кількості необхідних експериментів від величини довірчої ймовірності ε і довірчого інтервалу Q для випадку частоти випадання випадкової події p = 0.5

Важливо: оцінювання ведуть по гіршій із частот. Це забезпечує достовірний результат відразу по всім найма характеристикам, що, моделі.

Примітка. Варто мати на увазі, що дана оцінка кількості експериментів по ЦПТ не єдина з існуючих. Відомі аналогічні близькі за змістом оцінки Бернуллі, Муавра-Лапласа, Чебишева.

Як пояснити, чому так дивно поводиться крива знятої експериментально статистичної характеристики (див. мал. 34.7 і мал. 34.8)? При великому n крива вкрай повільно підходить до щирому значенню, хоча спочатку (при малих n) процес іде з великою швидкістю — ми швидко входимо в область наближеної відповіді (більші ε), але повільно наближаємося до точної відповіді (малі ε).

Наприклад, допустимо, що ми провели N випробувань. Випадань події в цих випробуваннях склало число N_1. Нехай імовірність випадання події близька до $N_1/N = 0.5$ або $N = N_1 \cdot 2$.

Припустимо, що ми хочемо провести ще одне випробування $(N + 1)$-е. Взявши відповідь (частота N_1/N) при N за 100%, оцінимо, наскільки відсотків зміниться відповідь після наступного досвіду? Складемо пропорцію:

N_1/N — 100%
$(N_1 + 1)/(N + 1)$ — X%

Звідси маємо: $X = (N_1 + 1) \cdot 100 \cdot N/(N_1 \cdot (N + 1))$, при $N_1 = N/2$ (імовірність 0.5) одержуємо, що $X = 100 \cdot (N + 2)/(N + 1)$.

І величина X утворить ряд: 150%, 133%, 125%, 120%, …, 100...1%, ..., ... -> 100%. Виходить, спочатку поліпшення відповіді на один додатковий експеримент склало 50%, на 2 - 33%, на 3 - 25%, на 4 - 20%, ..., на 100-м - усього на 0,1%.

Видно, що поліпшення точності на кожен новий експеримент (значення X) спочатку дуже гарне, а потім — незначне, після 100 експериментів ця величина міняється всього на частки відсотка розраховуючи на один додатковий експеримент! Підсумок: зміна оцінки, заснованої на сумі, після серії досвідів перестає сильно мінятися!!!

Ітоги. Важливо.

1. Як відповідь статистичного експерименту приймається частість p появи деякої вихідної події, що є оцінкою ймовірності. Чим більше експериментів n, тим ближче частість p до ймовірності P, а експериментальна відповідь до теоретичного.
2. Частості p, близькі за значенням до 0 або 1, більше кращі в змісті інформативності, чим частості близькі до 0.5, які мало інформативні й дають максимально непевну відповідь.
3. У моделюванні важливою метою є зниження дисперсії відповіді, розкиду вихідної величини моделі відносно частості. Дійсно, якщо розкид випадкової величини m_2 малий, те обчислена відповідь досить достовірна.

   Якщо в ряді випадкової величини зустрічаються значення досить вилучені друг від друга (див. мал. 34.2), то m_2' велика, і відповідь недостатньо визначена.
4. Статистична відповідь оцінюється не тільки значеннями частості й розкиду, але й точністю, роль якої грає **довірча ймовірність** Q і заданий довірчий інтервал ε. Ці величини пов'язані з розкидом m_2.

5. Необхідна кількість статистичних експериментів n залежить від заданої точності (Q, ε) і характеристик процесу (частості p і розкиду m_2).

   Підвищення вимог по точності, погані характеристики істотно підвищують витрати на дослідження моделі, збільшуючи число експериментів.