



МЕХАНИКО-
МАТЕМАТИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ. СЕМИНАРЫ

ШКЛЯЕВ
АЛЕКСАНДР ВИКТОРОВИЧ

МЕХМАТ МГУ

КОНСПЕКТ ПОДГОТОВЛЕН
СТУДЕНТАМИ, НЕ ПРОХОДИЛ
ПРОФ. РЕДАКТУРУ И МОЖЕТ
СОДЕРЖАТЬ ОШИБКИ.
СЛЕДИТЕ ЗА ОБНОВЛЕНИЯМИ
НА [VK.COM/TEACHINMSU](https://vk.com/teachinmsu).

ЕСЛИ ВЫ ОБНАРУЖИЛИ
ОШИБКИ ИЛИ ОПЕЧАТКИ,
ТО СООБЩИТЕ ОБ ЭТОМ,
НАПИСАВ СООБЩЕСТВУ
[VK.COM/TEACHINMSU](https://vk.com/teachinmsu).



БЛАГОДАРИМ ЗА ПОДГОТОВКУ КОНСПЕКТА
СТУДЕНТА ФАКУЛЬТЕТА ФКИ МГУ
БАРМИНА МАКСИМА АЛЕКСАНДРОВИЧА



Содержание

Семинар 1. Вероятностное пространство	6
Пример построения вероятностного пространства	6
Выбор шаров из урны	7
Упорядоченный выбор с возвращением	7
Упорядоченный выбор без возвращения	8
Неупорядоченный выбор без возвращения	8
Неупорядоченный выбор с возвращением	9
Интерпретация исходов	9
Семинар 2. Условные вероятности	12
Определение условной вероятности	12
Формула полной вероятности	14
Формула Байеса	15
Независимость	16
Схема независимых испытаний	17
Схема Бернулли	18
Семинар 3. Случайные величины	19
Случайные величины и их распределение	19
Сигма–алгебра	21
Маргинальное и совместное распределения	22
Независимые случайные величины	24
Формула свёртки	26
Некоторые базовые распределения	27
Семинар 4. Математическое ожидание	29
Определение математического ожидания	29
Свойства математического ожидания	32
Дисперсия	34
Свойства дисперсии и ковариации	35
Семинар 5. Производящие функции	39
Производящая функция случайной величины	39
Свойства производящей функции случайной величины	40
Производящие функции векторов	44
Свойства производящей функции вектора	45
Семинар 6. Общее вероятностное пространство	49
Построение общего вероятностного пространства	49

Пересечение множеств	50
Свойства вероятностных мер	52
Продолжение меры	54
Борелевская сигма–алгебра	56
Геометрическая вероятность	57
Семинар 7. Случайные величины в общем случае	59
Случайные величины и их распределение	59
Функция распределения	59
Свойства функции распределения	63
Дискретное распределение	64
Абсолютно–непрерывный случай	65
Случайные векторы	67
Формула свёртки	68
Семинар 8. Математическое ожидание в общем случае	69
Общее определение математического ожидания	69
Интеграл Римана–Стилтьеса	70
Абсолютно–непрерывный случай	72
Смешанное распределение	73
Интегрирование по частям	74
Свойства математического ожидания	75
Дисперсия	76
Семинар 9. Случайные векторы	79
Случайный вектор и его распределение	79
Абсолютно–непрерывный случай	84
Плотности подвекторов	86
Преобразование плотности при гладких заменах координат	87
Независимость компонент вектора	90
Математические ожидания функций от векторов	91
Семинар 10. Характеристические функции	92
Математическое ожидание комплекснозначной случайной величины	92
Определение и существование характеристической функции	92
Примеры характеристических функций	93
Свойства характеристической функции	95
Единственность и непрерывность характеристической функции	97
Формулы обращения	98

Семинар 11. Сходимости	101
Сходимость почти наверное	101
Сходимость по вероятности	102
Сходимость в среднем порядка p	103
Сходимость по распределению	104
Семинар 12. Законы больших чисел и предельные теоремы	107
Схема Бернулли	107
Законы больших чисел (ЗБЧ)	110
Центральные предельные теоремы (ЦПТ)	112

Семинар 1. Вероятностное пространство

Пример построения вероятностного пространства

Пример 1.1. Будем выбирать два множества с возвращением. Можно представить себе это так: на карточках написаны числа от 1 до N , мы наугад выбираем несколько карточек в первое множество, возвращаем назад, перемешиваем, а затем ещё раз выбираем наугад несколько карточек уже для второго множества.

Описать вероятностное пространство Ω можно большим количеством способов, однако подавляющее большинство этих способов неудобны для дальнейшего решения задач. Например, если взять $\Omega = \{(a_1, a_2)\}$, где a_1, a_2 — множества. Это будет абсолютно верно и абсолютно бессмысленно, потому что пользоваться этим мы не сможем (например, не определены их пересечения).

Гораздо удобнее задавать множество вектором длины N , где на i -ом месте стоит 0, если элемент не попадает в множество и 1, если попадает:

$$\Omega = \{(a_1, \dots, a_N), a_i \in \{0, 1\}\}.$$

При таком задании легко определить, какие элементы входят в множество, а какие нет.

В нашем примере вы выбираем два множества, поэтому элементом нашего пространства будут два вектора, или, что тоже самое, матрица $2 \times N$:

$$\Omega = \left\{ \begin{pmatrix} a_1 & \dots & a_N \\ b_1 & \dots & b_n \end{pmatrix}, a_i \in \{0, 1\} \right\}.$$

В первой строчке содержится информация о том, какие элементы мы взяли в первое множество, а во второй — во второе.

Теперь удобно формализовывать события. Например, рассмотрим событие $A = \text{«два множества не пересекаются»}$. В матрице это соответствует отсутствию столбца $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, то есть

$$A = \left\{ \begin{pmatrix} a_1 & \dots & a_N \\ b_1 & \dots & b_N \end{pmatrix}, a_i \cdot b_i \neq 1 \ \forall i \right\}.$$

Обсудим базовые задачи, с которыми будем работать в дальнейшем:

Выбор шаров из урны

Рассмотрим урну, в которой лежат шары с номерами $1, 2, \dots, N$. Эксперимент заключается в вытягивании k шаров и фиксации последовательности вытягиваемых номеров.

Упорядоченный выбор с возвращением

Пример 1.2. Слово «с возвращением» означает, что шары вытягиваются один за другим, а после каждого вытягивания мы записываем его номер и возвращаем в урну. Слово «упорядоченное» означает, что нам важно, в каком порядке мы вытаскивали шары. Мы можем описать соответствующее задаче пространство элементарных исходов множеством упорядоченных наборов:

$$\Omega_1 = \{(i_1, \dots, i_k), i_j \in \{1, \dots, N\}\}.$$

Множество событий — это множество всех подвекторов:

$$\mathcal{F}_1 = 2^{\Omega_1}.$$

Все наборы из k шаров равновероятны, поэтому

$$P_1((i_1, \dots, i_k)) = \frac{1}{N^k},$$

так как множество возможных исходов можно представить в виде k -мерной таблицы $\{1, \dots, N\} \times \dots \times \{1, \dots, N\}$, в которой N^k ячеек. Можем также представить множество всех векторов в виде дерева (см. рис. 1.1), на i -м слое которого мы будем определять, какой из N шаров мы вытягиваем на i -м шаге, тогда итоговый вектор будет задаваться вершиной на нижнем слое дерева. Так как на каждом новом слое такого дерева вершин в N раз больше, чем на предыдущем, всего их также N^k .

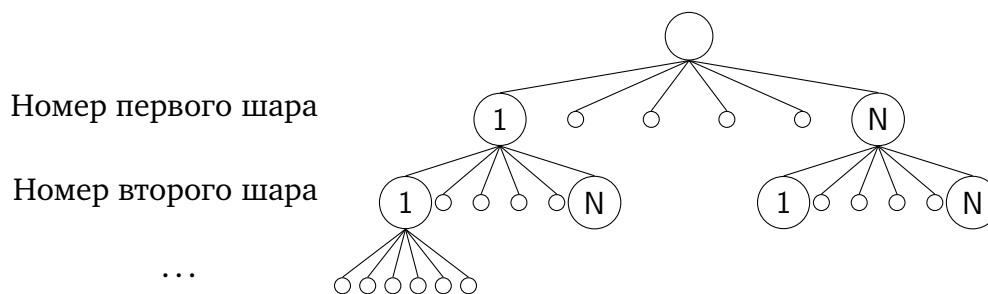


Рис. 1.1: Дерево, соответствующее векторам исходов упорядоченного выбора с возвращением

Упорядоченный выбор без возвращения

Пример 1.3. В отличие от предыдущего примера, вытасенные шары на каждом шаге будут откладываться. Соответствующее пространство элементарных исходов будет состоять из векторов длины k с попарно различными целочисленными компонентами:

$$\Omega_2 = \{(i_1, \dots, i_k), i_j \in \{1, \dots, N\}, i_j \neq i_\ell\}; \mathcal{F}_2 = 2^{\Omega_2}.$$

Выпадение каждого из оставшихся шаров на i -м шаге равновероятны, поэтому и вероятность любой раскладки шаров будет равна

$$P_2((i_1, \dots, i_k)) = \frac{1}{N(N+1)\dots(N-k+1)} = \frac{1}{A_N^k},$$

так как $|\Omega_2| = A_N^k$. Понять это можно, представив исходы в виде дерева. По сравнению с деревом, изображенным на рис. 1.1, поменяется количество шагов на каждом слое — у каждой вершины k -го поколения будет $N - k$ потомков в следующем.

Неупорядоченный выбор без возвращения

Пример 1.4. Теперь исходом эксперимента будет являться неупорядоченный набор из k чисел от 1 до N без повторений (можем считать, что он упорядочен в порядке, который мы сами зададим).

$$\Omega_3 = \{i_1, \dots, i_k\}, i_j \in \{1, \dots, N\}, i_j \neq i_\ell; \mathcal{F}_3 = 2^{\Omega_3}.$$

В этом случае также все возможные наборы из k шаров равновероятны, и вероятность каждого элементарного исхода равна

$$P_3(\{i_1, \dots, i_k\}) = \frac{1}{C_N^k},$$

где C_N^k — количество k -элементных подмножеств N -элементного множества. Если мы сравним с предыдущим пространством, то каждому исходу $\{i_1, \dots, i_k\}$ в Ω_3 соответствуют все возможные $k!$ исходов в Ω_2 . Эти исходы задают все возможные перестановки этих элементов:

$$C_N^k = \frac{A_N^k}{k!} = \frac{N!}{k!(N-k)!}.$$

Неупорядоченный выбор с возвращением

Пример 1.5. По аналогии, мы можем убрать условие $i_j \neq i_\ell$ из Ω_3 , однако тогда не совсем понятно, как определять множества, соответствующие наборам, в которых встречаются одинаковые числа (во множествах не принято дублирование элементов).

Мы рассмотрим вместо этого пространство другого типа, описывающее, сколько шаров с каждым номером от 1 до N мы вытащили. Это задаётся вектором длины N , у которого сумма элементов равна k :

$$\Omega_4 = \{(i_1, \dots, i_N), i_j \in \{0, \dots, k\}, i_1 + \dots + i_N = k\}; \mathcal{F}_4 = 2^{\Omega_4}.$$

Чтобы найти количество исходов и вероятности в пространстве, рассмотрим следующую задачу: возьмем k шаров и добавим к ним еще $N-1$ абстрактный шар. Затем выберем из этих $N+k-1$ шаров $N-1$ и покрасим их в черный цвет (см. рис. 1.2). Тогда чёрные шарики разобьют оставшиеся k шаров на N групп, которые будут соответствовать N элементам вектора. Каждая такая конструкция задаст ровно один элемент исходного пространства, а всего таких конструкций C_{N+k-1}^{N-1} . Тогда

$$P_4((i_1, \dots, i_k)) = \frac{1}{C_{N+k-1}^{N-1}}.$$

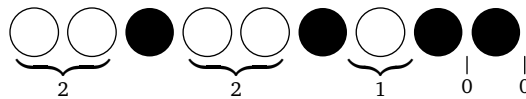


Рис. 1.2: Пример раскраски для неупорядоченного выбора с возвращением при $N = k = 5$, соответствующей вектору $(2, 2, 1, 0, 0)$

Интерпретация исходов

Обратите внимание, что один и тот же эксперимент можно задавать совершенно разными пространствами, потому что интерпретировать исход можно по-разному.

Пример 1.6. Рассмотрим пространства, описанные в примерах 1.3 и 1.4. В первом мы получаем исход, вытаскивая без повторений шары из урны с учётом порядка; во втором — без учёта порядка. При этом эти модели согласованы и все события, которые можно описать во обоих пространствах, будут иметь одну и ту же вероятность в обоих пространствах, поэтому без разницы, какое из этих двух пространств использовать в конкретной задаче.

Например, возьмём 3 шара из 7 без возвращения и посчитаем вероятность того, что шара с номером 5 нет: $A = \{\text{шара с номером 5 нет}\}$. Тогда в первом пространстве:

$$\Omega_1 = \{(i_1, \dots, i_3), i_j \in \{1, \dots, 7\}, i_j \neq i_\ell\} \implies |\Omega_1| = 7 \cdot 6 \cdot 5;$$

$$A_1 = \{(i_1, i_2, i_3), i_j \in \{1, \dots, 4, 6, 7\}, i_j \neq i_\ell\} \implies |A_1| = 6 \cdot 5 \cdot 4.$$

Поэтому

$$P_1(A_1) = \frac{|A_1|}{|\Omega_1|} = \frac{4}{7}.$$

Во втором пространстве:

$$\Omega_2 = \{(i_1, \dots, i_3), i_j \in \{1, \dots, 7\}\} \implies |\Omega_2| = C_7^3;$$

$$A_2 = \{(i_1, \dots, i_3), i_j \in \{1, 2, \dots, 4, 6, 7\}\} \implies |A_2| = C_6^3.$$

Поэтому

$$P_2(A_2) = \frac{\frac{6 \cdot 5 \cdot 4}{3!}}{\frac{7 \cdot 6 \cdot 5}{3!}} = \frac{4}{7}.$$

При этом важно не выпасть в другой мир при смене вероятностного пространства.

Пример 1.7. Если рассматривать пространства из примеров 1.2 и 1.5 классическими, они будут из разного мира.

Например, возьмем 2 шарика из 3 и посчитаем вероятность того, что оба вытасненных шара имеют номер 1 в каждом из пространств. В первом пространстве:

$$\Omega_1 = \{(i, j), i, j \in \{1, \dots, 3\}\} \implies |\Omega_1| = 3^2 = 9;$$

$$P_1((1, 1)) = \frac{1}{9}.$$

Во втором пространстве:

$$\Omega_2 = \{(i, j), i, j \in \{1, \dots, 3\}\} \implies |\Omega_2| = C_{3+2-1}^{3-1} = C_4^2 = 6.$$

Тогда в классической модели:

$$P_2(\{1, 1\}) = \frac{1}{6}.$$

Вероятности получились разные, потому что классическая модель, которую мы приписали второму исходу — это совсем другая реальность, в которой

все раскладки равновероятны, хотя в реальности интуитивно понятно, что шансы вытянуть $\{1, 1\}$ в два раза меньше, чем шансы вытянуть $\{1, 2\}$.

Здесь всегда помогает представление об общей физической модели. Ясно, что первая модель соответствует физической реальности — вытаскиваем оба шарика наугад, полученные раскладки равновероятны. А во второй модели мы вытаскиваем множество итоговых маркеров шариков, которые не должны быть равновероятны.

Чтобы не попадать в такие ловушки, нужно сначала выбрать классическое пространство или пространство с прослеживаемой симметрией, соответствующее эксперименту.

Упражнение 1.1. На экзамене студент Иванов знает первые m билетов из n . N студентов по очереди тянут билеты. Постройте вероятностное пространство и найдите вероятность того, что Иванов, стоя в очереди k -м, вытянет знакомый билет, если билеты тянутся а) с возвращением; б) без возвращения.

Ответ: каким бы по очереди студент не вытягивал экзаменационный билет, шансы вытянуть один из m знакомых билетов одни и те же у каждого студента.

Семинар 2. Условные вероятности

Определение условной вероятности

Нужно иметь в виду, что вероятность, с одной стороны, величина объективная — доля того, сколько раз случится событие, если мы много раз будем его повторять, а с другой стороны, совершенно не объективная. Представим себе следующий эксперимент:

Пример 2.1. Рассмотрим урну, в которой 12 красных шаров с номерами от 1 до 12 и 13 синих шаров с номерами от 13 до 25. Мы вытаскиваем шарик. Вероятностное пространство, соответствующее задаче — классическое, так как есть симметрия:

$$\Omega = \{1, \dots, 25\}.$$

Вероятность того, что вытащили шарик с номером 12:

$$P(12) = \frac{1}{25}.$$

В момент, когда мы достали шарик, сторонний наблюдатель увидел, что шар красный, поэтому для него вероятностное пространство и вероятность события A будут уже другим:

$$\tilde{\Omega} = \{1, \dots, 12\}; \quad \tilde{P}(12) = \frac{1}{12}.$$

Другой сторонний наблюдатель может увидеть, что цифры две и что шар красный. Поэтому у него будет всего 3 исхода:

$$\tilde{\tilde{\Omega}} = \{10, 11, 12\}; \quad \tilde{\tilde{P}}(12) = \frac{1}{3}.$$

Хоть все наблюдатели находятся в одном мире, из-за того, что они обладают разной информацией, у них разные вероятности. Мы одно и то же событие можем интерпретировать совершенно по-разному, с разными Ω в зависимости от того, что мы знаем про исход.

Нужно научиться работать в вероятностном пространстве, получая новую информацию и не перестраивая каждый раз пространство (в классическом пространстве перестроить легко — была симметрия исходов и осталась, если симметрии нет, то это сделать затруднительно).

На рис. 2.1 изображено Ω со всеми возможными исходами. Пусть случилось событие B — шар красный. Интересующее нас событие A — выпало

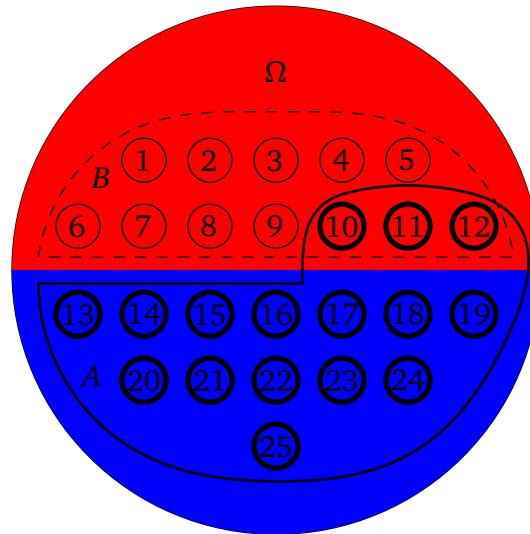


Рис. 2.1: Графическое представление пространства элементарных исходов из примера 2.1

двухзначное число. Тогда новая вероятность события A при условии B выражается через обычные вероятности (числитель и знаменатель разделили на мощность Ω):

$$\tilde{P} = P_B(A) = \frac{|A \cap B|}{|B|} = \frac{P(A \cap B)}{P(B)}.$$

Такая вероятность называется *условной*.

Определение 2.1. Условная вероятность A при условии B :

$$P(A|B) = \frac{P(AB)}{P(B)}, \quad (2.1)$$

где $P(AB)$ — вероятность пересечения A и B . Смысл этого выражения — вероятность того, что случится A , если мы знаем, что уже произошло B .

Условная вероятность возникает в задачах, если необходимо учесть дополнительную известную информацию. Важно не путать её с вероятностью пересечения, так как $P(AB)$ — вероятность того, что случится A и B сразу, если мы не знаем, что произошло B , а $P(A|B)$ — вероятность того, что случится A , если мы знаем, что произошло B .

Пример 2.2. Пусть человек летает в самолёте с бомбой, потому что думает, что вероятность того, что в самолёте окажется две бомбы маленькая. Это было бы правдой, если бы он рассматривал не условную вероятность,

а вероятность пересечения. В тот момент, когда он взял с собой в самолёт бомбу, он уже должен считать вероятность того, что в самолёте окажется вторая при условии, что там уже есть первая, а эта вероятность точно такая же, как если бы он не взял бомбу.

Формула полной вероятности

Из определения условной вероятности нетрудно заметить, что

$$P(AB) = P(A | B)P(B). \quad (2.2)$$

Так как физически условная вероятность $P(A | B)$ — это вероятность A в мире, где случилось B , зачастую бывает проще найти её, чем вероятность пересечения $P(AB)$.

Пример 2.3. Пусть в урне 15 белых и 13 чёрных шаров. Рассмотрим события A — «второй шар белый», B — «первый шар белый», \bar{B} — «первый шар чёрный». Найдём вероятность того, что мы вытащим 2 белых шара $P(AB)$.

Для этого сначала найдём вероятность $P(B) = \frac{15}{28}$. Теперь очень легко найти

$P(A | B) = \frac{14}{27}$, потому что вытащив первый белый шар, мы попали в мир, в котором осталось 14 белых и 13 чёрных шаров и в нём мы ищем вероятность вытащить белый шар. Поэтому по формуле (2.2) находим

$$P(AB) = P(A | B)P(B) = \frac{14}{27} \cdot \frac{15}{28}.$$

Теперь хотим найти вероятность, что второй шар белый $P(A)$. Мы не можем понять, как вытащить второй шар до того, пока не узнаем, какой первый, потому что мы попадём в разные миры. Рассмотрим оба случая:

$$P(A) = P(AB) + P(A\bar{B}).$$

Что-то из этих двух событий случится. Воспользуемся снова формулой (2.2):

$$P(A) = P(AB)P(B) + P(A\bar{B})P(\bar{B}) = \frac{14}{27} \frac{15}{28} + \frac{15}{27} \frac{13}{28} = \frac{15}{28}.$$

Этот трюк называется *формулой полной вероятности* и несмотря на его примитивность, он лежит в основе теории вероятностей довольно до глубоких вещей. Сформулируем её в общем виде:

Теорема 2.1 (формула полной вероятности). Рассмотрим разбиение пространства элементарных исходов $B_1, \dots, B_n: B_i \cap B_j = \emptyset, i \neq j, B_1 \cup \dots \cup B_n = \Omega$

(см. рис. 2.2). Тогда для любого события A :

$$P(A) = \sum_{i=1}^n P(A | B_i) P(B_i). \quad (2.3)$$

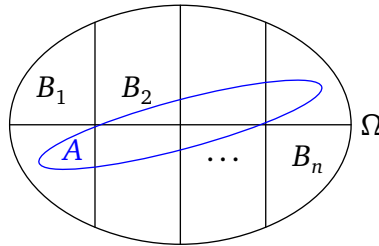


Рис. 2.2: Графическое представление разбиения Ω на множества B_i

В итоге, в задачах, где есть неопределённый первый шаг, мы можем его зафиксировать и перейти в более простые задачи на каждом этапе, в которых этой неопределённости нет. Зачастую решить несколько простых задач бывает значительно проще, чем одну сложную.

Пример 2.4. Есть монетка, у которой с одной стороны нарисован орёл, а про вторую сторону мы не знаем. Можем считать, что у нас есть две монетки, среди которых мы случайно выбираем: у одной нарисованы орёл и решка, а у второй два орла. Возьмём эту монету и бросим трижды. Найдём вероятность, что всё время выпадали орлы. События $A = \{3 \text{ орла}\}$; $B_1 = \{\text{монета с двумя орлами}\}$; $B_2 = \{\text{монета с орлом и решкой}\}$. B_1, B_2 — это разбиение Ω , потому что других монет по условию не бывает. Распишем $P(A)$ по формуле (2.3):

$$P(A) = \underbrace{P(A | B_1)}_{=1} \underbrace{P(B_1)}_{=\frac{1}{2}} + \underbrace{P(A | B_2)}_{=\frac{1}{8}} \underbrace{P(B_2)}_{=\frac{1}{2}} = \frac{9}{16}.$$

Формула Байеса

В жизни обычно всё устроено наоборот: мы не знаем, какие вероятности у монеток, а хотим узнать, какая у нас монетка. Будем много раз подбрасывать монетку, чтобы определить, какая она была (просто так посмотреть мы не можем).

Пример 2.5. В условии примера 2.4 хотим посчитать вероятность того, что монета с двумя орлами, зная что при трёх подбрасываниях выпало три орла.

В терминах наших событий это $P(B_1 | A)$. Распишем это выражение, используя (2.1), (2.2) и (2.3):

$$P(B_1 | A) = \frac{P(B_1 A)}{P(A)} = \frac{P(A | B_1)P(B_1)}{\sum P(A | B_i)P(B_i)} = \frac{1 \cdot \frac{1}{2}}{1 \cdot \frac{1}{2} + \frac{1}{8} \cdot \frac{1}{2}} = \frac{8}{9}.$$

Происходит то, что мы интуитивно понимаем: чем больше орлов выпадает, тем больше мы верим в то, что момента с двумя орлами, а как только выпадает хотя бы решка, мы сразу понимаем, что у монеты не может быть два орла. Такая формула называется *формула Байеса*. Сформулируем её в общем виде:

Теорема 2.2 (формула Байеса). Пусть B_i — разбиение. Тогда

$$P(B_1 | A) = \frac{P(A | B_1)P(B_1)}{\sum_{i=1}^n P(A | B_i)P(B_i)}.$$

Чтобы не запутаться во всех этих формулах, нужно сначала аккуратно расписать какие условные и какие прямые вероятности даны в задаче.

Пример 2.6. Пусть есть некоторая болезнь и некоторое тестирование на неё. Рассмотрим события A — «тест говорит, что болен»; B_1 — «болен»; B_2 — «здоров». Пусть мы знаем следующие вероятности: $P(B_1) = \frac{1}{100}$; $P(B_2) = \frac{99}{100}$; $P(A | B_1) = 0,99$; $P(A | B_2) = 0,02$. Мы сдали тест и он говорит, что мы больны, хотим узнать, какова вероятность, что мы больны на самом деле $P(B_1 | A)$:

$$P(B_1 | A) = \frac{0,99 \cdot 0,01}{0,99 \cdot 0,01 + 0,02 \cdot 0,99} = \frac{1}{3}.$$

Человеческий мозг, с одной стороны, постоянно занимается оценкой вероятностей, а с другой, очень плохо это делает, поэтому данный результат кажется довольно неожиданным.

Независимость

Определение 2.2. События A и B **независимы**, если

$$P(A | B) = P(A). \quad (2.4)$$

Если домножить обе части на $P(B)$, получим

$$P(AB) = P(A)P(B). \quad (2.5)$$

В некотором смысле, формула (2.5) более правильная, потому что в ней $P(B)$ может быть 0, так как событие вероятности 0 не зависит от чего угодно.

Обратите внимание, что формулу (2.5) можно применять только для независимых событий, которые сами по себе встречаются редко.

Пример 2.7. Рассмотрим колоду из 52 карт. Есть события A — «взять даму»; B — «взять пика». Проверим эти события на независимость. Дадим номера картам, чтобы выписать наши множества: $\Omega = \{1, \dots, 52\}$; $A = \{1, 14, 27, 40\}$; $B = \{1, \dots, 13\}$. Тогда $P(A) = \frac{4}{52} = \frac{1}{13}$; $P(B) = \frac{13}{52} = \frac{1}{4}$; $P(AB) = \frac{1}{52}$. Формула (2.5) верна, значит, события A и B независимы.

Пример 2.8. Добавим в колоду из примера 2.7 двух джокеров. Теперь у нас 54 карты. Изменим немного события: A — «карта сойдёт за даму» (дама и джокер); B — «карта сойдёт за пика» (пика и джокер). Теперь события A и B стали зависимыми: $P(A) = \frac{6}{54}$; $P(B) = \frac{15}{54}$; $P(AB) = \frac{3}{54}$. При этом $P(A)P(B) = \frac{90}{54^2} < P(AB)$. Удобно посмотреть на это в терминах условной вероятности: $P(A | B) > P(A)$, то есть вероятность встретить даму среди пик выше, чем даму среди всех карт. Это происходит из-за джокеров, которые в данном примере выступают как скрытый фактор, нарушающий независимость.

Схема независимых испытаний

Пример 2.9. Представим, что мы бросили несимметричную монетку. Вероятностное пространство броска — $\Omega_1 = \{1, 0\}$; $P_1(1) = p$, $P_1(0) = 1 - p = q$. Затем мы бросили эту же монетку снова. Вероятностное пространство второго броска — $\Omega_2 = \{1, 0\}$; $P_2(1) = p$, $P_2(0) = q$. Хотим объединить эти броски в одно пространство, где мы два раза бросаем монетку. Новым пространством будет $\Omega = \Omega_1 \times \Omega_2 = \{(i, j), i \in \{0, 1\}, j \in \{0, 1\}\}$ — *декартово произведение* исходных пространств. Если мы считаем, что испытания были независимыми, то $P(i, j) = P_1(i)P_2(j)$. В нашем случае $P(0, 0) = q^2$; $P(0, 1) = qr$; $P(1, 0) = pq$; $P(1, 1) = p^2$.

Точно так же устроена общая схема *модели независимых испытаний*, состоящие из большего числа испытаний (испытания не обязательно должны быть одинаково распределёнными): берём испытания сами по себе, перемножаем декартово их пространства и вероятности каждого векторного исхода получаем как произведение вероятностей для каждой компоненты.

Пример 2.10. Вытянуть подряд две карты из одной колоды — не модель независимых испытаний, потому что если мы не возвращаем первую карту,

то нет явно заданного пространства для второй карты. Если мы возвращаем карту в колоду, то это модель независимых испытаний.

Схема Бернулли

Определение 2.3. В качестве Ω возьмём векторы длины n из 0 и 1:

$$\Omega = \{(i_1, \dots, i_n), i_j \in \{0, 1\}\}.$$

В модели n испытаний с двумя возможными исходами, при этом успех имеет вероятность p , а неуспех — $1 - p$. Тогда

$$P((i_1, \dots, i_n)) = p^{\sum_{j=1}^n i_j} (1-p)^{n-\sum_{j=1}^n i_j},$$

потому что когда мы складываем i_j мы как раз считаем число 1. Такая схема независимых испытаний называется *схемой Бернулли*. Одно подбрасывание монетки или другое аналогичное испытание — *испытание Бернулли*.

Рассмотрим событие $A_k = \{\text{единиц ровно } k\}$. Найдём его вероятность:

$$P(A_k) = \sum_{|A_k|} p^k q^{n-k} = C_n^k p^k q^{n-k},$$

где $q = 1 - p$. Эта формула задаёт *биномиальное распределение*. Оно показывает какова вероятность того, что случилось ровно k успехов, если мы провели n испытаний. Заметьте, что модель уже не классическая, вероятности исходов разные, но нам повезло и в событие входят все исходы с одинаковой вероятностью.

Семинар 3. Случайные величины

Случайные величины и их распределение

Вероятностное пространство можно задавать по-разному, в то время как для всех моделей бывают одинаковые объективные показатели.

Пример 3.1. В комнате летает множество молекул, в результате что-то происходит. Движение молекул можно представить по-разному, померить скорость каждой молекулы мы тоже не сможем, их слишком много. Характеристика скорости молекул вместе во всей комнате — температура.

Если мы измеряем числовую характеристику Ω , то это общая шкала, так как числа — нечто объективное, а исходы субъективны. В конечном итоге, нас интересует следующее понятие:

Определение 3.1. Отображение $X : \Omega \rightarrow \mathbb{R}$ называется **случайной величиной**.

Мы Ω перевели в число. Ω была у всех разная, число у всех одинаковое. Вся случайность величины заключается именно в Ω . По существу, это функция, но так как Ω мы не знаем, появляется неопределенность. Когда мы дойдём до общих случайных величин, определение будет более строгое. Но пока Ω конечна или счётна, никаких дополнительных предположений нам не требуется.

Пример 3.2. Рассмотрим 2 броска симметричной монетки.

$$\Omega = \{(i_1, i_2), i_1, i_2 \in \{0, 1\}\}.$$

$$P(\omega = (i_1, i_2)) = \frac{1}{4},$$

так как монетка симметричная. Рассмотрим случайную величину $X_1(\omega)$ — результат первого броска. Она задаётся следующей формулой:

$$X_1(i_1, i_2) = i_1.$$

Аналогично $X_2(\omega)$ — результат второго броска:

$$X_2(i_1, i_2) = i_2.$$

Также рассмотрим $X_3(\omega)$ — общее число орлов:

$$X_3(i_1, i_2) = i_1 + i_2.$$

Базовая характеристика, на которую мы будем смотреть, называется распределением случайной величины.

Определение 3.2. Распределением случайной величины X называют меру

$$P_X(A) = P(\omega : X(\omega) \in A),$$

где $A \subseteq \mathbb{R}$.

То есть на прямой появляется новая мера, которая каждому подмножеству прямой сопоставляет вероятность тех ω , при которых X попало в множество A .

Вероятностные меры в простом дискретном случае можно задать с помощью вероятности разных возможных исходов:

$$P_X(A) = \sum_{x_i \in A} P(\omega : X(\omega) = x_i),$$

где $\mathcal{X} = \{x_1, \dots\} = \text{Im} X$.

Итак, чтобы задать распределение случайной величины, нужно задать \mathcal{X} — те значения, которые может принимать случайная величина, то есть образ нашего отображения. И для каждого из элементов этого образа задать вероятность того, что наша величина равна именно этому. После этого вероятность попасть в множество это просто $\sum X(\omega)$ по этому множеству.

По существу, каждая величина создает новое вероятностное пространство \mathcal{X} с вероятностью $P_X(A)$.

Пример 3.3. Посмотрим на примере случайных величин $X_1(\omega)$, $X_2(\omega)$, $X_3(\omega)$ из примера 3.2 как это работает.

$$X_1 : \mathcal{X} = \{0, 1\};$$

$$P_{X_1}(0) = \frac{1}{2}; P_{X_1}(1) = \frac{1}{2}.$$

Величина X_2 другая, она характеризует другой итоговый эксперимент, они с X_1 не совпадают, но распределения у них одинаковые. Мы будем часто с таким встречаться.

$$X_3 : \mathcal{X} = \{0, 1, 2\};$$

$$P_{X_3}(0) = \frac{1}{4}; P_{X_3}(1) = \frac{1}{2}; P_{X_3}(2) = \frac{1}{4}.$$

Соответственно, чтобы задать распределение величины, нужно задать вероятности всех исходов.

Теперь допустим, мы захотели что-то про случайную величину посчитать. Например, найдём вероятность события, что $X_3(\omega)$ чётна:

$$P(\omega : X_3(\omega) : 2) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

Точно также можем посчитать вероятность любого события про случайную величину. Так что мы задаём вероятности значения возможной случайной величины, а исходя из этого задаются вероятности исходов. Не так важно, какая была случайная величина, как она именно зависела от ω , сколько важно, какое у неё распределение с точки зрения этого. Соответственно, каждая случайная величина индуцирует тройку $(\mathcal{X}, 2^{\mathcal{X}}, P_X)$. Можно сказать, что это вероятностное пространство, заданное случайной величиной. Если мы хотим изучать величину X и ничего более, то тогда мы можем изучать только это пространство. Поэтому мы будем говорить, что величина имеет какое-то распределение и не уточнять, на каком Ω она задана, потому что это не важно при работе с этой случайной величиной.

Задать распределение можно несколькими способами, в частности, можно использовать табличку со всеми возможными значениями величины и соответствующими им вероятностями (например, для X_3 это будет таблица 3.1).

0	1	2
$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Таблица 3.1: Распределение случайной величины X_3 из примера 3.3

Вообще, для распределения формально нужно задавать всю меру, но задаётся эта мера вероятностями отдельных исходов, и поэтому можно задать такие вероятности либо таблицей, либо напрямую.

Упражнение 3.1. Закон Бенфорда говорит, что распределение первой цифры случайно выбранного числа, будет иметь распределение $P(X = x) = c \log_{10}(1 + x^{-1})$, $1 \leq x \leq 9$. Найти c , $P(X \leq k)$, $P(X : 2)$.

Указание: константу c можем определить исходя из того, что сумма всех вероятностей равна 1. Не забывайте про свойства логарифмов.

Сигма-алгебра

Введём ещё одно важное понятие:

Определение 3.3. σ -алгеброй (или алгеброй, если величина конечна), порождённой случайной величиной, называют набор событий $\{\omega : X(\omega) \in A\}$

для всех возможных $A \subseteq \mathbb{R}$.

То есть это прообраз всевозможных множеств с прямой. Для многих A эти множества одни и те же, мы их считаем один раз.

Пример 3.4. Вернёмся к величинам X_1 и X_3 из примера 3.2.

Напомним, что $X_1(i_1, i_2) = i_1$. Наша сигма-алгебра будет состоять из событий

$$D_i = \{\omega : X(\omega) = x_i, x_i \in \mathcal{X}\},$$

где D_i — множество тех ω , при которых $X(\omega)$ принимает значение x_i . Такие множества, а также все возможные их объединения дают сигма-алгебру, порождённую величиной. Для нашей случайной величины $\mathcal{X} = \{0, 1\}$ и

$$D_0 = \{(0, 0), (0, 1)\};$$

$$D_1 = \{(1, 0), (1, 1)\}.$$

Соответственно, сигма-алгебра, порождённая случайной величиной состоит из D_0 , D_1 , из объединения D_0 и D_1 и из пустого множества.

По сути, сигма-алгебра слепила исходы, на которых случайная величина приняла одно и то же значение.

Напомним, что $X_3(i_1, i_2) = i_1 + i_2$. В данном случае

$$D_0 = \{(0, 0)\};$$

$$D_1 = \{(1, 0), (0, 1)\};$$

$$D_2 = \{(1, 1)\}.$$

Опять же, исходы каким-то образом склеились. Чаще всего у нас часть исходов в вероятностном пространстве склеиваются, чтобы дать один результат, в данном случае, этот результат — это суммарное число орлов. Соответственно, распределение случайной величины — вероятности D_0 , D_1 , D_2 . В каком-то смысле порождающие сигма-алгебры D_i осуществляют склейку между исходным пространством и самой случайной величиной.

Маргинальное и совместное распределения

Пример 3.5. Пусть есть две случайные величины с известным распределением: $P(X = 0) = \frac{1}{2}$, $P(X = 1) = \frac{1}{2}$; $P(Y = 0) = \frac{1}{2}$, $P(Y = 1) = \frac{1}{2}$ (мы постепенно перестаём писать ω для краткости, но всегда их здесь подразумеваем). Из такого задания распределений мы не можем найти $P(\omega : X(\omega) = Y(\omega))$. Приведём примеры отображений X и Y , для которых эта вероятность будет разной:

I. $\omega = (i_1, i_2); X(\omega) = i_1, Y(\omega) = i_1 \implies P(\omega: X(\omega) = Y(\omega)) = 1;$

II. $\omega = (i_1, i_2); X(\omega) = i_1, Y(\omega) = 1 - i_1 \implies P(\omega: X(\omega) = Y(\omega)) = 0;$

III. $\omega = (i_1, i_2); X(\omega) = i_1, Y(\omega) = i_2 \implies P(\omega: X(\omega) = Y(\omega)) = \frac{1}{2}.$

Поэтому распределения X и Y (*маргинальные распределения*) не определяют вероятностей, где X и Y фигурируют вместе. Значит, чтобы задать *совместное распределение* X и Y , то есть уметь искать вероятности про X и Y вместе, нужно задавать что-то большее, чем просто одномерное распределение.

Определение 3.4. Совместное распределение X_1, \dots, X_n — набор вероятностей того, что все величины одновременно приняли какие-то фиксированные значения:

$$P(\omega: X_1(\omega) = i_1, \dots, X_n(\omega) = i_n, i_1 \in \mathcal{X}_1, \dots, i_n \in \mathcal{X}_n).$$

В английском языке для совместного распределения используется термин *probability mass function* (*функция масс*).

Теперь мы можем посчитать вероятность попадания вектора (X_1, \dots, X_n) в множество A :

$$P(\omega: (X_1(\omega), \dots, X_n(\omega)) \in A) = \sum_{(i_1, \dots, i_n) \in A} P(\omega: X_1(\omega) = i_1, \dots, X_n(\omega) = i_n).$$

Принципиальных изменений не произошло. Теперь мы рассматриваем новое пространство, где все случайные величины находятся вместе. Итогом опыта будет вектор (X_1, \dots, X_n) , а новую \mathcal{X} задаёт декартово произведение $\mathcal{X}_1, \dots, \mathcal{X}_n$. И в этом пространстве мы уже можем посчитать вероятность любого события, просто сложив вероятности всех подходящих исходов.

Совместное распределение в двухмерном случае можно задать таблицей.

Пример 3.6. Построим таблицы совместного распределения для случайных величин, приведённых в примере 3.5:

I. $X(\omega) = Y(\omega) = i_1$ — таблица 3.2.

$X \backslash Y$	0	1
0	$\frac{1}{2}$	0
1	0	$\frac{1}{2}$

Таблица 3.2: Совместное распределение X, Y при $X(\omega) = Y(\omega) = i_1$

Можем заметить некоторые свойства таблицы. Например, что сумма всех чисел равна 1. Мы знаем, что найти вероятность вектора — сложить все подходящие исходы, поэтому чтобы найти распределение X , нужно найти сумму по всем строчкам: $P(X = 0) = \frac{1}{2}$, $P(X = 1) = \frac{1}{2}$, а чтобы найти распределение Y , нужно найти сумму по всем столбцам: $P(Y = 0) = \frac{1}{2}$, $P(Y = 1) = \frac{1}{2}$. Также сумме двух диагональных элементов соответствует $P(X = Y) = 1$.

II. $X(\omega) = i_1$, $Y(\omega) = 1 - i_1$ — таблица 3.3.

$X \setminus Y$	0	1
0	0	$\frac{1}{2}$
1	$\frac{1}{2}$	0

Таблица 3.3: Совместное распределение X , Y при $X(\omega) = i_1$, $Y(\omega) = 1 - i_1$

Отсюда видим, что X и Y точно так же распределены, каждый из них равен 0 или 1 в вероятность $\frac{1}{2}$, но вероятность события $X(\omega) = Y(\omega)$, соответствующая сумме диагональных элементов, равна 0.

III. $X(\omega) = i_1$, $Y(\omega) = i_2$ — таблица 3.4.

$X \setminus Y$	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$

Таблица 3.4: Совместное распределение X , Y при $X(\omega) = i_1$, $Y(\omega) = i_2$

Здесь каждому варианту соответствует ровно один исход, например, если $X = 0$ и $Y = 0$, значит, исход $(i_1, i_2) = (0, 0)$, вероятность его исхода равна $\frac{1}{4}$. В этом случае сумма диагональных элементов $P(X = Y) = \frac{1}{2}$.

Соответственно, когда мы знаем распределения величин X и Y , мы знаем сумму по строкам и сумму по столбцам, но это не даёт нам восстановить таблицу. У таблиц 3.2 – 3.4 сумма по строкам и сумма по столбцам всё время $\frac{1}{2}$, а таблички разные. Поэтому нельзя задав распределения X и Y узнать распределение X и Y вместе, это нечто большее.

Независимые случайные величины

Определение 3.5. Величины X , Y называются **независимыми**, если

$$P(\omega : X(\omega) \in A, Y(\omega) \in B) = P(\omega : X(\omega) \in A)P(\omega : Y(\omega) \in B) \quad \forall A, B \subseteq \mathbb{R}.$$

Соответственно, если независимость есть, тогда легко найти совместное распределение по одномерному, потому что это просто произведение вероятностей.

Замечание 3.1. X, Y — независимы \iff

$$\iff P(\omega : X(\omega) = x, Y(\omega) = y) = P(\omega : X(\omega) = x)P(\omega : Y(\omega) = y).$$

Отличие в том, что множества A и B можно рассматривать только одноточечные, это гораздо удобнее, потому что не нужно возиться с разными трудностями.

Замечание 3.2. Если совместное распределение X и Y представимо в виде произведения двух функций, одна из которых зависит только от x , а вторая только от y , то этого уже достаточно, чтобы величины были независимы.

С точки зрения таблицы независимость означает, что вероятность попасть в каждую ячейку должна быть равна произведению суммы по строке и суммы по столбцу. Таким образом, условие независимости — очень жёсткое условие на таблицу распределения, она должна быть фиксированная, тогда как любые числа, в сумме дающие 1 задают распределение. Можно прямо так Ω и построить — эти числа приписать исходам.

Определение 3.6. X_1, \dots, X_n — независимые случайные величины, если

$$\begin{aligned} P(\omega : X_1(\omega) \in A_1, \dots, X_n(\omega) \in A_n) &= P(\omega : X_1(\omega) \in A_1) \dots P(\omega : X_n(\omega) \in A_n) \iff \\ &\iff \exists g_1, \dots, g_n : P(\omega : X_1(\omega) = x_1, \dots, X_n(\omega) = x_n) = g_1(x_1) \dots g_n(x_n) \iff \\ &\iff P(\omega : X_1(\omega) = x_1, \dots, X_n(\omega) = x_n) = P(\omega : X_1(\omega) = x_1) \dots P(\omega : X_n(\omega) = x_n) \end{aligned}$$

Если вспомните определение независимости n событий, то там были некоторые оговорки, что для любого поднабора вероятность пересечения должна быть равна произведению вероятностей. Тут таких оговорок нет, потому что если мы хотим рассматривать поднабор, можно просто часть A_i положить \mathbb{R} . Функции g_i автоматически окажутся пропорциональны $P(\omega : X_i(\omega) = x_i)$, можно убедиться в этом, просто просуммировав.

Поскольку распределение случайных величин — это вероятность каких-то событий, то те «трюки», которым мы научились при работе с событиями, верны и здесь.

Пример 3.7. Пусть мы бросили шестигранный кубик, он выпал на какой-то X , а затем мы бросили X симметричных монет. Число орлов на них в

сумме — это Y . Хотим найти распределение Y . Начнём с того, что построим множество значений Y :

$$\mathcal{Y} = \{0, \dots, 6\}.$$

Y — агрегированная величина, он не может учитывать X , так как X случайный. Все вероятности явно считать не будем, найдём какую-нибудь конкретную, например, $P(\omega: Y(\omega) = 3)$. Сложность в том, что мы не знаем, на сколько кубик выпал, поэтому не знаем, сколько бросали монет. Надо весь наш мир случайностей разбить на несколько кусков и в каждом из них задачу решить, это нам позволяет сделать формула полной вероятности (2.3).

$$P(\omega: Y(\omega) = 3) = \sum_{k=1}^6 P(Y = 3 | X = k) P(X = k) = \sum_{k=3}^6 C_k^3 \left(\frac{1}{2}\right)^k \cdot \frac{1}{6}.$$

Условная вероятность — это вероятность в том мире, где мы уже точно знаем, что бросаем k монет. А вероятность, что из k брошенных монет будет ровно 3 орла, известна — это схема Бернулли; вероятность выпадения конкретного k равна $\frac{1}{6}$. Дальше нужно посчитать эту сумму, и мы найдём распределение Y .

Поэтому те же самые приёмы, которые мы использовали для событий, остаются для случайных величин: неизвестно что-то — дофиксируй; неудобно работать с Y , когда не знаешь X — рассмотри формулу полной вероятности по значениям X .

По сути, можно было найти совместное распределение X, Y , а затем просуммировать по X . На самом деле $P(Y = 3 | X = k)P(X = k)$ и есть совместное распределение и мы производили суммирование по столбцу, чтобы найти вероятность Y .

Формула свёртки

Чтобы найти распределение суммы двух величин, нужно просуммировать табличку по диагональке, где $X + Y = \text{const}$:

$$P(X + Y = x) = \sum_i P(X = x_i, Y = x - x_i).$$

Если величины X, Y независимые, то тогда формула упрощается до:

$$P(X + Y = x) = \sum_i P(X = x_i)P(Y = x - x_i).$$

Это обычно и называют **формулой свертки**.

Некоторые базовые распределения

Ну и напоследок несколько базовых распределений:

1) *Бернуллиевское распределение* $Bern(p)$. У него

$$\mathcal{X} = \{0, 1\}; P_X(0) = 1 - P_X(1) = 1 - p.$$

Физически *распределение Бернулли* соответствует броску монетки, p — это какой-то параметр этого распределения, например, вероятность орла.

2) *Биномиальное распределение* $Binom(n, p)$. У него

$$\mathcal{X} = \{0, \dots, n\}; P_X(k) = C_n^k p^k (1 - p)^{n-k}.$$

Физически такое распределение соответствует количеству успехов в n испытаниях.

3) *Геометрическое распределение* $Geom(p)$. У него

$$\mathcal{X} = \{0, 1, \dots\}, P_X(k) = p(1 - p)^k.$$

Физически такое распределение соответствует числу неудач до первого успеха (или числу выпавших решек до первого орла). Есть разночтения по поводу того, что именно называть геометрическим. Иногда называют число бросаний до первого орла, тогда чуть изменится запись вероятности, поэтому эта формула может отличаться в разных источниках. Всего есть 4 принятых формы записи.

4) *Пуассоновское распределение* $Poiss(\lambda)$. У него

$$\mathcal{X} = \{0, 1, \dots\}, P_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Это предельное, в некотором смысле, распределение. Оно само по себе встречается редко, а вот в качестве приближения для биномиального — часто. Если взять в биномиальной схеме $p = \frac{\lambda}{n}$ и перейти к пределу, получим именно пуассоновское распределение. Но оно более тонкое и не требует, чтобы все события были одинаковой вероятности, и довольно удачно аппроксимирует все нечастые события в окружающей жизни. *Теорема Пуассона*, закон редких событий, будет рассмотрена в нашем курсе позднее.

Пример 3.8. В реальной жизни одно из первых применений *распределения Пуассона* было примерно в наполеоновские времена, когда считали количество прусских солдат, убитых ударом копыта лошади в лоб. Ясно, что это

не самый популярный вид смерти у прусских солдат, но солдат много, лошадей много, и время от времени кто-то погибает. Это типичная история, как возникает пуассоновское распределение: есть событие, которое само по себе маловероятное, но таких событий происходит много, и за долгое время какое-то количество будет реализовываться.

Пример 3.9. Взяв произвольную машину, её вероятность попадания в аварию сегодня, очень маленькая. Но машин в рамках, например, района, очень много и скорее всего кто-то туда всё-таки попадёт. Это история, похожая на пуассоновское распределение. Скорее всего, с какой-то погрешностью, вероятность того, что аварий будет именно k близка к пуассоновскому распределению с каким-то параметром λ .

Поэтому это распределение не такое простое, чтобы можно было его показать на шариках или монетках, но это распределение следующего этапа — предельное. Для нас такие распределения очень важны, потому что вся красота в теории вероятностей начинается именно с пределов.

5) *Равномерное распределение* $R\{1, \dots, N\}$ (можно задать его и на любом другом множестве мощности N). У него

$$\mathcal{X} = \{1, \dots, N\}, P_X(k) = \frac{1}{N}.$$

Это распределение описывает номер шара, вытасченного из урны с N шарами.

Упражнение 3.2. Изучите отрицательное биномиальное $NegBinom(r, p)$ и гипергеометрическое $Hypergeom(N, D, p)$ распределения самостоятельно.

6) *Пономиальное распределение* $Polynom(n, p_1, \dots, p_n)$, $\sum p_k = 1$. У него

$$\mathcal{X} = \{(i_1, \dots, i_k), \sum i_k = n, i_j \geq 0, i_j \in \mathbb{Z}\};$$

$$P_X((i_1, \dots, i_k)) = \frac{n!}{i_1! \dots i_k!} p_1^{i_1} \dots p_n^{i_n}.$$

В иностранной литературе полиномиальное распределение часто называют *мультиномиальное распределение*.

Семинар 4. Математическое ожидание

Определение математического ожидания

Пусть X — случайная величина на (Ω, \mathcal{F}, P) . Как всегда, Ω у нас конечное или счётное множество, поэтому можем считать, что $\Omega = \{\omega_1, \omega_2, \dots\}$.

Определение 4.1. Математическим ожиданием $X(\omega)$ называют

$$EX(\omega) = \sum_i X(\omega_i)P(\omega_i), \quad (4.1)$$

если этот ряд сходится абсолютно.

Иногда вместо обозначения EX используют обозначение MX . Они произошли от аббревиатур *expectation* и *moment* соответственно.

Нам важно, чтобы ряд сходился абсолютно, иначе по теореме Римана мы можем переставить элементы местами и получить наперёд заданную сумму, но никакая характеристика случайной величины не должна зависеть от того, в каком порядке мы пронумеровали исходы, поэтому проще выбросить этот случай и не сталкиваться с изменением суммы при перестановке мест слагаемых.

В классическом пространстве математическое ожидание будет просто средним арифметическим всех значений случайной величины, в неклассическом можно мысленно представить, что оно похоже на классическое, только исходы агрегированы в группы.

Пример 4.1. Если один исход имеет вероятность $\frac{2}{3}$, а другой $\frac{1}{3}$, то можно считать, что это два исхода по $\frac{1}{3}$ и ещё один исход в $\frac{1}{3}$. Тогда получится, что это также соответствует среднему арифметическому.

Если поиграться немного с классическими пространствами, можно понять, что это та самая формула, которую мы в голове представляем, когда говорим про среднее значение величины.

Немного доклассифицируем происходящее, введя следующие величины:

$$X^+(\omega) = \max(X(\omega), 0);$$
$$X^-(\omega) = -\min(X(\omega), 0).$$

Тогда $X(\omega) = X^+(\omega) - X^-(\omega)$. У X^+ и X^- ряд всегда сходится абсолютно, потому что там только положительные значения. Там проблем со сходимостью быть не может — ряд либо сходится, либо расходится в $+\infty$.

Определение 4.2.

- Если EX^+ и EX^- конечны, то говорят, что EX **конечно** и

$$EX = EX^+ - EX^-.$$

- Если одно из EX^+ и EX^- конечно, а другое бесконечно, то говорят, что математическое ожидание либо $+\infty$, либо $-\infty$ соответственно (**бесконечно**).
- Если оба EX^+ и EX^- бесконечны, тогда EX **не существует**.

Несложно доказывается формула

$$EX = \sum_{x_k} x_k P(X = x_k). \quad (4.2)$$

Для этого нужно подставить в (4.1) и сгруппировать одинаковые слагаемые. Различие здесь только в том, что здесь на разных ω_i может быть одно и то же x_k , но если мы возьмём все $\omega_i : X(\omega_i) = x_k$, то можно просуммировать их вероятность и получить $P(X = x_k)$.

У нас есть два взгляда на случайные величины: иногда удобно смотреть с точки зрения вероятностного пространства, иногда удобно смотреть с точки зрения распределения. С точки зрения математического ожидания им соответствуют формулы (4.1) и (4.2).

Пример 4.2. Пусть $X \sim \text{Bern}(p)$. Это случайная величина, принимающая значение 1 с вероятностью p и значение 0 с вероятностью $1 - p$. Тогда её математическое ожидание

$$EX = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Пример 4.3. Пусть $X \sim \text{Binom}(n, p)$. Тогда $P(X = k) = C_n^k p^k (1 - p)^{n-k}$, $k = 0, \dots, n$. Математическое ожидание в таком случае равно

$$\begin{aligned} EX &= \sum_{k=0}^n C_n^k p^k (1 - p)^{n-k} k = \sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^k (1 - p)^{n-k} = \\ &= np \cdot \underbrace{\sum_{k=1}^n C_{n-1}^{k-1} p^{k-1} (1 - p)^{n-1-(k-1)}}_{=(p+1-p)^{n-1}=1 \text{ — бином Ньютона}} = np. \end{aligned}$$

Не забывайте под знаком суммы вероятность домножить на значение случайной величины, иначе у вас сумма ряда будет равна 1.

Если возьмём заданную функцию от вектора $g: \mathbb{R}^k \rightarrow \mathbb{R}$ и заданные случайные величины X_1, \dots, X_k , вообще говоря, зависимые. Хотим посчитать $Eg(X_1, \dots, X_k)$. Можем представить, что X_1, \dots, X_k своим совместным распределением задают новое пространство, в котором живут X_1, \dots, X_k . В этом пространстве исходы — всевозможные векторы, а вероятности этих исходов — распределение случайного вектора. Соответственно, по формуле (4.2)

$$Eg(X_1, \dots, X_k) = \sum_{\vec{x}_i \in \mathcal{X}} g(x_{i,1}, \dots, x_{i,k}) P(X_1 = x_{i,1}, \dots, X_k = x_{i,k}).$$

Сумма идёт по всем возможным векторам x_i , которые вообще бывают значениями случайного вектора.

Принцип тот же самый, что и был, но немного поменялась формулировка. По сути мы здесь говорим следующее — рассматриваем возможные значения вектора в качестве элементарных исходов, а в качестве вероятности этих элементарных исходов — дискретное совместное распределение вектора. Тогда мы должны умножить значение случайной величины $g(X_1, \dots, X_k)$ на вероятность исхода. Эта формула очень удобная, потому что нам не нужно искать распределение этой функции, перестраивать табличку распределений и думать, какие значения принимает функция, вместо этого можем воспользоваться исходным распределением. Соответственно, имея таблицу распределения вектора, мы сможем легко посчитать математическое ожидание любой функции от него: нужно взять вероятность ячеек, умножить на значение функции в этой ячейке и просуммировать по всем ячейкам.

В принципе, подсчёт математического ожидания случайного вектора или случайной величины с известным распределением — это проблема подсчёта какой-то суммы. Тут бывает полезно использовать некоторые специфические приёмы для подсчёта сумм. Рассмотрим один из таких приёмов.

Пример 4.4. Пусть $X \sim Geom(p)$. Найдём EX :

$$EX = \sum_0^{\infty} k p (1-p)^k.$$

Чтобы посчитать эту сумму, рассмотрим вспомогательный ряд

$$\varphi_x(s) = \sum_{k=0}^{\infty} s^k \cdot p (1-p)^k,$$

где s — вспомогательный параметр. Сумма такого ряда называется *производящей функцией*, с которой мы будем работать на следующем семинаре. Это

степенной ряд, будем считать, что мы знаем, что он равномерно сходится, тогда $\varphi'(1) = EX$. Это удобный приём, потому что посчитать $\varphi_x(s)$ гораздо легче — это бесконечная геометрическая прогрессия:

$$\varphi_x(s) = p \sum_{k=0}^{\infty} ((1-p)s)^k = \frac{p}{1-(1-p)s}.$$

Продифференцировав по s , получим

$$EX = \varphi'(1) = -\frac{-p(1-p)}{(1-(1-p)s)^2} \Big|_{s=1} = \frac{1-p}{p}.$$

В каком-то смысле можно сказать, что мы просуммировали этот ряд по Абелю–Пуассону. Этот метод бывает часто полезен, потому что в вероятностях нередко встречаются выражения, которые при умножении на s^k суммируются хорошо, а при умножении на k — не очень.

Свойства математического ожидания

Рассмотрим несколько свойств математического ожидания. В рамках данного семинара мы будем в основном концентрироваться на одном из свойств, потому что это один из самых популярных способов как считать математическое ожидание.

Ещё раз отметим эту формулу, но уже выпишем ее отдельно для одномерного случая:

$$1) \quad E g(X) = \sum_{x_k} g(x_k) P(X = x_k).$$

Если нам надо посчитать математическое ожидание функции от случайной величины, мы можем сразу складывать значения функции на вероятности случайной величины. Почему так, мы уже обсуждали. Это очень удобно и при решении задач обычно проще перемножить вероятность того, что X равно чему-то и значение функции, чем искать распределение $g(X)$.

Мы не будем вписывать все существующие свойства математического ожидания, а сконцентрируемся на тех, которые нам сейчас наиболее важны

$$2) \quad E(X+Y) = EX + EY; \quad E(cX) = cEX \quad \text{— линейность математического ожидания.}$$

Это выполнено, если хотя бы два из трёх математических ожиданий конечны. Здесь очень важно то, что мы про X, Y не говорим, зависимы ли

они, одинаково ли распределённые ли они, потому что это неважно. С точки зрения определения через распределения (4.2) это свойство совершенно не очевидно, потому что распределение $X + Y$ может быть как угодно связано с распределением X и Y , потому что они, вообще говоря, зависимы. А с точки зрения исходного определения (4.1) это очевидное свойство, потому что ряд в левой части выражения можно разбить на два из правой части. Это оказывается фурором, потому что раньше ничего такого мы не видели: искать распределение суммы — ужасно тяжёлое занятие, а искать математическое ожидание суммы — ужасно приятное, даже если величины зависимы или их очень много.

Пример 4.5. Если $X \sim \text{Binom}(n, p)$, то тогда $X = X_1 + \dots + X_n$, где $X_i = 1$, если i -ая монетка выпала на орла и $X_i = 0$, если i -ая монетка выпала на решку. Линейность математических ожиданий позволяет сказать, что $EX = EX_1 + \dots + EX_n = np$, так как $EX_i = p$ (см. пример 4.2). Мы уже получали данный результат через бином в примере 4.3, но этот способ подсчёта математического ожидания гораздо проще.

В связи с этим, приобретает особую важность умение распознать в случайной величине сумму каких-то величин. Это особенно хорошо делать, если случайная величина называется «количество» чего-то, потому что количество мы хорошо раскладываем в сумму: мы когда пальцы загибаем, чтобы что-то посчитать, мы как раз это и делаем. Рассмотрим ещё один, менее тривиальный, пример:

Пример 4.6. Есть N ячеек и n шаров. Мы берём шарики по очереди и бросаем наугад в какую-то ячейку. Ячейки большие, могут хоть все шарики попасть в одну. X — число занятых ячеек (ячеек, в которые хоть кто-то попал). В такой задаче довольно сложно искать распределение X , а математическое ожидание X найти просто, потому что $X = X_1 + \dots + X_N$, где $X_i = 1$, если i -ая занята и $X_i = 0$, если не занята. Эти величины, безусловно, зависимы. Например, если $n = 1$, то если одна ячейка занята, то остальные пустые. Но это неважно, потому что в свойстве линейности нет никакого условия на независимость. С учётом симметрии задачи,

$$EX = \sum_{i=1}^N EX_i = N EX_1.$$

Посчитаем EX_1 :

$$EX_1 = P(X_1 = 1) = 1 - P(X_1 = 0) = 1 - \left(\frac{N-1}{N}\right)^n.$$

Итого получается

$$EX = N \left(1 - \left(\frac{N-1}{N} \right)^n \right).$$

Это очень популярный подход подсчёта математического ожидания, потому что не нужно искать сложное распределение сумм, а затем сворачивать сумму, состоящую из них.

Еще одно свойство математического ожидания, которое пригодится нам позднее:

3) Если X_i — независимые, то

$$EX_1 \dots X_n = EX_1 \dots EX_n.$$

Дисперсия

Определение 4.3. Дисперсией случайной величины X называется

$$DX = E(X - EX)^2 = \sum_{x_k} (x_k - EX)^2 P(X = x_k). \quad (4.3)$$

Механически она представляет собой момент инерции точек x_k относительно центра масс. В некотором смысле, это характеристика, насколько далеко случайные величины растянуты вокруг EX . Из определения сразу следует, что дисперсия — величина неотрицательная, потому что является средним неотрицательной случайной величины.

Пример 4.7. Кто-то может сказать, что в среднем он зарабатывает в день 100 рублей, при этом совершенно непонятно, в каких пределах может меняться зарплата. Может, в какие-то дни теряется миллион рублей, а в другие получается миллион рублей и в среднем выходит 100 рублей в день. А может, зарплата меняется от 99 до 101 рубля. Во втором случае гораздо большая стабильность, а в другом случае меньшая. Поэтому кроме среднего хотелось бы понимать, насколько случайная величина вокруг этого среднего разбросана.

Ближе к концу семестра мы научимся с помощью этого получать довольно интересные оценки, а в самом конце сформулируем центральную предельную теорему, где эти понятия соединятся и очень важную роль окажут.

Раскроем квадрат в формуле (4.3), получим

$$DX = EX^2 - 2EXEX + (EX)^2 = EX^2 - (EX)^2. \quad (4.4)$$

Не стоит путать EX^2 и $(EX)^2$, это разные вещи — среднее квадрата величины и квадрат среднего. Так как дисперсия очевидно не отрицательная, то $EX^2 \geq (EX)^2$.

Теперь поймём, как дисперсию считать — по формулам (4.3) и (4.4):

Пример 4.8. Посчитаем дисперсию для $X \sim \text{Bern}(p)$. Так как X^2 принимает значения только 0 и 1 с теми же вероятностями, что и X , то $EX^2 = EX = p$. Тогда

$$DX = EX^2 - (EX)^2 = p - p^2 = p(1 - p).$$

Такая дисперсия удовлетворяет каким-то представлениям о разбросе даже с точки зрения здравого смысла: если p очень маленькое, то и разброс маленький, потому что если p почти ноль — орлов почти нет, мы почти точно знаем, что величина равна среднему, то есть нулю. Точность будет высокой, а погрешность маленькой.

Свойства дисперсии и ковариации

Опять же, для подсчёта дисперсии нам удобно использовать разные свойства. Прямой линейности у дисперсии нет, но есть что-то похожее:

$$1) D(cX) = c^2DX.$$

Это легко получить, если подставить cX в формулу (4.3). Если по ошибке вынести c вместо c^2 , получится отрицательная дисперсия, чего быть не может.

$$2) D(X + Y) = DX + DY + \underbrace{2E(X - EX)(Y - EY)}_{\text{cov}(X,Y) \text{ — ковариация}}.$$

Строго говоря, название «ковариация» — это некий курьёз, как и всякий раз бывает, когда в разное время переводят термины разные люди. Потому что в английском всё понятно: дисперсия называется *variance*, а ковариация — *covariance* и ясно, что между ними есть связь. А по-русски дисперсию как вариацию переводить не захотели, потому что слово вариация имеет очень много смыслов, вы знаете это из анализа, и назвали дисперсией, по смыслу тоже самое. Но тогда надо было сделать кодисперсию вроде бы, но вот здесь решили перевести ковариация. В итоге они потерялись, но на самом деле это дисперсия и кодисперсия. Возникшая ковариация нарушает линейность. Ясно, что в общем случае формула точно такая же:

$$D(X_1 + \dots + X_n) = \sum_i DX_i + 2 \sum_{i < j} \text{cov}(X_i, X_j).$$

Мы пишем $i < j$ чтобы посчитать каждую пару ровно один раз. В каком-то смысле у нас что-то типа линейности сохранилось, но только в каком-то смысле.

Теперь свойство 3 математического ожидания приобретает для нас важную роль, потому что получается:

3) Если X, Y — независимы, то $\text{cov}(X, Y) = E(X - EX) \cdot E(Y - EY) = 0$.

В среднем величина, из которой вычли среднее, равна нулю. Поэтому для независимых величин ковариация равна нулю. Соответственно в этом случае дисперсия суммы просто равна сумме дисперсий, если величины независимые. Но если величины зависимые, то увы и ах, может быть не равна сумме дисперсий, но может быть и равна, это не критерий, это только достаточное условие. Бывает, что величины зависимые, а ковариация всё равно нулевая.

$$4) \text{cov}(X_1 + \dots + X_k, Y) = \sum_{i=1}^k \text{cov}(X_i, Y).$$

Ковариация уже линейна линейная в полном смысле этого слова, если быть точным, билинейная. Мы пишем по первому аргументу, но ковариация симметричная, поэтому можно также написать по второму аргументу. Поэтому с ковариацией в этом плане всё просто: ковариация суммы равна сумме ковариаций. Из этих соображений можно уже работать и с ковариациями, и с дисперсиями суммы. Посмотрим на примере, как это работает:

Пример 4.9. Найдём дисперсию X из примера 4.6 про ячейки. Напомним, $X = X_1 + \dots + X_n$, X_i — заполнена ли i -ая ячейка. Все дисперсии X_i равны между собой, потому что ячейки симметричные, в какую именно бросать не важно. Тоже самое можно сказать про $\text{cov}(X_i, X_j)$, потому что какую пару ячеек не рассматривай, задача абсолютно симметричная, соседние ячейки никак не связаны. Получается, что

$$DX = \sum_{i=1}^n DX_i + 2 \sum_{i < j} \text{cov}(X_i, X_j) = nDX_1 + n(n-1) \text{cov}(X_1, X_2).$$

Первый и второй мы взяли совершенно произвольные, просто чтобы их зафиксировать. В примере 4.6 мы уже определили, что $X_1 \sim \text{Bern}\left(1 - \left(\frac{N-1}{N}\right)^n\right)$, потому что он принимает значения 0 и 1, поэтому из примера 4.8

$$DX_1 = \left(\frac{N-1}{N}\right)^n \left(1 - \left(\frac{N-1}{N}\right)^n\right).$$

Теперь посчитаем ковариацию. Мы не можем сказать, что X_1 и X_2 независимы, поэтому придётся считать. Если раскрыть скобки в определении ковариации, получится такая же формула, как для дисперсии:

$$\text{cov}(X_1, X_2) = EX_1X_2 - EX_1EX_2.$$

Мы знаем $EX_1 = EX_2 = 1 - \left(\frac{N-1}{N}\right)^n$, поэтому второе слагаемое будет квадратом этого выражения. Разберёмся с X_1X_2 :

$$X_1X_2 = \begin{cases} 0, & \text{если хотя бы одна ячейка свободна;} \\ 1, & \text{если обе ячейки заняты.} \end{cases}$$

Проще всего посчитать, когда обе ячейки заняты, по формуле включений–исключений:

$$P(X_1X_2 = 1) = 1 - P(X_1 = 0) - P(X_2 = 0) + P(X_1 = 0, X_2 = 0).$$

Когда обе ячейки заняты, это означает, что не должна быть свободна первая и не должна быть свободна вторая, но когда мы складываем эти две вероятности, нужно вычесть вероятность пересечения, чтобы получить вероятность объединения. По сути мы воспользовались тем, что пересечение дополнений событий это объединение их самих. Вероятности, что свободна определённая ячейка мы уже знаем. Найдём вероятность, что обе ячейки свободны: каждый брошенный шарик может попасть в $N - 2$ ячейки из N возможных, тогда

$$P(X_1 = 0, X_2 = 0) = \left(\frac{N-2}{N}\right)^n.$$

Теперь нужно собрать всё вместе и получится дисперсия, равная огромной сумме, но мы смогли её посчитать. Это уже немало, потому что, опять же, распределения нам недоступны, и считать распределения напрямую мы бы не смогли.

Упражнение 4.1. Досчитайте пример самостоятельно.

Почему жизнь всё ещё хороша, когда мы работаем с дисперсией? С матожиданиями было совсем хорошо — мы работали с каждой X_i в отдельности, это было просто. С дисперсией мы работаем с каждой X_i в отдельности и ещё с каждой парой X_i , но только с парой, а чтобы изучать, сколько занятых ячеек, нужно работать со всеми возможными ячейками вместе, всё-таки с парами работать гораздо проще, чем со всеми вместе, как мы, собственно, и увидели.

Упражнение 4.2. Попробуйте записать вероятность того, что 5 ячеек заняты.

Свободны — легко, а вот заняты — тяжело. Там будет очень большая формула, поэтому получается очень удобный подход, не такой удобный, как для математического ожидания, но тоже рабочий, а в каких-то случаях такой же, как для математического ожидания.

Пример 4.10. Вернёмся к биномиальной величине. Пусть $X \sim \text{Binom}(np)$. Напомним, что биномиальное — сколько орлов выпало при n бросаниях монетки с вероятностью орла p . Мы X разложили в сумму X_i , где X_i — выпал ли орёл при i -м броске. DX_1 мы уже находили в примере 4.8. Ковариация X_i и X_j равна нулю, потому что X_i и X_j независимы. Тогда можем найти дисперсию X :

$$DX = DX_1 + \dots + DX_n + 2 \sum_{i < j} \text{cov}(X_i, X_j) = nDX_1 = np(1-p).$$

Таким образом, мы смогли по щелчку пальца посчитать дисперсию для биномиальной величины.

Семинар 5. Производящие функции

Производящая функция случайной величины

Определение 5.1. Формальный ряд $\varphi(s) = \sum_{n=0}^{\infty} a_n s^n$ — производящая функция последовательности $\{a_n\}$.

В принципе, не обязательно даже, чтобы этот ряд где-то сходился. В нуле он сходится заведомо и можно просто рассматривать такой формальный ряд, но мы будем рассматривать более удобную для нас штуку.

Определение 5.2. $\varphi_X(s)$ — производящая функция случайной величины X , если φ_X — производящая функция последовательности

$$\{P(X = k), k \geq 0\}.$$

Здесь $X \in \mathbb{N} \cup \{0\}$ — неотрицательная целочисленная случайная величина.

То есть мы берём такую последовательность и строим по ней производящую функцию. Иначе говоря,

$$\varphi_X(s) = \sum_{k=0}^{+\infty} P(X = k) s^k = E s^X,$$

где s — какой-то параметр. Можно смотреть на это как на определение, если вас смутили предыдущие тезисы. Мы с производящими функциями последовательности работать не будем.

Упражнение 5.1. Выведите формулу Бине для n -го числа Фибоначчи с помощью производящей функции последовательности.

Нам нужны только производящие функции случайных величин. В отличие от просто производящих функций мы точно можем сказать, при каких s эта функция заведомо определена: при $s = 1$ $\varphi_X(1) = 1$, значит, радиус сходимости как минимум 1. На самом деле, поскольку члены положительны, можно сказать, что в отрезке $[-1, 1]$ он сходится, но нас будет интересовать только отрезок $[0, 1]$. Вот получается такая функция. Зная $P(X = k)$, можем построить $\varphi_X(s)$; зная $\varphi_X(s)$, $s \in [0, 1]$, можем восстановить вероятности, разложив функцию в ряд с центром в нуле любым доступным способом и посмотрев на коэффициенты, например,

$$P(X = k) = \frac{f^{(k)}(0)}{k!}.$$

По крайней мере, что это легальный способ хранить нашу случайную величину таким вот образом. Почему это удобно — это вопрос другой.

Свойства производящей функции случайной величины

Давайте поймём, какими свойствами обладает φ как функция: она выпуклая, немонотонная, стартует из какой-то неотрицательной точки и приходит в точку $(1, 1)$ со всеми вытекающими. Так что сама по себе функция $\varphi_X(s)$ довольно симпатичная (см. рис. 5.1).

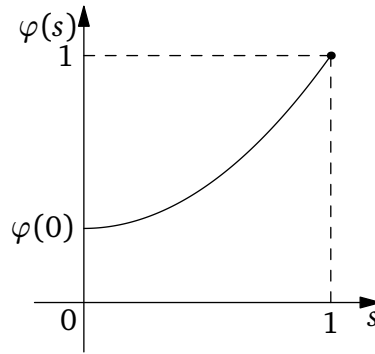


Рис. 5.1: Общий вид графика производящей функции

Найдём производную производящей функции в произвольной точке s :

$$\varphi'(s) = \sum_{k=0}^{\infty} s^{k-1} k P(X = k).$$

Воспользуемся теоремой Абеля, что если такой степенной ряд рассмотреть на границе круга сходимости в точке $s = 1$, если он сходится вообще, то

$$\varphi'(1) = \sum_{k=0}^{\infty} k P(X = k) = EX.$$

Поэтому получается такая удобная формула. Если быть более точным, то математическое ожидание существует тогда и только тогда, когда производящая функция дифференцируема в точке 1. Точно так же, что будет, если мы посчитаем, например, $\varphi_X''(1)$:

$$\varphi_X''(1) = EX(X - 1).$$

Соответственно точно также k -ая производная:

$$\varphi_X^{(k)}(1) = EX(X - 1) \dots (X - k + 1).$$

Выражение в правой части называется *факториальным моментом*. Выразить отсюда что-то, например, дисперсию, не очень удобно, но тем не менее можно:

$$DX = \varphi''(1) + \varphi'(1) - (\varphi'(1))^2.$$

Так или иначе, первые k факториальных моментов определяют и k обычных моментов EX, \dots, EX^k .

Пример 5.1. Пусть $X \sim \text{Bern}(p)$. Найдём его производящую функцию:

$$\varphi_X(s) = \mathbb{E}s^X = \sum s^k \mathbb{P}(X = k) = 1 - p + ps.$$

Соответственно можем отсюда найти легко

$$EX = \varphi'(1) = p.$$

Пример 5.2. Возьмём распределение поинтереснее: $X \sim \text{Poiss}(\lambda)$. Его производящая функция:

$$\varphi_X(s) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} s^k = e^{-\lambda} e^{-\lambda s} = e^{\lambda(s-1)}.$$

Тогда

$$\begin{aligned} EX &= \lambda e^{\lambda(s-1)} \Big|_{s=1} = \lambda; \\ EX(X-1) &= \lambda^2 \implies DX = \lambda^2 + \lambda - \lambda^2 = \lambda. \end{aligned}$$

Упражнение 5.2. Возьмём производящую функцию $\varphi_X(s) = c(1+s^2)^2$. а) Восстановите распределение X ; б) Найдите $\varphi_{2X+4}(s)$.

Указание: а) Чтобы восстановить распределение, можно продифференцировать, а можно разложить в ряд (в данном случае просто раскрыть скобки):

$$\varphi_X(s) = c(1 + 2s^2 + s^4).$$

Получается, что ненулевые вероятности $\mathbb{P}(X = 0) = c$, $\mathbb{P}(X = 2) = 2c$, $\mathbb{P}(X = 4) = c$. Осталось найти c , его можно найти из того, что сумма вероятностей равна 1 или из того, что $\varphi_X(1) = 1$ (это одно и то же). Поэтому $c \cdot 4 = 1 \implies c = \frac{1}{4}$.

б) Распишем по определению: $\varphi_{2X+4}(s) = \mathbb{E}s^{2X+4} = s^4 \mathbb{E}(s^2)^X = s^4 \cdot \varphi_X(s^2)$. Поэтому линейную комбинацию всегда легко выразить таким образом, тут никаких особенных хитростей нет. Если записать через ряд, мы бы увидели тоже самое.

Теперь поймём зачем нам производящие функции нужны. Хорошо, моменты можно с их помощью считать, но моменты и без них можно считать. Ключевым свойством для нас является следующее:

Утверждение 5.1. Пусть X_i — независимые величины, тогда

$$\varphi_{X_1+\dots+X_n}(s) = \varphi_{X_1}(s) \dots \varphi_{X_n}(s).$$

ДОКАЗАТЕЛЬСТВО. Если X_i независимы, то и функции от них тоже независимы (в частности, e^{X_i}), поэтому

$$\varphi_{X_1+\dots+X_n}(s) = \mathbb{E}s^{X_1+\dots+X_n} = \mathbb{E}s^{X_1} s^{X_2} \dots s^{X_n} = \mathbb{E}s^{X_1} \mathbb{E}s^{X_2} \dots \mathbb{E}s^{X_n}. \quad \blacksquare$$

Получается, что производящая функция — очень удобный аппарат для работы с суммой величин. Для величин просто мы знаем формулу свёртки, она довольно некрасивая, нужно сворачивать объёмные суммы, а с производящими функциями всё понятно. Поэтому возникает такой удобный вариант: задать распределение производящей функцией, перейти к распределению суммы, а потом вернуться назад к распределениям.

Пример 5.3. Пусть $X_i \sim \text{Poiss}(\lambda_i)$, независимые. Рассмотрим случайную величину $X_1 + \dots + X_n$. Её производящая функция в силу нашего свойства равна произведению производящих функций, каждую из которых мы уже находили в примере 5.2:

$$\varphi_{X_1+\dots+X_n}(s) = \varphi_{X_1}(s) \dots \varphi_{X_n}(s) = e^{\lambda_1(s-1)} \dots e^{\lambda_n(s-1)} = e^{(\lambda_1+\dots+\lambda_n)(s-1)}.$$

Заметим, что полученное выражение — производящая функция пуассоновского распределения с параметром $\lambda_1 + \dots + \lambda_n$. Поскольку степенной ряд однозначно восстанавливает коэффициенты, то производящая функция чего-то пуассоновская, то и сама величина пуассоновская, поэтому получаем, что $X_1 + \dots + X_n \sim \text{Poiss}(\lambda_1 + \dots + \lambda_n)$. И вот мы с вами просто и безболезненно доказали, что сумма независимых пуассоновских величин — пуассоновская с суммой параметров, гораздо приятнее, чем сидеть и считать какие-то свёртки.

Единственной проблемой иногда оказывается, что произведение производящих функций особо не разложится, но тогда мы хотя бы как-то зададим распределение.

Ещё одно замечательное свойство производящей функции:

Пусть X_i — независимые одинаково распределённые, N — независимая от них, все они целочисленные. Рассмотрим случайную величину $Y = \sum_{i=1}^N X_i$ — сумма случайного числа независимых случайных величин.

Пример 5.4. Вы можете такую случайную величину легко встретить, например, представив себе, как устроена какая-нибудь эпидемия Covid, когда есть больные, они заражают других больных в случайных количествах и так это всё продолжается. И вот если сейчас у нас больны N человек, это случайная величина, каждый из них заражает сколько-то, получается в сумме как раз

какая-то такая конструкция, на каждом шаге количество заболевших будет связано с предыдущим какой-то такой формулой.

Заметим, что распределение Y искать проблематично: надо по формуле полной вероятности фиксировать N , при этом у нас получится довольно сложная конструкция. Эту вероятность будет очень неудобно считать, потому что будет k -кратная свёртка, а по k мы ещё потом суммируем. А с точки зрения производящих функций всё получается очень удобно. Оказывается, что верна такая формула:

$$\varphi_Y(s) = \varphi_N(\varphi_{X_1}(s)). \quad (5.1)$$

Это композиция производящих функций. Это очень важная формула, она лежит в основе теории ветвящихся процессов. Мы пишем φ_{X_1} , потому что все X_i одинаково распределены.

Докажем формулу (5.1):

Доказательство. Чтобы расписать Y , нужно зафиксировать N , потому что иначе неудобно. Далее распишем вероятность по формуле полной вероятности:

$$\varphi_Y(s) = \sum_{k=0}^{\infty} P(Y = k) s^k = \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} P(Y = k \mid N = m) P(N = m) s^k \ominus$$

Ряды со знакоположительными членами, поэтому можем в любом порядке суммировать, всё сходится абсолютно. Воспользуемся также сразу тем, что событие $N = m$ не зависит от X_i по условию. Немного схитрим, и сначала подставим $N = m$ в правую часть, а затем скажем, что события независимы, но они действительно независимы, справа N , слева X_i .

$$\ominus \sum_{m=0}^{\infty} P(N = m) \underbrace{\sum_{k=0}^{\infty} P\left(\sum_{i=1}^m X_i = k \mid N = m\right) s^k}_{= \varphi_{\sum_{i=1}^m X_i}(s) = \varphi_{X_1}^m(s)} = \sum_{m=0}^{\infty} P(N = m) \varphi_{X_1}^m(s) = \varphi_N(\varphi_{X_1}(s)).$$

■

Вот такой нехитрый трюк, но даёт очень красивую формулу, очень удобную, если мы работаем с суммой случайного числа случайных слагаемых.

Третье полезнейшее для нас свойство производящих функций заключается в том, что

Теорема 5.1. $P(X_n = k) \rightarrow P(X = k) \forall k \iff \varphi_{X_n}(s) \rightarrow \varphi_X(s) \forall s \in [0, 1].$

То есть сходимость вероятностей (мы потом назовём её сходимостью по распределению) равносильна сходимости производящих функций. Это удобно, потому что, например, с суммами опять удобно работать.

Пример 5.5. Представьте себе, что $X_i \sim \text{Bern}\left(\frac{\lambda}{n}\right)$ — независимые одинаково распределённые. Рассмотрим $S_n = \sum_{i=1}^n X_i$. Производящая функция S_n выглядит следующим образом:

$$\varphi_{S_n}(s) = \prod_{i=1}^n \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n} s\right) = \left(1 - \frac{\lambda(1-s)}{n}\right)^n \rightarrow e^{-\lambda(1-s)} = e^{\lambda(s-1)}$$

Это пуассоновская производящая функция (см. пример 5.2), таким образом мы доказали замечательный факт, что вероятность $S_n = k$ стремится к вероятности, что пуассоновская величина равна k . Это, конечно, было просто доказать, просто явно выписав формулу, потому что у нас биномиальное распределение, но ровно то же доказательство практически, с маленьким исправлением, подходит в случае, когда λ разные, тогда распределение S_n посчитать практически невозможно, а предел его посчитать легко. Этот факт называется *теорема Пуассона*: мы берём много редких событий, у каждого вероятность очень маленькая, и оказывается, что общее число успехов сходится к пуассоновскому распределению. Опять же, такие факты очень удобно доказывать.

Производящие функции векторов

Когда мы работаем с векторами, нам нельзя задавать только маргинальное распределение, только распределение компонент, нам надо задать всё распределение в целом. Мы помним, что знать сумму по строкам или сумму по столбцам недостаточно, надо знать всю табличку. Соответственно, нужно ввести производящую функцию нескольких случайных величин. Она называется **производящей функцией случайного вектора** X_1, \dots, X_n и вводится следующим образом:

$$\varphi_{X_1, \dots, X_n}(s_1, \dots, s_n) = \mathbb{E} s_1^{X_1} s_2^{X_2} \dots s_n^{X_n} = \sum_{k_1=0}^{\infty} \dots \sum_{k_n=0}^{\infty} s_1^{k_1} \dots s_n^{k_n} \mathbb{P}(X_1 = k_1, \dots, X_n = k_n).$$

Здесь $s_i \in [0, 1]^n$. Такая производящая функция задаёт распределение всего вектора.

Ясно, что по дискретному распределению вектора $\mathbb{P}(X_1 = k_1, \dots, X_n = k_n)$ мы можем восстановить производящую функцию $\varphi_{\vec{x}}(s_1, \dots, s_n)$. И наоборот,

зная производящую функцию, можем восстановить вероятности. Чтобы их узнать, нужно продифференцировать по каждой переменной в нужном количестве (под нулевой производной понимаем саму функцию):

$$P(X_1 = k_1, \dots, X_n = k_n) = \frac{\partial^{k_1}}{\partial s_1^{k_1}} \dots \frac{\partial^{k_n}}{\partial s_n^{k_n}} \varphi_{\vec{X}}(s_1, \dots, s_n) \Big|_{s_1=0, \dots, s_n=0}$$

Не так уж приятно это делать, но тем не менее можно, можно просто взять и разложить функцию в ряд и отсюда получить всё, что нам нужно.

Чтобы найти производящую функцию подвектора, например, $\varphi_{X_1}(s) = E s^{X_1}$, зная производящую функцию вектора $\varphi_{X_1, \dots, X_n}(s_1, \dots, s_n) = E s_1^{X_1} \dots s_n^{X_n}$, нужно вместо s_1 подставить s , а вместо остальных s_i подставить 1:

$$\varphi_{X_1}(s) = \varphi_{X_1, \dots, X_n}(s, 1, \dots, 1).$$

Поэтому зная совместную производящую функцию нескольких величин, мы легко можем вывести производящую функцию одной или нескольких величин (может оставлять несколько s_i , а в остальные места ставить 1). Если подумать, можно производящую функцию суммы некоторых или всех X_i восстановить.

Свойства производящей функции вектора

Теперь про свойства. Кроме вероятностей, мы можем искать моменты с помощью производящих функций.

Пример 5.6. Найдём

$$EXY = \sum_{k, l} kl P(X = k, Y = l)$$

с помощью

$$\varphi_{X, Y}(s, t) = \sum_{k, l} P(X = k, Y = l) s^k t^l:$$

$$EXY = \varphi''_{st}(1, 1).$$

Если мы продифференцируем сначала по s , получится k , затем — по t , получится l : как раз kl на какую-то штуку, если положим $s = 1$, $t = 1$, эта штука пропадёт.

Поэтому зная совместную производящую функцию, мы можем искать ковариацию, смешанные факториальные моменты, математическое ожидание

различных произведений случайных величин (в том числе степени случайных величин). Общая формула для факториальных моментов через производящую функцию:

$$\begin{aligned} EX_1(X_1 - 1) \dots (X_1 - k_1 + 1) X_2 \dots (X_2 - k_2 + 1) \dots X_n \dots (X_n - k_n + 1) &= \\ &= \frac{\partial^{k_1}}{\partial s_1^{k_1}} \dots \frac{\partial^{k_n}}{\partial s_n^{k_n}} \varphi_{X_1, \dots, X_n}(1, \dots, 1). \end{aligned}$$

Моменты получаются факториальными и некрасивыми, тем не менее, опять можно попытаться выразить через них обычные моменты. Отличие от формулы для вероятности в том, что мы считаем производную в точке $(1, \dots, 1)$ вместо точки $(0, \dots, 0)$. Поэтому если мы умеем дифференцировать совместную производящую функцию, то мы можем выуживать любые смешанные моменты. Это хлопотно, но тем не менее, это возможно. Эта формула больше большая и страшная, чем действительно содержательно сложная.

Теперь рассмотрим несколько независимых векторов одинаковой длины k : $\vec{X}_1, \vec{X}_2, \dots, \vec{X}_n$. Их производящая функция вместе равна

$$\varphi_{\vec{X}_1, \dots, \vec{X}_n}(\vec{s}_1, \dots, \vec{s}_n) = \mathbb{E} s_{1,1}^{X_{1,1}} \dots s_{1,k}^{X_{1,k}} s_{2,1}^{X_{2,1}} \dots s_{2,k}^{X_{2,k}} \dots s_{n,1}^{X_{n,1}} \dots s_{n,k}^{X_{n,k}} \ominus$$

Вот такая вот страшная штука, но за счёт независимости каждое из произведений $s_{i,1}^{X_{i,1}} \dots s_{i,k}^{X_{i,k}}$ не зависит от другого такого же произведения, потому что это всё функции независимых векторов $\vec{X}_1, \dots, \vec{X}_n$, поэтому это равно произведению производящих функций отдельных компонент:

$$\ominus \varphi_{\vec{X}_1}(\vec{s}_1) \dots \varphi_{\vec{X}_n}(\vec{s}_n),$$

так что если векторы независимые, их совместную производящую функцию можно найти произведением производящих функций, каждую в своей точке. В частности, если у нас есть вектор из n независимых компонент, его производящая функция будет равна просто произведению производящих функций и наоборот: если совместная производящая функция распалась на произведение производящих функций, каждая своей переменной, то наш вектор состоит из независимых компонент.

Ну и наконец, заключительное свойство, похожее на предыдущее: если сложить независимые векторы \vec{X}_i , то производящая функция суммы будет равна произведению производящих функций каждого из векторов:

$$\varphi_{\vec{X}_1 + \dots + \vec{X}_n}(\vec{s}) = \varphi_{\vec{X}_1}(\vec{s}) \dots \varphi_{\vec{X}_n}(\vec{s}).$$

Это так, потому что, опять же, функции $s_i^{X_{1,i}}, \dots, s_i^{X_{n,i}}$ независимые. Разница с предыдущей формулой в том, что так у каждой X были свои индексы, а

здесь мы берём вектор суммы, и аргумент у них у всех одинаковый, поэтому важно не путать эти формулы.

Если продолжать дальше, то можно убедиться, что остальные два свойства также сохраняются для производящих функций векторов: если складывать их в случайном количестве, будет композиция производящих функций; если переходить к пределу производящих функций, то это равносильно сходимости всех возможных вероятностей. Нам это уже не понадобится.

Резюмируем: производящие функции — удобный способ характеризовать распределение с очень узких позиций, например, с позиции суммирования величин, с позиции перехода к пределу. Удобно тем, что производящая функция суммы — это произведение производящих функций для независимых величин, гораздо удобнее формул свёртки. Поэтому можно переходить в производящие функции, с ними как-то работать, а потом возвращаться назад так же, как мы делали с пуассоновским распределением.

Есть ещё несколько задач, где это удобно. Прежде всего, это сложение независимых величин; сложение независимых величин в случайных количествах; переход к пределу. Вокруг этих трёх вещей мы и будем виться.

Нельзя забывать, что производящая функция — это прежде всего математическое ожидание, и с ней можно играть в те же самые трюки, в которые мы умеем играть с математическим ожиданием.

Пример 5.7. Пусть $X \sim \text{Geom}(p)$. Попробуем найти производящую функцию X с помощью сведения задачи к самой себе, некой рекуррентной конструкции. Воспользуемся физическим смыслом геометрического распределения: введём события $A = \{1 \text{ бросок — орёл}\}$, $\bar{A} = \{1 \text{ бросок — решка}\}$. Тогда по формуле полной вероятности

$$E s^X = E(s^X | A)P(A) + E(s^X | \bar{A})P(\bar{A}) \ominus$$

В том мире, где случилось A (первым броском выпал орёл), $X = 0$, поэтому $E(s^X | A) = 1$, если случилось \bar{A} , то мы фактически начинаем эксперимент заново, новый эксперимент ничем не отличается от такого же эксперимента с начала, но одна решка уже есть:

$$\ominus 1 \cdot p + E s^{X+1} (1 - p).$$

В итоге получается, что

$$\varphi_X(s) = p + (1 - p)s \varphi_X(s).$$

Решив такое уравнение, получим $\varphi_X(s) = \frac{p}{1 - (1 - p)s}$.

И вот мы нашли производящую функцию, никак не работая с суммами и рядами. Этот подход мы будем называть *рекуррентным подходом* подсчета производящей функции, иногда он бывает очень удобным. Если вас пугают условные производящие функции, смотрите на них как на производящую функцию случайной величины X в том мире, где случилось событие A .

Семинар 6. Общее вероятностное пространство

Построение общего вероятностного пространства

Мы начинаем большой новый блок. В каком-то смысле, будем повторять всё то же, что мы уже делали, только называть это новыми словами. Шаг за шагом пойдём по тому же самому пути, только уже не будем предполагать конечность или счётность пространства. Специфика такова, что теория, которую мы с вами будем разбирать, по большей части действительный анализ. Чтобы понять, почему оно так странно устроено, нам нужно поспотыкаться на каждом камушке, на котором человечество поспотыкалось, пока оно к этому шло, иначе всё будет очень искусственно. Мы увидим, что начав строить вероятностное пространство естественно, нам придётся сделать те допущения и ограничения, которые человечество сделало, и которые рассказываются в действительном анализе. Поэтому мы будем в несколько итераций идти, совершать ошибки и начинать сначала, пытаясь их исправить.

Введём вероятностное пространство (Ω, \mathcal{F}, P) , где Ω — пространство элементарных исходов, $\mathcal{F} \subseteq 2^\Omega$ — пространство событий, $P: \mathcal{F} \rightarrow [0, 1]$ — вероятностная мера (вероятность).

Ω у нас имеет какую-то достаточно общую структуру, например, Ω может быть отрезком $[0, 1]$ — типовое пространство элементарных исходов. Про пространство событий пока будем предполагать, что это 2^Ω , а если не получится, будем думать, что с этим делать, а P — какая-то конструкция на \mathcal{F} .

Раньше мы брали $\mathcal{F} = 2^\Omega$, P определяли на каждом $\omega \in \Omega$, а P от множества называли суммой P от элементов. Теперь такой подход не сработает, потому что мы не умеем считать сумму несчётного числа слагаемых. Допустим, мы бросили точку на отрезок наугад, вероятность, что она попадёт в конкретный исход ω , из общих соображений, равна нулю. С другой стороны, мы должны приписать вероятность каждому исходу, чтобы сумма у всего отрезка получилась 1, а при нулевых вероятностях исходов сумма будет 0. Мы не смогли построить пространство, поэтому нам нужен принципиально другой подход.

Попробуем P определить сразу на всём \mathcal{F} , если это получится. По крайней мере, надо понять, что от \mathcal{F} мы хотим. Раньше всё было просто — мы задавали P на элементарных исходах, а на остальных событиях считали сумму, а теперь мы хотим задать P на всём \mathcal{F} и явно не абы как. Например, нам вряд ли понравится, если $P\left(\left[0, \frac{1}{2}\right)\right) = \frac{1}{2}$; $P\left(\left[\frac{1}{2}, 1\right)\right) = 1$; $P\left(\left[0, 1\right)\right) = 1$, потому что эта

мера не аддитивна: мы хотим, чтобы вероятность попадания в объединение двух непересекающихся множеств равнялась сумме вероятностей этих множеств, это соответствует нашему представлению о вероятности. Дополнительно нам хочется добавить мере какую-то непрерывность, чтобы если множества как-то приближались к множествам, то и вероятность у них тоже приближалась, но пока непонятно, что это значит: у нас нет окрестности множества, нет метрики на множествах.

Пересечение множеств

И вот тут на помощь приходит очень удобная операция, которая называется **пересечением**. Она удобная тем, что если множества A_i вложенные, то $\bigcap A_i$ — это как раз что-то такое, что мы могли бы назвать пределом. И наоборот, если множества расширяются, то $\bigcup A_i$ — то самое, что мы назвали бы их пределом. Научимся пересекать счётное и даже несчётное количество разных сообществ множеств. Для этого вспомним, что такое пересечение:

$$\bigcap_{\alpha} A_{\alpha} = \{\omega : \omega \in A_{\alpha} \forall \alpha\}.$$

Пример 6.1. Допустим, у нас есть плоскость, на ней есть полоска от $\frac{1}{n}$ до $-\frac{1}{n}$ по вертикали и во всю ширь по горизонтали (см. рис. 6.1), такую полоску назовём множеством A_n .

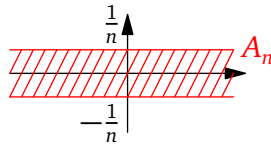


Рис. 6.1: Полоска на плоскости из примера 6.1

По рисунку видим, что пересечение всех возможных полосок A_n — ось x :

$$\bigcap_n A_n = \{(x, 0)\}_{x \in \mathbb{R}} = A.$$

Проверим наше предположение по определению. Для этого надо показать, что это множество лежит во всех A_n , и ничего больше во всех A_n не лежит:

- 1) $A \subseteq A_n \forall n$ — очевидно: ось явно лежит в каждой полосе при любом n ;
- 2) Пусть $\omega \notin A$. Покажем, что $\omega = (x, y)$, $y \neq 0$. Знаем, что $(x, y) \notin A_n$ при $|y| > \frac{1}{n}$ или $n > \frac{1}{|y|}$. Поскольку $y \neq 0$, то такое число существует, значит, с этого момента мы перестанем попадать в полоску и не будем лежать в пересечении.

Вещи довольно очевидные, но чтобы аккуратно их проводить, нужно наработать некоторый опыт. Мы обычно так и доказываем, что пересечение имеет какой-то вид. Попробуем ещё что-нибудь сделать:

Пример 6.2. Рассмотрим круг S . Вокруг него описали треугольник t . Под треугольником мы понимаем не границу, а фигуру с внутренностью, причём замкнутую. T — множество всех возможных описанных треугольников вокруг этой окружности. Очевидно, что это множество несчётное, но это нас не очень пугает. Пересечением всех таких треугольников будет наш круг:

$$\bigcap_{t \in T} t = S.$$

Докажем это. В одну сторону очевидно — каждая точка круга лежит в каждом элементе пересечения. Чтобы доказать в другую сторону, нужно взять какую-то точку вне круга. Соединим её с центром нашей окружности и проведём перпендикуляр в точке пересечения с окружностью, затем построим какой-нибудь треугольник со стороной на этой прямой. Тогда заведомо эта точка не лежит внутри треугольника, потому что она его стороной отделяется от центра окружности (см. рис. 6.2).

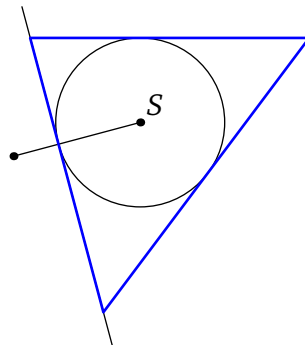


Рис. 6.2: Описанный вокруг окружности треугольник, в который заведомо не попадает заданная точка

Так как для любой точки вне окружности мы смогли построить треугольник, который не будет её содержать, других точек, кроме точек круга, в пересечении нет. Единственное, нужно уточнить, круг получился открытым или замкнутым. Раз треугольники были замкнутыми, то и круг будет замкнутым, потому что каждая точка границы действительно лежала в таком треугольнике. Если бы треугольники были открытыми, то и круг был открытым, потому что у треугольника, который касается окружность в этой точке, эта точка не входит в границу, а значит, любая точка окружности в какой-то треугольник не входит, и в пересечение не попадёт.

Вместе с пересечением мы будем работать и с объединением. Вы уже догадываетесь, что объединение множеств — то, что лежит хотя бы в одном множестве.

Упражнение 6.1. Даны полуинтервалы $[a - \frac{1}{n}, b + \frac{1}{n})$. Найдите их пересечение по всем $n \in \mathbb{N}$.

Указание. Заметим, что совершенно не важно, включена ли точка $b + \frac{1}{n}$ или нет:

$$\bigcap_n \left[a - \frac{1}{n}, b + \frac{1}{n} \right) = [a, b].$$

Такой отрезок получится, потому что каждая точка отрезка попадёт в каждый полуинтервал и ни одна точка извне этого отрезка не попадёт в пересечение: если она справа b , то она больше какого-то $b + \frac{1}{n}$; если она слева a , то она меньше какого-то $a - \frac{1}{n}$.

Поэтому наш «как-бы предел», пересечение вложенных множеств, обладает таким замечательным свойством, что оно умеет из полуинтервалов и интервалов делать отрезки, то есть оно создаёт некоторые новые конструкции.

Свойства вероятностных мер

Теперь попробуем в несколько итераций построить саму модель. Напомним, что мы сначала будем делать неправильно, сталкиваясь с тем, что это не работает, начинать с нового, так что дочитайте до конца и не забывайте как не надо.

Первая модель, которую мы хотим построить, выглядит так: возьмём Ω , 2^Ω и определим на них меру $P : 2^\Omega \rightarrow [0, 1]$. Потребуем от P три свойства:

- 1) $P(\Omega) = 1$ (мера, обладающая таким свойством, называется *вероятностной мерой*);
- 2) $P(A+B) = P(A \sqcup B) = P(A) + P(B)$ ($A \sqcup B$ — *объединение непересекающихся множеств A и B*);
- 3) *непрерывность снизу*: $P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} P(A_n)$, где $A_n \supseteq A_{n+1}$.

То есть если множества стягиваются, то вероятность множества, которое лежит в них всех (пересечения) должна быть равна пределу вероятности. С помощью свойства непрерывности можно получать меры для каких-то множеств, не похожих на те, у которых мы знаем, например, из упражнения 6.1, меры отрезков по мерам полуинтервалов. Можно вместо условия 3 требовать *непрерывность сверху* (эти условия легко выводятся друг из друга):

3') непрерывность сверху: $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} P(A_n)$, где $A_n \subseteq A_{n+1}$.

Также можно вместо условия 3 потребовать более привычную вам σ -аддитивность, которая также эквивалентна обоим приведённым ранее условиям (частичные суммы множеств будут расширяющимися множествами):

3'') σ -аддитивность:

$$P(A_1 + A_2 + \dots) = \sum_{i=1}^{\infty} P(A_i).$$

Обычно меру определяют σ -аддитивной, но с физических представлений о том, что такое вероятность, приятнее смотреть на свойство непрерывности, которое более естественно.

Оказывается, что в самой естественной конструкции меры есть несколько проблем. Первая проблема — непонятно, как вообще эту P задать: мы должны задать какую-то меру P на каждом подмножестве, например, отрезка, потом каждое возможное множество разбить в счётное объединение и показать, что вероятность кусков в сумме даст именно наше множество. А вторая проблема ещё хуже — это вообще нельзя сделать. Рассмотрим пример, полезный для осознания зачем нам нужна возня с σ -алгебрами:

Пример 6.3. Возьмём полуинтервал $[0, 1)$ и $2^{[0,1]}$. Возьмём $P: 2^{\Omega} \rightarrow [0, 1]$, которую мы хотим определить так, чтобы мера отрезка была равна длине этого отрезка: $P([a, b]) = b - a$, $0 \leq a \leq b \leq 1$, то есть мы просто хотим определить длину. Дополнительно мы хотим, чтобы мера была инвариантной относительно сдвигов, то есть если мы сдвинем множество по циклу (если что-то вышло за пределы отрезка, вернём это в начало отрезка), то мера не должна поменяться: $P(A + c \bmod 1) = P(A) \forall c$. Кажется, что мера таким свойством обладать должна, иначе это какая-то неправильная длина. Возьмём множество A_0 , неконструктивное, обладающее двумя свойствами:

- 1) $\forall x \in \mathbb{R} \exists y \in A_0: x - y \in \mathbb{Q}$;
- 2) $\forall x, y \in A_0, x - y \notin \mathbb{Q}$.

Получилось такое странное множество, но оно имеет право на существование. Возьмём $[0, 1) \cap \mathbb{Q}$ и пронумеруем его элементы: $\{0 = q_0, \frac{1}{2} = q_1, \dots\}$. Положим $A_i = A_0 + q_i \bmod 1$, то есть возьмём исходное множество и будем его двигать на все возможные рациональные числа, получая какие-то множества на отрезке. Сформулируем несколько тезисов:

- 1) Каждая точка из $[0, 1]$ хотя бы в одно A_i попадает по свойству 1, потому что каждая точка из отрезка $[0, 1]$ отличается от какой-то точки из A_0

на рациональное число, тогда при этом сдвиге мы попадём в A_i ;

- 2) A_i не пересекаются, потому что если бы два множества пересеклись, то какой-то элемент x в A_i , и этот же элемент x в A_j , а значит, соответствующие элементы в A_0 отличаются на рациональное число.

Поэтому $A_0 + \dots + A_i + \dots = [0, 1)$. По свойству σ -аддитивности получается, что $\sum_{i=0}^{\infty} P(A_i) = 1$, с другой стороны, по нашему свойству, у всех A_i одинаковая вероятность, но так не бывает, чтобы ряд из одинаковых чисел равнялся 1. Получается, что даже на таком уровне, когда мы просто хотим построить длину на всех подмножествах $[0, 1)$, из-за таких странных множеств у нас всё сломалось.

Теперь у нас есть два варианта: либо мы не будем больше требовать, чтобы длина была инвариантна относительно сдвига, но тогда даже страшно представить, что получится, мы же вводили длину, она была инвариантна, неясно, почему всё сломалось; либо мы должны запретить определять меру таких множеств, чтобы сделать меру хорошей.

Второй метод кажется разумнее, поэтому нам нужно сузить пространство множеств и не позвать туда те множества, которые ведут себя неприлично.

Для этого мы говорим, что будем рассматривать тройку (Ω, \mathcal{F}, P) , где Ω — какое-то пространство, \mathcal{F} — σ -алгебра на Ω , $P : \mathcal{F} \rightarrow [0, 1]$, обладающее введёнными ранее свойствами. σ -алгебру мы потребовали ну хотя бы потому, что наша мера σ -аддитивна, а чтобы она была σ -аддитивна, нужно, чтобы дополнения и счётные объединения лежали в нашем пространстве.

И мы видим, что это вынужденная мера: мы хотим сократить пространство всех событий, просто чтобы убрать вот такие вот множества. Мы просто выделили некоторое множество \mathcal{F} и сказали, что это хорошие множества, мы можем на них определить нашу меру, а на всех остальных нам не важно, чему она равна, мы её не будем там определять.

Продолжение меры

Вопросы «Откуда взять σ -алгебру?» и «Как задавать P ?» мы пока так и не решили. Обрисуем кратко траекторию, которая поможет нам понять общую философию траектории, которая приводит к ответам на эти вопросы.

Давайте сначала попробуем определить площадь на подмножествах квадрата $[0, 1) \times [0, 1)$. Начнём с того, что площадь прямоугольника (без верхней и

правой границ, но с нижней и левой границами и линиями, параллельными сетке) равна произведению сторон:

$$P([a, b] \times [c, d]) = (d - c)(b - a), \quad 0 \leq c < d < 1, \quad 0 \leq a < b < 1.$$

Сразу по свойствам аддитивности мера определилась на так называемых *простых фигурах* — тех, которые можно сложить из прямоугольников с линиями, параллельными сетке. Дополнение до какого-то количества прямоугольников также назовём простой фигурой, они могут быть с дырками. Простые фигуры образуют алгебру, то есть объединение двух простых фигур снова простая фигура, пересечение точно также и дополнение. При этом σ -алгебру они не образуют: если начать объединять неконечное количество простых фигур, можно получить что-то совсем новое, например, треугольник или круг. Поэтому дальше мы хотим совершить этот переход, нам просто его необходимо совершить. При этом возникает два важных вопроса: «Что попадёт в σ -алгебру, если там лежит простая фигура?» и «Определится ли там корректно площадь?». То есть мы сейчас начнём делать счётные пересечения, счётные объединения и ждать, пока не получится σ -алгебра.

Здесь возникает два важных тезиса:

- 1) Пересечение любого числа σ -алгебр — σ -алгебра.

Значит, существует минимальная σ -алгебра, содержащая данную систему множеств (в данном случае система простых фигур, которую мы будем называть \mathcal{S}). Называть мы такую алгебру будем $\sigma(\mathcal{S})$, то есть это пересечение всех σ -алгебр, содержащих \mathcal{S} :

$$\sigma(\mathcal{S}) = \bigcap_{\mathcal{F} \supseteq \mathcal{S}} \mathcal{F}$$

Мы бы хотели распространить меру с \mathcal{S} на $\sigma(\mathcal{S})$. Интуитивно кажется, что $\sigma(\mathcal{S})$ — все возможные объединения, пересечения и дополнения множеств из \mathcal{S} . На деле всё устроено гораздо хуже, но, по крайней мере, на $\sigma(\mathcal{S})$ нам меру перенести придётся, потому что если мы хотим определить её на какой-то σ -алгебре, эта σ -алгебра должна содержать \mathcal{S} , а значит, и $\sigma(\mathcal{S})$ по определению $\sigma(\mathcal{S})$. Поэтому на $\sigma(\mathcal{S})$ нам придётся меру как-то определить.

- 2) Есть такая замечательная теорема Каратеодори:

Теорема 6.1 (Каратеодори). Если мера P σ -аддитивна на алгебре \mathcal{S} , то существует единственная мера \tilde{P} на $\sigma(\mathcal{S})$ такая, что $\tilde{P}(A) = P(A) \forall A \in \mathcal{S}$.

Иначе говоря, мера единственным образом продолжается с алгебры на σ -алгебру, если мы определили меру на алгебре, и там она была σ -аддитивна.

Так как каждую простую фигуру, разбивая на куски, которые тоже простые фигурки, мы всегда получали, что вероятность большой простой фигурки равна сумме вероятностей кусков, нашу меру можно продлить на всю $\sigma(S)$. Как продлить — тайна великая, это вам в действительный анализ. Но, тем не менее, теорема для нас важна тем, что если мы определим площадь на всех прямоугольниках, она перенесётся на все простые фигуры, а оттуда как-то волшебным образом перенесётся на всю минимальную σ -алгебру, содержащую простые фигуры. σ -аддитивность в теореме можно заменить на непрерывность, это одно и то же.

Борелевская сигма-алгебра

Остаётся вопрос, что такое минимальная σ -алгебра, содержащая простые фигуры и чему же в итоге наша мера оказалась на этой σ -алгебре равна.

Определение 6.1. Минимальная σ -алгебра, содержащая все открытые подмножества Ω , называется **борелевской**. Обозначение: $\mathcal{B}(\Omega)$.

Множество Ω должно обладать топологией, чтобы было понятие открытого множества, или метрикой, которая эту топологию индуцирует.

На данном этапе мы с вами будем в основном работать с борелевской σ -алгеброй на прямой (содержит все отрезки или интервалы) или с борелевской σ -алгеброй на плоскости (содержит все круги).

Мы оставим без доказательства, что σ -алгебра, содержащая все простые фигуры, является борелевской σ -алгеброй прямоугольника. Нужно показать, что каждое открытое можно получить из наших прямоугольников. Это некоторый технический факт, будем считать его известным.

Борелевская σ -алгебра устроена сложно, принцип, что мы сейчас возьмём счётное объединение прямоугольников, как-то их опишем, и у нас сразу получится описать все множества борелевской σ -алгебры, не работает. Будет получаться такая история, которую вы, наверное, встречали в своей жизни: вы позвали на вечеринку своих гостей, те позвали своих друзей, те позвали своих друзей. В итоге пришёл кто-то, кто вообще непонятно откуда взялся, но он пришёл. Но зато не пришли те самые множества, про которые мы говорили в начале. Они в нашу σ -алгебру не попали, это нас очень радует, поэтому мера у нас определилась корректно. Но при этом попали какие-то странные множества, но это не так страшно, потому что нам важно, что интересующие нас множества попали в борелевскую σ -алгебру.

σ -алгебра получилась богатая, меру на ней мы до конца не знаем, но обычно у того, чего мы хотим, меру определять можем. Вы так делали, напри-

мер, с площадью. Вы научились считать площадь треугольника, площадь круга, отсюда ещё что-то поняли, потом доказали, что площадь каких-то фигур — это интеграл, но по-прежнему остались какие-то очень странные множества, у которых мы площадь считаем плохо.

И в целом наш принцип всегда такой — мы задаём меру на простых порождающих, на том, что называется полукольцо, с этого полукольца мы распространяем её на алгебру, с алгебры проверяем теорему Каратеодори, после чего распространяем на σ -алгебру, а на σ -алгебре уже живём с готовой мерой. Она переопределена, в том смысле, что она определена даже на том, что нас совсем не интересует, но это, к сожалению, пережиток того, что мы хотим всё-таки меру на σ -алгебре определить.

Эту философию для себя нужно уяснить, иначе вы всё время будете спотыкаться на этом. Ничего больше мы сделать особо и не могли, мы просто действовали так, как, кажется, нужно действовать, а получилась такая, довольно громоздкая, конструкция с непонятными борелевскими σ -алгебрами. Важно разобраться, как устроены σ -алгебры, понять, почему пересечение σ -алгебр — σ -алгебра, почему объединение — нет, что попало в σ -алгебру в каком-то простом случае, и так далее.

Геометрическая вероятность

Теперь важный частный случай, который мы сегодня будем использовать — это **геометрическая вероятность**.

В качестве Ω возьмём какое-то $A \subseteq \mathbb{R}^n$, у этого A определена *мера Лебега*. Возьмём $\mathcal{F} = \mathcal{B}(A)$, $P(B) = \frac{\lambda(B)}{\lambda(A)}$, $B \subseteq A$, λ — мера Лебега (длина, площадь, объём).

Получается что-то типа классической модели такой — вероятность попасть в множество пропорциональна, раньше было количеству элементов в множестве, а теперь площади множества. Это аналог классической модели в плане изотропности: как бы мы не расположили множество площади $\frac{1}{2}$, вероятность попасть в него $\frac{1}{2}$. В классической модели у нас была задача посчитать мощность множества и поделить его на мощность A , а здесь задача посчитать площадь множества, или объём множества и поделить на объём исходный. Это соответствует простым идеям о броске точки в множество, когда мы ищем вероятность попадания в какое-то подмножество.

Пример 6.4. Вася и Аня договорились встретиться на остановке. Они приходят в случайный момент от 2 до 3 часов независимо друг от друга. Каждый

готов ждать другого 10 минут. Найдём вероятность, что они встретятся.

Начнём с Ω . Описание этого опыта — два числа в диапазоне $[2, 3]$. Значит, это точка квадрата:

$$\Omega = \{(x, y), x \in [2, 3], y \in [2, 3]\}.$$

\mathcal{F} , соответственно, $\mathcal{B}([2, 3] \times [2, 3])$. И, наконец, вероятность множества — это просто его площадь: $P(A) = S(A)$, потому что $S(\Omega) = 1$. Поймём, что означает наше событие:

$$A = \{\text{встретились}\} = \{(x, y) : |x - y| \leq \frac{1}{6}\}.$$

Философский вопрос, строгое ли неравенство, но это, конечно, не важно. Построим рисунок: нам подходят $x - \frac{1}{6} \leq y \leq x + \frac{1}{6}$, эта область показана на рис. 6.3.

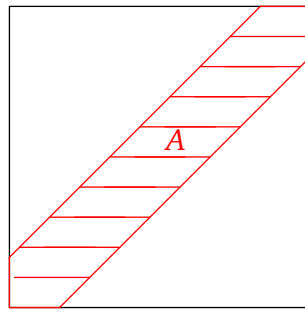


Рис. 6.3: Область, соответствующая событию A , показана красным цветом

Проще всего посчитать площадь A , перейдя к дополнению: вычтем два крайних треугольника из квадрата. Можно посчитать треугольники, но ещё приятнее заметить, что если убрать A , то всё сложится в квадрат со стороной $\frac{5}{6}$, поэтому получается, что

$$S(A) = 1 - \left(\frac{5}{6}\right)^2 = \frac{11}{36}.$$

Если бы они не ждали друг друга, тогда бы получилась вероятность того, что точка попадёт на диагональ, а площадь диагонали равна нулю, поэтому вероятность будет равна нулю. В классическом случае такое тоже могло быть — никто не запрещал положить в исход вероятность 0. Этот исход существует, он может настать, просто вероятность у него 0.

Семинар 7. Случайные величины в общем случае

Случайные величины и их распределение

Определение 7.1. Случайной величиной X называют отображение $X: \Omega \rightarrow \mathbb{R}$ такое, что $X^{-1}(B) \in \mathcal{F} \forall B \in \mathcal{B}(\mathbb{R})$.

Условие $X^{-1}(B) \in \mathcal{F} \forall B \in \mathcal{B}(\mathbb{R})$ называется *измеримостью*.

Оно берётся из того, что мы хотим быстро переходить от величины на Ω к распределению. Для этого нужно, чтобы эта мера была определена. Например, если мы хотим найти вероятность попадания в отрезок $P(\omega: X(\omega) \in [a, b])$, нужно, чтобы множество $\{\omega: X(\omega) \in [a, b]\}$ было в \mathcal{F} . А это, прямо по определению, полный прообраз отрезка, то есть уже прообраз отрезка $X^{-1}([a, b])$ должен лежать в \mathcal{F} . Отрезки порождают борелевскую σ -алгебру, а прообразы сохраняют все операции, поэтому прообраз любого борелевского тоже будет лежать в \mathcal{F} . Получается, что если мы не потребуем измеримости, то не сможем посчитать вероятность попадания в самые простые множества, так что это условие очень важно.

Как обычно, мы хотим от величины, по возможности, сбежать, и начать заниматься распределением.

Определение 7.2. Распределением случайной величины X называют меру $P_X(B) := P(\omega: X(\omega) \in B)$, $B \in \mathcal{B}(\mathbb{R})$.

Происходит почти то же самое, что было в дискретном случае: мы перешли из Ω в \mathbb{R} , это наше новое пространство, на нём появилась мера, заданная нашей случайной величиной X , и эта мера задаёт, соответственно, вероятности разных событий, уже на прямой.

Функция распределения

Задавать распределения на всех борелевских множествах слишком сложно, нам нужен другой способ задания распределения.

Возьмём вероятность P_X от луча:

$$P_X((-\infty, x]) = P(\omega: X(\omega) \leq x) = F_X(x) \text{ — функция распределения.}$$

Тезис в том, что зная функцию распределения, мы, в принципе, знаем распределение, потому что можем найти вероятность полуинтервала как разность вероятностей двух вложенных лучей (см. рис. 7.1):

$$P_X((a, b]) = F_X(b) - F_X(a).$$

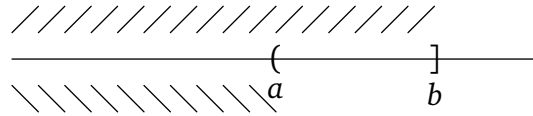


Рис. 7.1: Полуинтервал является разностью двух вложенных лучей

Позже мы убедимся, что вероятность на полуинтервалах задана корректно, поэтому она сама собой распространится на борелевскую σ -алгебру.

Поэтому вместо меры достаточно задать функцию распределения, а с ней гораздо удобнее работать, потому что это обычная функция числового аргумента. Неудобство в том, что не всегда так уж просто выразить через функцию распределения вероятность попасть в конкретное множество, но, тем не менее, давайте потренируемся.

Пример 7.1. Вероятность попадания в интервал:

$$P_X((a, b)) = P(X < b) - P(X \leq a).$$

Если мы объединим замкнутые лучи $(-\infty, b - \frac{1}{n}]$, то все точки до b у нас в них попадут, а сама точка b уже ни в один не попадёт. По свойству непрерывности меры, это расширяющиеся вложенные множества, а значит, вероятность их объединения — предел вероятностей, или, другими словами, предел слева функции распределения:

$$P_X((-\infty, b)) = P_X\left(\bigcup_n (-\infty, b - \frac{1}{n}]\right) = \lim_{n \rightarrow \infty} F_X\left(b - \frac{1}{n}\right) = F_X(b - 0).$$

Так что вероятность попасть в интервал — это

$$P_X((a, b)) = F_X(b - 0) - F_X(a).$$

Аналогично найдём вероятность отрезка:

$$P_X([a, b]) = F_X(b) - F_X(a - 0).$$

На прямой нам особо ничего, кроме отрезков, интервалов, лучей и их объединений, не интересно, поэтому, в принципе, всё, что захотим, мы отсюда выразить сможем.

Теперь попробуем посмотреть, как по заданной случайной величине искать её функцию распределения. Начнём с какой-нибудь простой величины:

Пример 7.2. $X(\omega) = 1 \forall \omega$, $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$, P — мера Лебега. Это простое стандартизированное пространство. Чтобы понять, как выглядит функция распределения такой случайной величины, построим её график (см. рис. 7.2).

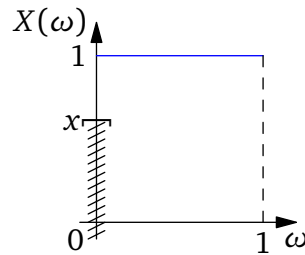


Рис. 7.2: График случайной величины из примера 7.2

Теперь мы строим множество от $-\infty$ до x и ищем прообраз случайной величины: фиксируем все ω , при которых наша величина попала в это множество по оси ординат, постепенно поднимаем x снизу вверх, и каждый раз смотрим, какое множество получается в прообразе:

$$X^{-1}((-\infty, x]) = \begin{cases} \emptyset, & x < 1; \\ \Omega, & x \geq 1. \end{cases}$$

Ему соответствует следующая функция распределения:

$$F_X(x) = \begin{cases} 0, & x < 1; \\ 1, & x \geq 1, \end{cases}$$

потому что функция распределения — это вероятность $X^{-1}((-\infty, x])$.

Теперь чуть-чуть усложним и ещё пару примеров посмотрим:

Пример 7.3. Рассмотрим $X(\omega) = \begin{cases} 0, & \omega \in [0, 1-p]; \\ 1, & \omega \in (1-p, 1]. \end{cases}$

График этой функции показан на рис. 7.3

Прообраз в этом случае выглядит так:

$$X^{-1}((-\infty, x)) = \begin{cases} \emptyset, & x < 0; \\ [0, 1-p], & 0 \leq x < 1; \\ \Omega, & x > 1. \end{cases}$$

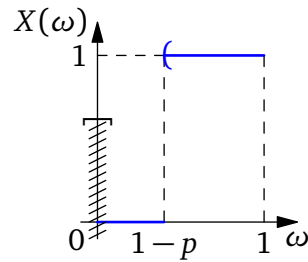


Рис. 7.3: График случайной величины из примера 7.3

Соответственно, функция распределения выглядит так:

$$F_X(x) = \begin{cases} 0, & x < 0; \\ 1-p, & 0 \leq x < 1; \\ 1, & x > 1, \end{cases}$$

так как у нас мера Лебега, где вероятность отрезка — это его длина.

Кстати, это распределение Бернулли, потому что эта случайная величина принимает значения 0 и 1 с вероятностями $1-p$ и p . Поэтому так выглядит функция распределения распределения Бернулли.

Не путайте функцию распределения и саму величину: сама величина могла бы быть задана на Ω по-другому, например, мы могли бы её определить от p до 1 нулём, а от 0 до p единицей, у неё функция распределения такая же.

Посмотрим также на непрерывный случай:

Пример 7.4. Рассмотрим $X(\omega) = \omega - \frac{1}{2}$. Её график показан на рис. 7.4.

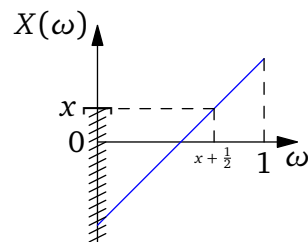


Рис. 7.4: График случайной величины из примера 7.4

Её прообразом будет

$$X^{-1}((-\infty, x]) = \begin{cases} \emptyset, & x < -\frac{1}{2}; \\ [0, x + \frac{1}{2}], & -\frac{1}{2} \leq x < \frac{1}{2}; \\ \Omega, & x \geq \frac{1}{2}. \end{cases}$$

Чтобы получить отрезок $[0, x + \frac{1}{2}]$, мы решили линейное уравнение $\omega - \frac{1}{2} = x$. Получается, что функция распределения будет выглядеть так:

$$F_X(x) = \begin{cases} 0, & x < -\frac{1}{2}; \\ x + \frac{1}{2}, & -\frac{1}{2} \leq x \leq \frac{1}{2}; \\ 1, & x > \frac{1}{2}. \end{cases}$$

Не важно, в какой из случаев включать равенство, потому что $x + \frac{1}{2} = 1$ при $x = \frac{1}{2}$. Выглядит функция распределения так, как показано на рис. 7.5. То, что она похожа на случайную величину — совпадение, так не всегда бывает. Например, если у нас функция, описывающая X , была бы выпуклая, то функция распределения была бы вогнутая.

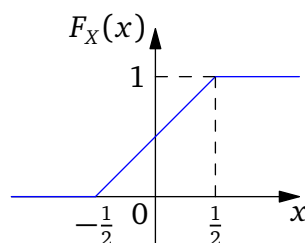


Рис. 7.5: Функция распределения случайной величины X из примера 7.4

Свойства функции распределения

Поговорим про свойства функции распределения $F_X: \mathbb{R} \rightarrow [0, 1]$:

- 1) $F_X(x) \rightarrow 1, x \rightarrow +\infty$, потому что работает свойство непрерывности меры: множество раздувается до всего пространства \mathbb{R} , мера всего \mathbb{R} равна 1;
 $F_X(x) \rightarrow 0, x \rightarrow -\infty$;
- 2) $F_X(x)$ монотонно не убывает;
- 3) $F_X(x)$ непрерывна справа, потому что пересечение $(-\infty, x + \frac{1}{n}]$ включает x . При этом она не непрерывна слева, потому что объединение $(-\infty, x - \frac{1}{n}]$ не будет включать x .

Вот этим трём свойствам точно должна функция распределения удовлетворять. Чтобы соответствующее ей распределение существовало, надо, чтобы теорема Каратеодори применилась. Для этого мера на алгебре, порождённой полуинтервалами, должна оказаться счётно-аддитивной, это нужно

как-то проверять, очень не хочется, поэтому за нас это сделали и доказали специальную теорему:

Теорема 7.1. Если F — какая-то функция, удовлетворяющая свойствам 1–3, то существует единственная мера P , такая, что $P((-\infty, x]) = F_X(x)$, то есть фактически существует распределение, соответствующее этой функции.

Очень удобно получается — чтобы задать распределение, нам всего-то нужно задать монотонную непрерывную справа функцию, идущую из 0 в 1, любая такая функция задаст какое-то распределение.

Дискретное распределение

Если величина принимает значения из какого-то конечного или счётного множества $X \in \{x_1, \dots, x_k, \dots\}$ с вероятностями $P_X(x_i) = p_i$, то можно задавать, как и раньше, дискретное распределение $P_X(x) = P(\omega: X(\omega) = x)$, и, соответственно, такая вероятность будет задавать функцию распределения

$$F_X(x) = \sum_{x_i \leq x} p_i.$$

Соответственно, наша функция распределения будет кусочно-постоянной. Если считаем, что X — упорядоченное конечное множество $\{x_1, \dots, x_m\}$, то график функции распределения выглядит как показано на рис. 7.6.

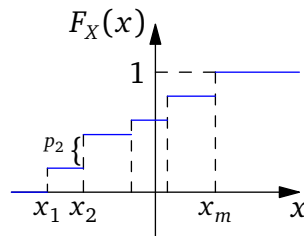


Рис. 7.6: Общий вид функции распределения дискретного распределения

Глядя на функцию распределения, можем определить, с какой вероятностью величина равна конкретному значению — это величина скачка в данной точке. Собственно, скачок функции распределения, как мы уже видели, это в точности свидетельство того, что конкретно у этой точки вероятность ненулевая. Таким образом, можем легко по функции распределения восстановить вероятности дискретного распределения.

Абсолютно–непрерывный случай

Второй интересный нам случай — случай, с которым удобно работать, *абсолютно–непрерывный случай*. Он предполагает, что существует такая функция $f_X \geq 0$, что выполнено тождество:

$$F_X(x) = \int_{-\infty}^x f_X(u) du \quad \forall x.$$

Неотрицательность f_X нужна, чтобы не нарушилась монотонность F_X . Соответственно, можно отсюда показать, что вероятность от любого множества:

$$P_X(B) = \int_B f_X(u) du.$$

Пока будем считать, что это риманов интеграл, под B подразумевается какое-то хорошее множество, отрезок или объединение отрезков, тогда верна такая форма. Естественно, функция $F_X(x)$ непрерывная, так что для дискретных распределений никаких плотностей не бывает. У непрерывных функций тоже не у всех бывает, но если бывает, то это очень удобно.

$f_X(x)$ называется **плотностью распределения**. Плотность можно выразить через функцию распределения:

$$f_X(u) = F'_X(u) = \lim_{\varepsilon \rightarrow 0} \frac{F_X(u + \varepsilon) - F_X(u)}{\varepsilon},$$

правда, функция распределения может быть недифференцируемой в каком-то количестве точек, но в этих точках мы полагаем плотность чему угодно равной, ясно, что изменение плотности в конечном числе точек никакого эффекта не окажет, потому что интеграл у нас не поменяется.

Соответственно, плотность физически характеризует удельную вероятность попасть куда-то недалеко возле точки, потому что это предел разности функции распределения, а это вероятность попасть в полуинтервал. Если плотность непрерывная в точке, то чем плотность больше, тем чаще мы попадаем в окрестность этой точки. Тут некоторое неудобство, что в отдельной точке плотность можно поменять как мы хотим, поэтому надо с некоторыми оговорками воспринимать наши слова.

Пример 7.5. Вернёмся к функции распределения из примера 7.4:

$$F_X(x) = \begin{cases} x + \frac{1}{2}, & x \in \left[-\frac{1}{2}, \frac{1}{2}\right]; \\ 0, & x < -\frac{1}{2}; \\ 1, & x > \frac{1}{2}. \end{cases}$$

Плотность у такого распределения будет

$$f_X(x) = \begin{cases} 0, & x < -\frac{1}{2}; \\ 0, & x > \frac{1}{2}; \\ 1, & x \in \left[-\frac{1}{2}, \frac{1}{2}\right]. \end{cases}$$

Такое распределение называется **равномерным** на отрезке $\left[-\frac{1}{2}, \frac{1}{2}\right]$, потому что плотность — это удельная вероятность попасть рядом с точкой, она у нас всё время одна и та же, то есть для нас все точки равноправны.

Соответственно, имея функцию распределения, мы можем дальше с ней работать, считать какие-то конструкции со случайной величиной:

Пример 7.6. Пусть X — случайная величина из отрезка $[0, \pi]$ с плотностью $f_X(x)$. Найдём плотность $\sin X$.

Удобнее работать не с плотностью, а с функцией распределения. Ясно, что на отрезке $[0, \pi]$ $\sin X$ принимает значения из $[0, 1]$. При $x \in [0, 1]$:

$$\begin{aligned} F_{\sin X}(x) &= P(\omega : \sin X(\omega) \leq x) = \\ &= P(\omega : X(\omega) \in [0, \arcsin x] \cup [\pi - \arcsin x, \pi]) = \\ &= F_X(\arcsin x) + 1 - F_X(\pi - \arcsin x). \end{aligned}$$

Мы не очень заботимся, открыто или замкнуто множество на границе, потому что у величины с плотностью функция распределения непрерывна, поэтому все эти -0 пропали, мы можем спокойно считать приращения функции состояния. Здесь мы также воспользовались тем, что по условию у нас $F_X(\pi) = 1$, $F_X(0) = 0$.

Чтобы найти плотность, возьмём производную от функции распределения:

$$\begin{aligned} f_{\sin X}(x) &= (F_X(\arcsin x))' + (1 - F_X(\pi - \arcsin x))' = \\ &= F_X'(\arcsin x) \cdot \frac{1}{\sqrt{1-x^2}} - F_X'(\pi - \arcsin x) \cdot \frac{-1}{\sqrt{1-x^2}} = \\ &= \frac{f_X(\arcsin x) + f(\pi - \arcsin x)}{\sqrt{1-x^2}}. \end{aligned}$$

Получили такую красивую формулу, которая связывает плотность синуса величины с плотностью самой величины. Это на самом деле якобиан какой-то замены, мы это обсудим с вами позже. Пока только принцип вывода таких формул: расписываем через исходную величину, выражаем через функцию распределения, дифференцируем, выражаем обратно через плотность.

Сверх этих двух стандартных случаев, дискретного и абсолютно–непрерывного, есть ещё разные, мы не будем на них останавливаться.

Пример 7.7. Представьте, что вы бежите на пару, подходите к дороге, и хотите её перейти, тогда распределение времени, которое вы потратите на переход дороги, будет смешанным, потому что если вы пришли и зелёный, то это просто будет время, за которое вы переходите дорогу, а если вы пришли и красный, то вы сначала подождёте случайное время, которое вам нужно, а потом перейдёте дорогу. Получится такое смешанное распределение, которое и разрывы имеет, и абсолютно–непрерывные куски.

Пример 7.8. Бывают распределения, которые тяжело представить: вы сидите и смотрите на стрелку часов, она вначале стоит на 2 часах, в конце на 3 часах, разрывов никаких не было, то есть скачков она не совершала, и при этом всегда стояла, кроме, может быть, множества меры ноль, то есть, в общем то, всегда, когда вы на неё посмотрите наугад, вы увидите, что она стоит, но при этом она была в 2, а оказалась в 3. В жизни такое редко встречается, но в математике такие сингулярные функции бывают.

Случайные векторы

Подробнее про случайные векторы мы поговорим в следующие разы, а пока просто дадим основные факты:

Определение 7.3. **Случайным вектором** будем называть набор случайных величин (X_1, \dots, X_n) на одном пространстве.

Более правильное определение — это измеримое отображение из Ω в \mathbb{R}^n , но мы сейчас не хотим на этом останавливаться, поэтому будем просто считать, что это набор случайных величин. Мы сегодня будем рассматривать только независимые величины, соответственно, X_1, \dots, X_n — **независимы**, если

$$P(\omega: X_1(\omega) \in B_1, \dots, X_n(\omega) \in B_n) = \prod_{i=1}^n P(\omega: X_i(\omega) \in B_i).$$

Слово «в совокупности» мы будем опускать, потому что *попарная независимость* — довольно экзотическая конструкция, мы не будем её использовать.

Для независимых величин X_1, \dots, X_n , имеющих плотность, выполнено:

$$P((X_1, \dots, X_n) \in B) = \int_{\vec{u} \in B} f_{X_1}(u_1) \dots f_{X_n}(u_n) du_1 \dots du_n,$$

Про это соотношение мы поговорим подробнее, когда будем говорить про векторы. Соответственно, теперь мы можем искать вероятности разных событий про несколько независимых величин:

Пример 7.9. Пусть X_1 и X_2 имеет плотность $f_X(x) = \begin{cases} 0, & x < 0; \\ 1, & x \in [0, 1]; \\ 0, & x > 1. \end{cases}$

Это *равномерное на отрезке* $[0, 1]$ распределение. Хотим найти вероятность того, что первая величина больше второй, если они независимы. Мы снова начинаем опускать ω из записи вероятностей.

$$P(X_1 > X_2) = \iint_{x_1 > x_2} f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \ominus$$

Не забывайте, что у нас есть ограничение: плотности ненулевые только при $x \in [0, 1]$. Поэтому получим довольно простую конструкцию:

$$\ominus \int_0^1 \int_0^{x_1} dx_2 dx_1 = \frac{x_1^2}{2} \Big|_0^1 = \frac{1}{2}$$

Из здравого смысла было понятно, что вероятность того, что первая правее второй такая же, как что вторая правее первой, но мы это посчитали честно через интеграл.

Формула свёртки

Если X_1, X_2 — независимы и абсолютно-непрерывны, то

$$F_{X_1+X_2}(x) = \iint_{x_1+x_2 \leq x} f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 = \int_{\mathbb{R}} \left(\int_{-\infty}^{x-x_1} f_{X_2}(x_2) dx_2 \right) f_{X_1}(x_1) dx_1$$

Теперь мы хотим это продифференцировать по x , получится

$$f_{X_1+X_2}(x) = \int_{\mathbb{R}} f_{X_2}(x-x_1) f_{X_1}(x_1) dx_1$$

Под интегралом монотонные функции, поэтому проблем с дифференцированием у нас нет. Получалась формула, очень похожая на то, что у нас было в дискретном случае, только интеграл, а не сумма. Эта формула называется **формулой свёртки**. Можно аккуратно изменить пределы интегрирования и получить формулы плотности $X_1 \cdot X_2$, $\sin X_1 + \sin X_2$, $X_1^{X_2}$ и так далее.

Семинар 8. Математическое ожидание в общем случае

Общее определение математического ожидания

Начнём с того, что обрисуем построение интеграла Лебега, которое делается в действительном анализе, чтобы у нас осталось впечатление, как это работает в нашем, чуть более простом, конечномерном случае.

Итак, как определить математическое ожидание:

- 1) Если $X \geq 0$ — неотрицательна и принимает конечное число значений $\{x_1, \dots, x_k\}$ с вероятностями p_1, \dots, p_k , то

$$EX = \sum_{i=1}^k x_i p_i.$$

- 2) Если $X \geq 0$ — любая неотрицательная, то можно представить её как монотонный предел неотрицательных величин $X_n \geq 0$ с конечным числом значений:

$$X = \lim_{n \rightarrow \infty} X_n, \quad X_n(\omega) \uparrow X(\omega).$$

Очень важно, чтобы предел был монотонным, в этом случае

$$EX = \lim_{n \rightarrow \infty} EX_n.$$

Так как величины монотонные, то и математические ожидания тоже монотонные, а у монотонной последовательности точно есть предел.

- 3) Если X — произвольная, то $X = X^+ - X^-$, где X^+ , X^- — положительная и отрицательная части соответственно. Напомним, $X^+ = \max\{X, 0\}$; $X^- = -\min\{X, 0\}$. И тогда мы определяем

$$EX = EX^+ - EX^-.$$

с теми же ограничениями, что и раньше: если две бесконечности, то мы говорим, что математического ожидания нет совсем.

Эта простая схема на деле содержит множество подводных камней. Приведём пример, показывающий, что сходимость нужна именно монотонная:

Пример 8.1. Пусть $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$, P — мера Лебега. Положим

$$X_n(\omega) = \begin{cases} 2^n, & \omega \in \left(0, \frac{1}{2^n}\right); \\ 0, & \text{иначе,} \end{cases}$$

то есть каждый раз она сужается, но повышается.

Как обычная числовая последовательность, $X_n(\omega) \rightarrow 0$ при каждом ω : в нуле она 0 по определению, а во всех остальных точках, рано или поздно, она оказывается больше, чем $\frac{1}{2^n}$, с этого момента последовательность X_n будет всё время 0 в этой точке. При этом EX_n к 0 не сходятся: $EX_n = 1$.

Поэтому из сходимости случайных величин не следует сходимость математических ожиданий. Это вы знаете из курса анализа — если функции сходятся, то интегралы совершенно не обязаны сходиться. В данном случае мы взяли вместо поточечной сходимости монотонную, она оказалась в этом плане удобная. И из-за этого же пришлось разбивать на случаи неотрицательной и произвольной, потому что к неотрицательной величине легко придумать монотонно сходящуюся, а к произвольной этого сделать не получится, если она уходит на $-\infty$, потому что мы фиксируем минимальное значение, и уже не можем его уменьшить.

Главное неудобство такого определения в том, что доказывая любое свойство математического ожидания, нам нужно будет через всю эту схему протаскивать.

Чтобы конструкция стала совсем легальной, нужно показать, что какая бы X_n не была выбрана, сходящаяся к X монотонно, предел будет один и тот же. Это доказывается не очень сложно технически, мы на этом не будем останавливаться. Важно, что общую конструкцию мы построили, и каждый раз, глядя на какое-то свойство математического ожидания, полезно мысленно протаскивать его через эту конструкцию.

Интеграл Римана–Стилтьеса

Для нас полезнее будет немного другое определение, у этого определения есть большое неудобство: нам нужно знать, как $X(\omega)$ устроено, а мы при этом верим, что математическое ожидание зависит только от распределения, поэтому мы бы хотели научиться по функции распределения восстанавливать математическое ожидание. Соответствующая конструкция называется **интеграл Римана–Стилтьеса**.

Пусть X имеет функцию распределения F , тогда рассмотрим

$$\int_a^b x dF_X(x) = \lim_{\text{diam } T \rightarrow 0} \sum_{i=0}^{n-1} \xi_i (F(x_{i+1}) - F(x_i)), \quad (8.1)$$

где $T = \{x_0 = a, \dots, x_n = b\}$, $\xi_i \in [x_i, x_{i+1}]$.

Один в один, как Римана, но там dx , поэтому F была тождественной функцией, а сейчас у нас какое-то dF . Вообще говоря, F не обязательно монотонную рассматривать, но нам будут, в основном, функции распределения интересны. Соответственно, интегралом по всей прямой, как обычно, называется предел, если он существует, интегралов по отрезкам:

$$\int_{\mathbb{R}} x dF_X(x) = \lim_{\substack{a \rightarrow -\infty \\ b \rightarrow \infty}} \int_a^b x dF_X(x).$$

Утверждается, что это в точности и совпадает с нашим математическим ожиданием. Начнём с математического ожидания дискретной случайной величины, свойства 1. Наша частичная сумма будет выглядеть так:

$$\sum_{i=1}^n \xi_i (F(x_{i+1}) - F(x_i)) \ominus$$

Значений у нас конечное количество, обозначим их y_1, \dots, y_k , а частичная сумма достаточно мелкая, поэтому можем считать, что ни на какой отрезок $[x_i, x_{i+1}]$ больше одного y_j не попало. Тогда получится, что большинство приращений — это просто 0; приращение F не 0 только на тех отрезках, которые включили в себя точку разрыва, поэтому получается, что это равно

$$\ominus \sum_{i=1}^k \xi_{n_i} (F(y_i) - F(y_i - 0)), \ominus$$

где n_i — номер того отрезка, на котором лежит y_i . Вот так у нас определяются приращения, а это в точности вероятность попасть в точку, а ξ при $\text{diam } T \rightarrow 0$ стремится в точности к y , поэтому это сойдётся в точности к тому, что надо:

$$\ominus \sum_{i=1}^k y_i p_i,$$

так что в дискретном случае наш интеграл Римана-Стилтьеса волшебным образом совпал с тем, что должно быть для случайной величины. Несколько изысканный способ, возможно, определять математическое ожидание дискретной величины, но, по крайней мере, легальный.

В произвольном случае произойдёт то, что написано в (8.1), это и есть последовательность простых случайных величин, которая монотонно сходится к F , если только делать разбиение правильным образом — это когда новое разбиение получается добавлением к старому разбиению ещё точек, и при

этом в качестве ξ_i берём инфимум значения x , то есть левую границу отрезка. Тогда получится, что на каждом отрезке мы случайную величину сделали только меньше. Поскольку разбиение каждый раз измельчается, то последовательность наших функций монотонна, соответственно, она действительно монотонно сходится к нашей случайной величине, потому что это просто один частный вид простых случайных величин, которые сходятся к нашей случайной величине, поэтому это то же самое, что математическое ожидание, то есть мы можем смело здесь написать E_X , иногда пишут $E_F X$, чтобы подчеркнуть, что это именно распределение F :

$$E_F X = \int_{\mathbb{R}} x dF_X(x) = \lim_{\substack{a \rightarrow -\infty \\ b \rightarrow \infty}} \int_a^b x dF_X(x).$$

Для величины, у которой знаки обоих видов, это очевидно получается из свойства 3. Теперь поймём как это считать для разных случаев.

Абсолютно-непрерывный случай

Пусть X — абсолютно-непрерывна с плотностью f_X . Тогда частичная сумма будет выглядеть так:

$$\lim_{\text{diam } T \rightarrow 0} \sum_{i=1}^n \xi_i (F(x_{i+1}) - F(x_i)) \ominus$$

Сделаем хитрый трюк: не будем ξ_i выбирать сразу, а решим позже. Попробуем упростить выражение в скобках в терминах плотности f_X :

$$F(x_{i+1}) - F(x_i) = f_X(\eta_i) \cdot (x_{i+1} - x_i).$$

Для каждого из отрезков $[x_i, x_{i+1}]$ мы выберем такую η_i . Здесь мы воспользовались теоремой Лагранжа. Функция распределения оказалась вот такой конструкцией, где плотность мы считаем интегрируемой по Риману сейчас, иначе у нас ничего не получится. Поскольку ξ_i не важно какую ставить, мы вместо неё тоже η_i поставим. Ну и тогда наше выражение сойдётся к

$$\ominus \lim_{\text{diam } T \rightarrow 0} \sum_{i=1}^n \eta_i f_X(\eta_i) \cdot (x_{i+1} - x_i) \rightarrow \int_a^b x f_X(x) dx.$$

В случае, когда у нас интеграл по всей прямой, мы должны перейти к пределу при $a \rightarrow -\infty$, при $b \rightarrow +\infty$, потому что у нас несобственный интеграл

так и определяется. Отсюда получается, что

$$EX = \int_{\mathbb{R}} x f_X(x) dx.$$

Здесь интеграл предполагается абсолютно сходящимся. Если интеграл абсолютно не сходится, то тогда и математического ожидания нет, или оно бесконечно. То есть в абсолютно-непрерывном случае надо только уметь считать обычный, риманов, интеграл.

Смешанное распределение

Чуть сложнее ситуация обстоит, если у нас распределение смешанное. Если F_X — дискретно-непрерывная, то есть у неё есть моменты разрывов, а между ними функция ведёт себя абсолютно-непрерывно, тогда математическое ожидание X будет равно

$$EX = \sum_i a_i \cdot (F(a_i) - F(a_i - 0)) + \int_{\mathbb{R}} x F'_X(x) dx.$$

Мы разбили функцию на участки непрерывности точками $\{a_1, \dots\}$, получится как-то так: мы должны сложить все точки разрыва, умноженные на высоту, а потом добавить интеграл от производной F , игнорируя её разрывы. В терминах прошлого семинара, это будет утверждение, что мы представили распределение в виде суммы абсолютно-непрерывного и дискретного. В целом, механизм точно такой же, доказывается точно также, как то, что мы делали, просто отдельно придётся рассмотреть те точки, в которых происходит разрыв, они ничего не значат с точки зрения интеграла Римана, потому что длины отрезков маленькие, а с точки зрения интеграла Римана-Стилтьеса значат, потому что у F большое приращение.

Посмотрим на пару примеров.

Пример 8.2. Если $X \sim R[a, b]$, то плотность выглядит

$$f_X(x) = \frac{1}{b-a} I_{x \in [a,b]}.$$

Мы по-взрослому написали, это сильно снижает количество ошибок: I — это индикатор, 1 если верно, 0, если неверно. Не забывайте про них, потому что самая популярная ошибка — это потеря отрезка $[a, b]$, после чего интегралы начинают расходиться.

Тогда математическое ожидание равно

$$EX = \int_{\mathbb{R}} x \frac{1}{b-a} I_{x \in [a,b]} = \int_a^b \frac{x}{b-a} dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}.$$

Если у вас получилась 1, подозреваем, что вы забыли умножить на x . Индикатор просто превратился в отрезок $[a, b]$ — мы интегрируем 0 вне отрезка $[a, b]$, поэтому можно выкинуть эту часть интеграла, интегрируем 1 по отрезку $[a, b]$, остаётся просто интеграл по $[a, b]$. Посчитали интеграл, подставили пределы интегрирования, получили ожидаемый ответ, что, конечно, математическое ожидание достигается прямо в серединке отрезка.

Упражнение 8.1. Пусть $F_X(x) = \frac{1}{3}I_{x \geq -1} + \frac{2}{3}I_{x \geq 1} + \frac{x}{2}I_{x \in [0,1]}$. Найти EX .
Указание. Построим график функции распределения — см. рис. 8.1.

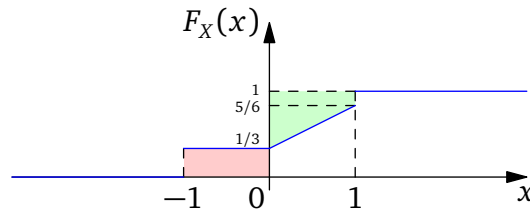


Рис. 8.1: График функции распределения из упражнения 8.1

Теперь берём интеграл:

$$\int x dF_X(x) = -1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{6} + \int_0^1 x \cdot \frac{1}{2} dx$$

Взяли отдельно точки разрыва: -1 и 1 , умножаем их на высоту разрыва, и добавляем интеграл на промежутке $[0, 1]$, где производная ненулевая. Остаётся только досчитать.

Интегрирование по частям

Ещё один полезный трюк для подсчёта математических ожиданий заключается в том, что можно интегрировать по частям, поэтому справедливы такие нехитрые формулы:

$$\int_0^{\infty} x dF_X(x) = - \int_0^{\infty} x d(1 - F_X(x)) \ominus,$$

потому что dF_X — это $-d(1 - F_X)$. Теперь возьмём интеграл по частям:

$$\ominus -x(1 - F_X(x)) \Big|_0^\infty + \int_0^\infty (1 - F_X(x)) dx \ominus$$

Первое слагаемое в 0 равно 0, в ∞ тоже 0, это можно доказать, если математическое ожидание конечно. Соответственно, остаётся такая штука:

$$\ominus \int_0^\infty (1 - F_X(x)) dx.$$

Точно также считается интеграл от $-\infty$ до 0:

$$\int_{-\infty}^0 x dF_X(x) = - \int_{-\infty}^0 F_X(x) dx.$$

Итого, если мы можем собрать вместе эти две формулы, вычтя их друг из друга. Получим, что на самом деле математическое ожидание можно посчитать так: посчитать интеграл от $1 - F$, то есть площадь над графиком, на положительной полуоси и вычесть интеграл от F , то есть площадь под графиком, на отрицательной полуоси. Получается, что на рис. 8.1 разность площадей **зелёной** и **красной** области — математическое ожидание. Такой симпатичный физический вариант подсчёта математического ожидания. Иногда этой формулой гораздо удобнее пользоваться, просто потому, что удобнее работать с $1 - F$ и F , а не с какими-то плотностями.

Свойства математического ожидания

Отметим, что свойства математического ожидания остаются практически такие же, как есть: остаётся линейность, математическое ожидание пересечения независимых — это произведение математических ожиданий, неравенство Йенсена, неравенство Коши–Буняковского, и всё остальное. Подробнее остановимся на предельных свойствах — двух теоремах, которые нам потом пригодятся.

Теорема 8.1 (о монотонной сходимости). Если $X_n(\omega) \uparrow X(\omega)$ и $EX > -\infty$, то $EX_n \rightarrow EX$.

Мы с вами это знали для простых неотрицательных величин, это верно и для произвольных величин. Поэтому бывает удобно такую конструкцию рассматривать.

Теорема 8.2 (о мажорируемой сходимости). Если $X_n(\omega) \rightarrow X(\omega)$ и $|X_n(\omega)| \leq Y(\omega)$, причём $EY < +\infty$, то $EX \rightarrow EY$.

Под \uparrow и \rightarrow понимаем обычную поточечную сходимость функции. Очень удобное свойство, и его полезно в голове держать. Мы с вами уже видели, что из обычной сходимости величин не следует сходимость математических ожиданий: либо нужно требовать монотонность, как в теореме 8.1, либо мажорируемость, как в теореме 8.2.

Дисперсия

Вторая часть семинара посвящена дисперсии. Но здесь первичным для нас фактом является то, что математическое ожидание функции $g(x)$ — это интеграл:

$$Eg(X) = \int_{\mathbb{R}} g(x) dF_X(x).$$

В определении, которое мы использовали, g должна быть кусочно-непрерывной. Вообще говоря, это не важно, но нужно вводить интеграл Лебега–Стилтьеса, мы с этим связываться не хотим, поэтому будем считать, что g — функция приличная.

Пример 8.3. Давайте посчитаем EX^2 , если $X \sim R[a, b]$:

$$\int_{\mathbb{R}} x^2 f_X(x) dx = \int_a^b x^2 \frac{1}{b-a} dx = \frac{b^2 + ab + a^2}{3},$$

поскольку у нас в данном случае $g(x) = x^2$, интеграл Римана–Стилтьеса, в данном случае, это просто $f_X(x) dx$, обычный интеграл Римана, то дальше мы подставляем, что такое f_X и получаем интеграл уже по отрезку.

Определение 8.1. Дисперсией случайной величины X называют

$$DX = E(X - EX)^2 = EX^2 - (EX)^2.$$

Пример 8.4. Для случайной величины из примера 8.3 ($X \sim R[a, b]$):

$$DX = \frac{b^2 + ab + a^2}{3} - \frac{(b+a)^2}{4} = \frac{4(b^2 + ab + a^2) - 3b^2 - 3a^2 - 6ab}{12} = \frac{(b-a)^2}{12}.$$

Базовый принцип здравости сохранён — дисперсия величины пропорциональна длине отрезков. Вроде понятно, что дисперсия равномерной случайной величины должна быть пропорциональна квадрату длины отрезка, и ничему иначе.

Зачастую, удобно, пользоваться первой формулой, не пренебрегайте ей, для некоторых распределений посчитать так дисперсию — это гораздо быстрее, а если вы будете её считать по второй формуле, вы сначала посчитаете первую формулу, подставите во вторую, раскроете, и долгим преобразованием сведёте обратно к тому, что было в начале. Поэтому будьте бдительны и не делайте лишнюю работу.

Полезно также пользоваться свойствами, свойства дисперсии у нас все сохранились с обычного случая:

- 1) $D(cX) = c^2DX$;
- 2) $D(X + c) = DX$;
- 3) если X, Y — независимые, то $D(X + Y) = DX + DY$;
- 4) $D(X + Y) = DX + DY + 2\text{cov}(X, Y)$.

Про дисперсию произведения ничего хорошего не скажем. Зачастую, по свойствам, можно посчитать математическое ожидание чуть приятней:

Пример 8.5. Если $X \sim R[a, b]$, как в примере 8.3, то $X = a + (b - a)Y$, где $Y \sim R[0, 1]$. Это и с физической точки зрения достаточно очевидно, и уж вдвойне очевидно с точки зрения формулы прямой. Соответственно, у нас получается:

$$DX = (b - a)^2 DY,$$

где DY просто проще считается, потому что это

$$DY = \int_0^1 x^2 dx - \left(\int_0^1 x dx \right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Так как-то приятнее считать, потому что всяких a и b нет уже, они все снаружи. Поэтому зачастую приятно воспользоваться свойствами распределения, вытащить оттуда часть параметров, и уже потом считать дисперсию и математическое ожидание у более простой случайной величины, это сокращает количество потраченных нервных клеток.

Нам также пригодится утверждение, что если у нас несколько независимых абсолютно-непрерывных случайных величин X_i , с плотностями f_{X_i} , то верна

такая формула:

$$E g(X_1, \dots, X_n) = \int_{\mathbb{R}^n} g(x_1, \dots, x_n) f_{X_1}(x_1) \dots f_{X_n}(x_n) dx_1 \dots dx_n,$$

то есть формула «математическое ожидание функции равно интегралу от функции по плотности» остаётся верной вот в таком специфическом виде и в многомерном случае тоже. Как это формулируется в самом общем виде, и почему это так, мы обсудим в следующем разделе.

Семинар 9. Случайные векторы

Случайный вектор и его распределение

Определение 9.1. Если X_1, \dots, X_n — случайные величины на одном и том же вероятностном пространстве Ω , но, возможно, с разными распределениями, то набор (X_1, \dots, X_n) называется **случайным вектором**.

С другой стороны,

Определение 9.2. Случайный вектор — это отображение $\vec{X}: \Omega \rightarrow \mathbb{R}^n$, которое обладает свойством измеримости, то есть $X^{-1}(B) \in \mathcal{F}$, $B \in \mathcal{B}(\mathbb{R}^n)$.

Мы с вами обсуждали, что эти два определения равносильны, проверять мы будем определение 9.1, а использовать — определение 9.2, потому что оно гарантирует, что мы можем посчитать вероятность любой функции от вектора, а определение 9.1 не гарантирует, потому что оно говорит про каждый X в отдельности. Теперь поговорим о том, что такое распределение:

Определение 9.3. Распределением называют набор вероятностей (вероятностную меру):

$$P(\omega: \vec{X}(\omega) \in B) = P_{\vec{X}}(B), \quad B \in \mathcal{B}(\mathbb{R}^n).$$

Когда мы говорим найти распределение, мы подразумеваем, что мы как-то охарактеризуем его меру, например, плотностью или функцией распределения. Обычно мы задаём меру P_X на простых порождающих. На прямой в качестве них берут лучи, порождающие функцию распределения, а здесь

$$F_{\vec{X}}(\vec{x}) = P(\omega: X_1 \leq x_1, \dots, X_n \leq x_n).$$

Для двухмерного случая можем изобразить квадрант (см. рис. 9.1), в который нужно попасть (включая его границу) — аналог одномерного луча.

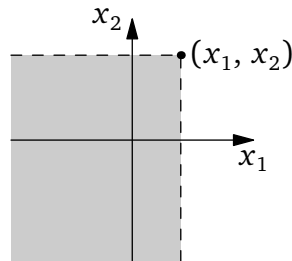


Рис. 9.1: Квадрант, соответствующий двумерной функции распределения

Зная такую вероятность, мы можем найти вероятность попасть в прямоугольник, а затем и в любое борелевское множество. Опять же, надо проверить теорему Каратеодори, но это вопрос к лекциям, возможно, даже по действительному анализу.

Найдём в двухмерном случае вероятность попасть в прямоугольник:

$$P(\vec{X} \in (a_1, b_1] \times (c_1, d_1]) = F_{\vec{X}}(b_1, d_1) - F_{\vec{X}}(b_1, c_1) - F_{\vec{X}}(a_1, d_1) + F_{\vec{X}}(a_1, c_1)$$

Получилось что-то типа формулы включения–исключения: мы из большого квадранта (на рис. 9.2 это вся закрашенная область) вычитаем боковые квадранты и возвращаем назад квадрант, который мы вычли два раза (на рис. 9.2 он показан **тёмно-красным**).

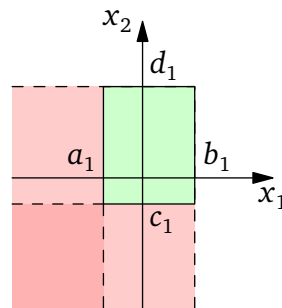


Рис. 9.2: Построение меры прямоугольника по мерам квадрантов

В многомерном случае формула выписывается аналогично, но выглядит более хитро: надо бежать по вершинам булева куба и, в зависимости от того, какая именно вершина, брать её либо с плюсом, либо с минусом:

$$P(\vec{X} \in (a_1, b_1] \times \dots \times (a_n, b_n]) = F_{\vec{X}}(b_1, \dots, b_n) - \sum_i F_{\vec{X}}(b_1, \dots, b_{i-1}, a_i, b_{i+1}, \dots, b_n) + \dots$$

Далее нужно всевозможные пары добавить, где F от всех b_i , но две из них a_i , потом вычесть все тройки, и так далее. Можно считать, что вы взяли булев куб, раскрасили его, начиная с правой верхней вершины: её покрасили в чёрный цвет, её соседей в белый, их соседей в чёрный, и так далее. И берём функцию распределения в каждой из вершин, считая, что чёрный цвет даёт плюс, а белый цвет даёт минус.

Не очень приятно, как можете заметить. Представьте, что надо найти вероятность попасть в шар, тогда придётся долго вписывать в шар параллелепипеды, параллелепипеды выражать таким образом, потом всё это складывать. Поэтому мы в основном будем использовать другие пути задания распределения.

Отметим, что функция распределения, как и прежде, обладает следующими свойствами:

- 1) $F_{\vec{X}}(\vec{x})$ монотонна по каждой переменной;
- 2) $F_{\vec{X}}$ непрерывна «сверху», то есть $F_{\vec{X}}(\vec{x}_m) \rightarrow F_{\vec{X}}(\vec{x})$, $x_{m,i} \rightarrow x_i + 0$, $m \rightarrow \infty$.

Причины этой непрерывности те же, что и раньше: если так приближаться, то квадранты будут вложенные, и их пересечение в точности будет равно нашему квадранту. Если мы, например, будем приближаться с другой стороны, то это уже не будут вложенные множества, но даже если бы они будут расширяющимися, их объединение будет равно квадранту без границы или нескольких границ, поэтому нарушится непрерывность.

- 3) $F_{\vec{X}}(x_1, \dots, x_n) \rightarrow 0$, если $\exists i: x_i \rightarrow -\infty$;
 $F_{\vec{X}}(x_1, \dots, x_n) \rightarrow 1$, если $\forall i: x_i \rightarrow +\infty$.

Оказывается, что, в отличие от одномерного случая, не все функции с такими свойствами являются функциями распределения, есть специальные хитрые примеры, например, такой:

Пример 9.1. Возьмём две точки $(0, 1)$ и $(1, 0)$, проведём через них отрезок, вертикальный и горизонтальный луч (см. рис. 9.3). Положим функцию равной 1 в зелёной области и на границе, и равной 0 — в красной области.

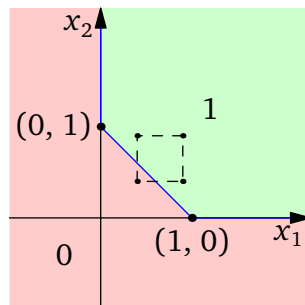


Рис. 9.3: Функция из примера 9.1

Эта функция обладает всеми свойствами 1–3, потому что если какой-то из x_i устремить в $-\infty$, то функция обнулится, ведь мы попадём в красную зону; если всё устремить в $+\infty$, функция станет 1; если взять по каждой из переменных, у нас функция монотонна, понятно, она сначала 0, потом 1 всегда становится; непрерывность «сверху» у неё есть просто потому, что мы границу включили в область с 1. Но это не является функцией распределения никакого распределения, потому что если мы возьмём квадратик, обозначенный на рис. 9.3 пунктиром, вероятность попасть в него $P = 1 - 1 - 1 + 0 = -1$.

Такого быть не может, значит, функцией распределения такая функция не является, хотя она нашими свойствами обладает.

Поэтому дописывают четвёртое свойство, которое на самом деле можно использовать вместо свойства 1:

4) Вероятность попасть в параллелепипед ≥ 0 .

То есть условие на то, что такая знакопеременная сумма функций распределения в разных точках, согласованных правильным образом, неотрицательна. Вот если такие 4 условия выполнены, тогда распределение точно существует и единственно, так что любая такая функция является функцией распределения. По сравнению с одномерным случаем, возникает неприятное условие на то, что надо проверять какие-то такие штуки. Скажем, что

Определение 9.4. Вектор \vec{X} **равномерно распределён** на множестве $S \subseteq \mathbb{R}^n$ ($\vec{X} \sim R(S)$), если

$$P(\vec{X} \in A) = \frac{V(A \cap S)}{V(S)}.$$

Если у нас есть множество S , на котором величина распределена, а мы считаем вероятность, что мы попадём в какое-то множество A , то часть $A \setminus S$ нас вообще не интересует, поэтому мы берём кусочек $A \cap S$, и считаем отношение объёмов $A \cap S$ и S , это и называется равномерным распределением, это соответствует нашему представлению о том, что такое равномерность. Первое, что мы должны научиться делать — это считать функции распределения по заданному распределению и по заданной величине. Давайте посчитаем функцию распределения для равномерного распределения на квадрате:

Пример 9.2. Пусть $\vec{X} \sim R[0, 1]^2$. Давайте зададим функцию распределения вектора \vec{X} . Мы должны перебрать всю плоскость, и для каждой точки плоскости посчитать вероятность попасть в квадрант с углом в этой точке. Если этот квадрант полностью расположен слева или снизу от квадрата, то пересечения с квадратом нет вообще, поэтому вероятность нулевая. Теперь следующий блок — если угол квадранта попал в квадрат, то пересечение будет прямоугольником со сторонами x и y , и площадь прямоугольника xy нужно поделить на площадь квадрата 1. В области над верхней стороной квадрата, часть, попавшая в квадрат — прямоугольник, у которого стороны 1 и x , площадь у него x . Аналогичный случай справа от правой стороны. Ну и, наконец, остаётся последний случай, когда $x > 1$, $y > 1$, там функция распределения равна 1.

Запишем найденную нами функцию распределения:

$$F_{\vec{X}}(\vec{x}) = \begin{cases} 0, & x < 0 \text{ или } y < 0; \\ xy, & x \in [0, 1], y \in [0, 1]; \\ x, & x \in [0, 1], y > 1; \\ y, & y \in [0, 1], x > 1; \\ 1, & x > 1, y > 1. \end{cases}$$

Вот такая получается нехитрая процедура, пока поупражняемся с ней.

Упражнение 9.1. Пусть $X, Y \sim \text{Bern}(p)$. Найти совместную функцию распределения $(X, X + Y)$.

Указание. Нарисуем, что такое $(X, X + Y)$ — оно принимает 4 различных значения: $(0, 0)$ с вероятностью $(1 - p)^2$; $(0, 1)$ с вероятностью $p(1 - p)$; $(1, 1)$ с вероятностью $p(1 - p)$ и $(1, 2)$ с вероятностью p^2 (см. рис. 9.4). Мы просто перебрали все значения X и Y и посмотрели, какие значения может принимать пара $(X, X + Y)$, она принимает эти 4 значения.

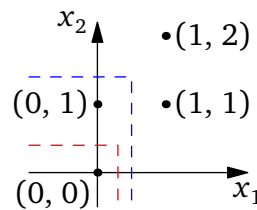


Рис. 9.4: Возможные значения пары $(X, X + Y)$, $X, Y \sim \text{Bern}(p)$

Дальше берём какой-нибудь квадрант, вероятность попасть в него — сумма вероятностей всех точек, которые в него попали, потому что распределение дискретное. Нужно аккуратно это выписать: если $x < 0$ или $y < 0$, то это 0; единственная точка $(0, 0)$ попадёт, если $x, y \geq 0, y < 1$ — любой такой квадрант (на рис. 9.4 изображён **красным** пунктиром) всегда будет включать только эту точку, при этом вероятность $(1 - p)^2$; пусть попало 2 точки, это могло произойти единственным образом, если попали точки $(0, 0)$ и $(0, 1)$ (соответствующий им квадрант изображён на рис. 9.4 **синим** пунктиром), это если $0 \leq x < 1$, а $y \geq 1$, в сумме эти точки дают вероятность $1 - p$; бывает, когда попадут 3 точки, но опять же, если $(1, 2)$ попадёт, то уже всё попадёт, то есть единственный вариант, когда попали 3 точки — это три точки без $(1, 2)$, вероятность будет $1 - p^2$, потому что не попадёт одна точка с вероятностью p^2 , это будет когда x уже хотя бы 1, а $1 \leq y < 2$, чтобы точка $(1, 2)$ ещё не попала, а остальные 3 уже попали; в оставшихся случаях вероятность равна 1.

Абсолютно–непрерывный случай

Две задачи, которые мы разобрали, отражают два типовых случая, в которых мы будем работать — дискретный и абсолютно–непрерывный. В одном случае мы считаем только отдельные точки, складываем их вероятности, а в другом случае мы считаем какие-то площади, или что-то там интегрируем. Поговорим подробнее про абсолютно–непрерывный случай.

Определение 9.5. Говорят, что вектор \vec{X} — **абсолютно непрерывный**, если

$$\exists f_{\vec{X}} : P(\vec{X} \in B) = \int_B \dots \int f_{\vec{X}}(\vec{x}) d\vec{x}. \quad (9.1)$$

Вообще говоря, это интеграл Лебега; если B — хорошее множество, по которому мы умеем интеграл Римана считать, то интеграл Римана. Очевидно, что $f_{\vec{X}}$ должна быть неотрицательной: $f_{\vec{X}} \geq 0$ и обладать свойством, что

$$\int_{\mathbb{R}^n} \dots \int f_{\vec{X}}(\vec{x}) d\vec{x} = 1.$$

Оказывается, что что-то является плотностью тогда и только тогда, когда эти два условия выполнены: оно неотрицательно и интеграл от него единичка, потому что можно просто меру явно задать формулой (9.1) и проверить, что эта мера счётно–аддитивная, это следует из свойства интеграла. Если по всему пространству проинтегрировать плотность, должна получиться вероятность попасть куда-то, она 1. В многомерном случае, в отличие от одномерного, плотность начинает играть решающее значение. Например, через плотность, в отличие от функции распределения, ничего не стоит найти вероятность попасть в шар: нужно просто проинтегрировать по шару заданную функцию. Так что получается очень полезный случай, правда, и семейство абсолютно–непрерывных распределений, стало гораздо уже, ведь, например, вектор с координатами (X, X) , где $X \sim R([0, 1])$, уже не будет абсолютно–непрерывным. Давайте пару примеров посмотрим:

Пример 9.3. Пусть $X, Y \sim R[0, 1]$, независимые. Это тоже самое, что бросить точку в квадратик, мы уже считали функцию распределения в примере 9.2. Напомним, что она получилась

$$F_{X,Y}(x, y) = \begin{cases} 0, & x < 0 \text{ или } y < 0; \\ xy, & x \in [0, 1], y \in [0, 1]; \\ x, & x \in [0, 1], y > 1; \\ y, & y \in [0, 1], x > 1; \\ 1, & x > 1, y > 1. \end{cases} \quad (9.2)$$

Мы это называли бросанием точки в квадрат, но это тоже самое, что независимо бросить две точки на отрезок. Найденная нами плотность, как минимум, должна быть такой, чтобы функция распределения в точке была равна интегралу по этому квадранту от плотности, потому что это частный случай: если множество B взять квадрантом, то получится, что функция распределения равна интегралу от квадранта, но если мы представим себе это в виде повторных интегралов, получается, что (напишем для двухмерного случая, точно также для любого другого):

$$F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) du dv.$$

Чтобы убить эти интегралы, нужно взять частную производную:

$$\frac{\partial}{\partial x} \frac{\partial}{\partial y} F_{X,Y}(x, y) = f_{X,Y}(x, y).$$

Если функция распределения дифференцируема, то у нас действительно получится плотность, и не важно в каком порядке дифференцировать. Ясно, что не всегда этот трюк удастся, например, $F_{X,Y}$ должна быть обязательно непрерывной. Получается, что в нашем примере плотность равна

$$f_{X,Y}(x, y) = \begin{cases} 1, & x \in [0, 1], y \in [0, 1]; \\ 0, & \text{иначе.} \end{cases}$$

Это соответствует нашему представлению об одномерном распределении: у равномерного распределения плотность должна быть одинаковой, ведь плотность — это предельная вероятность попасть куда-то в окрестность точки: если мы B начнём стягивать в точку, и делить это на объём B , то, для непрерывной плотности, это как раз сойдётся к плотности. Полезно задуматься над тем, является ли действительно эта функция плотностью, нужно проверить, обладает ли она этими двумя свойствами: функция распределения будет совпадать с (9.2), и это правда будет плотность распределения.

Иногда может оказаться, что у плотности интеграл не 1, это значит, что мы потеряли разрыв или ещё что-то, следовательно, плотности нет, в смысле, что производная не совпадает с плотностью. Посмотрим на второй пример:

Пример 9.4. Представим, что мы в примере 9.3 взяли не X, Y , а X, X . Тогда

$$F_{X,X}(x, y) = P(X \leq x, X \leq y) = \begin{cases} 0, & x < 0 \text{ или } y < 0; \\ x, & 0 < x < 1, x < y; \\ y, & 0 < y < 1, y < x; \\ 1, & \text{иначе,} \end{cases}$$

поскольку вероятность, что мы меньше x и y — это вероятность, что мы меньше меньшего из x и y ; поскольку $X \sim R[0, 1]$, то $P(X < x) = x$.

Частные производные этой функции равны 0 во внутренности каждой из областей, где-то эта производная вообще не определена, главное, что она определена с точностью до множества меры 0. Получили, что $\frac{\partial}{\partial x} \frac{\partial}{\partial y} F = 0$, значит, действительно, плотности у нашего распределения нет, потому что если бы она была, то вероятность попасть в отрезок, соединяющий точки $(0, 0)$ и $(1, 1)$, была бы равна интегралу по этому множеству от плотности, но у нас это 0, потому что площадь отрезка 0, а вероятность попасть в такой отрезок 1, потому что больше никуда у нас величина попадать не умеет, у неё же координаты X, X .

Поэтому все такие распределения, которые живут на кривой или на другом множестве меры 0, остаются сингулярными. На прямой множество меры 0 было изысканным объектом, на плоскости множество меры 0 — это многие объекты, к которым мы привыкли, например, одномерные множества в двухмерном пространстве. Все соответствующие распределения сразу перестают быть абсолютно-непрерывными, у них плотности нет. Это полезно зафиксировать, и в голове держать.

Теперь несколько ключевых фактов для нас, как работать с плотностями. Пусть у нас плотность есть, мы будем стараться так же, как в одномерном случае, задавать вектор с помощью плотности и работать дальше с этой плотностью, в основном, будем работать с абсолютно-непрерывным вектором. Значит, зафиксируем несколько важных фактов:

Плотности подвекторов

Если \vec{X} имеет плотность $f_{\vec{X}}$, то X_1 имеет плотность

$$f_{X_1}(x_1) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{\vec{X}}(x_1, y_2, \dots, y_n) dy_2 \dots dy_n. \quad (9.3)$$

Нужно проинтегрировать по всем «лишним» переменным, мы обозначили их другой буквой, чтобы не путаться. Чтобы понять, откуда берётся такая формула, выразим функцию распределения \vec{X} через плотность:

$$F_{X_i}(x_i) = P((X_1, \dots, X_n) \in (-\infty, x_1] \times \mathbb{R} \times \dots \times \mathbb{R}) \ominus$$

Выпишем это через плотность вектора, сразу записав повторные интегралы

вместо многократного:

$$\ominus \int_{-\infty}^{x_1} \int_{\mathbb{R}} \dots \int_{\mathbb{R}} f_{\vec{X}}(y_1, \dots, y_n) dy_1 \dots dy_n.$$

Чтобы теперь найти плотность, нужно взять производную, первый интеграл по верхнему пределу уберётся, и как раз получится искомая формула (9.3). Так что никакой магии здесь нет, но появляется такая удобная формула, как искать маргинальное распределение, зная совместное. Точно также можно найти плотность любого подвектора: нужно проинтегрировать по «лишним» переменным. Проще, конечно, по функции распределения вектора искать функцию распределения подвектора, это просто надо просто подставить $+\infty$ вместо ненужных компонент, там ничего не надо интегрировать, с плотностями придётся интегрировать, но зато с ними работать удобнее во всём остальном.

Преобразование плотности при гладких заменах координат

Пусть $g \in C^1(\mathbb{R}^n)$ — гладкое взаимно однозначное отображение из \mathbb{R}^n в \mathbb{R}^n , тогда плотность $g(\vec{X})$ в точке \vec{u} — это плотность \vec{X} в точке $g^{-1}(\vec{u})$, умноженная на модуль якобиана g^{-1} от \vec{u} :

$$f_{g(\vec{X})}(\vec{u}) = f_{\vec{X}}(g^{-1}(\vec{u})) \cdot |J_{g^{-1}}(\vec{u})|. \quad (9.4)$$

Примерно обрисую откуда такая формула берётся, чтобы было легче её вспоминать. Считаем, что плотность непрерывная, то есть мы в какой-то точке непрерывности плотности, тогда плотность $f_{g(\vec{X})}(\vec{u})$ равна интегралу от плотности по множеству $B(\vec{u})$, делить на объём $V(B(\vec{u}))$, при B , стягивающемся в точку (под диаметром мы понимаем диаметр множества):

$$f_{g(\vec{X})}(\vec{u}) = \lim_{\text{diam } B \rightarrow 0} \frac{\int_{B(\vec{u})} f_{g(\vec{X})}(\vec{v}) d\mathbf{v}}{V(B(\vec{u}))} \ominus$$

Взяли какое-то множество объёмное, например, шар, вокруг точки \vec{u} , посчитали интеграл по шару, разделили на объём шара, у нас это будет, по теореме о промежуточном значении, равно значению плотности в промежуточной точке, в силу непрерывности, это будет сходиться к значению плотности в точке, а интеграл в числителе — это вероятность $g(\vec{X})$ попасть в $B(\vec{u})$:

$$\ominus \lim \frac{P(g(\vec{X}) \in B(\vec{u}))}{V(B(\vec{u}))} = \lim \frac{P(\vec{X} \in g^{-1}(B(\vec{u})))}{V(B(\vec{u}))}, \ominus$$

поскольку у нас функция взаимно однозначная. $g^{-1}(B(\vec{u}))$ — это какая-то окрестность, но уже не шарообразная, её как-то деформировало g^{-1} , но от точки $g^{-1}(\vec{u})$. Заметим, что мы делим не на тот объём, чтобы получилась плотность \vec{X} : мы делим на объём $B(\vec{u})$, а надо делить на объём прообраза $B(\vec{u})$, и поэтому мы запишем это так:

$$\Leftrightarrow \lim \frac{\underbrace{P(\vec{X} \in \tilde{B}(g^{-1}(\vec{u})))}_{\rightarrow f_{\vec{X}}(g^{-1}(\vec{u}))}}{\underbrace{V(\tilde{B})}_{\rightarrow |J_{g^{-1}}(\vec{u})|}} \cdot \frac{V(\tilde{B})}{V(B)}.$$

Первый множитель сходится к плотности \vec{X} в точке $g^{-1}(\vec{u})$, а второй множитель есть предел, во сколько раз изменился объём окрестности точки при гладком преобразовании, это и есть модуль определителя матрицы Якоби, поскольку у нас объём без знака.

Вот и всё доказательство нашей теоремы. Доказывают её обычно не так, а просто через интеграл, но понимать приятнее всего так, понятно откуда взялся кусок с якобианом: он взялся из-за того, что мы смотрим вероятность попасть в точку вместе с окрестностью, у нас есть масштабный коэффициент, который отвечает за объём окрестности, и приходится добавлять дополнительный множитель, когда g этот объём меняет. За счёт этого у нас интеграл от плотности останется 1; если бы мы потеряли этот множитель, у нас интеграл от плотности перестал бы быть 1, а вот за счёт этого дополнения как раз всё хорошо. Давайте посмотрим, как это работает:

Пример 9.5. Пусть у (X, Y) уже знакомая нам из примера 9.3 равномерная плотность:

$$f_{X,Y}(x, y) = \begin{cases} 1, & x \in [0, 1], y \in [0, 1]; \\ 0, & \text{иначе.} \end{cases} = I_{x,y \in [0,1]}$$

Давайте потихонечку взрослеть, будем записывать это в терминах индикаторов, это удобно в данном случае, потому что это позволит нам не запутаться. Рассмотрим вектор $g(X, Y) = (U, V) = \left(XY, \frac{X}{Y}\right)$. Сначала найдём обратное отображение g^{-1} , то есть запишем систему и решим её относительно x, y :

$$\begin{cases} u = xy \\ v = \frac{x}{y} \end{cases} \Leftrightarrow \begin{cases} x = \sqrt{uv} \\ y = \sqrt{\frac{u}{v}} \end{cases}$$

Чтобы решить систему, нужно перемножить уравнения и извлечь корень, мы здесь пользуемся тем, что $x, y \in [0, 1]$, u, v тоже положительные, поэто-

му можно точно сказать, что x — это положительный корень. Действительно, функция на нашей области взаимно однозначна. Само отображение мы нашли, теперь ищем матрицу Якоби этого самого обратного отображения:

$$J_{g^{-1}}(u, v) = \begin{pmatrix} \frac{\sqrt{v}}{2\sqrt{u}} & \frac{\sqrt{u}}{2\sqrt{v}} \\ \frac{1}{2\sqrt{uv}} & \frac{-\sqrt{u}}{2\sqrt{v^3}} \end{pmatrix}.$$

Нас интересует не матрица Якоби, а только её определитель:

$$|J_{g^{-1}}| = \left| -\frac{\sqrt{uv}}{4\sqrt{uv^3}} - \frac{\sqrt{u}}{4\sqrt{uv^2}} \right| = \frac{1}{2v}.$$

Чтобы найти плотность U, V в точке (u, v) , надо подставить в старую плотность точку прообраза, здесь нам очень пригодится наш индикатор, а дальше умножаем на модуль определителя матрицы Якоби:

$$f_{U,V}(u, v) = I_{\sqrt{uv} \in [0,1], \sqrt{\frac{u}{v}} \in [0,1]} \cdot \frac{1}{2v}$$

Теперь выражение под индикатором попытаемся упростить. Корни меньше 0 не очень любят бывать, поэтому это не очень важное условие, а вот важно, что они ≤ 1 . Получается, что $uv \leq 1$ и $\frac{u}{v} \leq 1 \iff u \leq v$. Естественно, u и v положительны, иначе мы бы не смогли извлечь корень. Получаем вот такой вот ответ: $0 \leq uv \leq 1$, $0 \leq u \leq v$, и вот на этом множестве у нас функция имеет вид $\frac{1}{2v}$, а на всём остальном это 0.

Заметьте, что если бы мы потеряли индикатор, и сказали, что плотность равна 1, получилась бы неинтегрируемая функция $\frac{1}{2v}$, которая ещё непонятно как зависит от u , а у нашей функции зависимость от u есть, но только в индикаторе. Так что не теряйте индикаторы и внимательно следите за множеством. Индикатор — очень удобная штука, в том плане, что мы функцию записали единым выражением, и нам в неё легче подставлять переменные и следить, что с ними произошло. В принципе, можно это делать без индикатора, просто аккуратно работая с областью определения.

На прямой мы так уже делали, только явно это не афишировали: мы брали функцию распределения, как-то с ней работали, дифференцировали, но на самом деле всегда в конце мы приходили к (9.4), только там модуль якобиана — это 1, делить на модуль производной.

Теперь ещё одно важное замечание к (9.4): если g — не взаимно однозначная, но у неё несколько прообразов, например, всегда 2 у каждой точки,

кроме, может быть, какого-то множества меры 0, тогда у нас просто вероятность бы распалась на две. Вспоминая, откуда у нас бралась формула, когда мы брали вероятность того, что $g(\vec{X})$ попала в окрестность точки, у нас бы это превратилось в вероятность того, что \vec{X} попал либо в первый прообраз, либо во второй прообраз, и в итоге у нас бы получилась сумма.

То есть, если прообразов несколько, мы разрезаем обратную функцию на несколько кусков $g_1^{-1}, g_2^{-1}, \dots, g_k^{-1}$, склейки областей нас не интересуют, мы верим, что это множество меры 0. Тогда получится, что плотность образа склеивается как сумма:

$$f_{g(\vec{X})}(\vec{u}) = \sum_{i=1}^k f_{\vec{X}}(g_i^{-1}(\vec{u})) |J_{g_i^{-1}}(\vec{u})|.$$

Например, в одномерном случае, функция $y = x^2$ склеивает положительную полуось и отрицательную, у неё есть два прообраза — $\pm\sqrt{x}$, вот мы каждый из этих прообразов отдельно рассматриваем и решаем ту же задачу, а потом складываем результат. Если бы мы в примере 9.5 взяли не равномерное распределение, а какое-нибудь другое, то прообраз был бы не единственный, потому что мы могли бы брать корень и минус корень, и тогда мы бы как раз решали две такие задачи, отдельно бы смотрели прообразы каждой ветви, считали бы якобиан, считали бы плотность, и складывали бы их, но ничего сложного в этом бы не было, просто бы чуть побольше пришлось писать.

Независимость компонент вектора

Пусть вектор \vec{X} — абсолютно-непрерывный, то есть у него есть плотность. Тогда компоненты вектора X_1, \dots, X_n независимы \iff совместная плотность равна произведению плотностей, на самом деле, достаточно даже, чтобы было произведение функций, каждая своей переменной:

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n g_i(x_i),$$

где $g_i(x_i)$ — некоторые функции. Если плотность распалась в произведение функций, каждая своей переменной, то величины независимы; если не распалась, то зависимы. Можно это обобщать до бесконечности: можно сказать, например, что X_1, \dots, X_m не зависит от X_{m+1}, \dots, X_n тогда и только тогда, когда вся функция распалась в произведение функций первых m переменных и оставшихся переменных.

Пример 9.6. Пусть X, Y — независимы, найдём плотность f_{X+Y} , то есть выведем заново формулу свёртки с помощью новых инструментов. Для этого

найдем плотность $f_{X+Y,X}(u, v)$, а затем сведём это обратно к $f_{X,Y}(x, y)$:

$$f_{X+Y,X}(u, v) = f_{X,Y}(v, u - v) \Leftrightarrow$$

Определителя матрицы Якоби даже не возникло, поскольку матрица преобразования верхнетреугольная, и у неё определитель 1. Плотность $f_{X,Y}(x, y)$ — это просто произведение плотностей $f_X(x)$ и $f_Y(y)$:

$$\Leftrightarrow f_X(v) f_Y(u - v).$$

Теперь проинтегрируем по v , по ненужной нам переменной:

$$f_{X+Y}(u) = \int_{\mathbb{R}} f_X(v) f_Y(u - v) dv.$$

Получилась в точности формула свёртки. Нужно овладеть тремя простыми инструментами: плотность независимых величин — это произведение плотностей; плотность при преобразовании меняется так-то; плотность подвектора — интеграл плотности вектора по лишним переменным. Если бы здесь было произведение, мы бы взяли плотность $f_{X,Y,X}$, тут уже якобиан бы возник, ну а дальше всё тоже самое. Мы с вами уже использовали немножко этой теории, когда говорили про вероятность попадания вектора с независимыми компонентами в множество — это интеграл по этому множеству от произведения плотностей, это было потому, что плотность совместная в данном конкретном случае была равна произведению плотностей. Можно сказать, что вероятность попасть в любое множество — это интеграл совместной плотности по этому множеству.

Математические ожидания функций от векторов

Ну и последнее, это полезная формула, которая позволяет считать математические ожидания векторов, такая же, как и в одномерном случае:

$$E(X_1, \dots, X_n) = \int_{\mathbb{R}^n} g(x_1, \dots, x_n) f_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1 \dots dx_n,$$

если (X_1, \dots, X_n) — абсолютно-непрерывный вектор. Это обобщение той формулы, которую мы уже видели для независимых величин на прошлом занятии, там у нас было, опять же, произведение плотностей, но это же верно и в зависимом случае.

Семинар 10. Характеристические функции

Математическое ожидание комплекснозначной случайной величины

Будем рассматривать случайную величину $X: \Omega \rightarrow \mathbb{C}$ по следующему принципу: это просто пара случайных величин $X_1 + iX_2$, где X_1, X_2 — обычные случайные величины, а i — мнимая единица.

Ровно тем же принципом, что и раньше, мы вводим математическое ожидание $EX = EX_1 + iEX_2$, ровно так же вводим независимость: *независимость комплекснозначных случайных величин* — это значит, что пара (X_1, X_2) не зависит от пары (Y_1, Y_2) , как векторы. То есть, в принципе, комплекснозначная величина отличается от вектора только тем, что мы ввели отдельные правила умножения, которые мы знаем для комплексных чисел.

Соответственно, нам понадобятся два важных свойства:

1) $E|X| \geq |EX|$.

Раньше это было простое утверждение, потому что там просто модуль был больше или равен самой величине, теперь у нас отношений порядка на величинах нет, это не много более хитрый факт. В терминах X_1, X_2 , это означает, что $E\sqrt{X_1^2 + X_2^2} \geq \sqrt{EX_1^2 + EX_2^2}$ для любых X_1, X_2 . Мы это доказывать не будем, но это следует из неравенства Йенсена для функции двух переменных, оно доказывается точно также, как для одной переменной.

2) $EXY = EXEY$ для независимых X, Y .

Это не совсем очевидно, потому что произведение комплексных величин — специфическая штука, и то, что это сохранится — нужно просто взять и расписать аккуратно. Мы на этом подробно останавливаться не будем.

Определение и существование характеристической функции

Определение 10.1. Характеристической функцией вещественнозначной случайной величины X называют функцию

$$\psi_X(t) = E e^{itX}.$$

Она похожа на производящую, только в производящей был s^X , а здесь вместо s e^{it} . Для дискретного распределения, если вы знаете производящую функцию, то характеристическая — это будет она же, в точке e^{it} .

Мы при этом подразумеваем совсем простую вещь, то есть на самом деле

$$E e^{itX} = E \cos(tx) + i E \sin(tx).$$

Первый вопрос, который мы должны задать с вами, когда определяем такой объект — «Существует ли у любой величины такое математическое ожидание?» Математического ожидания у случайной величины может не быть, если у неё положительная часть бесконечна или отрицательная часть бесконечна, но у нас здесь как положительная, так и отрицательная части и у вещественной, и у мнимой частей ограничены 1, значит, всё у нас ограничено, и оба математических ожидания существуют и конечны.

Получается, что характеристическая функция существует у любой величины, мы бы с удовольствием производящую ввели для любой величины, но математическое ожидание s^X может расходиться, а тут такой проблемы нет, это заведомо существующий объект. Позже мы покажем, что комплексная экспонента обладает теми же свойствами, что s^X , прежде всего мультипликативностью, что делает такую конструкцию очень удобной.

Примеры характеристических функций

Пример 10.1. $X \sim \text{Bern}(p)$, у неё характеристическая функция

$$\psi_X(t) = (1 - p) + p e^{it},$$

поскольку случайная величина e^{itX} принимает значения 1 и e^{it} . Заметим что для дискретных величин можно подставить вместо s в производящую функцию e^{it} , и получить характеристическую, потому что формулы совпадают.

Пример 10.2. $X \sim \text{Geom}(p)$, тогда

$$\psi_X(t) = \sum_{k=0}^{\infty} (1-p)^k p e^{itk} \ominus$$

p выносится, остаётся геометрическая прогрессия, можно честно её посчитать: нужно расписать вещественную и мнимую часть, и изысканно посчитать суммы, по аналогии с суммой $\cos(tk)$, когда мы домножали на $\sin(\frac{x}{2})$ и переходили к разности синусов. Но мы сделаем вид, что для дискретного случая мы знаем, что ряды сходятся к тому же самому, получается, опять же, то же, что было в случае производящей функции, только с e^{it} :

$$\ominus p \frac{1}{1 - (1-p)e^{it}} = \frac{p}{1 - (1-p)e^{it}}.$$

Теперь давайте посмотрим на какой-нибудь экспоненциальный случай:

Пример 10.3. Пусть $X \sim \exp(1)$, тогда характеристическая функция будет:

$$\psi_X(t) = \int_0^{\infty} e^{itx} e^{-x} dx \ominus$$

Возникает соблазн собрать экспоненту, но мы так не умеем делать: если мы сделаем замену, то новая переменная уже не будет лежать от 0 до ∞ , она будет лежать на какой-то повернутой прямой, для работы с ней нужно некоторое владение комплексным анализом. Поэтому сегодня мы просто разбиваем на вещественную и мнимую часть и интегрируем их по-отдельности:

$$\ominus \underbrace{\int_0^{\infty} \cos(tx) e^{-x} dx}_{=I} + i \underbrace{\int_0^{\infty} \sin(tx) e^{-x} dx}_{=J} \ominus$$

Посчитаем оба интеграла по частям:

$$I = \int_0^{\infty} \overbrace{\cos(tx)}^{f(x)} \overbrace{e^{-x}}^{g'(x)} dx = -e^{-x} \cdot \cos(tx) \Big|_0^{\infty} - t \int_0^{\infty} \sin(tx) e^{-x} dx = 1 - tJ;$$

$$J = \int_0^{\infty} \sin(tx) e^{-x} dx = -e^{-x} \cdot \sin(tx) \Big|_0^{\infty} + t \int_0^{\infty} \cos(tx) e^{-x} dx = tI.$$

Значит, $I = 1 - t^2 I \implies I = \frac{1}{1+t^2}$, $J = \frac{t}{1+t^2}$. Подставляем в $\psi_X(t)$, получаем:

$$\ominus \frac{1+it}{1+t^2} = \frac{1}{1-it}.$$

Получилось в точности то же самое, что было бы, если мы в самом начале взяли, сделали замену и проинтегрировали, но мы так не умеем.

Мы узнали, что чтобы найти характеристическую функцию какой-то величины, просто считаем, по-честному, без манипуляций с комплексными экспонентами, разбивая на косинус и синус, либо сумму, либо интеграл; если мы хотим по характеристической функции найти случайную величину, мы пытаемся представить её в виде суммы комплексных экспонент или интеграла.

Свойства характеристической функции

Теперь поговорим, зачем характеристическая функция нужна и какие у неё свойства. Итак, начнём со свойств:

- 1) $\psi_X(0) = 1$, $|\psi_X(t)| \leq 1$, поскольку по свойству 1 комплекснозначного математического ожидания, $|\mathbb{E} e^{itX}| \leq \mathbb{E} |e^{itX}| = 1$;
- 2) ψ_X непрерывна на всей прямой \mathbb{R} , более того она равномерно непрерывна, то есть $\forall \varepsilon > 0 \exists \delta: \forall t_1, t_2: |t_1 - t_2| < \delta \implies |\psi_X(t_1) - \psi_X(t_2)| < \varepsilon$.

Важно, что она даже равномерно непрерывна на всей прямой. По сути, надо как-то переставить интеграл с пределом, это можно сделать с помощью теоремы Лебега о мажорируемой сходимости, возиться с этим не будем.

- 3) Если $\exists \mathbb{E} X^k$, то $\exists \psi^{(k)}(t)$ на всей прямой, причём $\psi^{(k)}(0) = i^k \mathbb{E} X^k$.

Если мы разрешим дифференцировать под знаком математического ожидания, не обосновывая законность этого, получим именно $(\mathbb{E} e^{itX})^{(k)} \Big|_{t=0} = i^k \mathbb{E} X^k$. Получается очень удобно, можно математические ожидания находить очень просто. Бывает, что это действительно очень быстрый способ найти момент.

- 3*) Если $\exists \psi^{(2k)}(0)$, то $\exists \mathbb{E} X^{2k}$, причём $\mathbb{E} X^j = i^{-j} \psi^{(j)}(0)$, $j \leq 2k$.

Помним, что математические ожидания обладают таким свойством: если есть $\mathbb{E} X^\alpha$, то есть $\mathbb{E} X^\beta$ при всех $\beta < \alpha$. Поэтому если $\exists \mathbb{E} X^{2k}$, то существуют и все моменты до, и их можно находить по этой формуле: $\mathbb{E} X^j = i^{-j} \psi^{(j)}(0)$, $j \leq 2k$. Важно, что производная именно чётная, потому что у математических ожиданий не бывает условной сходимости, а у пределов бывает сходимости, когда они знакопеременно сходятся. Заметьте, что если есть чётная производная в нуле, то есть математическое ожидание, то есть чётная производная везде, это довольно характерное свойство характеристической функции.

Мы нашли первую причину, для чего нужна характеристическая функция — научившись дифференцировать, мы можем быстро считать все моменты.

- 4) $\psi_{aX+b}(t) = \mathbb{E} e^{i(aX+b)t} = e^{ibt} \psi_X(at)$.

Удобно таким образом легко перепараметризовывать характеристическую функцию при линейных преобразованиях, мы это с вами будем использовать, когда будем заниматься центральной предельной теоремой.

- 5) $\psi_X(-t) = \psi_{-X}(t) = \overline{\psi_X(t)}$.

Первое равенство следует из свойства 4, второе прямо из определения, потому что сопряжение $\mathbb{E} e^{itX}$ — это $\mathbb{E} e^{-itX}$.

- 6) ψ_X — вещественна $\iff X \stackrel{d}{=} -X$.

В силу свойства 5, это равносильно тому, что ψ_X — вещественна, когда у X и $-X$ одна и та же характеристическая функция, то есть когда распределение симметрично. В дискретном случае это значит, что вероятности одинаковые у точек, идущих симметрично от нуля, в абсолютно-непрерывном это значит, что плотность симметрична около нуля, в общем случае это какое-то условие на функцию распределения.

7) Если X_1, \dots, X_n — независимые, то

$$\psi_{X_1+\dots+X_n}(t) = E e^{it(X_1+\dots+X_n)} = \psi_{X_1}(t) \dots \psi_{X_n}(t).$$

Вот оно — ключевое свойство, из-за которого так любят характеристические функции, из-за того же, из-за чего производящие — очень просто и удобно работать с суммой, не надо считать свёртки, возиться с ними, посчитали характеристическую функцию, перемножили, и готово. Кроме случаев, когда можно просто посчитать распределение с помощью характеристической функции, бывает также иногда, что дело даже не в том, чтобы посчитать, а в том, чтобы доказать что-то про него.

8) Если X_i — независимые одинаково распределённые, а N — независимая от них целочисленная неотрицательная, то

$$\psi_{X_1+\dots+X_N}(t) = \varphi_N(\psi_X(t)).$$

То же свойство, что было у производящей, только там была производящая от производящей, а тут производящая от характеристической.

Тут уже пошли изысканные свойства, ключевые свойства — это 3, 4, 7.

9) Если N — целочисленная неотрицательная, а X_i — какие-то, не зависящие от N , но могут быть зависимые друг с другом, то

$$\psi_{X_N}(t) = \sum_i \psi_{X_i}(t) \cdot P(N = i).$$

То есть производящая функция смеси, если мы разыгрываем внешний случайный инструмент, а потом выбираем из наших величин ту, у которой номер такой, как выдаст, — это сумма характеристических функций каждой величины с весами, равными вероятностям, это прямо из формулы полной вероятности следует, если расписать.

Пример 10.4. Пусть $X_1 \sim \mathcal{N}(a_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(a_2, \sigma_2^2)$, они независимы. Можем просто взять и посчитать характеристическую функцию, допустим, мы

это уже сделали:

$$\begin{aligned}\psi_{X_1}(t) &= e^{it a_1} e^{-\frac{t^2 \sigma_1^2}{2}}; \\ \psi_{X_2}(t) &= e^{it a_2} e^{-\frac{t^2 \sigma_2^2}{2}},\end{aligned}$$

значит, характеристическая функция суммы выглядит вот так:

$$\psi_{X_1+X_2}(t) = e^{it(a_1+a_2)} e^{-\frac{t^2(\sigma_1^2+\sigma_2^2)}{2}},$$

значит, сумма тоже будет нормальная: $X_1 + X_2 \sim \mathcal{N}(a_1 + a_2, \sigma_1^2 + \sigma_2^2)$.

Здесь мы, правда, невольно воспользовались теоремой единственности (если мы знаем характеристическую функцию, то распределение восстанавливается единственным образом). Если бы такой теоремы не было, то мы бы сказали, что это распределение с той же характеристической функцией, что и нормальное, но вдруг оно другое.

Единственность и непрерывность характеристической функции

Сформулируем теорему единственности:

Теорема 10.1 (единственности). Если $\psi_X(t) = \psi_Y(t) \forall t$, то $F_X(x) = F_Y(x) \forall x$, значит, и распределения тоже совпадают.

Здесь, правда, ломается такое свойство, что, у производящей функции, например, достаточно было, чтобы не на всём промежутке совпадали производящие функции, а только на каком-нибудь там подотрезке, у характеристических функций всё хуже, можно показать, что существуют две различные характеристические функции, полностью совпадающие на отрезке, например, $[-a, a]$, a может быть выбран любым, но за этими пределами одна ноль, а вторая периодически повторяется. Такое бывает, поэтому у характеристических функций, к сожалению, чтобы проверить их совпадение, нужно проверять на всей прямой. Тем не менее, факт такой есть, он для нас ключевой, тем самым можем исследовать характеристические функции, например, доказать, что характеристическая функция суммы имеет такой вот вид, и после этого уверенно сказать, что мы знаем распределение, у которого такая характеристическая функция, значит, это оно и есть.

Теорема 10.2 (непрерывности).

- 1) Пусть $F_{X_n}(x) \rightarrow F_X(x) \forall x$: $F_X(x-0) = F_X(x)$, $n \rightarrow \infty$. Тогда $\psi_{X_n}(t) \rightarrow \psi_X(t) \forall t$.
- 2) Пусть $\psi_{X_n}(t) \rightarrow g(t)$, где g непрерывна в нуле, тогда $F_{X_n}(x) \rightarrow F_X(x) \forall x$.

То есть сходимость функции распределения, правда, не совсем поточечная, а вот такая, слегка изменённая, почему она такая изменённая, мы обсудим на следующем семинаре, она равносильна сходимости характеристических функций, в частности, если X_n — дискретные, то это означает, что $P(X_n = k) \rightarrow P(X = k)$, и очень удобно, вместо того, чтобы проверять, что $P(X_n = k) \rightarrow P(X = k)$, можно проверять, что характеристические функции сходятся.

Формулы обращения

Мы умеем по распределению находить характеристическую функцию. Теперь нам нужен какой-то инструмент, чтобы из характеристической функции находить распределение, потому что мы не всегда знаем, какому распределению такая функция соответствует.

Есть два основных случая, когда мы можем это можно сделать, к сожалению, для произвольного распределения это довольно громоздко:

1) *Целочисленные случайные величины.* Для целочисленных случайных величин характеристическая функция имеет вот такой вот вид:

$$\psi_X(t) = \sum_{n=-\infty}^{\infty} P(X = n) e^{itn},$$

видим, что это 2π -периодичная функция. Поэтому если характеристическая функция оказалась 2π -периодичной, то можем найти распределение, разложив в ряд Фурье:

$$P(X = n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itn} \psi_X(t) dt.$$

Это формула обращения для дискретных распределений, в частности, у того же $\cos(t)$ мы могли не сами придумать, как разложить его, а просто взять интеграл, они были бы нули всегда, кроме n , равного 1 и -1 .

2) *Абсолютно-непрерывный случай* (величины с плотностью). У нас

$$\psi_X(t) = \int_{\mathbb{R}} e^{itx} f_X(x) dx. \quad (10.1)$$

Здесь тоже есть своя формула обращения — формула обратного преобразования Фурье, тогда

$$f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \psi_X(t) dt. \quad (10.2)$$

Заметим, что формулы (10.1) и (10.2) очень похожи, в частности, если нам повезло и характеристическая функция вдруг окажется, с точностью до константы, плотностью, то можно легко для такой плотности найти характеристическую функцию, и наоборот.

Отметим, что у целочисленных величин 2π -периодичная характеристическая функция, а у абсолютно-непрерывных она стремится к 0 на бесконечности, поэтому можно безошибочно отделить эти два случая друг от друга, правда, не от остальных, однако если у нас сингулярный случай, плотность из формулы обращения не получится.

Пример 10.5. Пусть $X: \psi_X(t) = \begin{cases} \frac{e^{it} - 1}{it}, & t \neq 0; \\ 1, & t = 0. \end{cases}$

Мы будем пользоваться формулой обращения (10.2), потому что наша характеристическая функция явно стремится к 0 на бесконечности, так что мы верим, что у нас есть плотность.

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \frac{e^{it} - 1}{it} dt = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{e^{-it(x-1)} - e^{-itx}}{it} dt = \\ &= \frac{1}{2\pi i} \int_{\mathbb{R}} \frac{\cos t(x-1) - \cos tx}{t} dt - \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\sin t(x-1) - \sin tx}{t} dt \end{aligned}$$

Подынтегральная функция в первом слагаемом нечётная, если доказать, что этот интеграл сходится, тогда получится, что он 0. Мы воспользуемся этим фактом без доказательства, этот интеграл сходится условно, показать это не очень просто, поскольку нельзя разбить его на два интеграла, но это возможно сделать с помощью признака Дирихле. Будем считать второй интеграл: его можно разбить на два, потому что интеграл от $\frac{\sin t}{t}$ сходится:

$$-\frac{1}{2\pi} \int_{\mathbb{R}} \frac{\sin t(x-1) - \sin tx}{t} dt = -\frac{1}{2\pi} \int_{\mathbb{R}} \frac{\sin t(x-1)}{t} dt + \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\sin tx}{t} dt$$

Посчитаем каждый такой интеграл в отдельности (помним, что $\int_{\mathbb{R}} \frac{\sin x}{x} dx = \pi$):

$$\int_{\mathbb{R}} \frac{\sin tu}{t} dt = \int_{\mathbb{R}} \frac{\sin tu}{tu} d(tu) = \begin{cases} \pi, & u > 0; \\ 0, & u = 0; \\ -\pi, & u < 0. \end{cases}$$

Посмотрим, чему же тогда равна плотность. Предположим, что $x > 1$, тогда $f_X(x) = 0$; предположим, что $x < 0$, тогда $f_X(x) = 0$. В этих случаях знак у них одинаковый, поэтому это просто одно и то же. Единственное, что бывает интересно, это когда $0 \leq x \leq 1$, причём в отдельных точках нам плотность не интересна, поэтому будем считать, что мы рассматриваем $0 < x < 1$, а что там в 0 и 1, нам, в общем-то, не очень интересно. Тут получается 1. Получили равномерную плотность:

$$f_X(x) = \begin{cases} 1, & x \in [0, 1]; \\ 0, & x \notin [0, 1]. \end{cases}$$

Ну и действительно получилась плотность, так что получается, что эта характеристическая функция была у равномерного распределения. Не очень приятно считать такую штуку, но, тем не менее, вполне себе доступно.

Семинар 11. Сходимости

Сходимость почти наверное

Определение 11.1. $X_n \rightarrow X$ почти наверное, если

$$P(\omega: X_n(\omega) \rightarrow X(\omega)) = 1, n \rightarrow \infty.$$

Фиксируем точку ω , смотрим, сходится ли X_n к X , потом смотрим, в каких точках сошлось, вот таких точек должно быть меры 1, то есть, почти все.

Пример 11.1. Пусть $X_n = \frac{1}{n}$, тогда $X_n \xrightarrow{\text{п.н.}} 0$. Действительно, фиксируем любую ω , при ней $X_n = \frac{1}{n} \rightarrow 0$, значит, всё хорошо.

Пример 11.2. Рассмотрим вероятностное пространство с $\Omega = [0, 1]$ и мерой Лебега и случайную величину $X_n = n \sin\left(\frac{\omega}{n}\right)$ на нём. При фиксированном ω $X_n \rightarrow \omega = X$. Фактически, мы зафиксировали ω , при фиксированном ω воспользовались стандартным приемом из математического анализа, перешли к пределу.

Это простой вид сходимости, но, тем не менее, он полезный по нескольким причинам. Одна из них — с ним очень удобно оперировать, потому что всегда можем зафиксировать ω и посмотреть, что произошло.

Лемма 11.1. Если $X_n \rightarrow X$ почти наверное, то $g(X_n) \rightarrow g(X)$ почти наверное для любой непрерывной функции g .

Доказательство. Есть множество $A = \{\omega: X_n(\omega) \rightarrow X(\omega)\}$, мы знаем, что $P(A) = 1$, тогда для любого $\omega \in A$ $g(X_n(\omega)) \rightarrow g(X(\omega))$, потому что это обычная числовая последовательность, она сходится, значит, непрерывная функция сохраняет эту сходимость. $P(\omega: g(X_n(\omega)) \rightarrow g(X(\omega))) \geq P(A) = 1$, ну а если мера хотя бы 1, то она ровно 1. ■

Аналогичным образом на этот случай переносятся все свойства, которыми обладают обычные числовые последовательности, это очень удобно. Есть даже специальные подходы, которые позволяют сбежать от какой-нибудь сходимости и привести её к сходимости почти наверное.

Сейчас сходимость выражена в терминах самих случайных величин, давайте выразим её в терминах распределений, что зачастую бывает удобно.

Сходимость $X_n(\omega) \rightarrow X(\omega)$ означает, что

$$\forall M \in \mathbb{N} \exists N \in \mathbb{N}: \forall n > N |X_n(\omega) - X(\omega)| < \frac{1}{M}.$$

Физический смысл квантора \forall — пересечение: если для любого x выполнено что-то, то соответствующее ω лежит в пересечении всех таких событий при всех x ; аналогично \exists — объединение. Используя это, событие $P(\omega: X_n(\omega) \rightarrow X(\omega)) = 1$ можно записать в виде

$$P\left(\bigcap_M \bigcup_N \bigcap_{n>N} \left\{|X_n(\omega) - X(\omega)| < \frac{1}{M}\right\}\right) = 1.$$

Вероятность счётного пересечения равна 1 только и только тогда, когда вероятность каждого из событий равна 1, поэтому выражение упрощается до

$$P\left(\bigcup_N \bigcap_{n>N} \left\{|X_n(\omega) - X(\omega)| < \frac{1}{M}\right\}\right) = 1.$$

Теперь у нас есть объединение расширяющихся событий, поскольку каждый раз мы пересекаем меньше событий. По свойству непрерывности меры это в точности предел вероятностей событий:

$$\lim_{N \rightarrow \infty} P\left(\sup_{n>N} |X_n - X| < \frac{1}{M}\right) = 1.$$

Это альтернативное определение сходимости почти наверное.

Если аналогичные рассуждения применить к критерию Коши, получим:

Теорема 11.1 (критерий Коши сходимости почти наверное). X_n сходится почти наверное тогда и только тогда, когда $\forall \varepsilon$:

$$P\left(\sup_{n,m>N} |X_n - X_m| < \varepsilon\right) \rightarrow 1, N \rightarrow \infty.$$

Сходимость по вероятности

Ослабим сходимость почти наверное, убрав супремум:

Определение 11.2. X_n сходится к X по вероятности (обозначение: $X_n \xrightarrow{P} X$), если $\forall \varepsilon > 0$:

$$P(|X_n - X| > \varepsilon) \rightarrow 0, n \rightarrow \infty.$$

Пример 11.3 (ступенька Рисса). Рассмотрим на пространстве $\Omega = [0, 1]$ с мерой Лебега такую последовательность: $X_1 = 1 \forall \omega$; $X_2 = 1$ на отрезке $[0, \frac{1}{2}]$ и 0 после; $X_3 = 1$ на отрезке $[\frac{1}{2}, 1]$ и 0 до (см. рис. 11.1); в следующий раз у нас будет 4 ступеньки ширины $\frac{1}{4}$ — от 0 до $\frac{1}{4}$, от $\frac{1}{4}$ до $\frac{1}{2}$, от $\frac{1}{2}$ до $\frac{3}{4}$ и от $\frac{3}{4}$ до 1; потом 8 ступенек ширины $\frac{1}{8}$, и так далее. $X_n \xrightarrow{P} 0$, потому что вероятность того, что X_i отличается от 0 — это ширина ступеньки, которая

с ростом n стремится к 0. Но при каждом ω на каждом пробеге ступеньки по отрезку, один раз будет 1, а остальные разы 0, значит, последовательность X_n содержит бесконечно много 1, значит, $X_n(\omega) \not\rightarrow 0 \forall \omega$, поэтому мера ω , при которых $X_n \rightarrow 0$, равна 0, так что последовательность не сходится почти наверное, но сходится при этом по вероятности.

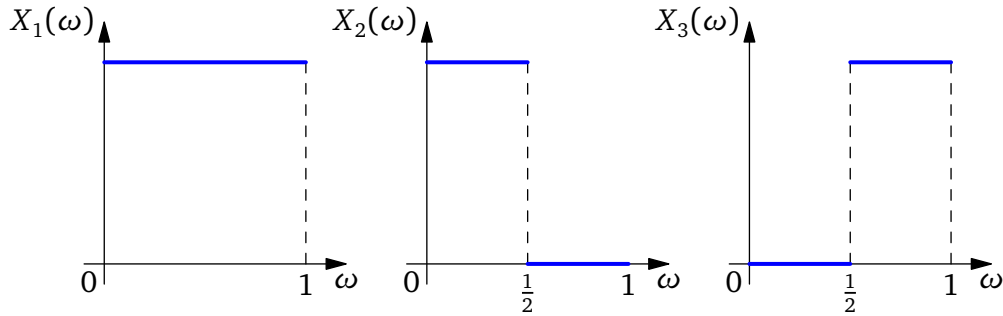


Рис. 11.1: Плавающие ступеньки Рисса

В каком-то смысле, все последовательности, сходящиеся по вероятности, но не сходящиеся почти наверное, так и устроены: место, где мы не сходимся к величине, хоть и маленькое, но всё время в разных местах. Отметим, что если сходимость вероятностей к 0 такая, что ряд из вероятностей сходится, то тогда из сходимости по вероятности следует сходимость почти наверное.

Теорема 11.2 (критерий Коши сходимости по вероятности). X_n сходится по вероятности тогда и только тогда, когда $\forall \delta > 0 \exists N \forall n, m > N$:

$$P(|X_n - X_m| > \varepsilon) < \delta.$$

Лемма 11.2. Если $X_n \xrightarrow{P} X$, $Y_n \xrightarrow{P} Y$, g — непрерывная из \mathbb{R}^2 в \mathbb{R} , то $g(X_n, Y_n) \rightarrow g(X, Y)$.

Мы рассмотрели две величины для удобства, их можно брать любое количество. Мы оставим эту лемму без доказательства.

Сходимость в среднем порядка p

Определение 11.3. X_n сходится к X в L^p , $p \geq 1$ (обозначение: $X_n \xrightarrow{L^p} X$), если $E|X_n|^p < +\infty$, $E|X_n - X|^p \rightarrow 0$.

Из сходимости почти наверное не следует сходимость в L^p , смотрите пример 8.1. Почти наверное такая последовательность сходится к 0, потому что она с какого-то момента 0 в каждой точке, а $EX^p = 1$ к 0 у неё не сходится.

Из сходимости в L^p также не следует сходимость почти наверное, смотрите пример 11.3. Ступенька Рисса в L^p сходится, потому что математическое ожидание у неё точно такое же, как вероятность.

Сформулируем три полезных неравенства:

1) (неравенство Маркова). Если $X \geq 0$, $EX < +\infty$, то $P(X \geq x) \leq \frac{EX}{x}$.

ДОКАЗАТЕЛЬСТВО. $EX = EXI_{X \geq x} + EXI_{X < x}$. Второе слагаемое неотрицательное, потому что $X \geq 0$, а первое устроено так, что $X \geq x$, поэтому можем оценить $EX \geq xEI_{X \geq x} + 0$ — это в точности то неравенство, которое мы хотим доказать. ■

Это очень простое, но довольно удобное неравенство, потому что с помощью математических ожиданий можно оценивать вероятности.

2) (неравенство Чебышёва). Если $DX < +\infty$, то $P(|X - EX| > \varepsilon) \leq \frac{DX}{\varepsilon^2}$.

Под вероятностью можно возвести в квадрат обе части, они неотрицательные, неравенство не поменяется, а после этого применить неравенство Маркова, обозначив $|X - EX|^2$ за Y . Аналогично доказывается и

3) $P(|X| > \varepsilon) \leq \frac{E|X|^p}{\varepsilon^p}$.

Получается, что из сходимости в L^p следует сходимость по вероятности.

Теорема 11.3 (критерий Коши сходимости в L^p). X_n сходится в L^p тогда и только тогда, когда $\forall \varepsilon > 0 \exists N \forall n, m > N$:

$$E|X_n - X_m|^p < \varepsilon.$$

Лемма 11.3. Если $X_n \xrightarrow{L^p} X$, $Y_n \xrightarrow{L^p} Y$, то $X_n + Y_n \xrightarrow{L^p} X + Y$.

Это следует из неравенства Минковского.

Сходимость по распределению

Определение 11.4. X_n сходится к X по распределению или слабо (обозначение: $X_n \xrightarrow{d} X$), если

$$F_{X_n}(x) \rightarrow F_X(x) \quad \forall x: F_X(x-0) = F_X(x).$$

Это сходимость другого рода, она не говорит про сходимость при каждом ω , она говорит про то, что функции распределения похожи, иногда это удобно.

Пример 11.4. Возьмём $X_n = \frac{1}{n\omega}$. Ясно, что $X_n \rightarrow 0$ почти наверное. А вот с

функцией распределения у неё не совсем всё хорошо:

$$F_{X_n}(x) = \begin{cases} 0, & x < \frac{1}{n}; \\ 1 - \frac{1}{nx}, & x \geq \frac{1}{n}, \end{cases}$$

поскольку $X_n \leq x \iff \omega \geq \frac{1}{nx}$. Если зафиксируем положительный x , а n устремим к ∞ , это сойдётся к

$$F_{X_n}(x) \rightarrow \begin{cases} 1, & x > 0; \\ 0, & x = 0; \\ 0, & x < 0. \end{cases}$$

И вот здесь поточечная сходимость сломалась: в $x = 0$ сходимости к функции распределения $X = 0$ нет, именно потому, что 0 — точка разрыва нашей функции распределения. Именно для борьбы с этим мы добавили в определение условие $\forall x: F_X(x - 0) = F_X(x)$.

Лемма 11.4. X_n сходится к X по распределению тогда и только тогда, когда $Eg(X_n) \rightarrow Eg(X)$ для любой непрерывной ограниченной функции $g \in CB$.

По теореме 10.2 это также равносильно $\psi_{X_n}(t) \rightarrow \psi_X(t) \forall t$.

Обратите внимание, что сами EX_n к EX сходятся не обязаны, поскольку функция $y = x$ неограниченная. Поэтому из сходимости по распределению сходимость в L^p не следует никоим образом.

Лемма 11.5. Если $X_n \xrightarrow{d} X$, то $f(X_n) \xrightarrow{d} f(X)$ для любой непрерывной f .

Это легко вывести из определения из леммы 11.4: функция $g(f(x))$ непрерывно ограничена, поскольку g непрерывно ограничена, а f непрерывна, поэтому условие на математические ожидания будет выполнено и для $f(X_n)$.

Однако, если $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$, то вовсе не обязательно, что $X_n + Y_n \xrightarrow{d} X + Y$. Проблема в том, что предельная величина не единственна, а все величины с распределением X и Y в сумме дадут слишком много всего.

Пример 11.5. Покажем, что если $X_n \xrightarrow{P} X$, а $Y_n \xrightarrow{P} 0$, то $X_n Y_n \xrightarrow{P} 0$. Рассмотрим такую вероятность, мы должны показать, что она стремится к 0:

$$P(|X_n Y_n| > \varepsilon) = P(|X_n Y_n| > \varepsilon, |Y_n| > \delta) + P(|X_n Y_n| > \varepsilon, |Y_n| \leq \delta).$$

Прямо из сходимости по распределению следует, что первое слагаемое стре-

мится к нулю. Посмотрим на второе:

$$\begin{aligned} P(|X_n Y_n| > \varepsilon, |Y_n| \leq \delta) &\leq P\left(|X_n| > \frac{\varepsilon}{\delta}\right) = \\ &= P\left(|X_n| > \frac{\varepsilon}{\delta}, |X_n - X| > \gamma\right) + P\left(|X_n| > \frac{\varepsilon}{\delta}, |X_n - X| \leq \gamma\right). \end{aligned}$$

Первое слагаемое стремится к 0 с ростом n при каждом фиксированном γ по определению сходимости по вероятности. Упростим второе слагаемое:

$$P\left(|X_n| > \frac{\varepsilon}{\delta}, |X_n - X| \leq \gamma\right) \leq P\left(|X| > \frac{\varepsilon}{\delta} - \gamma\right).$$

По сути, это неравенство треугольника на отрезке. Это выражение можно сделать меньше любого наперёд заданного числа ε_1 , если δ взять сильно меньше ε . Поэтому при достаточно больших n

$$P(|X_n Y_n| > \varepsilon) < 3\varepsilon_1,$$

а значит, она стремится к 0, поскольку ε_1 произвольно.

Общий принцип простой: разрезаем на куски, пользуемся предыдущими утверждениями, их немного, поэтому сводить к ним довольно просто: нужно по очереди отбрасывать куски, где кто-то от кого-то сильно отклоняется.

Лемма 11.6. Если величина сходится по распределению к константе, то она и по вероятности к ней тоже сходится.

Это несложно доказывается аналитически.

Обобщим все утверждения и построим карту сходимостей — см. рис. 11.2.

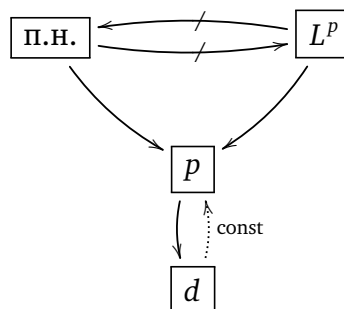


Рис. 11.2: Карта сходимостей

Семинар 12. Законы больших чисел и предельные теоремы

Схема Бернулли

Пример 12.1. Пусть вероятность того, что человек левша, равна 0.01. Мы умеем считать явно, например, что вероятность, что среди 100 человек не больше 2 левшей: $P = 0.99^{100} + 0.99^{99} \cdot 0.01 \cdot 100 + C_{100}^2 \cdot 0.99^{98} \cdot 0.01^2$.

Глядя на это выражение, сложно сказать, чему примерно оно равно. Даже если мы его упростим, мы не имеем представления о порядке величины 0.99^{98} . Когда индексы у биномиальных коэффициентов большие, с ними не очень удобно работать, поэтому на помощь приходит следующая теорема:

Теорема 12.1 (Пуассона). Пусть $X_{n,i}$, $i \leq n$ — независимые $Bern(p_{n,i})$. При этом мы считаем, что $\max_{i \leq n} p_{n,i} \rightarrow 0$, а $\sum_{i=1}^n p_{n,i} \rightarrow \lambda$. Тогда

$$P\left(S_n = \sum_{i=1}^n X_{n,i} = k\right) \rightarrow \frac{\lambda^k e^{-\lambda}}{k!}.$$

Ранее мы говорили, что отсюда распределение Пуассона и взялось: оно хорошо аппроксимирует схему Бернулли, в которой вероятность каждого успеха маленькая, а сумма при этом всё-таки большая, за счёт того, что количество большое. У этой теоремы есть большой плюс: в ней не требуется одинаковая распределённость. Доказывается эта теорема через сходимость производящих функций.

Пример 12.2. Применим теорему к примеру 12.1, получим, что

$$P(S_n \leq 2) \approx P(Z \leq 2) \textcircled{=}$$

100, конечно, к ∞ не стремится, но мы будем считать, что оно достаточно большое; $\lambda = \sum p_{n,i} = 100 \cdot 0.01 = 1$, то есть наше выражение равно:

$$\textcircled{=} \left(1 + \frac{1}{1!} + \frac{1}{2!}\right) e^{-1} = \frac{5}{2e} \approx \frac{10}{11}.$$

Здесь гораздо проще представить, насколько это число близко к 1.

Помочь оценить погрешность того, что мы считаем 100 достаточно большим, нам может следующая теорема:

Теорема 12.2 (Ле Кама). $|P(S_n \in A) - P(Z \in A)| \leq 2 \sum_{i=1}^n p_{n,i}^2 \quad \forall A, Z \sim Poiss\left(\sum_{i=1}^n p_{n,i}\right)$.

Так что погрешность в этой теореме можно явно выписать. Есть очень красивое доказательство данного факта, но оно не входит в данный курс.

Пример 12.3. В примере 12.2 погрешность получается $\sum p_{n,i}^2 = 0.02$, то есть 2%. В целом, такая погрешность сойдёт, мы $\frac{5}{2e}$ в уме лучше и не оценим, чем на 2%.

Мы смогли аппроксимировать вероятность и посчитать погрешность, при этом дополнительный плюс в том, что это работает даже если $p_{n,i}$ разные, когда мы явное распределение выписать не можем совсем, потому что сумму разнораспределённых индикаторов красиво не посчитать.

Если у нас события имеют высокую вероятность (например, обычно считается, вероятность, что человек мальчик 55%), то тогда получится, что наша формула будет не очень, потому что $\sum p_{n,i}^2$ не будет маленькой. Тут пуассоновская аппроксимация не годится, она годится только когда np_i^2 мало, иначе нам нужна другая теорема.

Будем считать, что $X_i \sim \text{Bern}(p)$ — одинаково распределённые, хотя это тоже не обязательно делать. $S_n = X_1 + \dots + X_n$. Можно явно выписать

$$P(S_n = k) = C_n^k p^k (1-p)^{n-k}.$$

Пример 12.4. Пусть у нас есть $n = 1000$ детей, а вероятность, что ребёнок является мальчиком равна $p = 0.55$. Вероятность, что мальчиков будет половина не очень приятно выписывается:

$$P(S_n = 500) = C_{1000}^{500} \cdot 0.55^{500} \cdot 0.45^{500} \ominus$$

В голове трудно представить, чему равно это число. Но опять же, тут есть удобная теорема, которая называется *локальной теоремой Муавра–Лапласа*. В одной из её формулировок она утверждает, что это выражение равно

$$\ominus (1 + o(1)) \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}},$$

где $o(1)$ равномерно мало по k : $|k - np| < r_n$, где $r_n = o(n^{2/3})$. То есть формула работает не везде, а только в окрестности точки np ширины примерно $n^{2/3}$. Теорема выводится из формулы Стирлинга: нужно расписать что такое C_n^k и p^k , разложить всё, что можно, и привести к такому виду. Нам она особо не понадобится, нам скорее понадобится то, что из неё получится, если просуммировать:

Теорема 12.3 (интегральная Муавра–Лапласа).

$$P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x\right) \rightarrow \Phi(x), \quad n \rightarrow \infty \quad \forall x,$$

где $\Phi(x)$ — функция распределения $Z \sim \mathcal{N}(0, 1)$.

В наших высоких терминах это означает, что $\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1)$.

Для нас важна сама теорема и с помощью неё мы можем посчитать, например, вероятность того, что мальчиков не больше, чем половина:

Пример 12.5. В примере 12.4 посчитаем вероятность, что из 1000 человек мальчиков не больше, чем девочек:

$$P(S_n \leq 500) = P\left(\frac{S_n - 550}{\sqrt{1000 \cdot 0.55 \cdot 0.45}} \leq \frac{500 - 550}{\sqrt{1000 \cdot 0.55 \cdot 0.45}}\right) \approx \Phi(-3).$$

Мы снова не можем устремить 1000 в бесконечность, но мы верим, что в этот момент вероятность уже примерно равна тому, что нам нужно. Проблема в том, что Φ в элементарных функциях не выражается, это

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt,$$

но, с другой стороны, она во всех задачах одна и та же, можно её один раз посчитать, составить табличку или запомнить, и дальше всегда использовать.

Из симметричности нормального распределения $\Phi(x) = 1 - \Phi(-x)$. Стандартные точки: $\Phi(2) \approx 0.975$, $\Phi(3) \approx 0.9985$, поэтому $P(S_n \leq 500) \approx 0.0015$.

Мы не очень заботимся о точности, потому что погрешность нашего приближения всё равно больше, чем это число. Соответственно, вероятность очень маленькая. Это частично контринтуитивно, потому что 550 в среднем должно быть мальчиков, ну могло бы уж на 50 поколебаться вправо-влево, но оказывается, что нет.

Поскольку $\Phi(-3)$ очень маленькое, а $\Phi(3)$ — очень большое, чаще всего, величина Z лежит от -3 до 3 , поэтому можно говорить, что S_n лежит от $np - 3\sqrt{np(1-p)}$ до $np + 3\sqrt{np(1-p)}$. Это *правило трёх сигм*, оно помогает быстро прикинуть, сколько может выпасть успехов в наборе испытаний.

Пример 12.6. В 100 подбрасываниях симметричной монетки практически наверняка будет от 35 до 65 орлов, хотя, казалось бы, всего-то 100 бросаний, должно флуктуировать довольно сильно, болтаться вправо-влево, но вот нет, оно довольно регулярно за счёт того, что порядок в знаменателе — корень, а в числителе порядок всё-таки n , это довольно плотно прижимается к среднему значению.

Результат, позволяющий оценить погрешность в этой теореме, также существует, мы о нём чуть позже поговорим в общем виде.

Теорема 12.4 (неравенство Берри–Эссеена).

$$|\mathbb{P}(S_n \leq x) - \Phi(x)| \leq \frac{0.5(p^2 + (1-p)^2)}{\sqrt{np(1-p)}}.$$

Это грубая оценка, но тем не менее, довольно удачная.

Пример 12.7. В примере 12.5 погрешность получается $\approx \frac{0.5}{\sqrt{1000}} \approx 0.01$. Поэтому такую точность высокую мы не сможем использовать, как дала нам теорема, но, по крайней мере, можем сказать, что вероятность достаточно маленькая, вероятность меньше 1% с небольшим, а дальше мы сказать не можем по нашей теореме.

Законы больших чисел (ЗБЧ)

Определение 12.1. X_n удовлетворяет закону больших чисел (ЗБЧ), если

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{P} C, \quad n \rightarrow \infty,$$

где C — некоторая константа.

Определение 12.2. X_n удовлетворяет усиленному закону больших чисел (УЗБЧ), если

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{\text{п.н.}} C, \quad n \rightarrow \infty,$$

где C — некоторая константа.

Поскольку сходимость по распределению и по вероятности к константе — это одно и то же, гипотетический ослабленный закон больших чисел будет совпадать с обычным законом больших чисел.

Теперь будем давать достаточные условия для закона больших чисел:

Теорема 12.5 (ЗБЧ Чебышёва). Пусть X_i — некоррелированные, то есть $\text{cov}(X_i, X_j) = 0$, $i \neq j$; $\sup DX_i < +\infty$. Тогда $X_n - EX_n$ удовлетворяет ЗБЧ с $C = 0$, то есть

$$\frac{X_1 + \dots + X_n - EX_1 - \dots - EX_n}{n} \xrightarrow{P} 0.$$

Достаточное условие, например, если величины независимые одинаково распределённые, тогда они некоррелированные из независимости, и из одинаковой распределённости, если у них дисперсия конечная, то получится сразу, что условие $\sup DX_i < \infty$ выполнено; EX_i при этом все одинаковые, получится, что просто сама последовательность X_i удовлетворяет ЗБЧ с $C = EX$.

В схеме Бернулли этот факт означает, что если сложить бернуллиевские величины и разделить на n , то это сходится к вероятности одной такой бернуллиевской величины. Это факт, который вы с детства знаете: если монетку много раз бросать, то доля орлов будет близка к вероятности орла. Заодно мы с вами только что оправдали эмпирическое определение вероятности: интуитивно мы понимаем, что вероятность — это доля раз, которое событие случится за n испытаний этого события, вот теперь нам закон больших чисел это оправдывает: если X_i — независимые одинаково распределённые бернуллиевские величины, то $\frac{X_1 + \dots + X_n}{n} \xrightarrow{p} p$, где p — вероятность успеха в этой бернуллиевской величине. Это и означает, что частота в схеме Бернулли сходится к вероятности, что мы изначально и хотели представлять себе определением вероятности. В итоге мы определением вероятности взяли другое, но оно сошлось, уже по теореме, по закону больших чисел, сошлось с нашим интуитивным пониманием.

Теорема очень простая, доказывается просто через неравенство Чебышёва в две строчки, но у него есть условие дисперсии, которое на самом деле не обязательно, поэтому сформулируем другой закон больших чисел:

Теорема 12.6 (ЗБЧ Хинчина). Пусть X_i — независимые одинаково распределённые, $EX_i = \mu$. Тогда

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mu,$$

то есть X_i удовлетворяют закону больших чисел с $C = \mu$.

Эта теорема похитрее, доказывается она через характеристические функции, как раз за счёт того, что можно требовать сходимость по распределению вместо сходимости по вероятности.

В некотором смысле, эта теорема морально устарела, потому что есть следующая теорема:

Теорема 12.7 (УЗБЧ Колмогорова). Пусть X_i — независимые одинаково распределённые, $EX_i = \mu$, тогда X_i удовлетворяют усиленному закону больших чисел с $C = \mu$.

На самом деле, даже закон больших чисел Колмогорова достаточно слабый, потому что можно сказать, что даже не важна независимость, и хватит попарной независимости. Это называется *усиленным законом больших чисел Этемади*. Таких законов больших чисел сотни и тысячи, люди придумывают разные достаточные условия, и они оказываются довольно широкими. Вот такие теоремы дают нам некоторые утверждения о сходимости. На самом деле, если посмотреть на усиленный закон больших чисел, то это совершенно удивительный результат, надо на него с правильной стороны посмотреть, чтобы поразиться какой он странный.

Пример 12.8. Представьте, что у вас есть $R \sim R[0, 1]$, запишем её в двоичной системе: $0.r_1r_2\dots$. Получается, что $r_i \sim \text{Bern}\left(\frac{1}{2}\right)$. Если мы возьмём случайное число на отрезке $[0, 1]$, то его двоичные цифры независимые и равновероятные, 0 или 1, из симметрии. В этом легко убедиться, выписав конкретные вероятности. Поэтому

$$P\left(\frac{r_1 + \dots + r_n}{n} \rightarrow \frac{1}{2}\right) = 1.$$

Казалось бы, если мы ткнём наугад в число на отрезке $[0, 1]$, его цифры должны прыгать как попало, а как бы не так, у него для всех чисел, кроме множества меры 0 последовательность цифр будет по Чезаро сходиться к $\frac{1}{2}$. Из соображений справедливости, кажется, числа должны чаще всего расходиться, чем сходиться куда-то, мы же ткнули в число наугад, но это не так, почти все числа устроены так, что у них примерно половина цифр всё время поддерживается 0, а половина — 1, а флуктуации на порядок меньше, чем n от этого числа.

Центральные предельные теоремы (ЦПТ)

Центральных предельных теорем также существует сотни разных версий, мы остановимся только на одной:

Теорема 12.8 (*центральная предельная теорема (ЦПТ)*). Пусть X_i — независимые одинаково распределённые, $EX_i = \mu$, $DX_i = \sigma^2 \in (0, \infty)$. Тогда

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1).$$

Одной из теорем такого рода была теорема 12.3. Там у нас X_i были бернуллиевские, μ было равно p , а $\sigma^2 = p(1-p)$. Ну а раз у нас сходимость по распределению есть, это означает, что функция распределения сходится:

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \rightarrow \Phi(x),$$

поскольку предельная функция непрерывна везде.

Это очень ценный факт, он гораздо ценнее закона больших чисел, потому что он позволяет приближать вероятности для левой величины: вместо вероятности, что левая величина попала от a до b , мы можем использовать, что правая, при этом у левой распределение мы не знаем, нужно свёртки считать, мы это делаем плохо, а сумму равномерных, например, мы посчитать не сможем вовсе, а нормальная величина ведёт себя прилично, у неё какая-то известная функция распределения, она не элементарная, но она одна и та же для всех распределений.

Это на самом деле удивительный факт, объяснить его на пальцах сложно, почему все распределения с конечной дисперсией решили так дружно к нормальному стягиваться, ну вот так устроен этот мир.

Вот такой красивый результат, из него вытекает много разных других красивых результатов, для себя этот факт надо зафиксировать, потому что, когда вы дойдёте до статистики, это станет очень важными фактами. Заодно мы с вами открыли, почему математическое ожидание и дисперсия так важны для нас, потому что есть утверждения такого рода, когда по сути мы можем описать вероятность, как сумма большого количества слагаемых себя ведёт, зная только μ и σ .

Есть даже такие результаты, которые называются *сильной аппроксимацией*, приведём их для эстетического чувства:

Можно подобрать такую последовательность нормальных, что последовательности будут близки:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx \frac{U_1 + \dots + U_n - n\mu}{\sigma\sqrt{n}},$$

в том смысле, что супремум модуля их разности стремится к 0 при разных условиях. Если пространство взять достаточно богатым, то можно прямо явно предъявить нормальные величины, которые, когда складываешь, получается то же самое, что и наши. И получается такой интересный закон *обезличивания распределения*: с течением времени нам уже не важно, кого мы складываем, не важно, как устроена его специфика, как устроены его вероятности и функция распределения. У нас только μ и σ , остальное всё не важно.



МЕХАНИКО-
МАТЕМАТИЧЕСКИЙ
ФАКУЛЬТЕТ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ