Тема 6. Важнейшие предельные теоремы

6.1. Независимость событий и случайных величин

Определение. *Условной вероятностью* события A при условии события B называется

$$\mathbf{P}(\mathrm{A}\,|\,\mathrm{B}) = \frac{\mathbf{P}(\mathrm{A}\bigcap\mathrm{B})}{\mathbf{P}(\mathrm{B})},$$

где предполагается, что $\mathbf{P}(\mathrm{B}) > 0.$

Пример 1. Из колоды, состоящей из 36 карт, наудачу выбирается одна карта. Рассмотрим события $\mathrm{A} = \{$вынута пика$\}$ и $\mathrm{B} = \{$вынута пика или трефа или бубна$\}$. Тогда

$$\mathbf{P}(\mathrm{A}\bigcap\mathrm{B}) = \mathbf{P}(\mathrm{A}) = 9/36 = 1/4, \qquad \mathbf{P}(\mathrm{B}) = 27/36 = 3/4, \qquad \mathbf{P}(\mathrm{A}\,|\,\mathrm{B}) = \frac{1/4}{3/4} = 1/3.$$

Определение. Событие A *не зависит* от события B, если $\mathbf{P}(\mathrm{A}\,|\,\mathrm{B}) = \mathbf{P}(\mathrm{A})$, что (при $\mathbf{P}(\mathrm{B}) > 0$) равносильно условию

$$\mathbf{P}(\mathrm{A}\bigcap\mathrm{B}) = \mathbf{P}(\mathrm{A})\cdot\mathbf{P}(\mathrm{B}),$$

которое симметрично относительно A и B. Его и будем считать определением *независимости* событий A и B, пригодным даже в случае $\mathbf{P}(\mathrm{A}) = 0$ или $\mathbf{P}(\mathrm{B}) = 0.$

Не перепутайте независимые события с несовместными, т. е. с непересекающимися множествами!

Например, событие $\mathrm{A} = \{$вынута пика$\}$ и событие $\mathrm{C} = \{$вынут туз$\}$ являются независимыми:

$$\mathbf{P}(\mathrm{A}\bigcap\mathrm{C}) = 1/36, \qquad \mathbf{P}(\mathrm{A}) = 9/36 = 1/4, \qquad \mathbf{P}(\mathrm{C}) = 4/36 = 1/9, \qquad 1/36 = (1/4)\cdot(1/9).$$

Однако если добавить в колоду джокера (карту без масти и наименования), то эти события уже станут формально зависимыми: $1/37 \neq (4/37)\cdot(9/37)$.

Теперь определим понятие независимости для случайных величин. Оно играет определяющую роль в теории вероятностей, выделяя вероятностные задачи из проблем теории меры и математического анализа. Даже существует такое шуточное определение:

теория вероятностей = ***теория меры*** + ***независимость***.

Определение. Дискретные случайные величины X и Y называются *независимыми*, если для любой пары их значений (x_i, y_j) выполняется равенство

$$\mathbf{P}(X = x_i, Y = y_j) = \mathbf{P}(X = x_i)\cdot\mathbf{P}(Y = y_j).$$

Это равенство также означает, что совместное распределение получается как произведение частных. Иначе говоря, для произвольных чисел x_i и y_j события $\{X = x_i\}$ и $\{Y = y_j\}$ являются независимыми.

Пример 2. Пусть в урне находятся m занумерованных шаров. Шары с номерами 1, 2, … , l белого цвета, а остальные шары — чёрного. Наудачу с возвращением извлекают шары. Появление белого шара назовём «успехом», чёрного — «неудачей». Тогда доля белых шаров в урне $p = l/m$ задаёт вероятность «успеха» при каждом извлечении шара. Положим $q = 1 - p.$

Определим случайную величину L_1 как число извлечений до первого «успеха» (включительно) и случайную величину L_2 как число извлечений после первого до второго «успеха» (включительно). Каждая из случайных величин может принимать счётное множество значений: 1, 2, … . Нетрудно понять, что совместное распределение случайных величин L_1 и L_2 имеет вид

$$\mathbf{P}(L_1 = i, L_2 = j) = \frac{(m-l)^{i-1} l\,(m-l)^{j-1} l}{m^{i+j}} = q^{i-1} p q^{j-1} p, \quad i, j = 1, 2, \ldots .$$

Суммируя эти вероятности по j от 1 до ∞ и учитывая, что сумма всех вероятностей геометрического распределения равна 1, находим частное распределение случайной величины L_1:

$$\mathbf{P}(L_1 = i) = \sum_{j=1}^{\infty} \mathbf{P}(L_1 = i, L_2 = j) = q^{i-1} p \sum_{j=1}^{\infty} q^{j-1} p = q^{i-1} p.$$

Аналогично получаем, что $\mathbf{P}(L_2 = j) = q^{j-1} p$. Заметим, что для произвольных i и j верно равенство

$$\mathbf{P}(L_1 = i, L_2 = j) = \mathbf{P}(L_1 = i) \cdot \mathbf{P}(L_2 = j),$$

что и доказывает независимость случайных величин L_1 и L_2.

Независимость дискретных случайных величин без труда обобщается на n-мерный случай.

Определение. Дискретные случайные величины $X_1, \ldots, X_n$ называются *независимыми*, если для произвольного набора их значений $x_1, \ldots, x_n$ выполняется равенство

$$\mathbf{P}(X_1 = x_1, \ \ldots, X_n = x_n) = \mathbf{P}(X_1 = x_1) \cdot \ \ldots \ \cdot \mathbf{P}(X_n = x_n) \equiv \prod_{i=1}^{n} \mathbf{P}(X_i = x_i).$$

Пример 3. Обобщим пример 2. Определим случайную величину L_i как число извлечений от $(i-1)$-го «успеха» (исключительно) до i-го «успеха» (включительно). Суммируя совместные вероятности как в примере 2, получим, что каждая случайная величина L_i имеет геометрическое распределение, и случайные величины $L_1, \ldots, L_n$ являются независимыми.

Пример 4. В условиях примера 2 обозначим через I_k индикатор «успеха» при k-м извлечении шара. Легко проверить, что $\mathbf{P}(I_k = 1) = p$, $\mathbf{P}(I_k = 0) = q$. Эти два равенства можно записать одной формулой:

$$\mathbf{P}(I_k = x) = p^x q^{1-x},$$

где переменная x принимает только значения 0 или 1. Случайные величины $I_1, I_2, \ldots$ называют ***испытаниями Бернулли*** с вероятностью «успеха» p или, кратко, ***схемой Бернулли***.

Докажем, что испытания Бернулли $I_1, \ldots, I_n$ являются независимыми случайными величинами. Действительно,

$$\mathbf{P}(I_1 = x_1, \ \ldots, I_n = x_n) = \frac{l^{x_1}(m-l)^{1-x_1} \ldots l^{x_n}(m-l)^{1-x_n}}{m^n} = p^{x_1} q^{1-x_1} \ldots p^{x_n} q^{1-x_n} = \prod_{i=1}^{n} \mathbf{P}(I_i = x_i).$$

Теперь дадим определение независимости случайных величин $X_1, \ldots, X_n$ в общем случае.

Определение. Компоненты случайного вектора $\boldsymbol{X} = (X_1, \ldots, X_n)$ называются *независимыми*, если для произвольных действительных чисел $x_1, \ldots, x_n$ выполняется равенство

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) = \mathbf{P}(X_1 \le x_1, \ldots, X_n \le x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \le x_i) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

Пример 5. Пусть точка $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ выбирается наудачу в n-мерном единичном кубе $[0, 1]^n$. Тогда для $0 \le x_1 \le 1, \ldots, 0 \le x_n \le 1$ совместная функция распределения $F_{\boldsymbol{Y}}(x_1, \ldots, x_n) = x_1 \cdot x_2 \cdot \ldots \cdot x_n$. При этом

$$F_{Y_i}(x_i) = F_{\boldsymbol{Y}}(1, \ldots, 1, x_i, 1, \ldots, 1) = x_i.$$

Следовательно, каждая случайная величина Y_i имеет равномерное распределение на отрезке [0, 1], и случайные величины $Y_1, \ldots, Y_n$ независимы.

Критерием независимости компонент случайного вектора $\boldsymbol{X} = (X_1, \ldots, X_n)$, имеющего плотность $f_{\boldsymbol{X}}(x_1, \ldots, x_n)$, служит выполнение для произвольных чисел $x_1, \ldots, x_n$ равенства

$$f_{\boldsymbol{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

Здесь $f_{X_i}(x)$ обозначает плотность случайной величины X_i.

Пример 6. Пусть случайные величины $T_1, \ldots, T_n$ одинаково распределены по показательному закону с параметром $\lambda > 0$ и независимы. Тогда их совместная плотность задаётся формулой

$$f_{T_1, \ldots, T_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} I_{\{x_i \ge 0\}} = \lambda^n e^{-\lambda \sum x_i} \prod_{i=1}^{n} I_{\{x_i \ge 0\}} = \lambda^n e^{-\lambda \sum x_i} I_{\{\min\{x_1, \ldots, x_n\} \ge 0\}},$$

где I_{A} обозначает индикатор множества A.

Свойства независимых случайных величин

1) Пусть $X_1, \ldots, X_n$ — независимые случайные величины, f и g — непрерывные функции от k и $n-k$ переменных соответственно. Тогда случайные величины $Y = f(X_1, \ldots, X_k)$ и $Z = g(X_{k+1}, \ldots, X_n)$ также являются независимыми.

2) Если случайные величины X и Y с конечными математическими ожиданиями независимы, то

$$\mathrm{M}XY = \mathrm{M}X \cdot \mathrm{M}Y.$$

Это свойство можно переформулировать так: ковариация независимых случайных величин равна 0. Обратное утверждение неверно (см. задачу 6.1).

3) Если случайные величины с конечными дисперсиями $X_1, \ldots, X_n$ независимы, то

$$\mathrm{D}(X_1 + \ldots + X_n) = \mathrm{D}X_1 + \ldots + \mathrm{D}X_n.$$

Это важное свойство вытекает из предыдущего свойства и свойства 3 ковариации из темы 5.

Вопрос 1. Как с помощью свойства 3 найти дисперсию суммы n испытаний Бернулли $I_1, \ldots, I_n$?

6.2. Формулы свёртки

Формулы свёртки позволяют вычислять распределение суммы независимых случайных величин. Сначала рассмотрим дискретный случай. Пусть дискретные случайные величины X и Y независимы, обозначим $Z = X + Y$. Тогда распределение случайной величины Z вычисляется с помощью ***дискретной формулы свёртки***:

$$\mathbf{P}(Z = z_k) = \sum_i \mathbf{P}(X = x_i)\, \mathbf{P}(Y = z_k - x_i). \qquad (1)$$

Доказательство. Используем счётную аддитивность вероятности и определение независимости:

$$\mathbf{P}(Z = z_k) = \sum_i \mathbf{P}(X = x_i,\ X + Y = z_k) = \sum_i \mathbf{P}(X = x_i,\ Y = z_k - x_i) = \sum_i \mathbf{P}(X = x_i)\, \mathbf{P}(Y = z_k - x_i).$$

Пример 7. Пусть I_1 и I_2 — испытания Бернулли из примера 4. Найдём распределение $Z = I_1 + I_2$:

$$p_k = \mathbf{P}(Z = k) = \sum_{i=0}^{1} \mathbf{P}(I_1 = i)\, \mathbf{P}(I_2 = k - i) = q\,\mathbf{P}(I_2 = k) + p\,\mathbf{P}(I_2 = k - 1).$$

Отсюда для возможных значений 0, 1, 2 случайной величины Z находим: $p_0 = q^2,\ p_1 = 2pq,\ p_2 = p^2$. Это распределение — частный случай для $n = 2$ биномиального закона (см. пример 2 из темы 2).

Вопрос 2. Какое распределение имеет случайная величина: а) $I_1 - I_2$; б) $I_1 + 2I_2$?

Теперь рассмотрим случай, когда случайные величины X и Y независимы и имеют плотности $f_X(x)$ и $f_Y(x)$ соответственно. Тогда $Z = X + Y$ имеет плотность, которая вычисляется согласно ***формуле свёртки для плотностей***:

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x)\, f_Y(z - x)\, dx. \qquad (2)$$

Отметим, что справа в формуле (2) стоит несобственный интеграл, зависящий от параметра z. В явном виде его удаётся взять только для некоторых простых плотностей.

Пример 8. Пусть случайные величины X и Y независимы и равномерно распределены на [0, 1]. Найдём, какими формулами задаётся плотность $Z = X + Y$ на разных промежутках действительной оси. Согласно формуле (2) имеем:

$$f_Z(z) = \int_{-\infty}^{+\infty} I_{[0,1]}(x)\, I_{[0,1]}(z - x)\, dx = \int_0^1 I_{[0,1]}(z - x)\, dx = \int_0^1 I_{[z-1,\,z]}(x)\, dx.$$

Отсюда видим, что $f_Z(z) = 0$ при $z \le 0$ и $z \ge 2$ (это понятно и без вычислений, поскольку Z принимает значения только из отрезка $[0, 2]$). Далее, для случая $0 \le z \le 1$ находим:

$$\int_0^1 I_{[z-1,\,z]}(x)\, dx = \int_0^z dx = z.$$

Наконец, для случая $1 \le z \le 2$ вычисляем:

$$\int_0^1 I_{[z-1,\,z]}(x)\, dx = \int_{z-1}^1 dx = 1 - (z - 1) = 2 - z.$$

График плотности $f_Z(z)$ имеет треугольный вид (рис. 1).

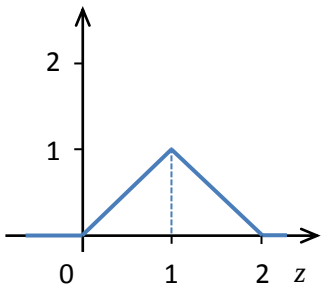

Рис. 1

6.3. Закон больших чисел

Одной из классических проблем теории вероятностей является изучение поведения при $n \to \infty$ распределения суммы $S_n = X_1 + \ldots + X_n$ независимых и одинаково распределённых слагаемых. Сложность проблемы связана с невозможностью явного вычисления n-кратных свёрток для произвольной плотности $f_X(x)$. Если слагаемые X_i положительны, то интуитивно понятно, что $S_n \to \infty$. С какой скоростью растёт сумма S_n при увеличении числа слагаемых n? В среднем она растёт как $\mathrm{M}S_n = n\,\mathrm{M}X_1$, т. е. линейно по n. Для обеспечения стабилизации разделим сумму S_n на n. Получим приближённое равенство $S_n / n \approx \mathrm{M}X_1$. В дальнейшем будем использовать стандартное обозначение $\overline{X}_n$ для *среднего арифметического* $(X_1 + \ldots + X_n)/n$. Как вместо приближённого равенства $\overline{X}_n \approx \mathrm{M}X_1$ сформулировать строгое утверждение? Ответ даёт следующая теорема.

Закон больших чисел. Пусть $X_1, X_2, \ldots$ — независимые и одинаково распределённые случайные величины с математическим ожиданием $\mu = \mathrm{M}X_1$. Тогда для любого действительного числа $\varepsilon > 0$

$$\mathbf{P}(|\overline{X}_n - \mu| > \varepsilon) \to 0 \text{ при } n \to \infty. \qquad (3)$$

Иначе говоря, распределение случайной величины $\overline{X}_n$ с ростом n «стягивается» к константе $\mu = \mathrm{M}X_1$. Утверждение (3) называется *сходимостью по вероятности* последовательности $\overline{X}_n$ к константе μ.

Вопрос 3. Положим $\sigma^2 = \mathrm{D}X_1$. Как выразить через σ стандартное отклонение $\sqrt{\mathrm{D}\overline{X}_n}$?

Предположим, что проводятся многократные независимые измерения X_i в одних и тех же условиях некоторого показателя μ со случайными ошибками E_i, т. е. $X_i = \mu + E_i$. Ошибки обычно можно рассматривать как независимые и одинаково распределённые с $\mathrm{M}E_i = 0$. По закону больших чисел при увеличении числа измерений n погрешность среднего арифметического $\Delta_n = |\overline{X}_n - \mu|$ будет по вероятности стремиться к нулю («*семь раз отмерь, один — отрежь*»).

6.4. Центральная предельная теорема

Какова типичная величина абсолютной погрешности Δ_n, если число измерений n велико? Оказывается, Δ_n имеет тот же порядок малости, что и стандартное отклонение $\sqrt{\mathrm{D}\overline{X}_n}$, т. е. $const / \sqrt{n}$. Более точный ответ даёт

Центральная предельная теорема. Пусть $X_1, X_2, \ldots$ — независимые и одинаково распределённые случайные величины с математическим ожиданием $\mu = \mathrm{M}X_1$ и дисперсией $0 < \sigma^2 = \mathrm{D}X_1 < \infty$. Тогда для произвольных действительных чисел $a < b$ имеет место сходимость при $n \to \infty$

$$\mathbf{P}(a \le \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \le b) \to \Phi(b) - \Phi(a), \text{ где } \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \qquad (4)$$

Сходимость (4) можно также записать в терминах сумм $S_n = X_1 + \ldots + X_n$:

$$\mathbf{P}\left(a \le \frac{S_n - \mu n}{\sigma\sqrt{n}} \le b \right) = \mathbf{P}\left(a \le \frac{S_n - \mathrm{M}S_n}{\sqrt{\mathrm{D}S_n}} \le b \right) \to \Phi(b) - \Phi(a).$$

Определение. $\Phi(x)$ из (4) называется *функцией распределения стандартного нормального закона*. Случайная величина Z с такой функцией распределения называется *стандартной нормальной.*

Известно, что $\Phi(x)$ нельзя выразить через элементарные функции в виде явной формулы. Её производная

$$\Phi'(x) \equiv \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

имеет форму «колокола». Величины некоторых площадей под графиком $\varphi(x)$ показаны на рис. 2.

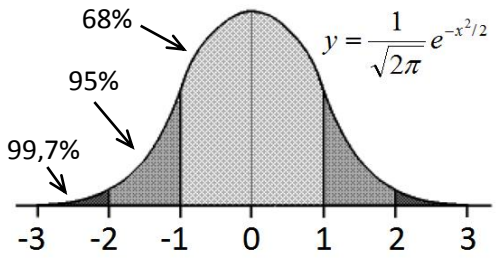

Рис. 2

Видим, что $\Phi(3) - \Phi(-3) \approx 0{,}997$, поэтому *стандартную нормальную случайную величину* Z, имеющую $\Phi(x)$ в качестве функции распределения, можно считать практически ограниченной.

Поскольку $\Phi(2) - \Phi(-2) \approx 0{,}95$, можно утверждать, что для достаточно больших n погрешность $\Delta_n = |\overline{X}_n - \mu|$ не будет превышать величину $2\sigma/\sqrt{n}$ с вероятностью приблизительно равной 0,95. Действительно, в силу центральной предельной теоремы имеем:

$$\mathbf{P}\left(\Delta_n \le 2\sigma/\sqrt{n}\right) = \mathbf{P}\left(|\overline{X}_n - \mu| \le 2\sigma/\sqrt{n}\right) = \mathbf{P}\left(-2 \le (\overline{X}_n - \mu)\sqrt{n}/\sigma \le 2\right) \to \Phi(2) - \Phi(-2) \approx 0{,}95.$$

Этот результат называют **«*правилом двух сигм*»**.

Определение. Если случайная величина Z имеет стандартное нормальное распределение, то случайная величина $X = \mu + \sigma Z$ называется *нормально распределённой с параметрами* μ *и* $\sigma > 0$ (обозн. $X \sim N(\mu, \sigma^2)$). Легко проверить, что $\mathrm{M}X = \mu$, $\mathrm{D}X = \sigma^2$, плотность $f_X(x) = \varphi((x-\mu)/\sigma)/\sigma$.

6.5. Теорема Пуассона (закон редких событий)

Напомним условия примера 1 из темы 2: «Среди m экзаменационных билетов l — лёгкие, n студентов по очереди наудачу с возвращением берут билеты». Пусть S_n — общее число лёгких билетов, вынутых всеми n студентами. В примере 2 из темы 2 было установлено, что случайная величина S_n имеет биномиальное распределение:

$$\mathbf{P}(S_n = i) = C_n^i\, p^i (1-p)^{n-i},$$ где $i = 0, 1, \ldots, n$; $p = l/m$ — доля лёгких билетов.

Как можно приближённо вычислить биномиальные вероятности $\mathbf{P}(S_n = i)$, если параметр p очень мал (скажем, $p < 0{,}01$), а параметр n, напротив, довольно велик (скажем, $n > 100$)? Проблема заключается в том, что в таком случае в биномиальном коэффициенте C_n^i будут присутствовать факториалы больших чисел, а величины p^i будут крайне малы даже при умеренных значениях i.

Пусть $\lambda = pn$. Предположим, что число λ не слишком велико (скажем, $\lambda < 10$). Тогда при фиксированном значении i для вероятности $\mathbf{P}(S_n = i)$ можно использовать приближение

$$\mathbf{P}(S_n = i) \approx \frac{\lambda^i}{i!} e^{-\lambda},$$

теоретическим обоснованием которого служит

***Теорема Пуассона*.** Если целое $i \geq 0$ фиксировано, $n \to \infty$, $p(n) \to 0$, $pn \to \lambda > 0$, то $\mathbf{P}(S_n = i) \to \frac{\lambda^i}{i!} e^{-\lambda}$.

Д о к а з а т е л ь с т в о.

$$\mathbf{P}(S_n = i) = \frac{n(n-1)\ldots(n-i+1)}{i!}\, p^i (1-p)^{n-i} = \frac{(pn)^i}{i!}(1-p)^n \cdot \left[\frac{n(n-1)\ldots(n-i+1)}{n^i}(1-p)^{-i}\right],$$

где $(1-p)^n \to e^{-\lambda}$ поскольку $n\ln(1-p) \to -\lambda$, а выражение в квадратных скобках

$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{i-1}{n}\right)(1-p)^{-i} \to 1$, так как каждый сомножитель стремится к 1.

Следующая таблица показывает точность пуассоновского приближения при n = 100 и p = 0,01:

i	0	1	2	3	4
$\mathbf{P}(S_n = i)$	0,366	0,370	0,185	0,061	0,015
$\lambda^i e^{-\lambda}/i!$	0,368	0,368	0,184	0,061	0,015

Пример 2. Вычисление страхового тарифа. Согласно статистике, накопленной страховой компанией, вероятность угона в течение года автомобиля определенной марки $p = 0{,}0123$. Из $n = 500$ страхуемых машин в следующем году предположительно будет угнано $\lambda = pn = 6{,}15$ автомобилей. Поскольку вероятность отдельного угона мала, а угоны предполагаются независимыми, то можно использовать пуассоновское приближение для распределения *числа ожидаемых угонов* K.

Проблема состоит в определении величины *страхового процента* δ, компенсирующего убытки в случае неожиданно большого числа угонов. Пусть C — *средняя цена автомобиля* данной марки, и в случае угона она выплачивается владельцу полностью. Тогда доход страховой компании равен δCn, а *страховые выплаты* составят сумму CK.

Зададим *коэффициент значимости* α — малую вероятность, которой мы готовы пренебречь (скажем, положим $\alpha = 0{,}05$), и вычислим значение δ из условия

$$\mathrm{P}(CK > \delta Cn) = \mathrm{P}(K > \delta n) = \alpha.$$

В частности, для приведённых выше значений p и n, используя встроенную в программу Excel функцию ПУАССОН (POISSON), находим, что $\mathrm{P}(K > 10) \approx 0{,}05$. Отсюда $\delta n = 10$ или $\delta = 10/500 = 2\%$.

6.6. Дивергенция Кульбака — Лейблера, перекрестная энтропия, взаимная информация

В заключение познакомимся с некоторыми полезными терминами из теории информации. В статье Kullback S., Leibler R. «*On Information and Sufficiency*» (1951) было предложено важное понятие, характеризующее степень различия между двумя вероятностными распределениями.

Определение. *Дивергенция Кульбака — Лейблера* между набором вероятностей $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$, $\sum p_i = 1$, и набором вероятностей $\boldsymbol{q} = (q_1, q_2, \ldots, q_n)$, $\sum q_i = 1$, задаётся следующей формулой:

$$D(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i}. \tag{5}$$

Дивергенцию Кульбака — Лейблера также называют *относительной энтропией* (*relative entropy*). (Напомним, что обычная энтропия $H(\boldsymbol{p})$ была определена ранее в разделе 2.5 темы 2.)

Строго говоря, $D(\boldsymbol{p}, \boldsymbol{q})$ не является расстоянием между $\boldsymbol{p}$ и $\boldsymbol{q}$. Во-первых, дивергенция не симметрична: вообще говоря, $D(\boldsymbol{p}, \boldsymbol{q}) \neq D(\boldsymbol{q}, \boldsymbol{p})$. Во-вторых, для неё не выполнятся *неравенство треугольника*: $D(\boldsymbol{p}_1, \boldsymbol{p}_2) + D(\boldsymbol{p}_2, \boldsymbol{p}_3)$ может быть меньше, чем $D(\boldsymbol{p}_1, \boldsymbol{p}_3)$. Однако всегда $D(\boldsymbol{p}, \boldsymbol{q}) \geq 0$, причём можно доказать, что $D(\boldsymbol{p}, \boldsymbol{q}) = 0$ тогда и только тогда, когда $\boldsymbol{p}$ и $\boldsymbol{q}$ совпадают.

Определение. *Перекрёстной энтропией* (*cross-entropy*) между $\boldsymbol{p}$ и $\boldsymbol{q}$ называется

$$H(\boldsymbol{p}, \boldsymbol{q}) = -\sum_{i=1}^{n} p_i \log_2 q_i. \tag{6}$$

Из определения энтропии $H(\boldsymbol{p})$ и формулы (5) вытекает тождество

$$H(\boldsymbol{p}, \boldsymbol{q}) = H(\boldsymbol{p}) + D(\boldsymbol{p}, \boldsymbol{q}).$$

Из него следует, что если требуется найти в некотором классе моделей набор вероятностей $\boldsymbol{q}$, ближайший в смысле дивергенции Кульбака — Лейблера к заданному набору вероятностей $\boldsymbol{p}$, то достаточно будет минимизировать функцию $H(\boldsymbol{p}, \boldsymbol{q})$ по аргументу $\boldsymbol{q}$.

Определение. *Взаимной информацией* (*mutual information*) между дискретными случайными величинами X и Y называется

$$I(X, Y) = \sum_i \sum_j r_{ij} \log_2 \frac{r_{ij}}{p_i q_j}. \tag{7}$$

Здесь $r_{ij} = \mathbf{P}(X = x_i, Y = y_j)$, $p_i = \mathbf{P}(X = x_i) = \sum_j p_{ij}$, $q_j = \mathbf{P}(Y = y_j) = \sum_i p_{ij}$.

Таким образом, $I(X,Y)$ представляет собой дивергенцию Кульбака — Лейблера между совместным распределением случайных величин X, Y и совместным распределением независимых случайных величин $\tilde{X}, \tilde{Y}$ с теми же самыми частными распределениями. Иначе говоря, $I(X,Y)$ — мера зависимости X и Y. В отличие от ковариации взаимная информация используется как характеристика не только линейной, но и произвольной связи случайных величин. Согласно определению $I(X,Y)=0$ для независимых случайных величин. Нетрудно убедиться, что $I(X,X)=H(\boldsymbol{p})$, поэтому энтропию $H(\boldsymbol{p})$ также называют *self-information*.

Понятие взаимной информации легко обобщается на случай нескольких случайных величин. В частности для трёх случайных величин X, Y, Z

$$I(X,Y,Z)=\sum_{i,j,k} s_{ijk}\log_2\frac{s_{ijk}}{p_i q_j r_k},$$

где $s_{ijk}=\mathbf{P}(X=x_i, Y=y_j, Z=z_k)$, $p_i=\mathbf{P}(X=x_i)=\sum_{j,k} s_{ijk}$, $q_j=\mathbf{P}(Y=y_j)=\sum_{i,k} s_{ijk}$, $r_k=\mathbf{P}(Z=z_k)=\sum_{i,j} s_{ijk}$.

Приведём пример использования понятия энтропии в теории кодирования информации и лингвистике из книги Лагутин М. Б. «Наглядная математическая статистика» (с. 173):

> **При передаче сообщений по каналу связи их необходимо записать в «двоичном коде». Если используется алфавит из N символов (букв), то для кодировки каждого символа потребуется (с точностью до 1) $\log_2 N$ «двоичных» символов 0 и 1. Например, для передачи текста на русском языке, состоящего из букв и пробелов, можно (при объединении «ь» и «ъ») каждый символ закодировать последовательностью из 0 и 1 длины $\log_2 32 = 5$. Для передачи текста из n символов алфавита понадобится код длины $n\log_2 N$.**
>
> **Для больших по объему сообщений можно существенно уменьшить эту величину, используя то, что разные символы алфавита встречаются в тексте с различными частотами (см. таблицу).**

—	о	е, ё	а	и	т	н	с
0,175	0,090	0,072	0,062	0,062	0,053	0,053	0,045
р	в	л	к	м	д	п	у
0,040	0,038	0,035	0,028	0,026	0,025	0,023	0,021
я	ы	з	ь, ъ	б	г	ч	й
0,018	0,016	0,016	0,014	0,014	0,013	0,012	0,010
х	ж	ю	ш	ц	щ	э	ф
0,009	0,007	0,006	0,006	0,004	0,003	0,003	0,002

«Рассказывают, что, создавая свой код, Морзе отправился в ближайшую типографию и подсчитал число литер в наборных кассах. Буквам и знакам, для которых литер в этих кассах было припасено больше, он сопоставил более короткие кодовые обозначения».

Аршинов М. Н., Садовский А. Е. «Коды и математика»

Если $p_1, \ldots, p_N$ — вероятности их появления, то в силу устойчивости частот среди сообщений длины n практически будут встречаться лишь сообщения, в которых каждый i-й символ алфавита будет появляться $\nu_i \approx np_i$ раз. Уточним это утверждение.

Допустим, что каждый символ сообщения появляется независимо от других с соответствующей вероятностью p_i. Для $\delta > 0$ обозначим через $A_{n,\delta}$ множество тех сообщений, у которых $\{|\nu_i - np_i| \leqslant \delta,\ i = 1, \ldots, N\}$. Их станем называть *типичными*, так как в силу закона больших чисел

$$\mathbf{P}\,(A_{n,\delta}) \geqslant 1 - \sum_{i=1}^{N} \mathbf{P}\left(\left|\frac{\nu_i}{n} - p_i\right| > \delta\right) \to 1 \quad \text{при } n \to \infty.$$

Пусть $M_{n,\delta}$ обозначает число «типичных» сообщений. При условии, что все $p_i > 0$ и $0 < \delta < 1$ из *теоремы Макмиллана* (см. [90, с. 64]) следует, что $\dfrac{1}{n}\log_2 M_{n,\delta}$ стремится к энтропии $H = -\sum p_i \log_2 p_i$ при $n \to \infty$.

Другими словами, число «типичных» сообщений не превосходит $2^{n(H+\varepsilon)}$, где $\varepsilon > 0$ сколь угодно мало. Каждому такому сообщению можно присвоить порядковый номер, для записи которого потребуется $n(H + \varepsilon)$ «двоичных» символов, и вместо сообщения передавать эту запись. Тем самым, с вероятностью близкой к 1, осуществляется сокращение длины сообщений с *коэффициентом сжатия* $\gamma_1 = H_1/H_0 \leqslant 1$, где $H_0 = \log_2 N$ и $H_1 = H$. Для русского алфавита на основе приведённой выше таблицы имеем $H \approx 4{,}35$, $\gamma_1 \approx 0{,}87$ (см. [92, с. 238]).

Для независимо появляющихся символов *невозможно* предложить способ кодирования (бесконечно большого текста), который давал бы большую экономию, чем γ_1 (см. [92, с. 200]). Однако, символы текста на русском языке, очевидно, зависимы: если очередная буквой является гласной, то следующая вероятнее всего окажется согласной; «ь» не может следовать ни за пробелом, ни за гласной; за буквой «и» после пробела часто следует еще один пробел; после сочетания «тс» естественно ожидать букву «я» и т. п. Эти наблюдения подсказывают разбить текст на блоки длины k и считать эти блоки символами нового алфавита.

Для подсчета частот двухбуквенных и трехбуквенных сочетаний Д. С. Лебедев и В. А. Гармаш использовали отрывок из романа «Война и мир» Л. Н. Толстого, содержащий около 30 000 букв (см. [92, с. 246]). На основе полученных данных были получены оценки соответствующих энтропий: $H_2 \approx 7{,}9$, $H_3 \approx 10{,}9$, что приводит к коэффициентам сжатия $\gamma_2 = H_2/(2H_0) \approx 0{,}79$ и $\gamma_3 = H_3/(3H_0) \approx 0{,}73$. Согласно [92, с. 245] коэффициент сжатия (бесконечно большого) текста не может быть меньше, чем $\gamma_\infty = \lim\limits_{k\to\infty} H_k/(kH_0)$. Лингвист Р. Г. Пиотровский (см. [92, с. 268]) оценил γ_∞ русских литературных текстов как 0,24, а деловых текстов — как 0,17.

К. Шеннон назвал величину $1 - \gamma_\infty$ *избыточностью языка*. Во многих случаях она полезна тем, что позволяет выявлять опечатки и восстанавливать пропуски. (О *кодах Хемминга*, умеющих исправлять подобные ошибки, можно почитать в [91, с. 288] или [92, с. 392].) Последовательность независимых *равновероятных*

символов, имеющая энтропию $H = \log_2 N$, несократима. Поскольку сильно сжатый текст похож на нее, практически невозможно восстановить в нем пропущенный или искаженный символ. Это обстоятельство нередко приводит к потере архивированных данных при возникновении дефектов на диске.

Литература

90. *Ширяев А. Н.* Вероятность. — М.: Наука, 1989
91. *Яблонский С. В.* Введение в дискретную математику. — М.: Наука, 1986
92. *Яглом А. М., Яглом И. М.* Вероятность и информация. — М.: Физматгиз, 1960

Задачи для решения на занятии

1) Из урны, содержащей 4 занумерованных шара, извлекают наудачу 2 шара. Пусть N_i — номер i-го шара, i = 1, 2. Являются ли случайные величины N_1 и N_2 независимыми, если шары извлекаются: а) с возвращением; б) без возвращения?

2) Доказать, что если событие A не зависит от события B, то:
а) A не зависит от $\overline{B}$; б) $\overline{A}$ не зависит от $\overline{B}$. (*Указание*. Используйте аддитивность вероятности.)

3) Пусть $I_1, I_2, \ldots$ — испытания Бернулли с вероятностью «успеха» p. Какое распределение имеет случайная величина: а) $X = 4I_1 + I_2$; б) $Y = 3I_1 + I_2$? Постройте график функции распределения.

Домашнее задание

Если первая буква вашей фамилии находится в диапазоне:

«**А – Е**», то «своими» являются задачи 6.1 и 6.5;
«**Ж – М**», то «своими» являются задачи 6.2 и 6.6;
«**Н – Р**», то «своими» являются задачи 6.3 и 6.7;
«**С – Я**», то «своими» являются задачи 6.4 и 6.8.

6.1) Пусть случайная величина X принимает значения 0, $\pi/2$, π с вероятностями 1/3 каждое. Доказать, что случайные величины $Y = \sin X$ и $Z = \cos X$ некоррелированы, но зависимы.

6.2) Найти формулу плотности суммы двух независимых и одинаково распределённых показательных случайных величин с параметром λ. (*Указание*. Используйте формулу (2).)

6.3) Пусть случайные величины N_1 и N_2 независимы и имеют пуассоновские распределения с параметрами λ и μ соответственно. Найти без знака суммы распределение случайной величины $N_1 + N_2$. (*Указание*. Используйте формулу (1) и бином Ньютона.)

6.4) Доказать свойство 2 независимых случайных величин в дискретном случае.

6.5) Пусть p = 0,002 — вероятность того, что изделие окажется бракованным. Найти приближённо с помощью теоремы Пуассона вероятность, что в партии из 1000 изделий будет не более трёх бракованных. (*Указание*. Просуммируйте вероятности и примените к ним теорему Пуассона.)

6.6) На фабрике 100 станков. С вероятностью 0,2 станок находится в ремонте. Из расчета на какое количество станков надо подавать энергию, чтобы с вероятностью 0,975 все исправные станки могли работать? (*Указание*. Примените центральную предельную теорему.)

6.7) Случайные величины T_1 и T_2 имеют одинаковое показательное распределение с параметром λ. Найти формулу и построить график плотности случайной величины $T_1 - T_2$. (*Указание*. Найдите формулу для плотности случайной величины $-T_2$ и используйте формулу (2). Рассмотрите два случая: $z < 0$ и $z \geq 0$.)

6.8) Случайные величины Y_1 и Y_2 имеют равномерное распределение на отрезке [0, 1] и независимы. Найти формулу и построить график плотности случайной величины $Y_1 - Y_2$. (*Указание*. Найдите формулу для плотности случайной величины $-Y_2$ и используйте формулу (2). Рассмотрите два случая: $-1 \leq z \leq 0$ и $0 \leq z \leq 1$.)