Комментарии к теме «*Слабая сходимость распределений*»

Практические занятия по теории вероятностей

кафедра статистического моделирования `http://statmod.ru`, матмех СПбГУ, 2018 г.

1 Определение и основные свойства

1.1 Общая постановка задачи

Пусть $\mathcal{P}_1$ и $\mathcal{P}_2$ — два распределения в $\mathbb{R}^d$ с борелевской σ-алгеброй $\mathcal{B}_d$. Интуитивно ясно, что какие-то $\mathcal{P}_1$ и $\mathcal{P}_2$ «похожи» друг на друга, а какие-то — совсем не похожи. Например, при маленьких ε равномерное распределение $\mathrm{U}(-\varepsilon,\varepsilon)$ на отрезке $[-\varepsilon,\varepsilon]$ «похоже» на распределение δ_0, сосредоточенное в нуле, распределение $\mathrm{U}(0,1+\varepsilon)$ близко к $\mathrm{U}(0,1)$, а распределение $\mathrm{U}(0,1)$ сильно отличается от $\mathrm{Exp}(\lambda)$ при любом $\lambda>0$.

Хочется формализовать эти понятия — «похоже» или «не похоже». Стандартный математический путь для этого — задание на множестве интересующих нас объектов метрики, измеряющей расстояние между объектами. Здесь, однако, не все так просто.

Даже в $\mathbb{R}$, где вроде бы есть единственная «естественная» метрика $\rho(x,y)=|x-y|$, иногда удобно использовать альтернативную метрику[1]

$$\rho_1(x,y)=\frac{|x-y|}{1+|x-y|}\,,$$

которая эквивалентна ρ в том смысле, что $x_n\xrightarrow{\rho_1}x$ тогда и только тогда, когда $x_n\xrightarrow{\rho}x$, но обладает тем свойством, что $\rho_1(x,y)<1$ при любых $x,y\in\mathbb{R}$. А можно рассматривать, например, дискретную метрику

$$\rho_2(x,y)=\begin{cases}1 & x\neq y,\\ 0 & x=y,\end{cases}$$

которая, конечно, уже не эквивалентна ρ.

При переходе к $\mathbb{R}^d$ ситуация еще усложняется. Скажем, в $\mathbb{R}^2$ можно задать уже как минимум 3 «естественные» (и эквивалентные) метрики — евклидову $\rho^{(2)}$, равномерную $\rho^{(\infty)}$ и метрику $\rho^{(1)}$:

$$\rho^{(2)}(\mathbf{x},\mathbf{y})=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2},\quad \rho^{(\infty)}(\mathbf{x},\mathbf{y})=\max(|x_1-x_2|,|y_1-y_2|),$$
$$\rho^{(1)}(\mathbf{x},\mathbf{y})=|x_1-x_2|+|y_1-y_2|,\quad \text{где}\ \ \mathbf{x}=(x_1,x_2),\ \mathbf{y}=(y_1,y_2).$$

В зависимости от обстоятельств оказывается удобно использовать одну или другую из них.

Итак, задание «естественной» метрики на некотором множестве объектов — задача не простая и не однозначная. В нашем случае она осложняется тем, что интересующие нас объекты сами по себе сложны — это распределения (то есть некоторые функции), имеющие областью определения борелевские подмножества $\mathbb{R}^d$.

Пожалуй, самое простое в этой ситуации — определить равномерную метрику на множестве Π_d распределений в $\mathbb{R}^d$ аналогично тому как определяется расстояние между двумя непрерывными функциями в пространстве $\mathbb{C}(0,1)$. Введем соответствующую терминологию и дадим соответствующее определение.

[1] Докажите, что это метрика.

1.2 Расстояние по вариации

Определение 1.1. (*Расстояние по вариации). Пусть* Π_d *— множество распределений, определенных на борелевской* σ*-алгебре* $\mathcal{B}_d$ *подмножеств* $\mathbb{R}^d$ *и* $\mathcal{P}, \mathcal{Q} \in \Pi_d$*. Расстоянием по вариации между* $\mathcal{P}$ *и* $\mathcal{Q}$ *называется число*

$$\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q}) = \sup_{B \in \mathcal{B}_d} \big|\mathcal{P}(B) - \mathcal{Q}(B)\big|. \tag{1.1}$$

Говорят, что последовательность $\mathcal{P}_n \in \Pi_d$ *сходится по вариации*[2] *к* $\mathcal{P} \in \Pi_d$*, если* $\rho_{\mathrm{var}}(\mathcal{P}_n, \mathcal{P}) \to 0$ *при* $n \to \infty$*. Этот факт кратко записывается как* $\mathcal{P}_n \overset{\mathrm{var}}{\to} \mathcal{P}$*.*

Замечание 1.1. 1. Формула (1.1) задает метрику в множестве Π_d.[3] 2. Ясно, что $\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q}) \le 1$.

Какими свойствами обладает расстояние по вариации? Один из ответов на этот вопрос дает теорема Шеффе.[4]

Теорема 1.1. (*Теорема Шеффе).*
1a. Если распределения $\mathcal{P}$ *и* $\mathcal{Q}$ *в* $\mathbb{R}^d$ *абсолютно непрерывны с плотностями* p *и* q*, то*

$$\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| dx. \tag{1.2}$$

1b. Если распределения $\mathcal{P}$ *и* $\mathcal{Q}$ *дискретны, сосредоточены на одном и том же не более чем счетном множестве и имеют таблицы распределений*

$$\mathcal{P}: \quad \begin{pmatrix} x_1 & \dots & x_\ell & \dots \\ p_1 & \dots & p_\ell & \dots \end{pmatrix} \quad \text{и} \quad \mathcal{Q}: \quad \begin{pmatrix} x_1 & \dots & x_\ell & \dots \\ q_1 & \dots & q_\ell & \dots \end{pmatrix},$$

то

$$\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2} \sum_\ell |p_\ell - q_\ell|. \tag{1.3}$$

2a. Пусть $\mathcal{P}_n$*,* $\mathcal{P}$ *абсолютно непрерывные распределения в* $\mathbb{R}^d$ *с плотностями* p_n *и* p*. Если* $p_n \to p$ *почти всюду по мере Лебега, то* $\mathcal{P}_n \overset{\mathrm{var}}{\to} \mathcal{P}$*.*
2b. Если $\mathcal{P}_n$*,* $\mathcal{P}$ *— дискретные распределения*[5] *с вероятностями* $p_\ell^{(n)}$ *и* p_ℓ *соответственно. Если* $p_\ell^{(n)} \to p_\ell$ *при всех* ℓ*, то* $\mathcal{P}_n \overset{\mathrm{var}}{\to} \mathcal{P}$*.*

Первое утверждение теоремы Шеффе показывает, что для абсолютно непрерывных распределений расстояние по вариации имеет очень простой вид: с точностью до множителя $1/2$ это всего лишь расстояние в $\mathbb{L}_1$ между плотностями соответствующих распределений.

С другой стороны, формула (1.2) дает возможность вычислять расстояние по вариации. Например, нетрудно посчитать, что расстояние по вариации между распределениями $\mathrm{U}(0, 1+\varepsilon)$ и $\mathrm{U}(0,1)$ при положительных[6] ε равно $\varepsilon/(1+\varepsilon)$, в время как $\rho_{\mathrm{var}}\big(\mathrm{U}(0,1), \mathrm{Exp}(\lambda)\big) = e^{-\lambda}$ при $\lambda \le 1$.[7]

Второе утверждение теоремы Шеффе позволяет доказывать (при благоприятных обстоятельствах) сам факт сходимости по вариации. Например, из того, что плотность

$$s_n(x) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\,\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

[2] В другой терминологии — *сильно сходится.*
[3] Докажите это.
[4] Здесь она формулируется отдельно для абсолютно непрерывных и дискретных распределений в $\mathbb{R}^d$. Более общий случай см. в [1, с. 306] или в [2, гл. III §9 лемма 2]
[5] Сосредоточенные на одном множестве и том же не более чем счетном множестве.
[6] А при отрицательных?
[7] А при $\lambda > 1$?

распределения Стьюдента S_n с параметром n (так называемым «числом степеней свободы») поточечно стремится[8] при $n \to \infty$ к плотности распределения стандартного нормального закона сразу же следует, что $\mathrm{S}_n \overset{\mathrm{var}}{\to} \mathrm{N}(0,1)$.[9]

Все это, несомненно, весьма привлекательные свойства. Рассмотрим, однако, расстояние по вариации (1.1) между распределениями $\mathcal{P} = \mathcal{P}_\varepsilon = \mathrm{U}(-\varepsilon, \varepsilon)$ и $\mathcal{Q} = \delta_0$. Оказывается, оно равно 1 для любого положительного ε. Действительно, взяв $B = \{0\}$, получим, что $\mathcal{P}_\varepsilon(B) = 0$, а $\mathcal{Q}(B) = 1$. Это выглядит странно: при $\varepsilon \downarrow 0$ распределения $\mathrm{U}(-\varepsilon, \varepsilon)$ все больше и больше концентрируются около нуля, а расстояние между этими распределениями и δ_0 остается равным 1!

Более того, расстояние по вариации между любым абсолютно непрерывным и любым дискретным распределением тоже равно единице (достаточно в качестве B взять носитель дискретного распределения).

И еще более того: если взять $\mathcal{P} = \mathcal{P}_\varepsilon = \delta_\varepsilon$ и $\mathcal{Q} = \delta_0$, то снова окажется, что $\rho_{\mathrm{var}}(\mathcal{P}_\varepsilon, \mathcal{Q}) = 1$ для любого сколь угодно малого $\varepsilon > 0$. Таким образом получается, что случайные величины $\xi_\varepsilon \equiv \varepsilon$ сходятся[10] при $\varepsilon \to 0$ к случайной величине $\xi \equiv 0$, а их распределения не сходятся.[11]

И еще один пример. Рассмотрим два дискретных равномерных распределения

$$\mathcal{P}_n : \begin{pmatrix} 0 & \dots & k/n & \dots & (n-1)/n \\ 1/n & \dots & 1/n & \dots & 1/n \end{pmatrix} \text{ и } \mathcal{Q}_{n-1} : \begin{pmatrix} 1/(n-1) & \dots & k/(n-1) & \dots & 1 \\ 1/(n-1) & \dots & 1/(n-1) & \dots & 1/(n-1) \end{pmatrix}.$$

При больших n оба распределения $\mathcal{P}_n$ и $\mathcal{Q}_n$ очень похожи на равномерное распределение $\mathrm{U}(0,1)$, в то время как $\rho_{\mathrm{var}}(\mathcal{P}_n, \mathcal{Q}_n) = 1$ для любого n.

Все это представляется неестественным и, следовательно, метрика Определения 1.1 является мало пригодной для описания близости распределений на всем множестве Π_d. Попробуем понять, как ее можно модифицировать, чтобы ликвидировать описанные неприятности.

Оказывается, имеет место следующее утверждение ([2, гл.III §9 лемма 1]).

Предложение 1.1.

$$\sup_{B \in \mathcal{B}_d} \left|\mathcal{P}(B) - \mathcal{Q}(B)\right| = \frac{1}{2} \sup_{|f| \le 1} \left| \int_{\mathbb{R}^d} f d\mathcal{P} - \int_{\mathbb{R}^d} f d\mathcal{Q} \right|, \tag{1.4}$$

где супремум берется по всем $\mathcal{B}_d$-измеримым функциям f, по модулю меньшим 1.[12]

Равенство (1.4) наводит на следующие соображения. Поскольку метрика $\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q})$ является слишком «жесткой» (она приписывает слишком большое отличие похожим распределениям), ее хочется «смягчить». Сделать это можно, сузив класс функций, по которым берется супремум в интегральном представлении (1.4) расстояния $\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q})$.

Как это естественно сделать? В множество рассматриваемых в (1.4) функций входят функции, сильно меняющиеся при любом сколь угодно малом изменении аргумента. Например, туда входят индикаторы измеримых множеств. Но именно индикаторы (которые фактически стоят в определении метрики $\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q})$) и доставили нам все неприятности! Давайте же наложим на множество функций ограничения, не позволяющие им сильно меняться при небольших изменениях аргументов!

Приведем формализацию этих соображений.

[8] Проверьте!

[9] В практической статистике принято считать, что при $n > 30$ различиями между распределениями S_n и $\mathrm{N}(0,1)$ можно пренебречь. Интересно, как такой выбор n описывается в терминах расстояния по вариации между этими распределениями?

[10] В любом разумном смысле.

[11] В этом примере расстояние по вариации аналогично дискретной метрике в $\mathbb{R}$.

[12] Исходя из равенства (1.4), за определение расстояния по вариации иногда берут $\sup_{|f| \le 1} \left|\int_{\mathbb{R}^d} f d\mathcal{P} - \int_{\mathbb{R}^d} f d\mathcal{Q}\right|$, то есть $2 \sup_{B \in \mathcal{B}_d} \left|\mathcal{P}(B) - \mathcal{Q}(B)\right|$.

Определение 1.2. (BL-метрика). Пусть Π_d — множество распределений, определенных на борелевской σ-алгебре $\mathcal{B}_d$ подмножеств $\mathbb{R}^d$ и $\mathcal{P}, \mathcal{Q} \in \Pi_d$. Тогда BL-*расстоянием* между $\mathcal{P}$ и $\mathcal{Q}$ называется число

$$\rho_{\mathrm{BL}}(\mathcal{P}, \mathcal{Q}) = \sup_{f \in \mathrm{BL}(1,1)} \left| \int_{\mathbb{R}^d} f d\mathcal{P} - \int_{\mathbb{R}^d} f d\mathcal{Q} \right|, \tag{1.5}$$

где $\mathrm{BL}(M, K)$ — множество $\mathcal{B}_d$-измеримых функций f, удовлетворяющих условию $|f| \leq M$ и условию Липшица с постоянной K.

Замечание 1.2. 1. Легко видеть, что формула (1.5) действительно определяет метрику.
2. Поскольку

$$\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q}) \leq 2\rho_{\mathrm{BL}}(\mathcal{P}, \mathcal{Q}),$$

то из сходимости по вариации следует сходимость распределений в BL-метрике. Далее мы увидим, что обратное, вообще говоря, неверно.

К сожалению, для BL-метрики не существует конструктивного представления в стиле теоремы Шеффе. Однако, если рассматривать не саму метрику $\rho_{\mathrm{var}}(\mathcal{P}, \mathcal{Q})$, а сходимость по этой метрике, то ситуация становится более оптимистичной. А именно, имеет место следующая теорема.

Обозначим $\mathbb{C}_b(\mathbb{R}^d)$ множество непрерывных ограниченных функций, определенных на $\mathbb{R}^d$.

Теорема 1.2. *Пусть распределения $\mathcal{P}$ и $\mathcal{P}_n\,(n \geq 1)$ определены на измеримом пространстве $(\mathbb{R}^d, \mathcal{B}_d)$. Тогда сходимость $\rho_{BL}(\mathcal{P}_n, \mathcal{P}) \to 0$ эквивалентна тому, что*

$$\int_{\mathbb{R}^d} f d\mathcal{P}_n \to \int_{\mathbb{R}^d} f d\mathcal{P} \tag{1.6}$$

для любой $f \in \mathbb{C}_b(\mathbb{R}^d)$.

Доказательство Теоремы 1.2 достаточно сложно[13] (в отличие от Теоремы 1.1 и Предложения 1.1), и поэтому саму сходимость (1.6) обычно берут в качестве определения, а метрику ρ_{BL} рассматривают только при доказательстве теоретических результатов. Начиная с этого места мы будем поступать именно таким образом.

1.3 Слабая сходимость

Определение 1.3. (*Слабая сходимость распределений). Говорят, что последовательность распределений $\mathcal{P}_n$, заданных на измеримом пространстве $(\mathbb{R}^d, \mathcal{B}_d)$, слабо сходится к распределению $\mathcal{P}$, если* (1.6) *выполняется для любой функции $f \in \mathbb{C}_b(\mathbb{R}^d)$.*

Факт слабой сходимости $\mathcal{P}_n$ к $\mathcal{P}$ записывается в виде $\mathcal{P}_n \Rightarrow \mathcal{P}$.[14]

Замечание 1.3. Из Теоремы 1.2, конечно же, следует, что сходимости $\mathcal{P}_n \Rightarrow \mathcal{P}$ и $\mathcal{P}_n \Rightarrow \mathcal{Q}$ влекут за собой равенство $\mathcal{P} = \mathcal{Q}$, то есть единственность слабого предела. Этот факт, однако, можно доказать и непосредственно, пользуясь только Определением 1.3.

Суммируем некоторые важные свойства слабой сходимости в следующем предложении. Рассмотрим последовательность распределений $\mathcal{P}_n$ и распределение $\mathcal{P}$, определенные на борелевской σ-алгебре $\mathcal{B}_d$ подмножеств $\mathbb{R}^d$. Обозначим соответствующие характеристические функции φ_n, φ и функции распределения — F_n, F.

[13] С ним можно ознакомиться в [2, гл.III §7].

[14] Иногда пишут $\mathcal{P}_n \xrightarrow{w} \mathcal{P}$. Используются и другие обозначения.

Предложение 1.2. *Эквивалентны следующие утверждения.*
1. $\mathcal{P}_n \Rightarrow \mathcal{P}$.
2. $F_n(x) \to F(x)$ *для любой точки* $x \in \mathbb{R}^d$*, являющейся точкой непрерывности функции* F.
3. $\varphi_n(t) \to \varphi(t)$ *для любого* $t \in \mathbb{R}^d$.
4. $\mathcal{P}_n(B) \to \mathcal{P}(B)$ *при любом* $B \in \mathcal{B}_d$ *таком, что* $\mathcal{P}(\partial B) = 0$*, где* ∂B *— граница множества* B.

Замечание 1.4. 1. Второй и третий пункт Предложения 1.2 дают возможность проверять слабую сходимость конкретной последовательности распределений.

Что касается четвертого пункта, он носит скорее теоретический характер. Например, сравнение этого пункта с определением сходимости по вариации еще раз показывает, что из сильной сходимости распределений следует слабая сходимость.

Действительно, согласно определению сходимости по вариации сходимость $\mathcal{P}_n(B) \to \mathcal{P}(B)$ должна выполняться[15] для любых измеримых B, в то время как при слабой сходимости на множества B накладываются ограничения.
2. В части достаточности утверждение об эквивалентности слабой сходимости и поточечной сходимости характеристических функций может быть усилено. А именно, можно доказать, что если $\varphi_n(t) \to \psi(t)$ для любого $t \in \mathbb{R}^d$ и ψ — непрерывная в нуле функция, то ψ является характеристической функцией некоторого распределения $\mathcal{Q}$ и, следовательно, $\mathcal{P}_n \Rightarrow \mathcal{Q}$.

Пример 1.1. Обсудим результат Предложения 1.2 на простейшем примере. Пусть $d = 1$, $a_n, a \in \mathbb{R}$, $a_n \neq a$ и $a_n \to a$. Положим $\mathcal{P}_n = \delta_{a_n}$ и $\mathcal{P} = \delta_a$.

1. Для начала заметим, что, поскольку $a_n \neq a$, то $\rho_{\text{var}}(\mathcal{P}_n, \mathcal{P}) = 1$ и $\mathcal{P}_n$ не сходится к $\mathcal{P}$ по вариации. В то же время для любой непрерывной ограниченной функции f
$$\int_{\mathbb{R}^d} f d\mathcal{P}_n = f(a_n) \to f(a) = \int_{\mathbb{R}^d} f d\mathcal{P},$$
то есть $\mathcal{P}_n \Rightarrow \mathcal{P}$ по определению.

2. В терминах характеристических функций доказательство слабой сходимости тоже просто:
$$\varphi_n(t) = e^{ita_n} \xrightarrow[n\to\infty]{} e^{ita} = \varphi(t).$$

3. С функциями распределения все похитрее, так как там присутствует условие на множество точек x, для которых должна быть сходимость. Рассмотрим 2 случая: $a_n \uparrow a$ и $a_n \downarrow a$. Так как
$$F_n(x) = \begin{cases} 0 & \text{при } x \le a_n, \\ 1 & \text{при } x > a_n \end{cases} \quad \text{и} \quad F_n(x) = \begin{cases} 0 & \text{при } x \le a, \\ 1 & \text{при } x > a, \end{cases}$$
то при $a_n \downarrow a$ имеет место сходимость $F_n(x) \to F(x)$ для любого $x \in \mathbb{R}$, а при $a_n \uparrow a$
$$F_n(x) \to F_*(x) = \begin{cases} 0 & \text{при } x < a, \\ 1 & \text{при } x \ge a, \end{cases}$$
что совпадет с $F(x)$ во всех точках, кроме точки $x = a$, которая как раз и является единственной точкой разрыва предельной функции распределения F.

   Если же взять такую последовательность a_n, что $a_n \downarrow a$ при четных n и $a_n \uparrow a$ при нечетных n, то, очевидно, $F_n(x) \to F(x)$ при $x \neq a$, в то время как последовательность $F_n(a)$ предела не имеет.

[15] Да еще и равномерно по B.

Этот пример показывает, что (при наличии слабой сходимости распределений) в точках разрыва предельной функции распределения со сходимостью функций распределения может быть все, что угодно. Если же предельная функция распределения непрерывна, то допредельные функции распределения сходятся к ней поточечно.[16]

4. Чем же отличается точка a от других точек на прямой? Перекидывая мостик к четвертому пункту Предложения 1.2, мы видим, что a — это единственная точка $x \in \mathbb{R}$, для которой $\mathcal{P}(\{x\}) > 0$.

   Более того, если рассмотреть всевозможные интервалы $B = \langle \alpha, \beta \rangle$,[17] то мы увидим, что $\mathcal{P}_n(B) \to \mathcal{P}(B)$ при условии, что $\alpha, \beta \neq a$. Но множество $\{\alpha, \beta\}$ — это и есть граница ∂B множества B. Таким образом мы приходим к иллюстрации (конечно, на простейшем примере и только для $B = \{\alpha, \beta\}$) четвертого пункта Предложения 1.2. То есть становится более-менее понятно, откуда берется условие $\mathcal{P}(\partial B) = 0$.

Посмотрим теперь на понятие слабой сходимости с точки зрения случайных величин (а не просто их распределений), ограничиваясь для простоты одномерным случаем $d = 1$.

Рассмотрим вероятностное пространство $(\Omega, \mathcal{F}, \mathrm{P})$ и пусть ξ_n, ξ — случайные величины, определенные на этом вероятностном пространстве. Обозначим $\mathcal{P}_n = \mathcal{L}(\xi_n)$, $\mathcal{P} = \mathcal{L}(\xi)$.

Естественно ожидать, что, если сами случайные величины близки в каком-то смысле как функции ω, то и распределения этих случайных величин должны быть близки. Посмотрим, как это можно формализовать в терминах слабой сходимости распределений.

Ясно, что в определении слабой сходимости (1.6) может быть переписано как

$$\mathrm{E} f(\xi_n) \to \mathrm{E} f(\xi), \quad f \in \mathbb{C}_b(\mathbb{R}). \tag{1.7}$$

Такая форма записи определения слабой сходимости наводит на мысль, что из сходимости $\xi_n \to \xi$ (понимаемой в некотором разумном смысле) должна следовать слабая сходимость распределений $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$.[18] Конечно, так и есть. Проиллюстрируем это с помощью следующей диаграммы.

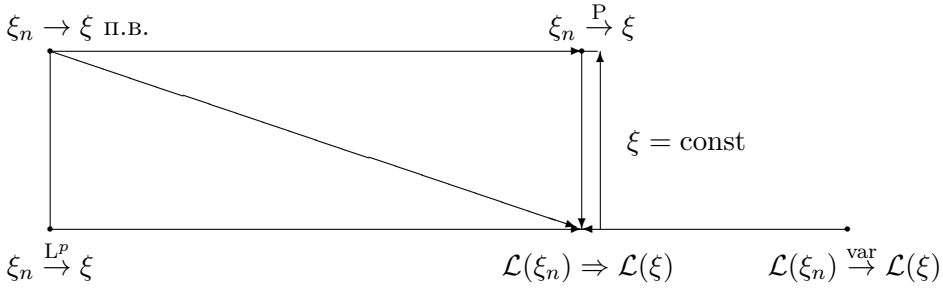

Рис. 1: Соотношения между разными видами сходимости

Рассматриваются 3 вида сходимости случайных величин: сходимость почти всюду[19], по вероятности[20] и в $\mathbb{L}^p$.[21] Диаграмма показывает, что каждая из этих сходимостей влечет за собой слабую

[16] Более того, в этом случае сходимость равномерна, то есть $\sup_x |F_n(x) - F(x)| \to 0$. Докажите это.

[17] Значки "$\langle$ " и " $\rangle$ " означают, что нам все равно, принадлежат концы интервала этому множеству или нет.

[18] Иногда вместо $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$ пишут $\xi_n \xrightarrow{d} \xi$ и говорят, что ξ_n сходится к ξ *по распределению*. При использовании такой терминологии нужно четко понимать, что речь идет не о сходимости случайных величин ξ_n к случайной величине ξ, а о сходимости их распределений. Например, в записи (1.7) ничего не мешает случайным величинам ξ_n и ξ быть заданным на разных вероятностных пространствах.

[19] Вероятность того, что $\xi_n \to \xi$, равна 1. Часто для такой сходимости применяю термин «почти наверное».

[20] То есть $\mathbb{P}(|\xi_n - \xi| > \varepsilon) \to 0$ для любого $\varepsilon > 0$.

[21] При $d = 1$ это означает, что $\mathbb{E}|\xi_n - \xi|^p \to 0$. Здесь, конечно, предполагается, что $\xi_n, \xi \in \mathbb{L}^p$. При $d > 1$ все аналогично.

сходимость соответствующих распределений.[22] Кроме того, для полноты на диаграмме изображен уже обсуждавшийся факт: из сильной сходимости распределений следует их слабая сходимость.

Наконец, вертикальная стрелочка от слабой сходимости к сходимости по вероятности с указанием $\xi = \mathrm{const}$ показывает, что в этом случае оба вида сходимости (то есть $\xi_n \xrightarrow{\mathrm{P}} a$ и $\mathcal{L}(\xi_n) \Rightarrow \delta_a$) эквивалентны.

При решении задач, связанных со слабой сходимостью распределений, эти соображения могут быть полезны. Действительно, пусть нам нужно доказать, что $\mathcal{P}_n \Rightarrow \mathcal{P}$. Если мы построим на некотором вероятностном пространстве такие случайные величины ξ_n, ξ, что $\xi_n \to \xi$ в одном из трех обсуждаемых смыслах, то мы автоматически получим, что $\mathcal{P}_n$ слабо сходится к $\mathcal{P}$. Такой прием носит название *метода одного вероятностного пространства*. Приведем один простой пример.

Пример 1.2. Путь распределения $\mathcal{P}_n$ определяются своими таблицами

$$\mathcal{P}_n : \quad \begin{pmatrix} 0 & \dots & k/n & \dots & (n-1)/n \\ 1/n & \dots & 1/n & \dots & 1/n \end{pmatrix}.$$

Докажем, что $\mathcal{P}_n \Rightarrow \mathrm{U}(0,1)$ четырьмя способами, три[23] из которых связаны с Предложением 1.2, а четвертое реализует идею метода одного вероятностного пространства.

1. Прежде всего, по определению слабой сходимости распределений нужно доказать, что для $f \in \mathbb{C}_b(0,1)$

$$\int_{\mathbb{R}} f d\mathcal{P}_n = \frac{1}{n}\sum_{k=0}^{n-1} f(k/n) \to \int_0^1 f(x)dx = \int_{\mathbb{R}} f d\mathcal{P}.$$

   Это, конечно, ни что иное, как свойство интеграла Римана.[24]

2. Если мы пользуемся вторым пунктом Предложения 1.2, то, взяв $x \in [0,1]$,[25] получаем, что

$$F_n(x) = \frac{\lceil nx \rceil}{n} \to x = F(x).$$

3. Если использовать для доказательства слабой сходимости характеристические функции, то при любом фиксированном t и достаточно большом n[26]

$$\varphi_n(t) = \int_{\mathbb{R}} e^{itx}\mathcal{P}_n(dx) = \frac{1}{n}\sum_{k=0}^{n-1} e^{itk/n} = \begin{cases} 1 & \text{при } \ t = 0, \\ \dfrac{e^{it}-1}{n(e^{it/n}-1)} & \text{при } \ t \neq 0 \end{cases} \xrightarrow[n\to\infty]{}$$

$$\xrightarrow[n\to\infty]{} \begin{cases} 1 & \text{при } \ t = 0, \\ \dfrac{e^{it}-1}{it} & \text{при } \ t \neq 0 \end{cases} = \int_0^1 e^{itx}dx = \int_{\mathbb{R}} e^{itx}\mathcal{P}(dx),$$

   где $\mathcal{P} = \mathrm{U}(0,1)$.

4. Воспользуемся теперь тем, что из сходимости случайных величин почти наверное следует слабая сходимость их распределений. А именно, рассмотрим $\xi \in \mathrm{U}(0,1)$ и положим $\xi_n = \lfloor n\xi \rfloor / n$.[27] Тогда при $k = 0, \dots, n-1$

$$\mathbb{P}\big(\xi_n = k/n\big) = \mathbb{P}\big(\lfloor n\xi \rfloor = k\big) = 1/n$$

[22] При желании это можно считать еще одним основанием называть такую сходимость «слабой».

[23] Как уже говорилось, четвертый пункт Предложения 1.2 мало пригоден для решения подобных задач.

[24] Здесь мы столкнулись с тем редким случаем, когда факт слабой сходимости может быть проверен прямо по определению.

[25] При других x сходимость функций распределения очевидна.

[26] Чтобы t/n не было целым числом при фиксированном $t \neq 0$.

[27] Тем самым ξ_n и ξ определены на одном вероятностном пространстве.

и, следовательно, $\mathcal{L}(\xi_n) = \mathcal{P}_n$. С другой стороны,

$$\frac{\lfloor n\xi \rfloor}{n} = \frac{n\xi - \{n\xi\}}{n} = \xi - \frac{\{n\xi\}}{n} \to \xi$$

при $n \to \infty$. Поэтому $\xi_n \to \xi$ п.в. и, следовательно, $\mathcal{P}_n \Rightarrow \mathcal{P}$.

Важным фактом, относящимся к слабой сходимости, является следующее утверждение, которое для простоты изложения мы сформулируем в одномерном случае и на языке случайных величин,

Предложение 1.3. *Пусть ξ_n, ξ — случайные величины*[28] *и $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$. Если функция $g : \mathbb{R} \mapsto \mathbb{R}$ является непрерывной, то $\mathcal{L}\big(g(\xi_n)\big) \Rightarrow \mathcal{L}\big(g(\xi)\big)$.*

Например, отсюда следует, что $\mathcal{L}\big(\sin(\xi_n)\big) \Rightarrow \mathcal{L}\big(\sin(\xi)\big)$, если только $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$.

Замечание 1.5. 1. Перенос Предложения 1.3 на многомерный случай не представляет труда:[29] слабая сходимость $\mathcal{L}\big(g(\xi_n)\big) \Rightarrow \mathcal{L}\big(g(\xi)\big)$ останется в силе, если $\xi_n, \xi \in \mathbb{R}^d$, $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$ и g — непрерывное отображение $\mathbb{R}^d \mapsto \mathbb{R}^k$.

Приведем пример использования многомерного аналога Предложения 1.3. Пусть $\xi_n = (\beta_n, \eta_n)^{\mathrm{T}}$ и $\xi = (\beta, \eta)^{\mathrm{T}}$ — двумерные случайные вектора, причем $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$. Тогда

$$\mathcal{L}(\beta_n) \Rightarrow \mathcal{L}(\beta), \quad \mathcal{L}(\beta_n + \eta_n) \Rightarrow \mathcal{L}(\beta + \eta), \quad \mathcal{L}(\beta_n \eta_n) \Rightarrow \mathcal{L}(\beta\eta)$$

и так далее.
2. Более сложным для доказательства представляется тот факт, что в Предложении 1.3 (и в его обобщении на многомерный случай) можно ослабить требование непрерывности g. Оказывается, что сходимость $\mathcal{L}\big(g(\xi_n)\big) \Rightarrow \mathcal{L}\big(g(\xi)\big)$ имеет место и в том случае, когда g является почти всюду непрерывным по предельной мере $\mathcal{P} = \mathcal{L}(\xi)$.

Например, если $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$, случайная величина ξ является абсолютно непрерывной, а функция g обладает лишь конечным числом точек разрыва, то все равно $\mathcal{L}\big(g(\xi_n)\big) \Rightarrow \mathcal{L}\big(g(\xi)\big)$.

1.4 Сходимость по вероятности

Пусть случайные величины ξ и ξ_n, $n \geq 1$ определены на одном вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$. Как уже говорилось, $\xi_n \xrightarrow{\mathrm{P}} \xi$, если (по определению)[30] при $n \to \infty$

$$\mathrm{P}(|\xi_n - \xi| > \varepsilon) \to 0 \tag{1.8}$$

для любого $\varepsilon > 0$.[31]

Из диаграммы 1 следует, что из сходимости по вероятности следует слабая сходимость, так что все свойства слабой сходимости имеют место и в этом случае. Например, согласно Предложению 1.3) для непрерывных функций g из сходимости $\xi_n \xrightarrow{\mathrm{P}} \xi$, вытекает, что $\mathcal{L}\big(g(\xi_n)\big) \Rightarrow \mathcal{L}\big(g(\xi)\big)$. Более аккуратный анализ показывает, что в этом случае $g(\xi_n) \xrightarrow{\mathrm{P}} g(\xi)$

А если $\xi \equiv a = \mathrm{const}$, то сходимость по вероятности $\xi_n \xrightarrow{\mathrm{P}} a$ эквивалентна слабой сходимости $\mathcal{L}(\xi_n) \Rightarrow \delta_a$.

В этом случае имеет место следующее полезное для решения задач утверждение.[32]

Предложение 1.4. *Пусть случайные величины ξ_n, ξ и η_n определены на одном вероятностном пространстве, причем $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$, а $\eta_n \xrightarrow{\mathrm{P}} a = \mathrm{const}$. Тогда $\mathcal{L}\big(g(\xi_n, \eta_n)\big) \Rightarrow \mathcal{L}\big(g(\xi, a)\big)$ для любой непрерывной функции $g : \mathbb{R}^2 \mapsto \mathbb{R}$.*

[28] Не обязательно определенные на одном вероятностном пространстве.
[29] И не только в смысле формулировки.
[30] Конечно, это определение ни что иное как определение сходимости по мере.
[31] Для случайных векторов $\xi, \xi_n \in \mathbb{R}^d$ в (1.8) нужно заменить $|\xi_n - \xi|$ на $\|\xi_n - \xi\|$.
[32] Оно снова будет сформулировано в одномерном варианте. Переформулируйте его в многомерном случае.

Из этого утверждения следует, например, что $\mathcal{L}(\xi_n + \eta_n) \Rightarrow \mathcal{L}(\xi)$, если только $\mathcal{L}(\xi_n) \Rightarrow \mathcal{L}(\xi)$ и $\eta_n \xrightarrow{\mathrm{P}} 0$. То есть случайные величины, по вероятности стремящиеся к нулю, играют роль бесконечно малых в обычном математическом анализе.

2 О законе больших чисел и центральной предельной теореме

Среди многих утверждений о слабой сходимости распределений исторически выделяются две группы утверждений, традиционно называемые Слабым Законом Больших Чисел и Центральной Предельной Теоремой. Дадим соответствующие определения.

Определение 2.1. Говорят, что последовательность случайных величин $\xi_1, \ldots, \xi_n, \ldots$, обладающих математическими ожиданиями $\mathrm{E}\xi_k = a_k$, удовлетворяет слабому[33] закону больших чисел (стандартная аббревиатура — СЗБЧ), если

$$\frac{\xi_1 + \ldots + \xi_n - (a_1 + \ldots + a_n)}{n} \xrightarrow[n\to\infty]{\mathrm{P}} 0. \tag{2.1}$$

Замечание 2.1. Поскольку (2.1) может быть переписано в виде

$$\mathcal{L}\left(\frac{\xi_1 + \ldots + \xi_n - (a_1 + \ldots + a_n)}{n}\right) \Longrightarrow \delta_0,$$

то утверждение, что какая-то последовательность случайных величин подчиняется СЗБЧ, является утверждением о слабой сходимости распределений. Но предельное распределение здесь является вырожденным.

Если все математические ожидания a_k одинаковы (обозначим их a), то (2.1) переписывается как

$$\frac{\xi_1 + \ldots + \xi_n}{n} \xrightarrow[n\to\infty]{\mathrm{P}} a. \tag{2.2}$$

Это выражение лучше поддается интерпретации: если последовательность $\{\xi_n,\, n \geq 1\}$ с $\mathrm{E}\xi_n = a$ удовлетворяет слабому закону больших чисел, то при больших n среднее арифметическое ξ_k с большой вероятностью близко к a.[34]

В случае, когда у случайных величин ξ_n имеются конечные вторые моменты, достаточным условием выполнения слабого закона больших чисел является сходимость $\mathrm{D}(\xi_1 + \ldots + \xi_n)/n^2 \to 0$. Этот факт легко выводится из неравенства Чебышева, которое утверждает, что для любого $\varepsilon > 0$ и любой случайной величины ξ с конечной дисперсией

$$\mathrm{P}\big(|\xi - \mathrm{E}\xi| > \varepsilon\big) \leq \frac{\mathrm{D}\xi}{\varepsilon^2}\ .$$

В частности, если случайные величины ξ_i независимы, то для выполнения (2.1) достаточно, чтобы

$$\sum_{i=0}^{n} \mathrm{D}\xi_i = o(n^2)$$

при $n \to \infty$. Если же случайные величины ξ_i не только независимы, но и одинаково распределены с математическим ожиданием $\mathrm{E}\xi_i = a$, то, согласно теореме Хинчина, Слабый Закон Больших Чисел (в традиционной форме (2.2)) выполняется автоматически, причем конечность дисперсии здесь не предполагается.

[33] В так называемом усиленном законе больших чисел рассматривается сходимость почти наверное, а не по вероятности.

[34] Физики описывают это свойство такими словами: среднее по времени сходится к среднему по пространству, имея в виду, что $a = \int_{\mathbb{R}} x\mathcal{P}_{\xi_k}(dx)$.

Определение 2.2. Говорят, что последовательность случайных величин $\xi_1, \ldots, \xi_n, \ldots$, обладающих математическими ожиданиями $\mathrm{E}\xi_k = a_k$ и конечными вторыми моментами, удовлетворяет центральной предельной теореме (стандартная аббревиатура — ЦПТ), если

$$\mathcal{L}\left(\frac{\xi_1 + \ldots + \xi_n - (a_1 + \ldots + a_n)}{B_n}\right) \Longrightarrow \mathrm{N}(0,1), \tag{2.3}$$

где $B_n^2 = \mathrm{D}(\xi_1 + \ldots + \xi_n)$.[35]

Структура формулы (2.3) очень проста. Из случайных величин ξ_k образуются последовательные суммы $S_n = \xi_1 + \ldots + \xi_n$, которые потом центрируются вычитанием математического $\mathrm{E}S_n$ и нормируются делением на $\sqrt{\mathrm{D}(S_n)}$. В результате полученная случайная величина $(S_n - \mathrm{E}S_n)/\sqrt{\mathrm{D}(S_n)}$ имеет нулевое среднее и единичную дисперсию. что совпадает с параметрами 0 и 1 нормального распределения в правой части (2.3).

Тем самым то, что в пределе получается именно *стандартное* нормальное распределение, можно считать естественным. А вот сама нормальность, конечно, не является очевидной.

Замечание 2.2. Легко видеть, что предельные соотношения (2.1) и (2.3) останутся в силе, если случайные величины ξ_k заменить на $\xi_k + c_k$, где c_k — некоторые постоянные. Поэтому при теоретических исследованиях обычно (и без ущерба для общности) считают, что $\mathrm{E}\xi_k = 0$.

Что касается ЦПТ, то наиболее популярной является теорема Поля Леви, утверждающая, что любая последовательность независимых одинаково распределенных величин, имеющих конечную ненулевую дисперсию, подчиняется Центральной Предельной Теореме. Сходимость (2.3) в этом случае приобретает вид

$$\mathcal{L}\left(\frac{\xi_1 + \ldots + \xi_n - na}{\sigma\sqrt{n}}\right) \Longrightarrow \mathrm{N}(0,1), \tag{2.4}$$

где $a = \mathrm{E}\xi_i$ и $\sigma^2 = \mathrm{D}\xi_i$.

Хорошо известен частный случай теоремы П. Леви, имеющий название интегральной теоремы Муавра-Лапласа. Пусть рассматриваемые в (2.4) независимые одинаково распределенные случайные величины ξ_i имеют распределение Бернулли $\mathrm{Ber}(p)$ с $0 < p < 1$. Тогда $a = p$, $\sigma^2 = p(1-p)$ и, если обозначить $k_n = \xi_1 + \ldots + \xi_n$, то окажется, что k_n — это число успехов в n испытаниях Бернулли с вероятностью успеха p. Тем самым (2.4) превращается в

$$\mathcal{L}\left(\frac{k_n - np}{\sqrt{np(1-p)}}\right) \Longrightarrow \mathrm{N}(0,1). \tag{2.5}$$

Поскольку функция распределения Φ распределения $\mathrm{N}(0,1)$ непрерывна на всей прямой, то, благодаря второму пункту Предложения 1.2 слабая сходимость (2.5) оказывается эквивалентной поточечной сходимости функций распределений[36]

$$\mathrm{P}\left(\frac{k_n - np}{\sqrt{np(1-p)}} < x\right) \to \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} dz.$$

Именно в такой форме обычно приводят интегральную теорему Муавра-Лапласа в элементарных учебниках.

[35] При этом подразумевается, что $B_n^2 > 0$ при всех достаточно больших n.

[36] Это, конечно, относится и к общей Центральной Предельной Теореме (2.4) или даже к (2.3). Более того, всегда, когда функции распределения F_n поточечно сходятся к непрерывной функции распределения F, эта сходимость является равномерной.

Имеет место и многомерный аналог теоремы Леви. Он утверждает, что для независимых одинаково распределенных случайных векторов $\overline{\xi}_1, \ldots, \overline{\xi}_n, \ldots \in \mathbb{R}^d$, все координаты которых обладают конечными вторыми моментами,

$$\mathcal{L}\left(\frac{\overline{\xi}_1 + \ldots + \overline{\xi}_n - n\overline{a}}{\sqrt{n}}\right) \Longrightarrow \mathrm{N}_d(\mathbf{0}, \Sigma),$$

где $\overline{a} = \mathrm{E}\overline{\xi}_k$, Σ — ковариационная матрица вектора $\overline{\xi}_k$, а $\mathrm{N}_d(0, \Sigma)$ — d-мерное гауссовское распределение с нулевым средним и ковариационной матрицей Σ, определяемое характеристической функцией

$$\varphi(\overline{t}) = e^{-(\Sigma\overline{t}, \overline{t})/2}\ , \quad \overline{t} \in \mathbb{R}^d.$$

Список литературы

[1] П. Биллингсли (1977), Сходимость вероятностных мер. М., Наука.

[2] А.Н. Ширяев (2004), Вероятность-1, М., Изд-во МЦНМО.