

КАЗАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Кафедра математической статистики

И.Н. Володин

ЛЕКЦИИ
ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ
И МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Допущено учебно-методическим советом по прикладной математике и информатике для студентов высших учебных заведений, обучающихся по специальности 010200 “Прикладная математика и информатика” и по направлению 510200 “Прикладная математика и информатика”.

Казань – 2006

УДК 336.71
ББК 00.000.00

Володин И.Н.

Лекции по теории вероятностей и математической статистике. - Казань: (Издательство), 2006. - 271с.

ISBN 0-00000-000-0

Учебник представляет стенографическую запись лекций, читаемых профессором И.Н. Володиным на факультете вычислительной математики и кибернетики Казанского университета в рамках специальности "Прикладная математика". Основное внимание уделяется математическим методам построения вероятностных моделей и статистическим выводам в рамках данных моделей. Предназначается для студентов и аспирантов, специализирующихся в области прикладной математики.

УДК 336.71
ББК 00.000.00

ISBN 0-00000-000-0

©И.Н.Володин

О Г Л А В Л Е Н И Е

Предисловие	3
<i>часть первая</i>	
теория вероятностей	
§1 Элементарная теория вероятностей	8
§2 Вероятностное пространство	21
§3 Условная вероятность и независимость событий	29
§4 Случайные величины и функции распределения	41
§5 Построение вероятностных моделей с помощью функций распределения	49
§6 Характеристики распределения случайной величины. Классификация распределений	58
§7 Предельные теоремы в схеме испытаний Бернулли. Нормальное распределение	76
§8 Векторные случайные величины. Независимость случайных величин	85
§9 Моментные характеристики многомерных распределений. Мультиномиальное и многомерное нормальное распределения	96
§10 Условное распределение вероятностей. Условное математическое ожидание	106
§11 Сходимость случайных величин и функций распределений	113
§12 Характеристические функции. Теоремы единственности и сложения	120
§13 Характеристические функции. Критерий слабой сходимости	132
§14 Предельные теоремы теории вероятностей	142

§15	Случайные процессы	150
-----	--------------------	-----

часть вторая
м а т е м а т и ч е с к а я с т а т и с т и к а

§1	Проблема статистического вывода	171
§2	Выборочные характеристики. Достаточные статистики	182
§3	Оценка параметров. Метод моментов	196
§4	Оценка параметров. Метод максимального правдоподобия	204
§5	Эффективность оценок	219
§6	Доверительные интервалы	226
§7	Статистическая проверка гипотез (критерии значимости)	239
§8	Равномерно наиболее мощные критерии	253
§9	Проверка модельных предположений. Критерии согласия	261

Предисловие

Данное учебное пособие представляет собой почти стенографическую запись лекций, читаемых мной на факультете вычислительной математики и кибернетики Казанского университета в течение двух семестров. Студенты получают специальность “Прикладная математика”, и это обстоятельство накладывает определенный отпечаток как на содержание курса, так и на форму его изложения. В части теории вероятностей основной упор делается на математические методы построения вероятных моделей и реализацию этих методов на реальных задачах естествознания и практической деятельности. Каждое семейство распределений, будь то пуассоновское, показательное, нормальное, гамма и т.д., вводится через рассмотрение некоторых реальных объектов, доставляющих систему математических постулатов, из которых путем аналитических выкладок определяются распределения числовых характеристик этих объектов. Математический аппарат теории вероятностей излагается только в том объеме, который позволяет корректно вводить новые вероятностные модели. Такой подход обеспечивает неформальное отношение к использованию методов математической статистики - осознанию того, что без построения вероятностной модели не представляется возможным судить о точности и надежности статистического вывода. Именно поэтому в разделе “Математическая статистика” основное внимание уделяется методам вычисления риска конкретных статистических правил и проблемам статистических решений с минимальным риском.

Вводя понятия условного математического ожидания и условного распределения вероятностей, я ограничился рассмотрением только случаев дискретного и непрерывного распределений, поскольку студентам факультета ВМК читается, как правило, весьма ограниченный курс теории меры и интеграла Лебега, – их редко знакомят с теоремой Радона–Никодима. Чтобы восполнить этот пробел, я привожу формулировку этой теоремы и определяю функцию плотности через производную Радона–Никодима, но, на мой взгляд, было бы крайне наивным полагать, что большинство моих слушателей воспримут строгое современное изложение концепции условного математического ожидания.

Учебное пособие содержит 25 лекций по теории вероятностей и 15 лекций по математической статистике. Каждая лекция слишком объемна для того, чтобы ее диктовать студентам; лекции рассчитаны на свободное изло-

жение материала с записью только определений и формулировок основных теорем. “Диктант” становится возможным, если вы располагаете дополнительно 7 лекциями.

Я приношу искреннюю благодарность профессору Дмитрию Михайловичу Чибисову и доценту нашей кафедры Сергею Владимировичу Симушкину. Их профессиональные замечания позволили устранить ряд огрехов и неточностей в доказательствах теорем и формулировках сложных понятий теории вероятностей и математической статистики; оформление лекций в электронной форме принадлежит С.В. Симушкину.

Л И Т Е Р А Т У Р А

1. Чистяков В.П. Курс теории вероятностей.– М.: Наука, 1982.
2. Ширяев А.Н. Вероятность.– М.: Наука, 1980.
3. Боровков А.А. Теория вероятностей.– М.: Наука, 1986.
4. Крамер Г. Математические методы статистики.– М.: Мир, 1975.

ЧАСТЬ ПЕРВАЯ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

“Пускай в данном случае вы не согласитесь мне дать гарантию, – но я ставлю вопрос шире: существует ли вообще, может ли существовать в этом мире хоть какое-нибудь обеспечение, хоть в чем-нибудь порука, – или даже сама идея гарантии неизвестна тут?”

В. Набоков. Приглашение на казнь

§1. Элементарная теория вероятностей

Лекция 1

Во многих областях человеческой деятельности существуют ситуации, когда определенные явления могут повторяться неограниченное число раз в одинаковых условиях. Анализируя последовательно результаты таких простейших явлений, как подбрасывание монеты, игральной кости, выброс карты из колоды и т.п., мы замечаем две особенности, присущие такого рода экспериментам. Во-первых, не представляется возможным предсказать исход последующего эксперимента по результатам предыдущих, как бы ни было велико число проведенных испытаний. Во-вторых, относительная частота определенных исходов по мере роста числа испытаний стабилизируется, приближаясь к определенному пределу. Следующая таблица служит подтверждением этого замечательного факта, составляющего основу аксиоматического построения теории вероятностей как математической дисциплины.

$N \setminus n$	10^2	10^4	10^6
1	41	4985	499558
2	48	5004	499952
3	44	5085	500114
4	52	4946	500064
5	58	4978	500183
6	52	4985	499533
7	45	5012	500065
8	50	4931	500317
9	52	5016	500449
10	45	4973	500704
Er	10^{-1}	10^{-2}	10^{-3}

Первый столбец этой таблицы указывает номер эксперимента; последующие столбцы содержат данные о количествах m выпадения герба в n ($= 10^2, 10^4, 10^6$) подбрасываниях (испытаниях) правильной симметричной монеты. Таким образом, проводилось три серии экспериментов с разным числом испытаний в каждой серии. Каждая серия состоит из десяти экспериментов с одним и тем же числом n подбрасываний монеты, что позволяет судить об изменчивости числа m выпадений герба от эксперимента к эксперименту внутри одной серии. Очевидна стабилизация

относительной частоты $p_n = m/n$ выпадений герба с ростом числа испытаний n , а также стремление p_n к величине $p = 1/2$. Можно даже высказать некоторое суждение об изменчивости этой частоты от эксперимента к эксперименту при фиксированном n : отклонение p_n от центра рассеивания, равного $1/2$, имеет порядок $n^{-1/2}$ (см. в связи с этим нижнюю строку таблицы).

Обнаруженные закономерности, распространенные на испытания с произвольным числом исходов, позволяют построить простейшую математическую модель *случайного эксперимента*.

Построение начинается с описания множества Ω всевозможных исходов ω , которые могут произойти в результате каждого испытания. Множество Ω называется *пространством элементарных исходов*, его точки (элементы) ω – *элементарными исходами* или *элементарными событиями*. Любое подмножество A пространства Ω (совокупность элементарных исходов ω) называется *событием*; пространство Ω также является событием, но имеющим особое название *достоверного события*. Говорят, что *произошло событие* A , если в испытании наблюдается элементарный исход $\omega \in A$.

В этом параграфе, посвященном так называемой *элементарной теории вероятностей*, будут рассматриваться только пространства Ω , состоящие из не более чем счетного числа элементов. Проиллюстрируем введенные понятия на ряде простейших примеров, относящихся к случайным испытаниям.

Пример 1.1. Подбрасывается правильная монета и регистрируется сторона (герб или решка) монеты, которая обращена к наблюдателю после ее падения. Пространство Ω состоит из двух точек: $\omega_1 = Г$ (выпал герб) и $\omega_2 = Р$ (выпала решка). Любое событие A в этом примере является либо элементарным, либо достоверным.

Пример 1.2. Правильная монета подбрасывается два раза или, что одно и то же, подбрасываются две монеты. Пространство Ω содержит четыре точки: ГГ, ГР, РГ, РР. Событие $A = \{ГР, РГ\}$ означает, что монеты выпали на разные стороны, и, очевидно, не является элементарным событием. Интересно, что на раннем этапе становления теории вероятностей событие A трактовалось как элементарное (то есть полагалось $\Omega = \{ГГ, A, РР\}$), и это приводило к вероятностной модели результатов испытаний двух правильных монет, которая противоречила наблюдаемой частоте элементарных исходов.

Пример 1.3. Бросается игральная кость и регистрируется число выпавших очков (номер грани игральной кости). Пространство элементарных исходов состоит из шести элементов $\omega_i = i$, $i = 1, \dots, 6$. Пример составного события: $A = \{2, 4, 6\}$ – выпало четное число очков.

Пример 1.4. Бросаются две игральные кости. Пространство элементарных исходов можно представить в виде матрицы $\Omega = \|(i, j)\|$, $i, j =$

1, ..., 6. Пример составного события: сумма очков больше 10; появление этого события возможно лишь при элементарных исходах (5,6), (6,5), (6,6).

Пример 1.5. Подбрасываются n монет. Пространство Ω содержит 2^n элементов; любой элементарный исход ω имеет вид “слова”, состоящего из букв Г и Р, например, РГГРР ... ГРГ. Пример составного события, состоящего из C_n^k элементарных исходов, – “выпало k гербов”.

Пример 1.6. Монета подбрасывается до первого появления герба. Пространство Ω состоит из счетного числа элементов вида $\omega_i = \text{Р} \dots \text{РГ}$, в которых начальные Р повторяются $i - 1$ раз, $i = 1, 2, \dots$. Пример составного события, осуществление которого сопряжено с появлением одного из четырех элементарных исходов, – “герб появился до пятого подбрасывания монеты”.

Пример 1.7. Наблюдатель фиксирует число метеоров, появившихся в заданном секторе небесного свода в течение фиксированного промежутка времени. Поскольку не представляется возможным ограничить сверху число возможных появлений метеоров, то естественно отождествить Ω , с множеством всех неотрицательных целых чисел $\{0, 1, 2 \dots\}$, то есть положить $\omega_k = k$. Пример составного события: $A = \{1, 2 \dots\}$ – “наблюдался по крайней мере один метеор”.

Если ограничиться рассмотрением пространств элементарных исходов, состоящих из не более чем счетного числа элементов, то построение вероятностной модели по существу состоит в задании *распределения вероятностей* на пространстве Ω , в соответствии с которым каждому элементарному исходу $\omega \in \Omega$ ставится в соответствие число $p(\omega)$, называемое *вероятностью* элементарного события ω . Постулируется, что $0 \leq p(\omega) \leq 1$, каково бы ни было $\omega \in \Omega$, и

$$\sum_{\omega \in \Omega} p(\omega) = 1.$$

Вероятность любого составного события A вычисляется по формуле

$$P(A) = \sum_{\omega \in A} p(\omega).$$

Число $P(A)$ интерпретируется как относительная частота появления события A в статистическом эксперименте, состоящем из достаточно большого числа испытаний. Опираясь на эту интерпретацию, легко построить распределение вероятностей в примерах 1.1–1.6.

Если подбрасывается “правильная” (симметричная) монета (см. пример 1.1), то естественно определить вероятности элементарных исходов из условия симметрии и положить $p(\Gamma) = p(P) = 1/2$, что блестяще подтверждается результатами статистических экспериментов (см. таблицу). Однако уже при подбрасывании двух монет (пример 1.2) у части неискушенных исследователей возникает желание нарушить условие симметрии и приписать исходам ГГ и РР меньшую вероятность, чем ГР или РГ. В истории стохастики известен также парадокс, основанный на некорректном определении пространства Ω , когда составное событие $A = \{\Gamma P, P\Gamma\}$ трактовалось как элементарное и, следуя “аксиоме симметрии”, утверждалось, что $p(\Gamma\Gamma) = p(P P) = p(A) = 1/3$. Поскольку результаты опытов противоречили такой вероятностной модели (наблюдения показывали, что $p(A) = 1/2$, $p(\Gamma\Gamma) = p(P P) = 1/4$), то указанный феномен объявлялся парадоксом теории вероятностей, над разрешением которого бились многие известные математики и естествоиспытатели, в том числе и великий Даламбер. Все разъяснилось только после четкого математического определения *независимости* событий. Мы познакомимся с этим фундаментальным понятием теории вероятностей несколько позднее, а пока, следуя принципу симметрии, припишем каждому из четырех элементарных исходов, наблюдаемых при подбрасывании двух монет, одну и ту же вероятность $1/4$. Как уже говорилось выше, эта вероятностная модель согласуется с результатами наблюдений частот элементарных исходов в соответствующем статистическом эксперименте, состоящем из большого числа испытаний двух правильных монет.

Чтобы закончить с испытаниями правильных монет, обратимся сразу к примеру 1.5, где элементарный исход формируется из результатов подбрасываний n монет. В этой ситуации убедить вышеупомянутого “неискушенного исследователя” в равновероятности всех элементарных исходов практически невозможно. Например, считается, что элементарный исход ГГГ-ГГГГГГГГ имеет значительно меньшую вероятность появления, чем исход РРГРГГГРРГ (здесь $n = 10$). Это чисто психологический феномен, связанный с неосознанной подменой этих двух элементарных исходов двумя составными событиями: A – все монеты выпали одной стороной (событие, состоящее из двух элементарных исходов) и B – хотя бы одна монета выпала не той стороной, что все остальные (событие, состоящее из $2^n - 2$ исходов). По этой же причине абсолютное большинство покупателей лотерейных билетов откажутся от билета, номер которого состоит из одинаковых

цифр, хотя, очевидно, все билеты имеют одинаковый шанс быть выигрышными. В последнем легко убедиться, наблюдая, как происходит розыгрыш лотерейных билетов, то есть как обеспечивается *равновероятность* билетов вне зависимости от их номеров. Итак, в примере 1.5 вероятностная модель определяется вероятностями $p(\omega) = 2^{-n}$, каково бы ни было $\omega \in \Omega$. В соответствии с этой, подтверждаемой реальными статистическими экспериментами, моделью вероятности упомянутых событий A и B равны соответственно $1/2^{n-1}$ и $1 - 1/2^{n-1}$. Например, при $n = 10$ $P(A) = 1/512$, а $P(B) = 511/512$, так что событие B происходит в 511 раз чаще, чем событие A .

Принцип “симметрии” также применяется и в построении вероятностной модели испытаний правильной кости (примеры 1.3 и 1.4). Естественно, все грани имеют одинаковую вероятность выпадения, в соответствии с чем $p(\omega) = 1/6$ в примере 1.3 и $p(\omega) = 1/36$ в примере 1.2, каково бы ни было $\omega \in \Omega$. Однако не следует излишне доверять этой модели на практике, когда вам придется играть в кости с приятелем или в казино. При раскопках египетских пирамид были найдены игральные кости со смещенным центром тяжести, так что еще за тысячелетия до нашей эры находились “весьма искушенные испытатели”, способные управлять частотой элементарных исходов.

Распределение вероятностей в примере 1.6 можно получить, используя те же рассуждения, что и в примерах 1.1, 1.2 и 1.5. Действительно, осуществление элементарного исхода ω_1 означает выпадение герба в однократном подбрасывании монеты, так что (см. пример 1.1) $p_1 = p(\omega_1) = 1/2$. Элементарный исход ω_2 совпадает с элементарным исходом РГ в примере 1.2, следовательно, $p_2 = p(\omega_2) = 1/4$. Наконец, при произвольном $n = 1, 2, \dots$, используя вероятность элементарного исхода РР, ..., РГ (первые $n - 1$ испытаний закончились выпадением решки, а при n -ом испытании выпал герб) в примере 1.5, получаем $p_n = p(\omega_n) = 2^{-n}$. Завершив построение вероятностной модели, убедимся в справедливости равенства

$$\sum_{n=1}^{\infty} p_n = \sum_{n=1}^{\infty} 2^{-n} = 1.$$

Итак, при построении вероятностных моделей в примерах 1.1–1.6 мы существенно использовали физическую природу объектов, с которыми проводились эксперименты, – монета и кость были “правильными” (симметричными), и только это свойство позволило нам приписать одинаковые веро-

ятности всем элементарным исходам. Если подбрасывается гнутая монета, то определение вероятности p выпадения герба, на основе уравнения, описывающего динамику полета вращающейся неправильной монеты и закономерности ее упругого столкновения с поверхностью, представляет собой весьма сложную и вряд ли разрешимую задачу. Следует также отметить, что если p известно, но не равно $1/2$, то мы не в состоянии найти распределение вероятностей в примерах 1.2 и 1.5 с многократным подбрасыванием монеты, пока не формализовано понятие независимости испытаний монеты.

Если теперь обратиться к примеру 1.7, то в свете вышесказанного становится понятным, что построение вероятностной модели численности метеоров невозможно без привлечения знаний об их распределении в околоземном пространстве, учета эффекта вращения Земли в интенсивности появления метеоров, разделения метеорных явлений на “потoki” и “спорадический фон”. Учитывая наши более чем скудные познания в теории вероятностей, следует признать, что решение этой задачи нам пока “не по зубам”. И все же, предвосхищая наши дальнейшие построения, наиболее любопытным и нетерпеливым сообщу, что после учета эффекта вращения Земли распределение метеоров в спорадическом фоне выражается формулой

$$p_k = p(\omega_k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, \dots$$

Мы завершим этот параграф решением некоторых задач, в которых используются модели, основанные на равновероятности элементарных исходов. Все эти задачи, так или иначе, сводятся к подсчету числа элементарных исходов, влекущих некоторое событие A ; определение вероятности этого события и составляет предмет задачи.

Задача 1.1. Бросаются две правильные кости. Найти вероятность того, что сумма выпавших очков больше 6.

В соответствии с распределением вероятностей, полученным в примере 4, все 36 элементарных исходов имеют одинаковую вероятность $1/36$, так что для решения задачи достаточно подсчитать число целых решений неравенства $x + y > 6$ или обратиться к матрице элементарных исходов $\|\omega_{i,j}\|$, выделив в ней элементы с $i + j > 6$, составляющие искомое событие

A ,

$$\begin{pmatrix} 11 & 12 & 13 & 14 & 15 & \mathbf{16} \\ 21 & 22 & 23 & 24 & \mathbf{25} & \mathbf{26} \\ 31 & 32 & 33 & \mathbf{34} & \mathbf{35} & \mathbf{36} \\ 41 & 42 & \mathbf{43} & \mathbf{44} & \mathbf{45} & \mathbf{46} \\ 51 & \mathbf{52} & \mathbf{53} & \mathbf{54} & \mathbf{55} & \mathbf{56} \\ \mathbf{61} & \mathbf{62} & \mathbf{63} & \mathbf{64} & \mathbf{65} & \mathbf{66} \end{pmatrix}$$

Число “благоприятных” для события A исходов (они выделены жирным шрифтом) равно $(1+6)6/2=21$, откуда $P(A) = 21/36 = 7/12$. Итак, если вам предложат играть в кости, где ставка идет на сумму очков больше 6 или на противоположное событие $i + j \leq 6$, то следует ставить на первое событие - в среднем один раз из двенадцати ставок вы будете получать дополнительный выигрыш по сравнению с вашим партнером по игре.

Задача 1.2. (*вероятностная задача Шевалье де Мере*). Один из создателей современной теории математической статистики Ю.Нейман утверждает, что основателями теории статистических решений следует считать тех азартных игроков, которые впервые стали рассчитывать шансы определенных ставок при игре в кости, карты и т.п., и в связи с этим излагает некоторые фрагменты из переписки Б.Паскаля с одним из таких игроков. Ниже приводится выдержка из вводного курса Ю.Неймана по теории вероятностей и математической статистике.

“В конце семнадцатого века один французский вельможа Шевалье де Мере, известный игрок в азартные игры, в частности в кости, заинтересовался возможностью вычислить математически, как следует делать ставки. Его интересовала игра, состоящая из 24 бросаний пары костей. По правилам игры ставить можно было или на появление “двойной шестерки” по крайней мере один раз в 24 бросаниях, или против этого результата.

Вычисления Шевалье де Мере привели его к заключению, что в длинном ряде игр “двойная шестерка” должна появляться (хоть один раз) более чем в пятидесяти процентах всех игр и что поэтому выгодно ставить на появление двойной шестерки. Хотя Шевалье де Мере был уверен в правильности своих вычислений, он сомневался в надежности математики и произвел очень длинный ряд опытов с бросанием костей (этот эмпирический метод применяется и теперь и носит название “метод Монте-Карло”). Оказалось, что частность двойной шестерки в ряду игр меньше пятидесяти процентов! Получив этот результат, де Мере рассвирепел и написал известному французскому математику Паскалю письмо, утверждающее, что

математика как наука никуда не годится, и пр. Письмо это было настолько яростным и вместе с тем забавным, что попало в историю!

Паскаль рассмотрел задачу Шевалье де Мере, и ответ его гласил: если кости “правильные”, то относительная частота игр с хотя бы одной двойной шестеркой равна 0.491. Таким образом, оказалось, что математические выкладки Шевалье де Мере были ошибочны, а его эмпирический результат согласуется с теорией” (конец цитаты).

Приведем решение задачи де Мере, данное Паскалем.

Пространство элементарных событий Ω в этой задаче состоит из 36^{24} равновероятных исходов. Следовательно, для решения задачи достаточно подсчитать число элементарных исходов, влекущих событие A : двойная шестерка появилась хотя бы один раз. Однако несомненно проще подсчитать число исходов для противоположного события A^c : ни одно из бросаний двух костей не закончилось появлением двойной шестерки. Очевидно, число таких исходов равно 35^{24} , откуда число исходов, благоприятствующих событию A , равно $36^{24} - 35^{24}$ и

$$P(A) = 1 - \left(\frac{35}{36}\right)^{24} \approx 0.491$$

Можно предположить, что де Мере напрямую, не зная, по всей видимости, формулы биномиальных коэффициентов, подсчитывал, сколько элементарных исходов благоприятствует однократному появлению двойной шестерки, потом двукратному, и так далее до 24, а потом сложил эти числа. Произвести все эти действия с многозначными числами и при этом не ошибиться вряд ли по плечу даже французскому вельможе! Из всей этой истории мы должны сделать один практически важный при решении задач вывод: переход к противоположному событию и использование очевидной формулы

$$P(A) = 1 - P(A^c)$$

может значительно упростить решение вероятностной задачи, связанной с комбинаторными выкладками.

Лекция 2

Задача 1.3. (*гипергеометрическое распределение вероятностей*). Существует довольно большой класс задач элементарной теории вероятностей, которые можно интерпретировать в рамках так называемой *урновой*

схемы: событие, вероятность которого необходимо вычислить, можно трактовать как результат случайного выбора шаров различной расцветки из урны. Простейшая из таких урновых схем состоит в следующем. Из урны, содержащей M черных и $N - M$ белых шаров, случайным образом отбирается n шаров. Какова вероятность, что выборка содержит t черных шаров (событие A)?

В этом эксперименте пространство элементарных событий состоит из C_N^n исходов (шары одинакового цвета не различаются), и случайность отбора означает, что элементарные исходы имеют одну и ту же вероятность $1/C_N^n$. Следовательно, решение задачи сводится к подсчету числа выборов из n шаров, которые содержат t черных и $n - t$ белых. Очевидно,

$$\max(0, n - (N - M)) \leq t \leq \min(n, M), \quad (1)$$

если объем выборки n превышает число черных шаров M , то мы не сможем выбрать более чем M черных, и если n больше, чем число белых шаров $N - M$, то число t черных шаров в выборке не может быть меньше $n - (N - M)$.

Из M черных шаров выбирается t шаров того же цвета, и число всевозможных способов такого выбора равно C_M^t . Аналогично, из $N - M$ белых шаров $n - t$ шаров того же цвета можно выбрать C_{N-M}^{n-t} способами. Следовательно, общее число исходов, благоприятствующих событию A , равно $C_M^t \cdot C_{N-M}^{n-t}$, и искомая вероятность

$$P(A) = \frac{C_M^t C_{N-M}^{n-t}}{C_N^n}. \quad (2)$$

Говорят, что формула (2) определяет *гипергеометрическое распределение* целочисленной случайной величины X , принимающей значение из области (1), – вероятность

$$p_t = P(X = t | N, M, n)$$

равна правой части (2) при любом t из области (1) и $\sum p_t$ по всем t , удовлетворяющим (1), равна 1.

Приведем несколько примеров на применения гипергеометрического распределения вероятностей.

1. *Выигрыш в лотерее Спортлото “6 из 49”*. В начале 70-х годов получила распространение разновидность лотереи, носящая название “спортлото”. Участник лотереи из 49 видов спорта, обозначенных просто цифрами, называет шесть. Выигрыш определяется тем, сколько наименований

он угадал из шести других наименований, которые были заранее выделены комиссией. Спрашивается, какова вероятность того, что участник угадает все шесть наименований, пять наименований и т.д.

Нетрудно видеть, что это есть не что иное, как задача о гипергеометрическом распределении, где $N = 49$, $M = 6$ (угадываемые номера – черные шары), $n = 6$ и $m (= 1, \dots, 6)$ – число угаданных номеров. Вероятность угадать m номеров равна

$$P(X = m | 49, 6, 6) = \frac{C_6^m C_{43}^{6-m}}{C_{49}^6}.$$

Например, вероятность максимального выигрыша ($m = 6$) равна

$$C_6^6 C_{43}^0 / C_{49}^6 = 1 / C_{49}^6 = 6! 43! / 49! \approx 7.2 \cdot 10^{-8}.$$

Это меньше одной десятиллионной(!) – шансы на выигрыш ничтожны.

2. Как вытащить “счастливый” билет на экзамене? Группа из N студентов сдает экзамен, на котором каждому студенту предлагается выбрать наугад один из N билетов. Студент Петров знает $M (< N)$ билетов, и считает, что если он пойдет сдавать экзамен первым, то шансов “вытянуть” счастливый билет у него несомненно больше, чем если он пойдет отвечать последним (его доводы в пользу этого – “все счастливые билеты будут разобраны”). Прав ли Петров?

Если Петров пойдет первым, то вероятность выбора счастливого билета равна, очевидно, M/N . Если же Петров идет последним, то мы можем при расчете вероятности воспользоваться гипергеометрическим распределением $P(X = M - 1 | N, M, N - 1)$ – предшествующие Петрову $N - 1$ студентов (объем выборки $n = N - 1$) должны выбрать ровно $m = M - 1$ счастливых билетов, и тогда Петрову, который сдает последним, достанется счастливый билет. Имеем

$$P(X = M - 1 | N, M, N - 1) = \frac{C_M^{M-1} C_{N-M}^{(N-1)-(M-1)}}{C_N^{N-1}} = \frac{C_M^{M-1}}{C_N^{N-1}} = \frac{M}{N},$$

так что шансы выбрать счастливый билет одинаковы. Нетрудно, производя аналогичные выкладки, убедиться, что шансы выбрать счастливый билет вообще не зависят от того, каким по счету придет Петров на экзамен, – они всегда одни и те же M/N .

3. *Оценка численности замкнутой популяции животных (метод максимального правдоподобия)*. Предыдущие два примера иллюстрировали применение вероятностной модели гипергеометрического распределения к решению, так называемых, *прямых* задач теории вероятностей: зная *параметры* модели N , M и n , мы определяли вероятности событий, связанных со значениями случайной величины X . Но в естественных науках (физика, биология, экономика и пр.) обычно приходится решать *обратные* задачи – наблюдая значения случайной величины X , исследователь стремится сделать определенное заключение о неизвестных параметрах вероятностной модели. Решением таких обратных задач занимается родственная теории вероятностей наука *математическая статистика*. Следующий пример иллюстрирует один из типичных методов решения задачи по оценке параметра вероятностной модели.

Проблема состоит в определении численности N рыб, живущих на момент наблюдения в замкнутом водоеме, скажем, в пруду рыбоводного хозяйства. Для определения (точнее, приближенной оценки) N исследователь отлавливает заданное количество M рыб, метит их каким-либо способом и возвращает в пруд. По истечении некоторого промежутка времени, когда, по его мнению, меченые рыбы “перемешались” с другими обитателями пруда, он снова отлавливает фиксированное количество n рыб (в математической статистике эта процедура называется извлечением выборки объема n из генеральной совокупности) и подсчитывает число t отмеченных рыб, попавших во второй улов. В рамках гипергеометрической модели такого эксперимента мы располагаем значениями параметров M и n , знаем результат t наблюдения случайной величины X , но не знаем значения параметра N гипергеометрического распределения $P(X = t | N, M, n)$.

Один из основных методов решения обратных задач теории вероятностей (задач математической статистики), который называется *методом максимального правдоподобия*, состоит в выборе такого значения \hat{N} параметра N , которое соответствует максимуму вероятности наблюдаемого исхода t в наблюдении X . Основной довод в пользу такого поведения статистика состоит в простом житейском наблюдении: если происходит какое-либо событие, то это событие должно иметь большую вероятность по сравнению с другими исходами статистического эксперимента.

Итак, метод максимального правдоподобия предлагает в качестве оценки неизвестного значения N (численности рыб в пруду) взять решения

следующей задачи на экстремум:

$$\hat{N} = \arg \max_N \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}.$$

Решить эту задачу можно с помощью определения значения N , при котором происходит смена неравенства

$$\frac{C_M^m C_{N-M}^{n-m}}{C_N^n} < \frac{C_M^m C_{N+1-M}^{n-m}}{C_{N+1}^n}$$

на обратное. Используя известную формулу для вычисления биномиальных коэффициентов, находим, что это неравенство эквивалентно $(N + 1)m > nM$, откуда получаем оценку максимального правдоподобия для численности рыб в пруду:

$$\hat{N} = \left[n \frac{M}{m} \right].$$

Легко заметить, что такая оценка согласуется с простыми рассуждениями типа “если при повторном отлове я обнаружил половину отмеченных рыб, то в пруду их в два раза больше, чем я поймал”.

Задача 1.4. Геометрические вероятности: задача о встрече. Два человека договорились встретиться в течение определенного часа. Предлагается, что момент прихода каждого из встречающихся не зависит от намерений другого и имеет “равномерное” распределение в назначенном промежутке встречи 60 минут (момент прихода случаен). Пришедший первым ждет другого только 10 минут, после чего уходит (встреча не состоялась). Какова вероятность встречи?

Это одна из типичных задач *геометрической вероятности*, для решения которой используется следующая математическая формализация понятия “случайности” момента прихода. Рассмотрим более общую (и более абстрактную) задачу. В евклидовом пространстве \mathbb{R}^n выделяется замкнутая область Ω конечной лебеговой меры $\mu(\Omega)$. На область Ω случайно бросается точка, и требуется определить вероятность того, что точка попадет в подмножество $S \subset \Omega$. Естественно формализовать понятие “случайности” в терминах независимости вероятности попадания точки в S от положения этого множества в области Ω и его конфигурации и постулировать, что искомая вероятность пропорциональна $\mu(S)$. В таком случае Ω играет роль пространства элементарных исходов, вероятность попадания точки в



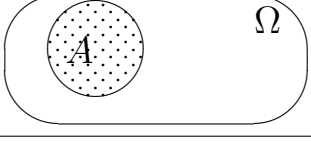
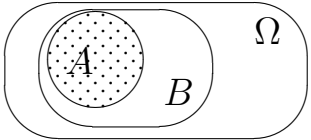
Ω должна равняться единице, так что вероятность попадания в множество S равна $P(S) = \mu(S)/\mu(\Omega)$.

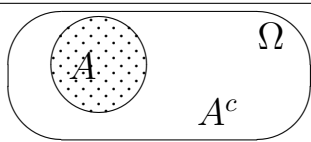
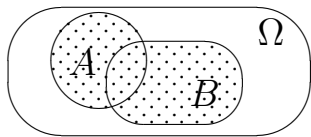
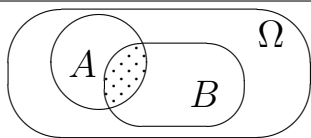
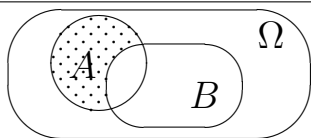
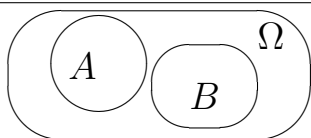
Используем этот метод в решении задачи о встрече. Здесь Ω – квадрат 60×60 , множество S – полоса вдоль диагонали квадрата, которую в декартовой системе координат можно задать в виде области $|x - y| \leq 10$. Очевидно, площадь этой области равна $60 \cdot 60 - 50 \cdot 50$, площадь квадрата – $60 \cdot 60$, откуда искомая вероятность встречи, равная отношению площадей, $P(S) = 1 - (50/60)^2 = 11/36$.

Основной вывод, который мы должны сделать из решения данной задачи, состоит в осознании невозможности определения вероятности на несчетных пространствах Ω посредством задания функции $p(\omega)$, как вероятности элементарного исхода $\omega \in \Omega$. Распределение вероятностей следует определять с помощью функций на подмножествах Ω , причем эти функции должны быть *нормированными мерами* – вероятность всего Ω должна равняться единице, и $P(S)$ должна обладать свойством счетной аддитивности.

§2. Вероятностное пространство

Аксиоматическое построение теории вероятностей начинается с формализации (описания) *пространства Ω элементарных исходов ω* некоторого статистического эксперимента. Определенные (см. ниже) подмножества пространства Ω называются *событиями*; говорят, что произошло событие $A (\subset \Omega)$, если статистический эксперимент закончился элементарным исходом $\omega \in A$. Над событиями A , как подмножествами пространства Ω , вводятся теоретико-множественные операции, вероятностная трактовка которых приводится в следующей таблице.

Теоретико-множественные объекты и операции	Вероятностная трактовка	Геометрическая интерпретация
Ω – множество	пространство элементарных исходов, <i>достоверное событие</i>	
ω – элемент Ω	элементарный исход эксперимента, элементарное событие	
A – подмножество множества Ω	событие	
\emptyset – пустое множество	<i>невозможное событие</i>	
$A \subset B$ – подмножество A есть часть (принадлежит) B	событие A влечет событие B	

A^c – дополнение подмножества A до Ω	событие A не произошло	
$A \cup B$ – объединение подмножеств A и B	Произошло по крайней мере одно из событий A или B	
$A \cap B$ – пересечение подмножеств A и B	Произошли одновременно оба события A и B	
$A \setminus B$ – разность: из подмножества A вычитается подмножество B	Произошло событие A , в то время как событие B не произошло	
$A \cap B = \emptyset$ – множества A и B не имеют общих точек (не пересекаются)	события A и B <i>несовместны</i>	

Если рассматривать введенные операции над множествами как алгебраические, то Ω выступает в роли “единицы” алгебры, а \emptyset – в роли ее “нуля”, что видно из следующих равенств:

$$\begin{aligned}
 A \subset \Omega, \quad \Omega^c = \emptyset, \quad \emptyset^c = \Omega, \quad (A^c)^c = A; \\
 A \cup \emptyset = A, \quad A \cup A = A, \quad A \cup \Omega = \Omega, \quad A \cup A^c = \Omega; \\
 A \cap \emptyset = \emptyset, \quad A \cap A = A, \quad A \cap \Omega = A, \quad A \cap A^c = \emptyset.
 \end{aligned}$$

Операции объединения и пересечения распространяются на любое, возможно бесконечное семейство $\{A_i, i \in I\}$ событий:

$$\begin{aligned}
 \bigcup_{i \in I} A_i & \quad - \text{ произошло по крайней мере одно из событий семейства } \{A_i, i \in I\}, \\
 \bigcap_{i \in I} A_i & \quad - \text{ произошли одновременно все события семейства } \{A_i, i \in I\}.
 \end{aligned}$$

Определение 2.1. Семейство событий $\{A_i, i \in I\}$ называется семейством *несовместных событий*, если $A_i \cap A_j = \emptyset$ при любых $i \neq j, i, j \in I$.

Если $A_i, i \in I$, несовместны, то вместо знака \bigcup используется знак “прямой суммы” \sum (или $+$):

$$\bigcup_{i \in I} A_i = \sum_{i \in I} A_i, \quad A \cup B = A + B.$$

Имеет место *правило двойственности*:

$$\left(\bigcup_I A_i\right)^c = \bigcap_I A_i^c, \quad \left(\bigcap_I A_i\right)^c = \bigcup_I A_i^c.$$

Напомним, что операции объединения и пересечения обладают свойствами коммутативности $A \cup B = B \cup A$, ассоциативности $(A \cup B) \cup C = A \cup (B \cup C)$ и дистрибутивности $B \cap (\cup_I A_i) = \cup_I (A_i \cap B)$. Отношение принадлежности $A \subset B$ порождает частичный порядок на подмножествах пространства Ω , так что отношение эквивалентности (равенства) $A = B$ двух событий означает, что одновременно $A \subset B$ и $B \subset A$. Введенные выше операции над множествами определяют структуру булевой алгебры: имеет место

Определение 2.2. *Булевой алгеброй* называется такой класс \mathcal{A} подмножеств Ω , что

- (A1) $\Omega \in \mathcal{A}$,
- (A2) $A \in \mathcal{A} \implies A^c \in \mathcal{A}$,
- (A3) $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$.

Лекция 3

Предложение 2.1. *Если \mathcal{A} – булева алгебра, то*

- (1) $\emptyset \in \mathcal{A}$,
- (2) $A_1, \dots, A_n \in \mathcal{A} \implies \bigcup_1^n A_i \in \mathcal{A}, \bigcap_1^n A_i \in \mathcal{A}$,
- (3) $A, B \in \mathcal{A} \implies A \setminus B \in \mathcal{A}$.

Доказательство. (1) Так как $\emptyset = \Omega^c$, то в силу (A1) и (A2) $\emptyset \in \mathcal{A}$.

(2) Используя метод индукции, легко показать, что для любого $n = 1, 2, \dots$ включение $\{A_i, i = 1, \dots, n\} \subset \mathcal{A}$ влечет $\bigcup_1^n A_i \in \mathcal{A}$ ($n - 1$ раз используется аксиома (A3) булевой алгебры). С другой стороны, из правила двойственности вытекает, что $\{A_i, i = \overline{1, n}\} \subset \mathcal{A} \implies \bigcap_1^n A_i \in \mathcal{A}$, ибо

$$\begin{aligned} \{A_i^c, i = \overline{1, n}\} \subset \mathcal{A} &\implies \bigcup_1^n A_i^c = \left(\bigcap_1^n A_i\right)^c \in \mathcal{A} \implies \\ \bigcap_1^n A_i &= \left[\left(\bigcap_1^n A_i\right)^c\right]^c \in \mathcal{A}. \end{aligned}$$

(3) Это свойство немедленно следует из очевидного равенства $A \setminus B = A \cap B^c$ (разность множеств означает, что ω одновременно принадлежит дополнению множества B и множеству A).

Таким образом, булева алгебра содержит “единицу” Ω , “ноль” \emptyset и замкнута относительно конечного числа операций объединения, пересечения и вычитания (взятия дополнения).

Примеры булевых алгебр. **1.** Самая “тонкая” булева алгебра: множество $\mathcal{P}(\Omega)$ всевозможных подмножеств пространства Ω , включая пустое множество \emptyset , как подмножество любого $A \in \Omega$. **2.** Самая “грубая” булева алгебра $\mathcal{A} = \{\emptyset, \Omega\}$. **3.** Булева алгебра, порожденная событием A : $\mathcal{A} = \{\emptyset, \Omega, A, A^c\}$.

Определение 2.3. Вероятностью P на булевой алгебре \mathcal{A} подмножеств Ω называется отображение \mathcal{A} в отрезок $[0; 1]$, обладающее следующими свойствами:

(P1) *нормируемость*: $P(\Omega) = 1$;

(P2) *конечная аддитивность*: если события A_1, \dots, A_n несовместны, то

$$P\left(\sum_1^n A_i\right) = \sum_1^n P(A_i);$$

(P3) *непрерывность*: если $\{A_n, n \geq 1\}$ – монотонно убывающая по включению последовательность элементов из \mathcal{A} и $\bigcap_1^\infty A_n = \emptyset$ (в этом случае пишут $A_n \downarrow \emptyset$, когда $n \rightarrow \infty$), то

$$\lim_{n \rightarrow \infty} P(A_n) = 0.$$

В следующем предложении предполагается, что все события принадлежат булевой алгебре \mathcal{A} .

Предложение 2.2. Вероятность P на булевой алгебре \mathcal{A} обладает следующими свойствами:

(1) $P(\emptyset) = 0$;

(2) $P(A^c) = 1 - P(A)$;

(3) если $A \subset B$, то $P(A) \leq P(B)$ (свойство монотонности) и $P(B \setminus A) = P(B) - P(A)$;

(4) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (свойство сильной аддитивности);

(5) $P(\bigcup_1^n A_i) \leq \sum_1^n P(A_i)$ (свойство полуаддитивности);

(6) если $A_n \downarrow A$ или $A_n \uparrow A$, то справедливо свойство непрерывности относительно монотонной сходимости

$$\lim_{n \rightarrow \infty} P(A_n) = P(A);$$

(7) если $\{A_n, n \geq 1\}$ – бесконечная последовательность несовместных событий, то имеет место свойство σ -аддитивности

$$P\left(\sum_1^\infty A_n\right) = \sum_1^\infty P(A_n);$$

(8) $P(\bigcup_1^\infty A_n) \leq \sum_1^\infty P(A_n)$ (свойство σ -полуаддитивности.)

Доказательство. (1) Используя в нужном месте аксиомы (P2) и (P1), получаем

$$1 = P(\Omega) = P(\Omega + \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset),$$

откуда $P(\emptyset) = 0$.

(2) Используя аксиому аддитивности (P2), имеем $1 = P(\Omega) = P(A + A^c) = P(A) + P(A^c)$, откуда $P(A^c) = 1 - P(A)$.

(3) Так как $B = A + (B \setminus A)$, то, в силу (P2), $P(B) = P(A) + P(B \setminus A)$, откуда $P(B \setminus A) = P(B) - P(A)$. Поскольку $P(B \setminus A) \geq 0$, то из последнего равенства вытекает свойство монотонности $P(A) \leq P(B)$.

(4) Легко видеть, что $A \cup B = A + (B \setminus (A \cap B))$, и поскольку $A \cap B \subset B$, то в силу аксиомы аддитивности и доказанного свойства (3) монотонности вероятности получаем

$$P(A \cup B) = P(A) + P(B \setminus (A \cap B)) = P(A) + P(B) - P(A \cap B).$$

(5) Доказательство проведем по индукции. При $n = 2$ из доказанного свойства (4) сильной аддитивности и положительности вероятности вытекает, что

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2).$$

Теперь, полагая, что доказываемое неравенство имеет место для некоторого целого n , убеждаемся, что оно справедливо для $n + 1$, используя представление

$$\bigcup_1^{n+1} A_i = \left(\bigcup_1^n A_i \right) \cup A_{n+1}.$$

(6) Если $A_n \downarrow A$, то $A_n \setminus A \downarrow \emptyset$, и требуемое свойство вытекает из представления $A_n = A + (A_n \setminus A)$ и аксиом аддитивности (P 2) и непрерывности (P 3) вероятности P . Случай $A_n \uparrow A$ рассматривается аналогично и при этом используется представление $A = A_n + (A \setminus A_n)$.

(7) Свойство σ -аддитивности вытекает из доказанного свойства (6) и свойства (P 2) конечной аддитивности:

$$\begin{aligned} P \left(\sum_1^\infty A_k \right) &= P \left(\lim_{n \rightarrow \infty} \uparrow \sum_1^n A_k \right) = \lim_{n \rightarrow \infty} P \left(\sum_1^n A_k \right) = \\ &= \lim_{n \rightarrow \infty} \sum_1^n P(A_k) = \sum_1^\infty P(A_k). \end{aligned}$$

(8) Используя, как и в (7), свойство (6), а также свойство полуаддитивности (5), получаем

$$\begin{aligned} P \left(\bigcup_1^\infty A_k \right) &= P \left(\lim_{n \rightarrow \infty} \uparrow \bigcup_1^n A_k \right) = \lim_{n \rightarrow \infty} P \left(\bigcup_1^n A_k \right) \leq \\ &= \lim_{n \rightarrow \infty} \sum_1^n P(A_k) = \sum_1^\infty P(A_k). \end{aligned}$$

Из доказанных свойств вероятности следует обратить особое внимание на свойство (7) σ -аддитивности. Дело в том, что это свойство часто кладется в основу определения вероятности вместо аксиом (P 2) и (P 3). Имеет место

Определение 2.4. Вероятностью P на булевой алгебре \mathcal{A} подмножеств Ω называется такое отображение \mathcal{A} в отрезок $[0; 1]$, что

(P 1) (нормируемость) $P(\Omega) = 1$,

(P 2') (σ -аддитивность) если объединение $\sum_1^\infty A_n$ счетного семейства $\{A_n, n \geq 1\}$ несовместных событий принадлежит булевой алгебре \mathcal{A} , то

$$P \left(\sum_1^\infty A_n \right) = \sum_1^\infty P(A_n).$$

Имеет место

Предложение 2.3. *Определения 2.3 и 2.4 вероятности P на булевой алгебре \mathcal{A} эквивалентны.*

Доказательство. То, что $(P2)$ и $(P3)$ влечет $(P2')$ было установлено в утверждении (7) предложения 2.2. Докажем обратное – σ -аддитивность влечет непрерывность P .

Пусть $A_n \downarrow \emptyset$; требуется доказать, что $P(A_n) \rightarrow 0$, когда $n \rightarrow \infty$. Представим A_n в виде объединения последовательности несовместных событий:

$$A_n = \sum_{k=n}^{\infty} (A_k \setminus A_{k+1}).$$

В силу аксиомы σ -аддитивности

$$P(A_n) = \sum_{k=n}^{\infty} P(A_k \setminus A_{k+1}).$$

Правая часть этого равенства представляет остаточный член сходящегося ряда

$$S = \sum_{k=1}^{\infty} P(A_k \setminus A_{k+1}),$$

поскольку из включения $A_{k+1} \subset A_k$ (последовательность $\{A_n, n \geq 1\}$ – монотонно убывающая) следует

$$S = \sum_{k=1}^{\infty} [P(A_k) - P(A_{k+1})] = P(A_1) \leq 1.$$

Итак, ряд сходится и, следовательно, его остаточный член $P(A_n) \rightarrow 0$, когда $n \rightarrow \infty$.

Определение 2.4 постулирует, что P есть нормированная счетно аддитивная мера на булевой алгебре \mathcal{A} подмножеств (событий) пространства элементарных исходов Ω . Поскольку нам придется довольно часто вычислять вероятности объединений бесконечного числа событий, а такие объединения не обязательно принадлежат \mathcal{A} , то естественно, как это принято в теории меры, расширить булеву алгебру \mathcal{A} , включив в нее пределы монотонных последовательностей ее элементов, после чего продолжить вероятность P на расширенный таким образом класс подмножеств Ω .

Определение 2.5. Совокупность \mathcal{A} подмножеств пространства Ω называется *булевой σ -алгеброй*, если

$$(A1) \quad \Omega \in \mathcal{A},$$

$$(A2) \quad A \in \mathcal{A} \implies A^c \in \mathcal{A},$$

$$(A3)_S \quad \{A_n, n \geq 1\} \subset \mathcal{A} \implies \bigcup_1^\infty A_n \in \mathcal{A}.$$

Используя правило двойственности по аналогии с доказательством предложения 2.1, легко убедиться, что булева σ -алгебра замкнута не только относительно объединения счетного числа своих элементов, но и относительно их пересечения.

Всегда существует хотя бы одна σ -алгебра подмножеств любого пространства Ω , например, таковой является совокупность $\mathcal{P}(\Omega)$ всевозможных подмножеств Ω , включая \emptyset (самая “тонкая” булева алгебра – см. пример 1). Вспоминая наши рассуждения о пополнении булевой алгебры пределами монотонных (по включению) последовательностей ее элементов, введем

Определение 2.6. Наименьшая σ -алгебра $\mathcal{B}(\mathcal{A})$, содержащая булеву алгебру \mathcal{A} , называется *σ -алгеброй, порожденной булевой алгеброй \mathcal{A}* .

Из курса анализа вам хорошо известна знаменитая теорема о продолжении меры. В терминах вероятностной меры она формулируется следующим образом

Теорема (о продолжении вероятности). *Любая вероятность P , заданная на булевой алгебре \mathcal{A} , имеет единственное продолжение на порожденную \mathcal{A} σ -алгебру $\mathcal{B}(\mathcal{A})$.*

Определение 2.7. Пара (Ω, \mathcal{A}) , состоящая из пространства элементарных исходов Ω и булевой σ -алгебры \mathcal{A} его подмножеств, называется *измеримым пространством*. Только элементы \mathcal{A} называются *событиями*, остальные подмножества Ω , не принадлежащие \mathcal{A} , называются *неизмеримыми подмножествами*. Наконец, триплет (Ω, \mathcal{A}, P) , в котором P – вероятность на σ -алгебре \mathcal{A} , называется *вероятностным пространством*.

§3. Условная вероятность и независимость событий

Лекция 4

Понятие вероятностного пространства играет фундаментальную роль в приложениях теории вероятностей, поскольку это – математическая формализация вероятностной модели. Зная распределение вероятностей, мы в состоянии оптимизировать свое поведение при “игре” с природой, производя “ставки” на те события из сигма-алгебры \mathcal{A} , которые обладают наибольшей вероятностью. Дальнейшая оптимизация такой игры обычно осуществляется за счет дополнительной информации, которой может располагать игрок, и учет такой информации осуществляется в терминах так называемой *условной вероятности*. Чтобы уяснить смысл этого нового для нас понятия, рассмотрим следующий простой пример.

Бросается правильная кость и нас интересует событие A : выпало 6 очков. Априори вероятность этого события равна $1/6$, но пусть мы располагаем дополнительной информацией, что выпало четное количество очков (событие B). В таком случае вероятность события A должна увеличиться, и для ее пересчета мы должны рассмотреть суженное пространство элементарных исходов $\Omega_B = \{2, 4, 6\}$. В соответствии с распределением вероятностей на исходном пространстве элементарных исходов $\Omega = \{1, 2, 3, 4, 5, 6\}$ вероятность $P(B)$ события B (или, что то же, вероятностная мера нового пространства элементарных исходов Ω_B) равна $1/2$. Условие “произошло событие B ” делает пространство элементарных исходов Ω_B достоверным событием, и, следовательно, мы должны приписать ему вероятность единица, а вероятности $p(2) = p(4) = p(6) = 1/6$ остальных исходов из Ω_B пронормировать – разделить на меру $P(B) = P(\Omega_B) = 1/2$. Таким образом, *условное распределение* на Ω_B следует вычислять по формуле $p(\omega | B) = P(\{\omega\} \cap B) / P(B)$. Итак, искомая вероятность события A при условии, что произошло событие B , (условная вероятность) равна $P(A | B) = P(A \cap B) / P(B)$.

Определение 3.1. *Условная вероятность события A относительно события B (более длинная и устаревшая терминология – вероятность A при условии, что произошло B) определяется формулой*

$$P(A | B) = \frac{P(A \cap B)}{P(B)},$$

если $P(B) \neq 0$.

Последнее условие, касающееся вероятности события B , является весьма важным в данном определении условной вероятности, и мы пока не располагаем техническими возможностями для определения условной вероятности относительно события с нулевой вероятностью его осуществления. Становление теории вероятностей как математической дисциплины во многом было связано с попыткой дать корректное определение условной вероятности относительно любого элемента σ -алгебры, и именно это удалось сделать Андрею Николаевичу Колмогорову в 20-х годах XX столетия, что привело к исключительно бурному развитию стохастических дисциплин и расширению области их применения.

Введем теперь одно из важнейших понятий теории вероятностей, которое, по существу, выделяет ее в самостоятельную дисциплину из общей теории меры. Если оказывается, что условная вероятность события A относительно события B равна безусловной вероятности события A , то есть

$$P(A | B) = P(A \cap B) / P(B) = P(A),$$

то естественно сказать, что A *не зависит* от B . Оказывается, что в таком случае и B не зависит от A , то есть события A и B взаимно независимы, поскольку

$$P(B | A) = P(A \cap B) / P(A) = P(A)P(B) / P(A) = P(B)$$

и, в силу независимости A от B (см. предыдущее равенство), $P(A \cap B) = P(A)P(B)$. Итак, мы пришли к следующему определению взаимной независимости событий.

Определение 3.2. События A и B называются *независимыми*, если

$$P(A \cap B) = P(A)P(B).$$

Легко понять, что несовместные события зависимы. Действительно, справедливо

Предложение 3.1. Если A , B – несовместные события, причем $P(A) > 0$ и $P(B) > 0$, то $P(A \cap B) \neq P(A)P(B)$ (события A и B зависимы).

Доказательство. Для несовместных событий вероятность их одновременного появления $P(A \cap B) = P(\emptyset) = 0$, и, в то же время, в силу ненулевой вероятности появления каждого из событий, $P(A)P(B) \neq 0$.

Приведем пример независимых событий.

Пример 3.1. Обратимся к эксперименту с двукратным подбрасыванием правильной монеты (см. пример 1.2), в котором пространство элементарных исходов $\Omega = \{\Gamma\Gamma, \Gamma P, P\Gamma, PP\}$ наделяется равномерным распределением вероятностей: $p(\omega) = 1/4$ при любом $\omega \in \Omega$. Покажем, что выпадение герба при втором подбрасывании не зависит от того, что герб выпал при первом бросании монеты. Рассмотрим два события: $A = \{\Gamma\Gamma, \Gamma P\}$ – при первом бросании появляется герб и $B = \{\Gamma\Gamma, P\Gamma\}$ – второе испытание монеты закончилось выпадением герба, и покажем, что эти события независимы. Действительно,

$$P(A) = 1/2, \quad P(B) = 1/2, \quad P(A \cap B) = P(\Gamma\Gamma) = 1/4 = P(A)P(B).$$

Распространим теперь понятие независимости на совокупности событий.

Определение 3.3. События семейства $\mathcal{C} = \{A_i, i \in I\}$ называются *независимыми в совокупности* или *совместно независимыми*, если

$$P\left(\bigcap_{j=1}^k A_{i_j}\right) = \prod_{j=1}^k P(A_{i_j}),$$

каков бы ни был конечный набор событий A_{i_1}, \dots, A_{i_k} , $k \geq 2$, из совокупности \mathcal{C} .

Покажем, что *попарная* независимость событий:

$$P(A_i \cap A_j) = P(A_i)P(A_j),$$

если $i \neq j$, не влечет, вообще говоря, совместную независимость событий A_1, \dots, A_n .

Пример 3.2 (*Пирамидка Бернштейна*). Правильная четырехгранная пирамида, которая при бросании с одинаковой вероятностью, равной $1/4$, падает на любую из четырех граней, раскрашивается в три цвета. Одна грань покрывается красным цветом (элементарный исход $\omega_1 = \text{к}$), другая – зеленым ($\omega_2 = \text{з}$), третья – синим ($\omega_3 = \text{с}$), а четвертая ($\omega_4 = \text{м}$) делится на три части, каждая из которых закрашивается своим цветом – красным, зеленым и синим.

Рассмотрим три события: $A = \{\text{к}, \text{м}\}$ – пирамида упала гранью, содержащей красный цвет; $B = \{\text{з}, \text{м}\}$ – зеленый цвет; $C = \{\text{с}, \text{м}\}$ – синий цвет. Каждое из этих событий содержит по два равновероятных исхода, поэтому

$P(A) = P(B) = P(C) = 1/2$. Если эти события независимы, то, согласно определению 3.3, должно выполняться равенство

$$P(A \cap B \cap C) = P(A)P(B)P(C) = 1/8.$$

Однако в нашем случае одновременное осуществление всех трех событий возможно лишь при появлении единственного элементарного исхода $\omega_4 = m$, так что

$$P(A \cap B \cap C) = p(m) = 1/4 \neq 1/8.$$

Итак, события A , B и C зависимы.

В то же время события A , B и C попарно независимы. Действительно,

$$P(A \cap B) = p(m) = 1/4 = P(A)P(B),$$

и точно такие же равенства справедливы для остальных пар событий.

Распространим теперь понятие независимости на классы событий. Фиксируем некоторое вероятностное пространство (Ω, \mathcal{A}, P) и введем

Определение 3.4. Булевы подалгебры (или σ -подалгебры) $\mathcal{A}_1, \dots, \mathcal{A}_n$ булевой σ -алгебры \mathcal{A} называются *независимыми в совокупности*, если для *любого* набора событий A_1, \dots, A_n из соответствующих алгебр выполняется равенство

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i). \quad (1)$$

Заметим, что в случае булевых подалгебр формула (1), определяющая их независимость, содержит события, взятые одновременно из *всех* алгебр, – не рассматриваются всевозможные различные наборы подалгебр. Такие наборы в (1) получаются автоматически, если некоторые из $A_i = \Omega$, а достоверное событие Ω принадлежит всем подалгебрам σ -алгебры \mathcal{A} .

Пример 3.3 (независимых булевых подалгебр). Определение независимости булевых алгебр позволяет дать строгое математическое обоснование независимости результата очередного испытания правильной монеты от того, какими исходами закончились предыдущие испытания, или от того, что будет в будущем. Рассмотрим, как и в примере 1.5, статистический эксперимент, в котором регистрируются результаты n испытаний правильной монеты. Пусть \mathcal{A} – булева алгебра всевозможных подмножеств пространства Ω элементарных исходов этого эксперимента, число которых равно 2^n .

В соответствии с вероятностной моделью, обоснование которой было дано в §1, распределение вероятностей $\{P(A), A \in \mathcal{A}\}$ на булевой алгебре \mathcal{A} определяется следующим образом: $P(A)$ равна числу элементарных исходов, содержащихся в событии A , поделенному на 2^n . Рассмотрим n подалгебр $\mathcal{A}_1, \dots, \mathcal{A}_n$ булевой алгебры \mathcal{A} , где \mathcal{A}_i порождается противоположными событиями: A_i – при i -ом испытании выпал герб и A_i^c – выпала решка, то есть

$$\mathcal{A}_i = \{A_i, A_i^c, \Omega, \emptyset\}, \quad i = 1, \dots, n.$$

Покажем, что эти подалгебры независимы в совокупности.

Пусть B_1, \dots, B_n – некоторый набор элементов (событий) из соответствующих подалгебр. Требуется показать, что

$$P\left(\bigcap_{i=1}^n B_i\right) = \prod_{i=1}^n P(B_i). \quad (2)$$

Каждое из событий A_i или A_i^c состоит из 2^{n-1} исходов, поэтому $P(B_i) = 1/2$, если $B_i = A_i$ или A_i^c ; $P(B_i) = 1$, если $B_i = \Omega$, и $P(B_i) = 0$, если $B_i = \emptyset$. Таким образом, (2) выполняется тривиальным образом, если хотя бы одно из $B_i = \emptyset$; достоверные события $B_i = \Omega$ при доказательстве (2) можно просто игнорировать, так что осталось убедиться в справедливости (2), когда все B_i равны A_i или A_i^c , $i = 1, \dots, n$. Но в таком случае $\bigcap_{i=1}^n B_i$ совпадает с одним из элементарных исходов, вероятность которого равна 2^{-n} (значение левой части (2)), и то же значение принимает правая часть (2), поскольку все $P(B_i) = 1/2$, $i = 1, \dots, n$.

Рассмотренный пример указывает нам путь к построению вероятностной модели статистического эксперимента с независимыми испытаниями “гнутой” монеты, для которой вероятность выпадения герба отлична от $1/2$. Естественно, такого рода испытания осуществляются в практической и научной деятельности не только сгнутой монетой – испытания с бинарными исходами имеют место при контроле качества (изделия могут быть кондиционными и дефектными), эпидемиологических исследованиях (выбранная особь из популяции инфицирована или нет) и т.п. Общая теория экспериментов с бинарными исходами была разработана в XVII веке И.Бернулли и поэтому названа его именем.

Схема испытаний Бернулли. Эксперимент состоит в наблюдении $n(\geq 1)$ однотипных объектов, каждый из которых независимо от остальных объектов с одинаковой вероятностью p может обладать определенным

признаком или нет. Если i -й объект обладает указанным признаком, то говорят, что i -е испытание завершилось успехом, и в журнале наблюдений против i -го объекта ставится цифра 1; отсутствие признака (неудача) отмечается цифрой 0, $i = 1, \dots, n$. Таким образом, результат эксперимента можно представить в виде последовательности x_1, \dots, x_n наблюдений случайных индикаторов X_1, \dots, X_n – случайных величин, принимающих значение 1 с вероятностью p и 0 с вероятностью $1 - p$. В этих обозначениях вероятность того, что при i -м испытании X_i приняло значение x_i (равное 0 или 1), можно представить формулой

$$P(X_i = x_i) = p^{x_i}(1 - p)^{1-x_i}, \quad i = 1, \dots, n.$$

Как и в испытаниях правильной монеты, пространство Ω элементарных исходов рассматриваемого эксперимента состоит из 2^n элементов вида x_1, \dots, x_n . Пусть \mathcal{A} – булева алгебра всевозможных подмножеств Ω и $\mathcal{A}_1, \dots, \mathcal{A}_n$ – подалгебры \mathcal{A} , причем \mathcal{A}_i порождается событием $X_i = x_i$ (=0 или 1), $i = 1, \dots, n$. Если нам *a priori* известно, что как наблюдаемые объекты, так и результаты наблюдений над ними не оказывают влияния друг на друга, то естественно формализовать эту априорную информацию в виде утверждения: “подалгебры $\mathcal{A}_1, \dots, \mathcal{A}_n$ независимы в совокупности”. В таком случае вероятность каждого элементарного исхода x_1, \dots, x_n совпадает с вероятностью одновременного осуществления n независимых событий $X_i = x_i$, $i = 1, \dots, n$ и, следовательно,

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i) = p^{\sum_1^n x_i} (1 - p)^{n - \sum_1^n x_i}.$$

Полученное распределение вероятностей на пространстве элементарных исходов Ω обладает одной интересной особенностью: C_n^m исходов, содержащих одно и то же количество

$$m = \sum_1^n x_i$$

успешных испытаний, обладают одинаковой вероятностью их появления, равной $p^m(1 - p)^{n-m}$. Рассмотрим в связи с этим случайную величину

$$X = \sum_1^n X_i,$$

результат наблюдения которой m трактуется как число успешных испытаний в эксперименте. На пространстве значений $m = 0, 1, \dots, n$ этой случайной величины получаем распределение вероятностей, которое называется *биномиальным распределением*

$$P(X = m | p, n) = C_n^m p^m (1 - p)^{n-m}.$$

Отметим, что биномиальное распределение служит аппроксимацией гипергеометрического распределения при больших значениях N и M (см. задачу 3 и формулу 2 в §1). Имеет место

Предложение 3.2. *Если в гипергеометрическом распределении $P(X = m | N, M, n)$ параметры $N \rightarrow \infty$, $M \rightarrow \infty$ и при этом $M/N \rightarrow p$, то для всех фиксированных n и m*

$$P(X = m | N, M, n) \rightarrow P(X = m | p, n) = C_n^m p^m (1 - p)^{n-m}.$$

Доказательство легко получить, используя следующие элементарные преобразования гипергеометрической вероятности:

$$\begin{aligned} P(X = m | N, M, n) &= \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} = \\ &= \frac{M!}{m!(M-m)!} \cdot \frac{(N-M)!}{(n-m)!(N-M-(n-m))!} \cdot \frac{n!(N-n)!}{N!} = \\ &= \frac{n!}{m!(n-m)!} \cdot [(M-m+1) \cdots (M-1)M] \cdot \\ &= \frac{[(N-M-(n-m)+1) \cdots (N-M-1)(N-M)]}{(N-n+1) \cdots (N-1)N} = \\ &= C_n^m \left[\left(\frac{M}{N} - \frac{m-1}{N} \right) \cdots \left(\frac{M}{N} - \frac{1}{N} \right) \frac{M}{N} \right] \cdot \\ &= \left[\left(1 - \frac{M}{N} - \frac{n-m-1}{N} \right) \cdots \left(1 - \frac{M}{N} - \frac{1}{N} \right) \left(1 - \frac{M}{N} \right) \right] \cdot \\ &= \left[\left(1 - \frac{n-1}{N} \right) \cdots \left(1 - \frac{1}{N} \right) \cdot 1 \right]^{-1}. \end{aligned}$$

Следующие две формулы условной вероятности играют важную роль при решении многих практических задач. Обе формулы связаны с так называемой *полной группой событий* $\{B_1, \dots, B_n\}$, которые несовместны ($B_i \cap B_j = \emptyset, i \neq j$) и в объединении дают все пространство элементарных исходов Ω . Говорят, что эта группа событий определяет *разбиение* Ω , так как $\Omega = \sum_1^n B_i$.

Предложение 3.3 (Формула полной вероятности). Для любого события A и полной группы событий $\{B_1, \dots, B_n\}$ справедлива формула

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

Доказательство немедленно следует из следующей цепочки равенств, в которой на последнем этапе используется формула условной вероятности:

$$\begin{aligned} P(A) &= P(A \cap \Omega) = P\left(A \cap \sum_1^n B_i\right) = \\ &P\left(\sum_1^n A \cap B_i\right) = \sum_1^n P(A \cap B_i) = \sum_1^n P(A|B_i)P(B_i). \end{aligned}$$

Предложение 3.4 (Формула Байеса). Для любого события A и полной группы событий $\{B_1, \dots, B_n\}$ справедлива формула

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

Доказательство. В силу формулы условной вероятности

$$P(A \cap B_k) = P(A | B_k)P(B_k),$$

поэтому

$$P(B_k | A) = \frac{P(A \cap B_k)}{P(A)} = \frac{P(A | B_k)P(B_k)}{P(A)}.$$

Подставляя в правую часть последнего равенства вместо $P(A)$ ее выражение по формуле полной вероятности, получаем искомую формулу Байеса.

З а м е ч а н и е. Вероятности $P(B_1), \dots, P(B_n)$ часто называют *априорными* вероятностями группы событий B_1, \dots, B_n , в то время как условные вероятности $P(B_1 | A), \dots, P(B_n | A)$ – *апостериорными*, полученными после дополнительного эксперимента, в котором произошло событие A . В связи с этим формула Байеса называется также формулой *обновления априорных вероятностей*.

Приведем несколько задач, решаемых с помощью полученных формул условной вероятности.

З а д а ч а 3.1. Из урны, содержащей 3 белых и 2 черных шара, наугад вынимают 2 шара и перекладывают в другую урну, содержащую 4 белых и 4 черных шара. Какова вероятность иметь белый шар при случайном выборе одного шара из второй урны после перекладывания?

Эта задача решается обычно с помощью формулы полной вероятности. Пусть A – событие, означающее отбор белого шара. Определим полную группу событий в соответствии с возможными результатами перекладывания: $B_1 = \{Б, Б\}$, $B_2 = \{Б, Ч\} + \{Ч, Б\}$, $B_3 = \{Ч, Ч\}$. Здесь первая буква в фигурных скобках указывает цвет шара (Б – белый, Ч – черный), который был вынут из первой урны первым, а вторая буква – цвет второго шара. Термин “наугад” означает, что вероятность вынуть шар определенного цвета равна отношению числа шаров этого цвета к общему числу шаров в урне. В таком случае, в соответствии с формулой условной вероятности, $P(B_1) = P(Б \cap Б) = P(\text{второй шар белый} | \text{первый шар белый}) \cdot P(\text{первый шар белый}) = (3/5) \cdot (2/4) = 3/10$. Аналогично, $P(B_2) = (3/5) \cdot (2/4) + (2/5) \cdot (3/4) = 6/10$ и $P(B_3) = (2/5) \cdot (1/4) = 1/10$. Условные вероятности события A вычисляются в соответствии с числом белых шаров во второй урне после добавления в нее двух шаров из первой урны: $P(A | B_1) = 6/10$, $P(A | B_2) = 5/10$, $P(A | B_3) = 4/10$. Формула полной вероятности дает

$$P(A) = \frac{6}{10} \cdot \frac{3}{10} + \frac{5}{10} \cdot \frac{6}{10} + \frac{4}{10} \cdot \frac{1}{10} = \frac{13}{25}. \quad (3)$$

З а м е ч а н и е к задаче 3.1. Следует обратить особое внимание на пространство Ω элементарных исходов в этой задаче, – глубоко заблуждается тот, кто наделяет Ω всего двумя элементами Б и Ч. Наш эксперимент состоял не только в отборе шара из второй урны – перед этим производился случайный отбор двух шаров из первой урны, и результат этого отбора

влият на условную вероятность выбора белого шара. Пространство Ω в действительности состоит из восьми элементов

$$\begin{array}{cccc} \text{ББ.Б} & \text{БЧ.Б} & \text{ЧБ.Б} & \text{ЧЧ.Б} \\ \text{ББ.Ч} & \text{БЧ.Ч} & \text{ЧБ.Ч} & \text{ЧЧ.Ч} \end{array} \quad (3)$$

Здесь первые две буквы до точки указывают цвет шаров, вынутых из первой урны, а буква после точки – цвет шара, вынутого из второй урны после перекладывания. Вычисления, проводимые в (3), представляют собой суммирование вероятностей элементарных исходов, указанных в первой строке таблицы.

Следующая задача удивительно точно иллюстрирует недоразумения, которые могут возникнуть из-за неправильной спецификации пространства элементарных исходов.

Задача 3.2. Экспериментатор располагает двумя парами шаров одинакового цветового состава БЧ и ЧЧ. Из каждой пары наугад выбирается по одному шару и бросается в урну, где лежит белый шар. Из трех шаров в урне наугад отбирается один. Какова вероятность, что вынут белый шар?

Мы снова находимся в ситуации, связанной с применением формулы полной вероятности, где полная группа событий соотносится с возможным составом урны: $B_1 = \text{БББ}$ (в урне 3 белых шара), $B_2 = \text{ББЧ} + \text{БЧБ}$ (в урне 2 белых) и $B_3 = \text{БЧЧ}$ (в урне 1 белый). Поскольку вероятность выбора шара определенного цвета из каждой пары равна $1/2$ и выбор в каждой паре осуществляется независимо от результата выбора в другой, то вероятности событий из полной группы вычисляются очень просто: $P(B_1) = P(B_3) = (1/2) \cdot (1/2) = 1/4$, $P(B_2) = (1/2) \cdot (1/2) + (1/2) \cdot (1/2) = 1/2$. Условные вероятности отбора белого шара при каждом фиксированном составе урны равны $P(A | B_1) = 1$, $P(A | B_2) = 2/3$, $P(A | B_3) = 1/3$. Теперь, используя формулу полной вероятности, находим $P(A) = 1 \cdot (1/4) + (2/3) \cdot (1/2) + (1/3) \cdot (1/4) = 2/3$. Если игнорировать процесс случайного формирования состава урны и считать, что мы имеем дело с двухточечным пространством элементарных исходов $\Omega = \{\text{Б}, \text{Ч}\}$, то приходим к парадоксальному выводу: состав урны всегда один и тот же – два белых и один черный!

Нетрудно понять, что в этой задаче пространство элементарных исходов то же, что и в предыдущей задаче 3.1 (дополнительный белый шар фиксирован и его можно не учитывать при определении Ω), и наши вычисления $P(A)$ состоят в суммировании вероятностей элементарных исходов первой строки в таблице, представляющей пространство Ω .

Задача 3.3 *Статистический контроль качества*. Формула Байеса играет большую роль в планировании процедур гарантийного контроля качества выпускаемой продукции. Производитель продукта должен выполнять определенные договорные обязательства перед потребителем, которые, так или иначе, сводятся к ограничениям на долю некондиционной продукции, поставляемой потребителю, или, что то же, доля кондиционной продукции должна быть достаточно высокой. Обеспечение этих ограничений достигается с помощью контроля (как правило, выборочного) производимой продукции. Пусть Q_{in} – доля кондиционной продукции среди изготавливаемой предприятием. Обычно эта доля называется *входным уровнем качества*, и необходимость контроля продукции обуславливается невысоким значением Q_{in} , которое не удовлетворяет потребителя. Если контроль продукции производится на основе обследования только ее части (так называемый *выборочный* или *статистический* контроль качества), то возникает вероятность принятия ошибочного решения о качестве контролируемого продукта: с некоторой вероятностью β процедура контроля может пропустить некондиционный продукт или, наоборот, с вероятностью α отклонить кондиционный. Вероятность β называется *риском потребителя*, а вероятность α – *риском изготовителя*. Существуют методы расчета этих рисков на основе вероятностной модели статистического контроля, с которыми мы познакомимся в курсе математической статистики. Зная значения Q_{in} , α и β , можно, используя формулу Байеса, вычислить *выходной уровень качества* Q_{out} – долю кондиционной продукции среди отсылаемой потребителю после контроля.

Пусть B_1 – событие, состоящее в том, что поступивший на контроль продукт кондиционен, а $B_2 = B_1^c$ – продукт “плохой”. В наших обозначениях $P(B_1) = Q_{in}$. Пусть, далее, A – утверждение о кондиционности продукта после его контроля. Тогда $Q_{out} = P(B_1 | A)$ – вероятность кондиционности продукта при условии, что он прошел контроль. Наконец, $P(A | B_1) = 1 - \alpha$ и $P(A | B_2) = \beta$. По формуле Байеса

$$Q_{out} = P(B_1 | A) = \frac{P(A | B_1)P(B_1)}{P(A | B_1)P(B_1) + P(A | B_2)P(B_2)} = \frac{(1 - \alpha)Q_{in}}{(1 - \alpha)Q_{in} + \beta(1 - Q_{in})}.$$

Проиллюстрируем расчеты, производимые по этой формуле, на основе конкретных числовых данных. Пусть предприятие работает из рук вон

плохо: $Q_{in} = 0.1$ (90% выпускаемой продукции не удовлетворяет нормам качества), но на предприятии существует довольно жесткий контроль, в котором риск потребителя $\beta = 0.01$, а риск изготовителя $\alpha = 0.1$. Тогда выходной уровень качества

$$Q_{out} = \frac{0.9 \cdot 0.1}{0.9 \cdot 0.1 + 0.01 \cdot 0.9} = \frac{10}{11} \approx 0.91,$$

и это совсем неплохо по сравнению с тем, что было до контроля.

§4. Случайные величины и функции распределения

Лекция 6

В применениях методов теории вероятностей исследователь чаще всего имеет дело с числовыми характеристиками наблюдаемого объекта, которые являются функциями элементарных исходов – состояний объекта. При использовании различных характеристик важным является то обстоятельство, что все они определены на одном и том же пространстве Ω , и если мы приступаем к построению вероятностной модели, на основании которой будет получено распределение наблюдаемой характеристики $X = X(\omega)$, то мы должны понимать, что это распределение индуцировано исходным распределением P на σ -алгебре \mathcal{A} подмножеств Ω . Напомним, что такого рода построения проводились при выводе гипергеометрического и биномиального распределений.

Итак, мы приступаем к теории распределений функций $X = X(\omega)$ на пространстве элементарных исходов, фиксируя некоторое вероятностное пространство (Ω, \mathcal{A}, P) . Областью значений функции X служит евклидово пространство \mathbb{R} , и это пространство является новым пространством элементарных исходов. Поскольку нас, в основном, будут интересовать вероятности попадания значений X в интервалы, то естественно рассмотреть борову σ -алгебру подмножеств \mathbb{R} , порожденную всевозможными интервалами на прямой \mathbb{R} . Как нам известно из общего курса анализа, такая σ -алгебра \mathcal{B} , состоящая из всевозможных объединений и пересечений счетного числа интервалов, называется *борелевским полем*, и для ее построения достаточно рассмотреть открытые интервалы вида $(-\infty, x)$.

Введем измеримое пространство $(\mathbb{R}, \mathcal{B})$ значений X и рассмотрим следующий, совершенно естественный метод “наведения” распределения P^X на \mathcal{B} посредством вероятности P на \mathcal{A} . Каждому борелевскому множеству $B \in \mathcal{B}$ сопоставим его *прообраз* $X^{-1}(B) = \{\omega : X(\omega) \in B\} \subset \Omega$. Если $X^{-1}(B) \in \mathcal{A}$, то естественно определить вероятность попадания значения X в B как $P^X(B) = P(X^{-1}(B))$. Функции, которые обладают свойством $X^{-1}(B) \in \mathcal{A}$ при любом $B \in \mathcal{B}$, называются *измеримыми*, и в дальнейшем будут рассматриваться только такие характеристики наблюдаемого объекта. Мы подошли к основному понятию теории распределений на подмножествах \mathbb{R} .

Определение 4.1. *Случайной величиной* $X = X(\omega)$ называется измеримое отображение измеримого пространства (Ω, \mathcal{A}) на борелевскую

прямую $(\mathbb{R}, \mathcal{B})$.

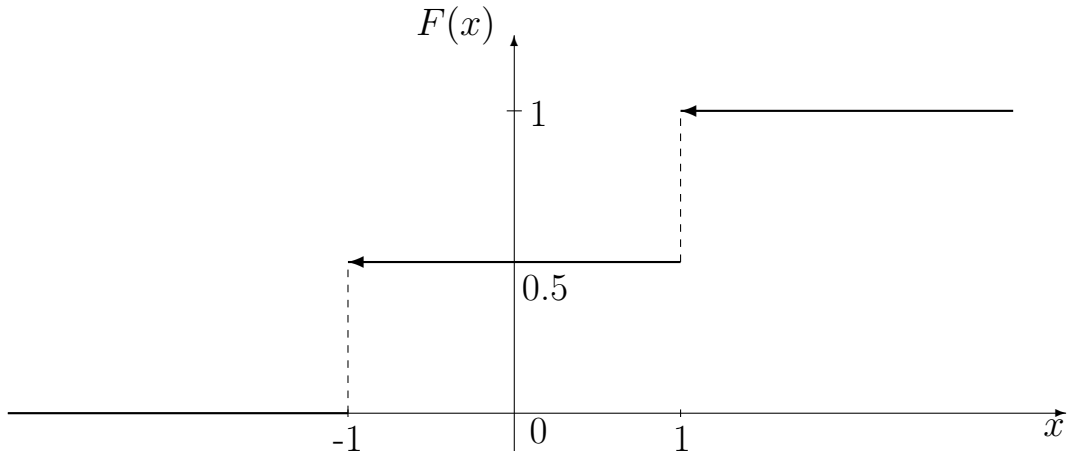
Легко понять, что, с точки зрения практических приложений, мы могли бы не обращаться к определению случайной величины как измеримой функции, а просто сказать, что сейчас мы займемся построением вероятностных моделей, в которых пространство элементарных исходов есть числовая прямая. Тем не менее, чтобы описать класс возможных распределений наблюдаемой случайной величины X иногда просто необходимо знать причину изменчивости состояний объекта (или инструмента исследования), которая обуславливает разные значения в повторных наблюдениях X .

Борелевское поле \mathcal{B} , на котором будет определяться распределение X , является чрезвычайно сложным объектом с точки зрения строения его элементов, поэтому задание функции $P(B)$, $B \in \mathcal{B}$ представляется совершенно неразрешимой проблемой. Однако мы знаем, что \mathcal{B} порождается интервалами вида $(-\infty, x)$ (событиями $X < x$), и это указывает простой путь к заданию распределения случайной величины X . Что если начать с задания вероятности только на событиях, порождающих \mathcal{B} , то есть с определения функции $F(x) = P(X < x)$, $x \in \mathbb{R}$, потом распространить ее аддитивным образом на булеву алгебру конечных объединений всевозможных интервалов на \mathbb{R} , показать, что полученная таким образом аддитивная функция на булевой алгебре обладает свойством непрерывности относительно монотонно убывающих последовательностей событий (является вероятностью), и, наконец, закончить построение вероятности на \mathcal{B} ссылкой на теорему об единственности продолжения вероятности с булевой алгебры объединений интервалов на порожденную этой алгеброй σ -алгебру борелевских подмножеств \mathbb{R} .

Мы приступаем к реализации этой программы и введем сначала

Определение 4.2. Функция $F(x) = P(X < x)$, определенная на всей числовой прямой \mathbb{R} , называется *функцией распределения* случайной величины X .

Пример 4.1. Пусть случайная величина X принимает с ненулевой вероятностью всего два значения: $x = -1$ с вероятностью $1/2$ и $x = +1$ с той же вероятностью $1/2$ (игра в орлянку со ставкой 1 рубль). Тогда функция распределения X имеет следующий вид.



Действительно, для любого $x < -1$ множество $(-\infty, x)$ не содержит значений X , которые она могла бы принять с положительной вероятностью, так что $F(x) = P(X < x) = 0$. Далее, $F(-1) = P(X < -1) = 0$, но если $-1 < x \leq +1$, то $F(x) = P(X = -1) = 1/2$. В области $x > +1$ содержатся все значения случайной величины X , которые она принимает с положительной вероятностью, поэтому $F(x) = 1$ при $x > +1$.

Исследуем некоторые особенности поведения функции F .

Предложение 4.1. *Функция $F(x)$, $x \in \mathbb{R}$ обладает следующими свойствами.*

$$(F1) \quad \lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1.$$

$$(F2) \quad F(x) \text{ — неубывающая функция } x \in \mathbb{R}.$$

$$(F3) \quad \text{Функция } F(x) \text{ непрерывна слева: } \lim_{x \rightarrow a-} F(x) = F(a).$$

(F4) *Вероятности попадания значений случайной величины X в интервалы на \mathbb{R} вычисляются по формулам*

$$P\{X \in [a, b]\} = F(b) - F(a), \quad P\{X \in [a, b]\} = F(b+) - F(a),$$

$$P\{X \in (a, b]\} = F(b+) - F(a+), \quad P\{X \in (a, b)\} = F(b) - F(a+).$$

(F5) *Функция $F(x)$ имеет не более чем счетное множество скачков.*

Доказательство. (F1). Рассмотрим последовательность событий

$$\{A_n = (-\infty, x_n), n \geq 1\}.$$

Если $x_n \searrow -\infty$ при $n \rightarrow \infty$, то, очевидно, $A_n \downarrow \emptyset$, и, аналогично,

$$A_n \uparrow R(= \Omega),$$

если $x_n \nearrow +\infty$. Так как

$$F(x_n) = P(X < x_n) = P(A_n),$$

то свойства (F1) вытекают из аксиомы непрерывности (P3).

(F2). Если $x_1 \leq x_2$, то $F(x_1) \leq F(x_2)$, так как

$$A_1 = (-\infty, x_1) \subset A_2 = (-\infty, x_2)$$

и, в силу свойства монотонности вероятности (см. (3) в предложении 2.2),

$$F(x_1) = P(A_1) \leq P(A_2) = F(x_2).$$

(F3). Пусть последовательность $x_n \uparrow x$ при $n \rightarrow \infty$, так что соответствующая последовательность событий

$$A_n = (-\infty, x_n) \uparrow A = (-\infty, x).$$

Используя свойство (P3) непрерывности P , получаем

$$F(x_n) = P(A_n) \rightarrow P(A) = F(x),$$

что, по определению, означает непрерывность слева функции $F(x)$.

(F4). Если $A \subset B$, то $P(B \setminus A) = P(B) - P(A)$ (см. (3) в предложении 2.2). Из этого свойства вероятности и только что доказанного свойства непрерывности вытекает, что, например, замкнутый интервал $[a, b] = (-\infty, b] \setminus (-\infty, a)$, и поскольку множество $(-\infty, a) \subset (-\infty, b]$, то

$$P\{X \in [a, b]\} = P\{X \in (-\infty, b]\} - P\{X \in (-\infty, a)\} =$$

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

В последнем равенстве мы использовали запись $F(b+)$ для выражения вероятности события $\{X \leq b\}$. Дело в том, что $P(X < b) = F(b)$, и если в точке b функция $F(x)$ имеет скачок, то его величина равна $F(b+) - F(b)$.

(F5). В этом пункте предложения утверждается, что все скачки (точки разрыва) функции $F(x)$ можно занумеровать. Поступим следующим образом: рассмотрим последовательность множеств $\{A_n, n \geq 1\}$, где A_n

есть множество точек разрыва функции $F(x)$ с величиной скачка, не меньшей $1/n$. Поскольку $0 \leq F(x) \leq 1$, то множество A_n конечно и содержит не более чем n точек. Следовательно, мы можем занумеровать все скачки функции $F(x)$ в порядке убывания их величины, осуществляя последовательную нумерацию точек множества A_1 , потом A_2 и так далее, возможно, до бесконечности, если число скачков $F(x)$ не конечно. При таком способе нумерации любому, сколь угодно малому по величине скачку функции $F(x)$ рано или поздно будет присвоен номер.

Итак, мы убедились, что функция распределения является хорошим и достаточно простым инструментом для вычисления вероятностей попадания значений случайной величины в интервалы на действительной прямой. Однако, если мы определим только вероятности элементов борелевского поля \mathcal{B} , имеющих вид интервалов, то сможем ли на основании их вычислять вероятности других событий из \mathcal{B} ? Ответ на этот вопрос дает

Теорема 4.1. Пусть функция $F(x)$, $x \in \mathbb{R}$, обладает свойствами

$$(F1) \quad \lim_{x \rightarrow -\infty} F(x) = 0, \quad \lim_{x \rightarrow +\infty} F(x) = 1;$$

$$(F2) \quad F(x) \text{ — неубывающая функция } x \in \mathbb{R};$$

$$(F3) \quad F(x) \text{ непрерывна слева: } \lim_{x \rightarrow a-} F(x) = F(a).$$

Тогда на борелевской прямой $(\mathbb{R}, \mathcal{B})$ существует единственная вероятность P , для которой $P\{(-\infty, x)\} = F(x)$ для всех $x \in \mathbb{R}$.

Доказательство. Функция $F(x)$ определяет функцию множеств P' на семействе \mathcal{C} открытых интервалов вида $C = C_x = (-\infty, x)$ посредством равенства $P'(C_x) = F(x)$, причем в силу свойства (F1),

$$P'(\Omega) = P'(\mathbb{R}) = 1.$$

Распространим эту функцию множеств на булеву алгебру $\mathcal{A} = \mathcal{A}(\mathcal{C})$, порожденную семейством \mathcal{C} . Элементы A булевой алгебры \mathcal{A} очевидно имеют вид

$$A = \sum_1^k [a_i, b_i),$$

и поэтому естественно положить (см. (F4) в предложении 4.1)

$$P'(A) = \sum_1^k [F(b_i) - F(a_i)].$$

Очевидно, функция множеств $P'(A)$ на булевой алгебре \mathcal{A} обладает такими свойствами вероятности, как нормируемость ($P1$) и конечная аддитивность ($P2$). Если мы покажем, что P' обладает свойством σ -аддитивности $P(2')$, то утверждение теоремы будет простым следствием общей теоремы о продолжении меры на порожденную булевой алгеброй σ -алгебру, ибо, как известно, борелевское поле порождается алгеброй \mathcal{A} (более того, — семейством \mathcal{C}).

Рассмотрим произвольную последовательность не пересекающихся множеств

$$A_n = \sum_{i=1}^{k_n} [a_{ni}; b_{ni}), \quad n = 1, 2, \dots,$$

для которой множество $A = \sum_1^{\infty} A_n$ принадлежит алгебре \mathcal{A} . По определению алгебры \mathcal{A} это означает, что множество A можно представить в виде конечного объединения интервалов, не имеющих точек соприкосновения: $A = \sum_1^m [c_j; d_j)$. После соответствующей перестановки интервалов внутри объединения множеств A_n , можно добиться для каждого из интервалов $[c_j; d_j)$ представления вида

$$[c_j; d_j) = \sum_1^{\infty} [a_{jk}; b_{jk}), \quad j = 1, \dots, m.$$

Таким образом, достаточно доказать, что для любых $c < d$

$$F(d) - F(c) = \sum_{j=1}^{\infty} [F(b_j) - F(a_j)], \quad (1)$$

если интервал

$$[c; d) = \sum_1^{\infty} [a_j; b_j).$$

Очевидно,

$$F(d) - F(c) \geq \sum_1^n [F(b_j) - F(a_j)],$$

ибо дополнение множества

$$\sum_1^n [a_j; b_j)$$

до интервала $[c, d)$ можно представить в виде конечного объединения не пересекающихся полуоткрытых интервалов. Устремляя n к бесконечности, получаем

$$F(d) - F(c) \geq \sum_1^{\infty} [F(b_j) - F(a_j)].$$

Покажем теперь, что имеет место противоположное неравенство, и, следовательно, справедливо равенство (1).

Предположим сначала, что $-\infty < c < d < \infty$. Выберем произвольное $\varepsilon > 0$. Исходный интервал $[c; d)$ сузим до замкнутого интервала $[c; d']$ так, чтобы $d' < d$ и $F(d') \geq F(d) - \varepsilon$. Этого всегда можно добиться в силу непрерывности слева функции F . Аналогично, каждый из интервалов $[a_n; b_n)$ расширим до открытого интервала $(a'_n; b_n)$ так, чтобы $a'_n < a_n$ и

$$F(a'_n) \geq F(a_n) - \varepsilon/2^n.$$

В результате получим покрытие

$$[c; d'] \subset \bigcup_{n=1}^{\infty} (a'_n; b_n)$$

ограниченного замкнутого множества семейством открытых интервалов.

В силу известной леммы Гейне-Бореля найдется конечное покрытие

$$[c; d'] \subset \bigcup_{i=1}^N (a'_{n_i}; b_{n_i}),$$

в котором $a'_{n_1} < c$, $b_{n_N} > d'$ и $b_{n_{i-1}} > a'_{n_i}$ для всех $i = 2, \dots, N$. Точки $b_{n_1}, \dots, b_{n_{N-1}}$ образуют разбиение интервала $[a'_{n_1}, b_{n_N})$, который содержит интервал $[c, d')$, и поэтому

$$\begin{aligned} F(d') - F(c) &\leq F(b_{n_N}) - F(a'_{n_1}) = F(b_{n_1}) - F(a'_{n_1}) + \\ &\sum_{i=2}^N [F(b_{n_i}) - F(b_{n_{i-1}})] \leq \sum_{i=1}^N [F(b_{n_i}) - F(a'_{n_i})] \leq \sum_{n=1}^{\infty} [F(b_n) - F(a'_n)]. \end{aligned}$$

Из построения интервалов следует, что

$$F(d) - F(c) \leq F(d') - F(c) + \varepsilon$$

и

$$F(b_n) - F(a'_n) \leq F(b_n) - F(a_n) + \varepsilon/2^n,$$

откуда

$$F(d) - F(c) \leq \sum_{n=1}^{\infty} [F(b_n) - F(a_n)] + 2\varepsilon. \quad (2)$$

Устремляя $\varepsilon \rightarrow 0$, получаем окончательное доказательство равенства (1) для конечных интервалов.

Для бесконечных интервалов вида $[c; \infty)$ достаточно, воспользовавшись свойствами функции F , рассмотреть конечный интервал $[c; d)$, удовлетворяющий условию $1 - F(d) \leq \varepsilon$. Открытое покрытие исходного интервала $[c; \infty)$ индуцирует естественным образом открытое покрытие интервала $[c; d)$, к которому применимы все предыдущие рассуждения, приводящие к неравенству (2). Очевидно, разность $1 - F(c)$ не превосходит правой части (2) с заменой 2ε на 3ε , что завершает доказательство теоремы.

Доказанная теорема позволяет нам вычислять вероятности событий с помощью интеграла Стильтьеса:

$$P(A) = \int_A dF(x), \quad A \in \mathcal{A}.$$

newpage

§5. Построение вероятностных моделей с помощью функций распределения

Лекция 7

В этом параграфе мы будем решать несколько практических задач на построение вероятностных моделей, цель которого состоит в спецификации наиболее узкого семейства возможных распределений наблюдаемой случайной величины X . Эти модели носят универсальный характер и применяются в различных областях науки и практической деятельности, поэтому целесообразно после решения каждой задачи рассмотреть возможные аналоги этих задач, приводящие к тем же вероятностным моделям. Каждой модели мы присвоим свое имя и аббревиатуру, содержащую “параметры” модели; значения параметров, как правило, неизвестны, и определение этих значений составляет предмет другой, родственной теории вероятностей, науки – математической статистики.

Собственно говоря, мы уже давно занимаемся построением вероятностных моделей с помощью функций распределений случайных величин, принимающих дискретный ряд значений, – речь идет о гипергеометрическом и биномиальном распределениях (см. §1 и §3). Вероятность, с которой случайная величина принимала конкретное целочисленное значение m , равна величине скачка функции распределения в точке $x = m$. С этих двух распределений мы начнем составление нашего каталога вероятностных моделей.

Гипергеометрическое распределение $GG(N, M, n)$. Исследуется конечная популяция, состоящая из N единиц, часть из которых (M единиц) помечены. Из популяции извлекается случайная выборка объема n , и с этой выборкой соотносится случайная величина X , наблюдаемое значение x которой указывает число помеченных единиц в выборке. Область значений X , которые она принимает с ненулевой вероятностью, составляют целочисленные точки отрезка

$$\mathcal{X} = [\max(0, n - (N - M)); \min(n, M)].$$

Функция распределения $F(x)$ равна нулю в области, лежащей слева от \mathcal{X} , и как только x выходит на правый конец отрезка \mathcal{X} , функция $F(x)$ принимает значение 1, которое сохраняется при всех x , лежащих справа от \mathcal{X} . Внутри отрезка \mathcal{X} функция распределения имеет ступенчатый вид, возрастая скачками в целочисленных точках $x = m$, и величина скачка определяется формулой (2) §1:

$$P(X = m) = F(m+) - F(m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}, \quad m \in \mathcal{X}.$$

Переменные N , M и n являются *параметрами* модели; в практических приложениях модели $GG(N, M, n)$ значение по крайней мере одного из параметров N или M неизвестно. Область возможных значений параметров составляет так называемое *параметрическое пространство* и обозначается обычно Θ . В данном случае Θ состоит из целочисленных значений параметров N , M и n , причем $N \geq 2$, $1 \leq M \leq N$, $1 \leq n \leq N$.

Итак, дискретная вероятностная модель гипергеометрического распределения полностью определяется “функцией скачков”

$$f(x | \theta), x \in \mathbb{R}, \theta = (N, M, n) \in \Theta,$$

которая принимает ненулевые значения $P(X = x)$ только в целочисленных точках x отрезка \mathcal{X} . Функция $f(\cdot | \theta)$ обычно называется *функцией плотности* распределения случайной величины X . Вероятность события вида $X \in B$ ($B \in \mathcal{B}$) вычисляется с помощью f по формуле

$$P(X \in B) = \sum_{x \in B} f(x | \theta),$$

в частности, функция распределения

$$F(x) = \sum_{t < x} f(t | \theta).$$

Биномиальное распределение $V(n, p)$. Рассматривается схема независимых испытаний, каждое из которых с некоторой вероятностью p может быть “успешным” (в результате испытания осуществилось некоторое событие A) или, с вероятностью $1 - p$, “неудачным”. Нас интересует распределение случайной величины X , результат x наблюдения которой регистрирует число успехов в n испытаниях Бернулли.

Как было установлено в §3, распределение X определяется функцией плотности $f(x | \theta)$, принимающей ненулевые значения

$$C_n^x p^x (1 - p)^{n-x} (= P(X = x))$$

только в точках $x = 0, 1, \dots, n$, в то время как двумерный параметр $\theta = (n, p)$ может изменяться в области $\Theta = \mathcal{N} \times [0; 1]$, где $\mathcal{N} = \{1, 2, \dots\}$ – множество натуральных чисел. Поведение биномиальной функции распределения аналогично поведению $F(x)$ в модели $GG(N, M, n)$, если считать, что отрезок $\mathcal{X} = [0; n]$.

В практических применениях биномиального распределения обычно неизвестно только значение параметра p – вероятности успеха в испытаниях Бернулли. Однако существуют ситуации, когда экспериментатор регистрирует только число успехов x , не имея сведений о числе испытаний n . Например, в исследованиях нервного синапса прибор регистрирует только общее напряжение электрического поля, и по величине этого напряжения определяется количество x пузырьков с ацетилхолином, освободившихся при раздражении нерва. Ни общее количество n пузырьков, ни вероятность p выброса из пузырька ацетилхолина, экспериментатору неизвестны, – проблема оценки параметров n и p составляет предмет исследования.

Особо следует отметить частный случай биномиального распределения с одним испытанием ($n = 1$) в схеме Бернулли. Это так называемое *двухточечное* распределение вероятностей $V(1, p)$ с функцией плотности

$$f(x | p) = p^x (1 - p)^{1-x}, \quad x = 0, 1.$$

В §3 было установлено, что модель $V(n, p)$ является “предельной” для модели $GG(N, M, n)$, когда размер N популяции неограниченно растет и число M помеченных единиц соизмеримо с N , то есть $M/N = p (= \text{const})$. Следующая вероятностная модель, имеющая широкие практические применения, является предельной для биномиальной модели, когда число проводимых испытаний n велико, а вероятность p успешного испытания чрезвычайно мала.

Распределение Пуассона $P(\lambda)$. При исследовании интенсивности радиоизлучения обычно регистрируется число x атомов радиоактивного элемента, распавшихся за единицу времени. Повторные наблюдения указывают на значительную изменчивость числа распавшихся атомов, и поэтому проблема стабильного, не зависящего от случайных флуктуаций, показателя интенсивности излучения должна решаться в рамках теории вероятностей.

Пусть X – случайная величина, которая наблюдается в эксперименте, n – число атомов, из которых состоит образец исследуемого радиоактивного элемента, p – вероятность, с которой возможен распад любого из атомов образца за время наблюдения. Существующая теория радиоактивного излучения утверждает, что атомы распадаются независимо друг от друга, и поэтому результат x , который фиксирует счетчик распавшихся атомов, можно трактовать как реализацию случайной величины X с биномиальным законом $B(n, p)$ распределения вероятностей. Легко понять, что расчет вероятностей исходов эксперимента по формуле

$$P(X = x) = C_n^x p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n$$

ряд ли возможен из-за непреодолимых технических сложностей, вызванных огромным значением n и ничтожно малым значением p . Поэтому возникает математическая проблема асимптотики биномиальных вероятностей, когда $n \rightarrow \infty$ и одновременно $p \rightarrow 0$. Решение проблемы дает

Предложение 5.1. Если $n \rightarrow \infty$, $p \rightarrow 0$ и при этом $np = \lambda$ ($= \text{const}$), то

$$P\{X = x | n, p\} \longrightarrow \frac{\lambda^x e^{-\lambda}}{x!}.$$

Доказательство. Предельное значение биномиальных вероятностей легко получить, если представить их в виде

$$P\{X = x | n, p\} = \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} =$$

$$\left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{x-1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^{-x} \cdot \left(1 - \frac{\lambda}{n}\right)^n \frac{\lambda^x}{x!}$$

и воспользоваться замечательным пределом

$$(1 - \lambda/n)^n \rightarrow e^{-\lambda}.$$

Этот асимптотический результат впервые был получен Пуассоном, и поэтому распределение вероятностей

$$P(X = x | \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, \dots, \quad (1)$$

называется *распределением Пуассона* и обозначается $P(\lambda)$. Правая часть (1) представляет ненулевые значения функции плотности $f(x | \lambda)$ распределения Пуассона, $\lambda (> 0)$ называется параметром *интенсивности потока Пуассона* – в терминах задачи с радиоактивным распадом λ равно среднему числу атомов, распавшихся за единицу времени. Функция распределения Пуассона равна нулю на отрицательной полуоси, а на положительной возрастает скачками в целочисленных точках $x = 0, 1, \dots$, величина которых равна правой части (1).

Трудно переоценить значимость закона Пуассона в различных проблемах естествознания. Это распределение используется при исследовании числа несчастных случаев на предприятиях, числа вызовов на телефонной станции; этому закону подчиняются метеорные явления, потоки транспорта, размеры очередей систем обслуживания и пр.

Равномерное распределение $U(a, b)$. На отрезок $[0; 1]$ “наугад” бросается точка, так что вероятность ее попадания в любой интервал $(\alpha; \beta) \in [0; 1]$ зависит только от длины $\beta - \alpha$ интервала и не зависит от его положения внутри отрезка $[0; 1]$. Экспериментатора интересует распределение случайной величины X , реализующей координату x точки после бросания.

Ключ к выводу функции распределения X указывает следующая эквивалентная формулировка условий эксперимента: интервалы одинаковой длины обладают одинаковой вероятностью попадания в них бросаемой точки. Если разделить отрезок $[0; 1]$ на n одинаковых частей, то для функции распределения X имеет место двусторонняя оценка:

$$\frac{[nx]}{n} \leq F(x) \leq \frac{[nx] + 1}{n},$$

где $[t]$ – целая часть t . Действительно, всем отрезкам, полученным в результате деления $[0; 1]$, соответствует одинаковая вероятность, равная $1/n$ попадания в них точки, так что вероятность $P(X < x) = F(x)$ можно оценить количеством отрезков длины $1/n$, покрывающих $[0; x]$. Устремляя теперь n к бесконечности, получаем, что $F(x) = x$, если $x \in [0; 1]$. Поскольку вероятность попадания точки во внешность отрезка $[0; 1]$ равна нулю, то $F(x) = 0$ при $x < 0$ и $F(x) = 1$ при $x > 1$.

Итак, мы построили вероятностную модель равномерного распределения $U(0, 1)$ на отрезке $[0, 1]$. Легко понять, что если аналогичный эксперимент проводится с отрезком $[0, b]$, то функция распределения на этом отрезке будет иметь вид $F(x) = x/b$, так как свойство линейности должно сохраняться в силу принципа случайности бросания точки на отрезок $[0, b]$, и, в то же время, $F(b+) = 1$. Наконец, если точка бросается на отрезок общего вида $[a, b]$, то $F(a) = 0$, $F(b+) = 1$, и поэтому $F(x) = (x - a)/(b - a)$. Таким образом, мы пришли к равномерному распределению $U(a, b)$ на отрезке $[a, b]$. Это распределение зависит от двумерного параметра $\theta = (a, b)$ с областью значений (параметрическим пространством)

$$\Theta = \{(a, b) \in \mathbb{R}^2 : a < b\}.$$

Равномерное распределение имеет интересную связь с последовательностью испытаний Бернулли. Если представить реализацию x случайной величины X с распределением $U(0, 1)$ в виде двоичной дроби, то ее дробная часть реализует последовательность индикаторов успеха в бесконечной последовательности испытаний Бернулли с $p = 1/2$. Легко проверить, что справедливо и обратное утверждение, что дает один из простейших способов генерирования случайных величин с равномерным законом распределения.

Лекция 8

Показательное распределение $E(\theta)$. Вы, наверное, обратили внимание, что большинство, по крайней мере, “серьезных” изделий, которые выпускают предприятия, снабжается гарантийным сроком службы t_0 , и если изделие отказывает до момента t_0 , то предприятие несет определенные убытки, связанные с ремонтом или заменой изделия. Естественно, долговечность x (или, как говорят англичане, “срок жизни” – lifetime) является реализацией случайной величины X , и только знание ее функции распределения $F(x)$ позволит предприятию установить тот гарантийный срок службы, который отвечает его финансовым возможностям по обеспечению ремонта или замены. Для расчета t_0 необходимо определиться с требуемой надежностью изделия P_0 – “средней” долей изделий, которые обязаны отработать гарантийное время. Зная надежность P_0 , мы находим гарантийный срок t_0 из уравнения

$$P(X \geq t_0) = 1 - F(t_0) = P_0.$$

В связи с этим функция $H(t) = 1 - F(t)$, $t \geq 0$, называется *функцией надежности*.

Обычно построение модели надежности изделия опирается на некоторые постулаты, связанные с функционированием изделия, его старением, износом, подверженностью ударным нагрузкам и т.п. Мы рассмотрим сейчас один из таких постулатов применительно к изделиям, которые отказывают не в силу процессов старения, а только по причине резко возросших (так называемых “ударных”) нагрузок на режим его работы. Естественно, в такой ситуации вероятность того, что изделие прослужит еще некоторое время t при условии, что оно уже отслужило срок s , не должна зависеть от s , то есть

$$P\{X \geq t + s \mid X \geq s\} = \frac{P(\{X \geq t + s\} \cap \{X \geq s\})}{P(X \geq s)} =$$

$$\frac{P(X \geq t + s)}{P(X \geq s)} = P(X \geq t).$$

Таким образом, функция надежности $H(t)$ изделия должна удовлетворять функциональному уравнению

$$H(t + s) = H(t)H(s), \quad t \geq 0, \quad s \geq 0. \quad (2)$$

Предложение 5.2. Если функция $H(t)$, $t \geq 0$ удовлетворяет краевым условиям

$$\lim_{t \rightarrow 0} H(t) = 1, \quad \lim_{t \rightarrow \infty} H(t) = 0$$

и непрерывна слева, то все решения уравнения (2) имеют вид

$$H(t) = e^{-\lambda t},$$

где $\lambda > 0$ – произвольный параметр.

Доказательство. Из уравнения (2) легко вывести, что для любого $c > 0$ и любого целого $n \geq 1$ имеет место соотношение

$$H(nc) = H^n(c). \quad (3)$$

Действительно, в силу (2), используя индукцию, получаем

$$H(nc) = H((n - 1)c + c) = H((n - 1)c)H(c) =$$

$$= H((n-2)c)H^2(c) = \dots = H^n(c).$$

Далее, для любых $c > 0$ и целого $m \geq 1$ справедливо равенство

$$H(c/m) = H^{1/m}(c), \quad (4)$$

которое немедленно следует из (3):

$$H(c) = H(mc/m) = H^m(c/m).$$

Соотношения (3) и (4) позволяют установить строгое неравенство $0 < H(1) < 1$. Действительно, если допустить противное: $H(1) = 0$, то в силу (4) для любого целого $m \geq 1$ получаем

$$H(1/m) = H^{1/m}(1) = 0.$$

Устремляя m к бесконечности и используя свойство непрерывности H в нуле, получаем противоречие

$$1 = H(0) = \lim_{m \rightarrow \infty} H(1/m) = 0.$$

Аналогично, если предположить, что $H(1) = 1$, то, в силу (3), для любого целого n $H(n) = H^n(1) = 1$ и, в то же время,

$$\lim_{n \rightarrow \infty} H(n) = 0.$$

Неравенство $0 < H(1) < 1$ означает, что существует такое $\lambda > 0$, что $H(1) = e^{-\lambda}$. Но тогда, в силу (3) и (4), для любых целых n и m имеем

$$H(n) = e^{-n\lambda}, \quad H(n/m) = H^{1/m}(n) = \exp\{-n\lambda/m\}.$$

Это означает, что наше предположение доказано для всех рациональных t . Любое другое значение t на положительной полуоси можно сколь угодно точно оценить снизу рациональным числом и затем воспользоваться непрерывностью слева $H(t)$ при переходе в оценке t к пределу.

Итак, мы нашли функцию распределения случайной величины X , реализующую долговечность изделия,

$$F(x) = 1 - H(x) = 1 - \exp\{-\lambda x\}$$

в области $x > 0$. Как будет показано в дальнейшем, это распределение тесно связано с распределением Пуассона и параметр λ , как и в модели

$P(\lambda)$, характеризует *интенсивность потока отказов*. Однако в теории вероятностей обычно модель показательного распределения параметризуется иным способом, через параметр $\theta = 1/\lambda$, который имеет смысл средней долговечности. Таким образом, показательное распределение $E(\theta)$, которое будет в дальнейшем рассматриваться, имеет функцию распределения $F(x) = 0$ при $x \leq 0$ и

$$F(x) = 1 - \exp\{-x/\theta\},$$

если $x > 0$.

Мы завершим этот параграф построением еще одной дискретной модели теории надежности, в которой прослеживаются первые, пока еще очень смутные, связи пуассоновского и показательного распределений.

Геометрическое распределение $\text{Geo}(p)$. При посадке воздушного лайнера возможен сильный удар о посадочную полосу, который может привести к разрушению шасси. Пусть p – вероятность грубой посадки; нас интересует вероятность того, что шасси не будет разрушено до момента t ($\gg 1$) (надежность шасси).

С подобной задачей мы имели дело в §1 (пример 6), когда определяли вероятность первого появления герба при n -м испытании правильной монеты ($p = 1/2$). В данном, более общем случае естественно воспользоваться предположением о независимости ситуаций, возникающих при каждой посадке лайнера. Пусть X – случайная величина, принимающая значения $x = 1, 2, \dots$, которые указывают момент разрушения шасси, точнее, номер посадки, которая оказалась грубой. Тогда событие $X = x$ состоит из $x - 1$ благополучных посадок и грубой посадки с номером x , откуда находим функцию плотности *геометрического* распределения $\text{Geo}(p)$:

$$f(x | p) = P(X = x) = (1 - p)^{x-1}p,$$

если $x \in \mathbb{N}$, и $f(x | p) = 0$ в остальных точках вещественной оси \mathbb{R} .

В дискретной функции надежности

$$H(t) = P(X \geq t) = \sum_{x=t}^{\infty} (1 - p)^{x-1}p = (1 - p)^{t-1}, t \geq 1$$

практический интерес представляют очевидно малые значения p и большие значения t . Найдем асимптотику $H(t)$, положив $p = \lambda/N$, $t = Nx$ и

устремив N к бесконечности. Имеем

$$H(Nx) = \left(1 - \frac{\lambda}{N}\right)^{Nx-1} \rightarrow e^{-\lambda x}.$$

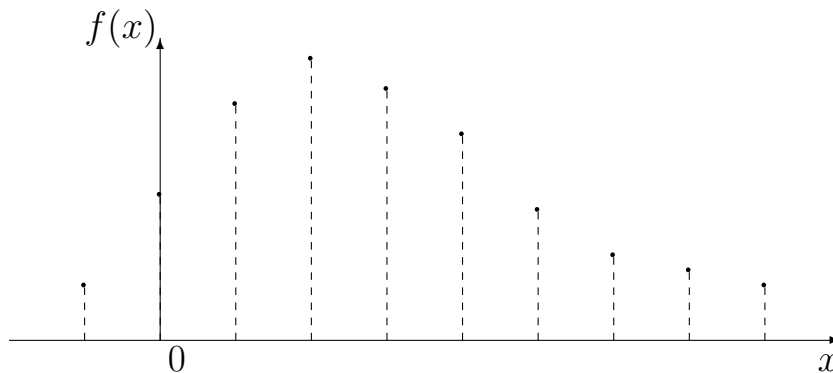
Итак, асимптотический анализ $H(t)$, аналогичный теореме Пуассона, привел нас к функции надежности показательного распределения.

Для того чтобы строить новые вероятностные модели, нам необходимо ближе познакомиться с числовыми и функциональными характеристиками распределений, которые постоянно используются на практике, когда возникает проблема сравнения распределений или характеристика их специфических особенностей. Этому вопросу посвящен следующий параграф.

§6. Характеристики распределения случайной величины. Классификация распределений

Мы построили шесть вероятностных моделей, и если пред нами стоит задача их классификации, то первая очевидная особенность, которой обладает каждое из распределений соответствующей случайной величины, это – непрерывность или разрывность функции распределения. Полученные семейства распределений можно разбить на два класса – *дискретный* и *непрерывный*.

Гипергеометрическое $GG(N, M, n)$, биномиальное $B(n, p)$, пуассоновское $P(\lambda)$ и геометрическое $Geo(p)$ распределения принадлежат к дискретному классу. При выводе этих распределений мы вполне могли бы ограничиться техникой элементарной теории вероятностей, поскольку пространства элементарных исходов (значений случайной величины X) состояли из конечного или счетного числа точек, и функции плотности $f(x | \theta)$ в области их ненулевых значений определяли вероятности каждого элементарного исхода $X = x$. Графическое изображение $f(x) = f(x | \theta)$ как функции x при каждом фиксированном θ позволяет наиболее полно представить картину общего распределения вероятностей и, одновременно, вызывает некоторые ассоциации с “нагруженным стержнем”, а также стремление характеризовать распределение масс по стержню такими механическими характеристиками, как центр тяжести, момент инерции, асимметрия и эксцесс в распределении масс и пр.



Прибегая к такой “механической” интерпретации распределения вероятностей, мы соотносим вероятность события $X \in B$ при любом $B \in \mathcal{B}$ с массой участка стержня B и вычисляем величину этой массы по формуле

$$P(B) = \sum_{x \in B} f(x).$$

Центр тяжести нагруженного стержня называется *средним значением* случайной величины X , обозначается $\mathbf{E}X$ и вычисляется как

$$\mathbf{E}X = \sum_{x \in \mathbb{R}} x f(x).$$

Момент инерции относительно точки $\mu = \mathbf{E}X$, равный

$$\mathbf{D}X = \sum_{x \in \mathbb{R}} (x - \mu)^2 f(x),$$

характеризует меру разброса (удаленности) отдельных точек нагружения от центра масс и поэтому в теории вероятностей называется *дисперсией* случайной величины X . Кроме стандартного обозначения $\mathbf{D}X$, за величиной дисперсии закреплен символ σ^2 , в то время как квадратный корень из дисперсии $\sigma = \sqrt{\mathbf{D}X}$ называется *стандартным отклонением* X .

Несомненный практический интерес представляет также точка достижения максимума функции $f(x)$, как наиболее вероятного значения X . Эта точка называется *модой* распределения X , и как-то так сложилось, что стандартного, наиболее распространенного обозначения у этой характеристики нет, разве лишь $\text{mod}(X)$.

Мы не будем торопиться с введением других характеристик распределения X , а также иллюстрировать вычисления $\mathbf{E}X$, $\mathbf{D}X$ и $\text{mod}(X)$ на конкретных распределениях и сначала попытаемся ввести аналоги этих характеристик для случайных величин с непрерывной функцией распределения.

К классу непрерывных распределений принадлежат равномерное $U(a,b)$ и показательное $E(\theta)$ распределения. При построении этих вероятностных моделей функция распределения играла определяющую роль и теорема 4.1 использовалась по существу.

Графическое изображение непрерывной функции распределения вряд ли стоит рассматривать как столь же наглядную иллюстрацию распределения вероятностей, как, например, график функции плотности (функции скачков) распределения дискретного типа. Это замечание в равной степени относится как к дискретному, так и непрерывному классу распределений. Графики возрастающих функций с областью значений в интервале $[0; 1]$ так похожи друг на друга, что их главная примечательность – точки перегиба “на глаз” определяются только при высоких художественных достоинствах графического изображения. Другое дело – производная функции, где эти точки перегиба превращаются в точки экстремума. С другой

стороны, производная функции распределения в непрерывном случае, так же как и функция скачков дискретного распределения, допускает механическую интерпретацию функции плотности единичной массы, “размазанной” по бесконечному стержню, и в рамках этой интерпретации мы снова можем рассматривать такие характеристики, как центр тяжести, момент инерции и тому подобное.

Итак, определим *функцию плотности* непрерывного распределения $F(x)$ как производную $f(x) = dF(x)/dx$, которая в нашем случае определяется почти всюду по мере Лебега, что, как будет в дальнейшем, вполне достаточно для вычисления характеристик непрерывного распределения. Так, для равномерного распределения $f(x) = f(x|\theta) = 0$ равна нулю вне сегмента $[a; b]$ и $f(x|\theta) = (b-a)^{-1}$, то есть постоянна на этом сегменте. В случае показательного распределения $f(x|\theta) = 0$ при $x < 0$,

$$f(x|\theta) = \frac{1}{\theta} \exp\left\{-\frac{x}{\theta}\right\},$$

если $x \geq 0$, и отнесение точки $x = 0$ к области нулевых значений функции f очевидно не изменит значений интегральных характеристик распределения; аналогичное заключение можно сделать и относительно конечных точек a и b равномерного распределения $U(a, b)$.

Функция распределения из непрерывного класса выражается через свою функцию плотности в виде

$$F(x) = \int_{-\infty}^x f(t)dt,$$

а вероятность “попадания” X в некоторое произвольное борелевское множество B (вероятность события B) записывается как

$$P(X \in B) = \int_B f(x)dx = \int_{\mathbf{R}} \mathbf{I}_B(x)f(x)dx,$$

где $\mathbf{I}_B(x)$ – индикаторная функция множества B . Естественно, в силу явной нерегулярности (разрывности и прочих пакостей) подынтегральных функций интегралы в этих формулах следует рассматривать как интегралы Лебега по лебеговой мере dx на борелевской прямой $(\mathbf{R}, \mathfrak{B})$.

Центр тяжести стержня с непрерывным распределением масс, которое определяется функцией плотности $f(x)$, вычисляется по известной нам из

курса математического анализа формуле

$$\mu = \mathbf{E}X = \int_{-\infty}^{\infty} x f(x) dx$$

и называется, как и в дискретном случае, *средним значением* случайной величины X . Точно так же момент инерции

$$\sigma^2 = \mathbf{D}X = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

называется *дисперсией* X , а σ – *стандартным отклонением*. Наконец, точка достижения максимума функции плотности:

$$\text{mod}(X) = \arg \max_{x \in \mathbf{R}} f(x) -$$

модой распределения X . Окрестность точки $\text{mod}(X)$ обладает наибольшей концентрацией вероятностной массы.

Лекция 9

Естественно, рассмотрев два основных класса распределений, мы могли бы теперь продолжить изучение характеристик распределений каждого типа, но возникает естественный вопрос, а существуют ли смешанные дискретно-непрерывные распределения или вообще распределения, не принадлежащие к изученным классам, и как тогда вычислять их средние значения и дисперсии?

Что касается дискретно-непрерывных распределений, то о существовании и практической ценности таких распределений свидетельствует следующая вероятностная модель теории надежности. Предположим, что предприятие выпускает изделия с показательным распределением долговечности, но в силу специфических дефектов производства каждое изделие с некоторой вероятностью p может быть “мертворожденным”, то есть отказать при его “включении”. В таком случае функция распределения долговечности в области $x > 0$ имеет вид (используется формула полной вероятности)

$$F(x) = p + (1 - p)(1 - \exp\{-x/\theta\}),$$

а средний срок службы

$$EX = 0 \cdot p + (1 - p)\theta^{-1} \int_0^{\infty} x \exp \{-x/\theta\} dx = (1 - p)\theta$$

(опять новая формула для вычисления характеристик распределения X !).

Дальше – больше, оказывается существует еще один тип распределений, вычисление характеристик которого вообще невысказимо вне рамок теории интеграла Лебега. Помните, мы говорили с вами о связи между схемой испытаний Бернулли с вероятностью успешного испытания $p = 1/2$ и равномерным распределением на отрезке $[0, 1]$? Оказывается, если вероятность успеха $p \neq 1/2$, то двоичная дробь, составленная из реализаций индикаторов успеха, представляет результат наблюдения случайной величины с весьма загадочной функцией распределения. Во-первых, эта функция почти всюду постоянна – производная от нее почти всюду по мере Лебега на $(\mathbb{R}, \mathcal{B})$ равна нулю. Тем не менее эта функция возрастает, непрерывна(!), но точки ее роста составляют счетное множество, имеющее, естественно, нулевую лебегову меру. Соответствующая этой функции распределения вероятностная мера P на борелевской прямой сингулярна относительно меры Лебега: если множество $B \in \mathcal{B}$ имеет нулевую лебегову меру, то отсюда не следует, что $P(B) = 0$.

Распределения такого вида, имеющие непрерывную функцию распределения, но сингулярные по отношению к мере Лебега, составляют класс *сингулярных* распределений. Легко понять, что явная запись таких распределений вряд ли возможна. В нашем примере с построением реализаций случайной величины X с помощью схемы Бернулли для функции распределения X составляется некоторое операторное уравнение, и если мы хотим рассчитать вероятности попадания X в интервалы на прямой, то придется использовать численные методы решения таких уравнений.

Итак, мы рассмотрели три типа распределений: дискретный, непрерывный и сингулярный. Удивительно то, что других типов не существует, о чем свидетельствует знаменитая

Теорема Лебега. *Любая функция распределения представима в виде суммы трех неотрицательных, неубывающих функций, одна из которых абсолютно непрерывна и имеет неотрицательную производную на множестве положительной лебеговой меры; вторая является ступенчатой и обладает не более чем счетным множеством точек разрыва (скачков); третья непре-*

рывает, но имеет не более чем счетное множество точек роста.

Доказательство этой теоремы выходит из рамок нашего общего курса теории вероятностей. В не столь отдаленные времена, когда на факультете ВМК занимались преподаванием фундаментальных наук, а не обучением примитивному ремеслу работы на компьютере, теорема Лебега доказывалась в общем курсе математического анализа. Из теоремы Лебега вытекает, что в чистом виде существует только три типа распределений, из которых два (непрерывный и дискретный) нам знакомы, а третий – сингулярный – загадочен, и мы пока не в состоянии представить себе, каким образом вычислять интеграл Лебега,

$$\mathbf{E}X = \int_{\mathbb{R}} x dP(x),$$

определяющий среднее значение случайной величины X с сингулярным распределением вероятностей $P(B)$, $B \in \mathcal{B}$.

Спешу обрадовать вас, что мы не будем рассматривать сингулярные вероятностные модели. Тем не менее существует весьма общий подход к определению *функции плотности* для любого, в том числе и смешанного, типов распределений, опираясь на который можно предложить некоторый общий метод определения и вычисления характеристик распределений различных типов. Этот подход указывает следующая, не менее знаменитая, чем теорема Лебега,

Теорема Радона–Никодима. Пусть на борелевской прямой $(\mathbb{R}, \mathcal{B})$ заданы вероятность P и сигма-конечная мера μ , причем P абсолютно непрерывна относительно μ , то есть $\mu(B) = 0$ влечет $P(B) = 0$. Тогда для почти всех по мере μ точек $x \in \mathbb{R}$ существует такая единственная неотрицательная функция $f(x)$, что

$$P(B) = \int_B f(x) d\mu(x), \quad \forall B \in \mathcal{B}. \quad (1)$$

Эта теорема, доказательство которой мы также опускаем (и не потому, что времени нет, а просто – знаний не хватает), позволяет ввести одно из центральных понятий теории вероятностей, постоянно используемое при построении вероятностных моделей.

Определение 6.1. Функция $f(x)$, определяемая соотношением (1) для почти всех по мере μ точек $x \in \mathbb{R}$, называется *функцией плотности* распределения вероятностей P по мере μ . Эта функция называется также

производной Радона–Никодима меры P по мере μ , и имеет место символическая запись $f(x) = dP/d\mu$.

В рамках этого определения введенная выше функция плотности непрерывного распределения есть производная Радона–Никодима вероятности P по мере Лебега $d\mu = dx$ на борелевской прямой. Так как вероятность P в соответствии с теоремой 4.1 определялась с помощью функции распределения $F(x)$, то мы использовали тот вариант производной Радона–Никодима, который совпадает с обычной производной функции $F(x)$, доопределяя эту функцию в точках, где производная не существует, таким образом, чтобы не возникали дополнительные разрывы. Что же касается дискретного случая, то здесь мы использовали производную Радона–Никодима по *считающей мере* μ : для любого $B \in \mathcal{B}$ мера $\mu(B)$ равна количеству точек с целочисленными координатами, которые принадлежат B . Например, борелевское множество $B = [-2.5; 5]$ содержит восемь точек с целочисленными координатами $-2, -1, 0, \dots, 5$, и поэтому $\mu(B) = 8$. В “дробных” точках $x \in \mathbb{R}$ мы полагали $f(x) = 0$, хотя могли бы выбирать любые другие значения при вычислении вероятностей по формуле (1). Дело в том, что при интегрировании по дискретной считающей мере интеграл Лебега от любой функции превращается в сумму значений этой функции в целочисленных точках, и (1) принимает известный нам из элементарной теории вероятностей вид

$$P(B) = \sum_{x \in B} P(X = x) = \sum_{x \in B} f(x).$$

Теперь мы обладаем общим подходом к определению характеристик распределения случайной величины X . Значительная часть из них определяется через интеграл Лебега по мере (вероятности) P от специально подобранных функций.

Определение 6.2. Пусть X – случайная величина с распределением P и $f(x)$ – функция плотности P по сигма-конечной мере μ . *Математическим ожиданием* любого измеримого отображения $g(X)$ борелевской прямой в себя (измеримой функции от случайной величины X) называется интеграл Лебега

$$\mathbf{E}g(X) = \int_{\mathbb{R}} g(x)dP(x) = \int_{\mathbb{R}} g(x)f(x)d\mu(x).$$

В частности, *математическое ожидание случайной величины X* вычис-

ляется по формуле

$$\mathbf{E}X = \int_{\mathbf{R}} x dP(x) = \int_{\mathbf{R}} x f(x) d\mu(x).$$

З а м е ч а н и е. В отечественной литературе по теории вероятностей (например, в учебнике А.А.Боровкова “Теория вероятностей”) математическое ожидание обозначается латинской буквой **M** а не **E**.

Моментные характеристики распределения случайной величины. Математическое ожидание функции $g(X) = (X - a)^k$ от случайной величины X , где k принимает только целочисленные значения $1, 2, \dots$, называется *моментом k -го порядка случайной величины X относительно точки a* . Если $a = 0$, то $\alpha_k = \mathbf{E}X^k$ называется просто *моментом k -го порядка* случайной величины X , а если $a = \mathbf{E}X (= \alpha_1)$, то момент $\mu_k = \mathbf{E}(X - \mathbf{E}X)^k$ называется *центральный момент k -го порядка*. Иногда, во избежание недоразумений, моменты α_k называются *нецентральными моментами*. Первый нецентральный момент $\alpha_1 = \mathbf{E}X$ называется *средним значением* или *математическим ожиданием* случайной величины X и обозначается обычно буквой μ . Вторым центральным моментом $\mu_2 = \mathbf{E}(X - \mu)^2$ называется *дисперсией* случайной величины X и обозначается или буквой σ^2 , или вводится оператор $\mathbf{D}X$. Напомним, что квадратный корень из дисперсии: $\sigma = \sqrt{\mathbf{D}X}$ мы договорились называть *стандартным отклонением X* . Поскольку σ имеет ту же размерность, что и наблюдаемая случайная величина X , то в практических приложениях в качестве меры “разброса” вероятностей используется обычно стандартное отклонение σ , а не дисперсия σ^2 . Для среднего и дисперсии X справедливо

Предложение 6.1. *Среднее значение $\mathbf{E}X$ и дисперсия $\mathbf{D}X$ обладают следующими свойствами:*

$$1^0. \mathbf{E}(aX + b) = a\mathbf{E}X + b \text{ для любых постоянных } a, b \in \mathbf{R},$$

2⁰. $\mathbf{D}(aX + b) = a^2\mathbf{D}X$ для любых постоянных $a, b \in \mathbf{R}$, то есть дисперсия инвариантна относительно сдвигов случайной величины X на постоянную величину;

$$3^0. \mathbf{D}X = \mathbf{E}X^2 - (\mathbf{E}X)^2 = \alpha_2 - \mu^2,$$

$$4^0. \inf_{a \in \mathbf{R}} \mathbf{E}(X - a)^2 = \mathbf{D}X, \text{ то есть } \arg \inf_{a \in \mathbf{R}} \mathbf{E}(X - a)^2 = \mathbf{E}X.$$

Доказательство.

1⁰. Данное утверждение есть простая констатация известного свойства линейности интеграла Лебега.

$$2^0. \mathbf{D}(aX + b) = \mathbf{E}(aX + b - a\mu - b)^2 = a^2\mathbf{E}(X - \mu)^2 = a^2\mathbf{D}X.$$

$$3^0. \mathbf{D}X = \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}(X^2 - 2X\mathbf{E}X + (\mathbf{E}X)^2) =$$

$$\mathbf{E}X^2 - 2\mathbf{E}X \cdot \mathbf{E}X + (\mathbf{E}X)^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2.$$

$$4^0. \mathbf{E}(X - a)^2 = \mathbf{E}((X - \mu) - (a - \mu))^2 =$$

$$\mathbf{E}((X - \mu)^2 - 2(a - \mu)(X - \mu) + (a - \mu)^2) =$$

$$\mathbf{E}(X - \mu)^2 - 2(a - \mu)\mathbf{E}(X - \mu) + (a - \mu)^2 = \mathbf{D}X + (a - \mu)^2 \geq \mathbf{D}X,$$

причем равенство достигается тогда и только тогда, когда $a = \mu = \mathbf{E}X$.

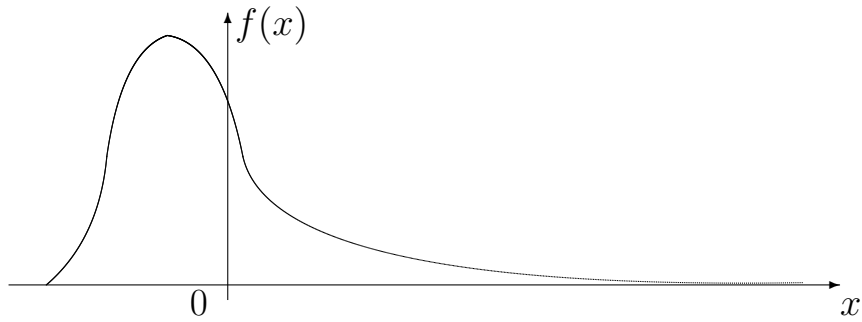
С моментами случайной величины X связаны две замечательные *характеристики формы* распределения X :

коэффициент асимметрии $\gamma_1 = \mu_3/\sigma^3$, и

коэффициент эксцесса $\gamma_2 = \mu_4/\sigma^4 - 3$.

Легко заметить по аналогии с доказательством пункта 2⁰ предыдущего предложения, что γ_1 и γ_2 *инвариантны относительно линейных преобразований случайных величин*, то есть X и $aX + b$ имеют одинаковые коэффициенты асимметрии и эксцесса при любых постоянных a и b .

Как и выше, мы будем называть *модой* распределения случайной величины X любую точку $\text{mod}(X)$ достижения локального максимума у функции плотности $f(x)$. Если мода единственна, то говорят, что распределение X *унимодально*. Когда график унимодальной кривой плотности имеет “длинный хвост” справа от моды (см. рисунок на этой странице), то в выражении μ_3 кубы положительных отклонений перевесят отрицательные кубы, и коэффициент асимметрии γ_1 будет положителен. Если же мода “свалена” вправо (длинный хвост слева от моды), то $\gamma_1 < 0$. Распределения с симметричной функцией плотности, как, например, биномиальное с $p = 1/2$ или равномерное $U(a,b)$, обладают нулевой асимметрией: $\gamma_1 = 0$.



Что же касается коэффициента эксцесса γ_2 , то его подлинный смысл мы поймем после знакомства в следующем параграфе с *нормальным распределением* на борелевской прямой, а пока только отметим, что положительный эксцесс говорит об излишней “пикообразности” – вытянутости вверх кривой плотности, в то время как отрицательное значение γ_2 указывает на более плоский характер вершины кривой плотности.

Лекция 10

Прежде чем перейти к примерам по вычислению моментных характеристик случайных величин, следует обратить внимание на то, что в рассмотренных нами вероятностных моделях существуют довольно крупные элементы, имеющие нулевую вероятность, например, во всех моделях $P(X \in (-\infty, 0)) = 0$. В связи с этим вводится понятие *носителя* распределения случайной величины, как замыкания множества $\{x \in \mathbb{R} : f(x) > 0\}$. Такое определение носителя не является достаточно общим и связано с мерой μ , по которой вычисляется плотность $f(x)$, но поскольку мы договорились рассматривать только дискретные и непрерывные распределения (μ – считающая мера или мера Лебега), то такое определение вполне работоспособно и позволяет легко найти носитель любого из шести известных нам распределений. Носитель распределения будет обозначаться рукописной буквой \mathcal{X} . Нетрудно понять, что при вычислении моментных и прочих интегральных характеристик распределения из области интегрирования можно убрать все точки, не принадлежащие \mathcal{X} , и при этом величина характеристики останется неизменной.

Пример 6.1 (биномиальное распределение $B(n, p)$). Носитель этого распределения $\mathcal{X} = \{0, 1, \dots, n\}$. Для вычисления первых двух моментов биномиального распределения воспользуемся методом “дифференцирова-

ния по параметру” и формулой бинома Ньютона:

$$\sum_{k=0}^n C_n^k a^k b^{n-k} = (a + b)^n.$$

По определению среднего значения

$$\begin{aligned} \mu = \mathbf{E}X &= \sum_{k=0}^n k C_n^k p^k (1-p)^{n-k} = p \left[\frac{d}{dx} \sum_{k=0}^n C_n^k x^k (1-p)^{n-k} \right]_{x=p} = \\ &= p \left[\frac{d}{dx} (x+1-p)^n \right]_{x=p} = pn(x+1-p)^{n-1} \Big|_{x=p} = np. \end{aligned}$$

Второй момент

$$\begin{aligned} \alpha_2 = \mathbf{E}X^2 &= \sum_{k=0}^n k^2 C_n^k p^k (1-p)^{n-k} = p \left[\frac{d}{dx} x \frac{d}{dx} \sum_{k=0}^n C_n^k x^k (1-p)^{n-k} \right]_{x=p} = \\ &= p \left[\frac{d}{dx} x \frac{d}{dx} (x+1-p)^n \right]_{x=p} = p \left[\frac{d}{dx} xn(x+1-p)^{n-1} \right]_{x=p} = \\ &= np \left[(x+1-p)^{n-1} + x(n-1)(x+1-p)^{n-2} \right]_{x=p} = np(1-p) + (np)^2, \end{aligned}$$

откуда дисперсия биномиального распределения

$$\sigma^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2 = np(1-p).$$

С помощью аналогичных, но более утомительных выкладок можно найти третий и четвертый моменты, а также коэффициенты асимметрии и эксцесса

$$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}, \quad \gamma_2 = \frac{1-6p(1-p)}{np(1-p)}.$$

Следовательно, биномиальное распределение “свалено” влево (– хвост справа длиннее) при $p < 1/2$, симметрично, как нам было известно ранее, при $p = 1/2$ и “свалено” вправо при $p > 1/2$. Коэффициент эксцесса положителен в области $p(1-p) < 1/6$, а наибольшее по абсолютной величине отрицательное значение $\gamma_2 = -2/n$, когда $p = 1/2$.

Мода $B(n, p)$ определяется как целочисленное x , при котором происходит смена неравенства

$$f(x | n, p) < f(x + 1 | n, p)$$

на обратное. Нетрудно убедиться, что это неравенство эквивалентно $x+1 < p(n+1)$, так что $\text{mod}(X)$ определяется через сравнение значений $f(x | n, p)$ при целых $x \geq 0$, ближайших к $p(n+1)$.

Пример 6.2 (*распределение Пуассона* $P(\lambda)$). Носитель распределения $\mathcal{X} = \{0, 1, \dots, \infty\}$ – точка $x = \infty$ должна быть включена в носитель по требованию замыкания множества вероятности единица. Моментные характеристики пуассоновского распределения можно рассчитать, используя тот же метод дифференцирования по параметру, но проще, вспомнив, что $P(\lambda)$ есть предел $B(n, p)$ при $n \rightarrow \infty$, $p \rightarrow 0$ и $np = \lambda$, перейти к этому пределу в моментных характеристиках биномиального распределения. В результате получаем

$$\mathbf{E}X = \mathbf{D}X = \lambda, \quad \gamma_1 = \lambda^{-1/2}, \quad \gamma_2 = \lambda^{-1},$$

а $\text{mod}(X) = [\lambda]$, поскольку асимметрия $P(\lambda)$ всегда положительна и график $f(x | \lambda)$ “свален” влево.

Следует обратить особое внимание на то, что у распределения Пуассона дисперсия совпадает со средним значением: $\mu = \sigma^2 = \lambda$.

Пример 6.3 (*равномерное распределение* $U(a, b)$). Носитель распределения $\mathcal{X} = [a; b]$. Модой распределения является любая точка интервала (a, b) , поскольку плотность $f(x) = (b-a)^{-1}$ постоянна на этом интервале.

Нетрудно убедиться, что если случайная величина X имеет распределение $U(0, 1)$, то $Y = (b-a)X + a$, $b > a$, распределена как $U(a, b)$. Это вытекает из-за следующего соотношения между функциями распределения случайных величин:

$$P(Y < x) = P((b-a)X + a < x) = P(X < (x-a)/(b-a)) = (x-a)/(b-a).$$

В силу этого для вычисления моментных характеристик $U(a, b)$ достаточно найти соответствующие характеристики $U(0, 1)$ и затем воспользоваться предложением 6.1.

Для распределения $U(0, 1)$ имеем

$$\mu = \mathbf{E}X = \int_0^1 x dx = 1/2, \quad \alpha_2 = \int_0^1 x^2 dx = 1/3,$$

откуда дисперсия $\sigma^2 = 1/3 - 1/4 = 1/12$. Следовательно, для распределения $U(a, b)$ (см. предложение 6.1)

$$\mu = a + (b-a)/2, \quad \sigma^2 = (b-a)^2/12.$$

Симметричное равномерное распределение $U(a, b)$ имеет нулевой коэффициент асимметрии, в то время как коэффициент эксцесса γ_2 отрицателен (не будем заниматься его вычислением).

Пример 6.4 (*показательное распределение* $E(\theta)$). Носитель распределения $X = [0, \infty]$ – расширенная положительная часть прямой \mathbb{R} . Наибольшее значение плотности

$$f(x) = \theta^{-1} \exp\{-x/\theta\}$$

достигается в точке $x = 0$, поэтому $\text{mod}(X) = 0$.

Моменты показательного распределения

$$\alpha_k = \theta^{-1} \int_0^{\infty} x^k \exp\{-x/\theta\} dx = \theta^k \int_0^{\infty} x^k e^{-x} dx = \Gamma(k+1)\theta^k = k!\theta^k,$$

откуда $\mu = \theta$, $\sigma^2 = \theta^2$ и стандартное отклонение $\sigma = \theta$ совпадает со средним значением.

Естественно, моментные характеристики далеко не универсальны, и можно привести примеры распределений, у которых существует ограниченное количество моментов, или не существует даже среднего значения. Мы приведем два из таких распределений, одно из которых может представлять некоторый практический интерес, а другое будет использоваться для иллюстраций различных патологий в теории статистического вывода; оба распределения занесутся в каталог вероятностных моделей.

Распределение Парето $\text{Par}(a, \alpha)$. Налоговые органы обычно интересуются распределением годовых доходов тех лиц, годовой доход которых превосходит некоторый предел a , установленный законами о налогообложении. Такого рода распределения иногда считают (к сожалению, без особого “экономического” обоснования) приближенно совпадающими с *распределением Парето*, вся вероятностная масса которого сосредоточена в области $x > a$ (носитель распределения $X = [a, \infty]$), и функция распределения на сегменте X равна

$$F(x) = 1 - \left(\frac{a}{x}\right)^\alpha, \quad x > a, \quad \alpha > 0.$$

Это распределение, зависящее от двумерного параметра $\theta = (a, \alpha)$ с параметрическим пространством $\Theta = \mathbb{R}_+ \times \mathbb{R}_+$, принадлежит непрерывному

типу; его функция плотности в области $x > a$ равна

$$f(x | \theta) = \frac{\alpha}{a} \left(\frac{a}{x}\right)^{\alpha+1}.$$

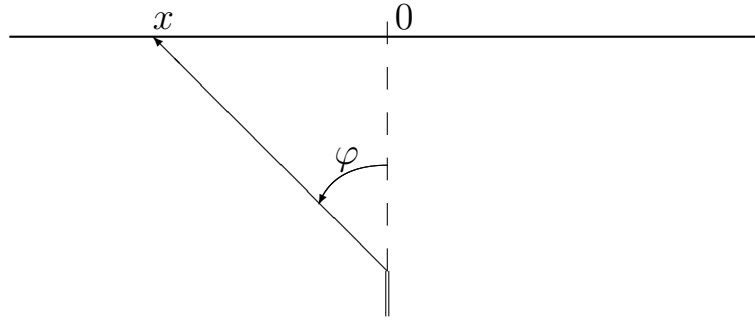
Момент k -го порядка у распределения Парето существует только при значениях параметра $\alpha > k$, например, неравенство $\alpha > 1$ гарантирует существование среднего значения, которое, как нетрудно подсчитать, равно $\alpha a / (\alpha - 1)$.

Если случайная величина X распределена по закону Парето, то, как легко видеть, $\ln X$ имеет показательное распределение, “сдвинутое вправо” на величину $\ln a$, так как

$$P(\ln X < x) = P(X < e^x) = F(e^x).$$

Это замечание объясняет, почему распределение Парето адекватно описывает распределение наблюдаемых доходов у лиц с высоким уровнем дохода. Вспомним постулат “отсутствия последействия”, приводящий к показательному распределению долговечности: вероятность того, что изделие прослужит промежуток времени, не меньший s , при условии, что оно уже отработало срок t , не зависит от величины t . В основу модели Парето положен тот же принцип, только в мультипликативной, а не в аддитивной, формулировке: *вероятность того, что доход отдельного лица увеличится не меньше, чем в s раз, при условии, что он уже достиг уровня t , не зависит от величины достигнутого уровня.* Это происходит, по-видимому, от того, что обладающий большими доходами стремится сохранить достигнутое положение и редко стремится вкладывать большие капиталы в новые отрасли с целью наращивания денежной массы. В таком случае изменчивость дохода за наблюдаемые периоды времени носит случайный характер и не связана с величиной капитала, которым располагают отдельные субъекты. В то же время у “предпринимателей” распределение доходов отлично от закона Парето. Это так называемое логарифмически нормальное распределение, с которым мы познакомимся несколько позже, освоив новые математические методы построения вероятностных моделей.

Распределение Коши $S(a, b)$. Орудие с вращающимся лафетом помещается на единичном расстоянии от стены, бесконечно уходящей в обе стороны.



Представим, что стена является действительной прямой \mathbb{R} с началом координат в основании перпендикуляра, опущенного из орудия на стену. Ствол орудия размещается параллельно стене с направлением выстрела в сторону отрицательной полуоси, лафет орудия начинает равномерно вращаться по ходу часовой стрелки, и прежде, чем ствол займет первое положение параллельное стене, в случайный момент времени происходит выстрел. Экспериментатора интересует распределение случайной величины X , реализация x которой совпадает с координатой точки попадания снаряда.

Пусть φ – случайная величина, соответствующая величине угла, между перпендикуляром к стене и положением ствола в момент выстрела. Нам будет удобнее измерять φ в пределах $[-\pi/2; \pi/2]$ и трактовать предположение о случайном моменте выстрела в терминах равномерного распределения φ на этом сегменте. Следовательно, функция распределения φ при $-\pi/2 \leq \varphi \leq \pi/2$ равна $F(\varphi) = (\varphi + \pi/2)\pi^{-1}$. Очевидно, координата точки попадания (см. рисунок) $X = \operatorname{tg} \varphi$, откуда искомая функция распределения

$$F(x) = P(X < x) = P(\operatorname{tg} \varphi < x) = P(\varphi < \operatorname{arctg} x) = \frac{1}{\pi} \left(\operatorname{arctg} x + \frac{\pi}{2} \right), \quad x \in \mathbb{R},$$

а функция плотности

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2}.$$

Сдвиг вправо на параметр a и выбор масштабного параметра b определяет то распределение, которому мы присвоим имя Коши и будем обозначать $C(a, b)$; его функция плотности

$$f(x | a, b) = \frac{1}{\pi b} \left[1 + \left(\frac{x - a}{b} \right)^2 \right]^{-1},$$

носителем распределения является расширенная числовая прямая $\mathcal{X} = \bar{\mathbb{R}} = [-\infty, +\infty]$.

Легко видеть, что распределение Коши не обладает даже конечным средним значением, не говоря о моментах более высокого порядка. Однако это распределение симметрично и имеет ярко выраженную моду, $\text{mod}(X)=a$, которая с успехом заменяет среднее значение как характеристику положения центра масс. В связи с этим полезно сделать замечание о среднем значении как характеристике положения: оно действительно играет свою роль только в случае симметричных распределений, но при больших абсолютных значениях γ_1 среднее перестает быть полезной характеристикой распределения, в то время как мода “всегда хороша”.

Какие же характеристики используются при описании распределений, у которых отсутствуют моменты?

Определение 6.3. Пусть функция распределения $F(x)$ случайной величины X строго возрастает в области всех значений своего аргумента, для которых $0 < F(x) < 1$. Тогда для любого $p \in (0; 1)$ корень $x_p = F^{-1}(p)$ уравнения $F(x) = p$ называется *p-квантилью* распределения X .

В том случае, когда $F(x)$ непрерывна, но не строго монотонна, так что уравнение $F(x) = p$ имеет много решений, в качестве *p-квантили* обычно берется наибольший или наименьший из корней этого уравнения, и выбор корня определяется существом рассматриваемой вероятностной проблемы. В случае же дискретного распределения это уравнение может вообще не иметь решений, и тогда в качестве *p-квантили* выбирается то значение x , для которого значение $F(x)$ ближе всего к заданному p .

Квантиль считается характеристикой *положения*, и с этой точки зрения особого внимания заслуживает квантиль $x_{0.5}$, которая разделяет всю вероятностную массу на две одинаковые половинки. Эта квантиль носит название *медианы* распределения и обычно обозначается буквой m . У симметричных распределений (биномиальное с вероятностью успешного испытания $p = 1/2$, равномерное и Коши) медиана совпадает с центром симметрии распределения, а при наличии среднего значения у симметричного распределения медиана $m = \mathbf{E}X$. Если p кратно 0.1, то квантиль называется *децилью*, а если $p = 1/4$ или $3/4$, то – *квартилью*.

С квантилями связаны также несколько характеристик *рассеяния* распределения вероятностей. Очевидно, интервал $(x_{1-p}; x_p)$ при достаточно близких к единице значениях p покрывает основную часть вероятностной

массы, и поэтому разность $x_p - x_{1-p}$, $p > 1/2$, служит характеристикой *толерантности* распределения случайной величины X . Если $p = 3/4$, то разность $x_{3/4} - x_{1/4}$ называется *семиинтерквартильной широтой* распределения X .

Лекция 11

Мы завершим этот параграф доказательством одного замечательного неравенства, играющего исключительную роль при доказательстве многих теорем (или, как часто говорят, “законов”) теории вероятностей. Это неравенство или, в большей степени, следствие из него связывает квантильные и моментные характеристики рассеяния распределения.

Предложение 6.2 (неравенство Чебышева). *Для любой неотрицательной измеримой функции $g(x)$ и любого $\varepsilon > 0$ имеет место неравенство*

$$P(g(X) > \varepsilon) \leq \frac{\mathbf{E} g(X)}{\varepsilon}.$$

Доказательство. Если $\mathbf{E} g(X) = +\infty$, то неравенство тривиально. В случае конечного математического ожидания

$$\mathbf{E} g(X) = \int_{\mathbf{R}} g(x) dP(x) = \int_{g(x) < \varepsilon} g(x) dP(x) + \int_{g(x) \geq \varepsilon} g(x) dP(x).$$

Если в правой части этого равенства первое слагаемое заменить нулем (оно неотрицательно), а во втором слагаемом под интегралом вместо $g(x)$ подставить его наименьшее значение ε , то получим оценку снизу

$$\mathbf{E} g(X) \geq \varepsilon \int_{g(x) > \varepsilon} dP(x) = \varepsilon P(g(X) > \varepsilon),$$

из которой немедленно следует неравенство Чебышева.

Следствие 6.1. *Для любой случайной величины X с конечным средним значением $\mathbf{E}X$ и любого $\varepsilon > 0$ имеет место неравенство*

$$P(|X - \mathbf{E}X| > \varepsilon) \leq \frac{\mathbf{D}X}{\varepsilon^2}. \quad (2)$$

Доказательство. Если дисперсия X не существует (равна бесконечности), то утверждение следствия тривиально. В случае $\mathbf{D}X < \infty$ достаточно заменить событие $|X - \mathbf{E}X| > \varepsilon$ на эквивалентное $|X - \mathbf{E}X|^2 > \varepsilon^2$ и применить неравенство Чебышева.

Доказанное неравенство часто используется на практике для универсальной характеристики толерантности распределений, обладающих конечным средним μ и конечной дисперсией σ^2 . Имеется в виду распространенное

Правило трех сигм. *Интервал с концами $\mu \pm 3\sigma$ содержит приблизительно 90% вероятностной массы распределения X .*

Действительно, если в неравенстве (2) положить $\varepsilon = 3\sigma$, то получим: $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 1 - P(|X - \mu| > 3\sigma) \geq 8/9 \approx 0.9$.

Так как правило 3σ носит универсальный характер, то оно дает в большинстве случаев слишком грубую оценку толерантности распределения. Например, можно доказать, что для симметричных распределений с конечным третьим моментом μ_3 справедливо *правило 2σ* : интервал с концами $\mu \pm 2\sigma$ содержит 90% вероятностной массы распределения.

В дальнейшем, чтобы не писать длинные названия рассмотренных нами распределений, мы будем указывать распределение X посредством ссылки на символ этого распределения, используя при этом знак эквивалентности, например, $X \sim B(n, p)$ означает, что X имеет биномиальное распределение.

§7. Предельные теоремы в схеме испытаний Бернулли.

Нормальное распределение

При выводе распределения Пуассона мы исследовали асимптотику биномиального распределения, когда $n \rightarrow \infty$, $p \rightarrow 0$, $np = \lambda (const)$. Существует, однако, широкий класс практических задач, в которых построение вероятностных моделей требует асимптотического анализа биномиального распределения при фиксированном $p \in (0; 1)$ и $n \rightarrow \infty$.

Пример 7.1 (*определение видимой звездной величины*). Наблюдения за изменением блеска небесных светил, в частности звезд, являются одной из важнейших задач практической астрономии. Только с помощью анализа таких наблюдений можно обнаружить *переменные* звезды, поставляющие информацию о расстояниях до отдаленных светил (цефеиды), а также об их массах, размерах и пр. (затменные переменные и спектрально-двойные звезды).

Величина блеска звезды определяется так называемой *видимой звездной величиной* – характеристикой светимости, пропорциональной количеству квантов света, исходящих от звезды и достигших прибора (электрического фотометра, фотографической пластинки и т.п.), который регистрирует поток лучевой энергии. С точки зрения проблемы построения вероятностной модели изменчивости в повторных наблюдениях блеска, мы имеем ту же картину, что и при измерениях интенсивности радиоактивного источника: каждый квант света с определенной вероятностью p достигает регистрирующего прибора, и общее количество регистрируемых квантов определяет результат наблюдения блеска звезды. Принципиальное различие с измерениями радиоактивности состоит в достаточно большом значении вероятности “успешного исхода” p , в то время как общее количество “испытаний” n (в данном случае – количество квантов, направленных на прибор) чрезвычайно велико. Таким образом возникает проблема асимптотического анализа биномиального распределения при фиксированном p и $n \rightarrow \infty$.

Пример 7.2 (*определение общего содержания серы в дизельном топливе*). Общее содержание серы служит одной из важных характеристик экологической чистоты дизельного топлива. Речь идет не об “элементарной сере” (процентном содержании химического элемента S, что с высокой степенью точности определяется с помощью спектрального анализа вещества), а способности элемента S при сгорании топлива соединяться с кислородом, образуя серный газ SO_2 . Именно этот газ через выхлопные трубы

машин попадает в среду нашего обитания и соединяется с водой, образуя серную кислоту H_2SO_4 . Ну, а что такое серная кислота, и что она может натворить с нашими легкими, вы знаете из школьного курса химии.

Итак, речь идет о химической активности серы, содержащейся в дизельном топливе в связанном виде. Анализ этой активности производится следующим образом. Берется определенное количество дизельного топлива, скажем 100 грамм, и сжигается в замкнутой колбе. Продукты сгорания частично выпадают в золу или в виде дыма по трубчатому отводу попадают в другую замкнутую колбу, наполненную водой. Серный газ соединяется с водой, образуя раствор серной кислоты. Титруя этот раствор определенным количеством щелочи, мы можем определить общее количество элемента серы, которое из дизельного топлива через сжигание и последующее соединение с кислородом и водой перешло в серную кислоту. Разделив это количество серы на вес анализируемой пробы топлива (100 грамм) и умножив результат на 100%, мы получим результат x нашего статистического эксперимента по наблюдению случайной величины X .

Повторные анализы аналогичных проб той же партии топлива, в тех же условиях эксперимента и на тех же приборах указывают на значительную изменчивость результатов каждого эксперимента. Метрологический анализ испытаний указывает на то, что эта изменчивость в первую очередь обусловлена случайным характером процессов “спекания” определенного количества серы с другими продуктами сгорания и выпадения их в золу, а также неполным соединением серного газа с водой. Грубо говоря, каждая молекула серы только с некоторой достаточно высокой вероятностью p может достичь своего конечного состояния в молекуле серной кислоты и внести свой вклад в результат x наблюдения X . Понятно, что количество n молекул серы в пробе топлива чрезвычайно велико. Следовательно, мы имеем дело с проблемой асимптотического анализа биномиального распределения при растущем числе испытаний n и постоянной вероятности успеха p .

Ограничимся рассмотрением этих двух примеров, из которых легко видеть, что существует обширнейший класс статистических экспериментов, связанных с наблюдением линейной функции от случайной величины с биномиальным законом распределения $B(n, p)$, в котором $p = \text{const}$, а n чрезвычайно велико. Проведем асимптотический анализ такой ситуации и начнем его с исследования асимптотического поведения X/n – частотной оценки вероятности p успешного испытания в схеме Бернулли. Тот факт, что при $n \rightarrow \infty$ относительная частота X/n стремится к p , в определенном

вероятностном смысле устанавливает один из основных законов теории вероятностей, открытый И. Бернулли в XVII веке.

Теорема 7.1. (Закон больших чисел Бернулли). Пусть $X \sim B(n, p)$. Тогда, каково бы ни было $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X}{n} - p\right| > \varepsilon\right) = 0.$$

Доказательство. Воспользуемся неравенством Чебышева в форме следствия 6.1, где в случае биномиального распределения $EX = np$ и $DX = np(1-p)$. Имеем

$$P\left(\left|\frac{X}{n} - p\right| > \varepsilon\right) \leq \frac{D(X/n)}{\varepsilon^2} = \frac{np(1-p)/n^2}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \rightarrow 0,$$

когда $n \rightarrow \infty$.

Закон больших чисел разъясняет природу стабилизации относительной частоты выпадения герба около значения $p = 1/2$, которую мы наблюдали на первой лекции по теории вероятностей. Действительно, в случайных экспериментах нельзя утверждать, что $|X/n - p| \leq \varepsilon$, начиная с некоторого n . Истина в том, что, начиная с некоторого n , это неравенство выполняется с любой, наперед заданной и сколь угодно близкой к единице вероятностью. Таким образом, мы должны сказать, что в данном случае наблюдается *сходимость по вероятности*, которая имеет совершенно другую природу, чем та сходимость, которую мы изучаем в курсе математического анализа.

Вывод закона больших чисел содержит также объяснение феномену, связанному с порядком $n^{-1/2}$ ошибки в приближении $p (= 1/2)$ величиной X/n . Действительно, в случае $p = 1/2$ стандартное отклонение $\sigma = \sqrt{D(X/n)} = (2\sqrt{n})^{-1}$, распределение случайной величины X/n симметрично, и в силу правила “двух сигм” интервал $0.5 \pm n^{-1/2}$ накрывает 90% центральной части области возможных значений X/n .

Естественно, если не делить X на n , то $X \rightarrow \infty$ по вероятности, когда $n \rightarrow \infty$. Но если X центрировать ее средним значением np и затем масштабировать стандартным отклонением, то построенная таким образом случайная величина $Y_n = (X - np) / \sqrt{np(1-p)}$ имеет при $n \rightarrow \infty$ невырожденное распределение. Вид этого распределения устанавливает знаменитая *предельная теорема Муавра–Лапласа* (XVIII век!). При доказательстве существенно используется следующий технический результат.

Лемма 7.1. Пусть $X \sim B(n, p)$, $n \rightarrow \infty$ и целое $k \rightarrow \infty$ так, что $1 > \hat{p} = k/n = O(1)$. Тогда

$$P(X = k) = f(k | n, p) = \frac{1}{\sqrt{2\pi n \hat{p}(1 - \hat{p})}} \exp\{-nH(\hat{p})\} (1 + O(n^{-1})),$$

где

$$H(x) = x \ln \frac{x}{p} + (1 - x) \ln \frac{1 - x}{1 - p}, \quad 0 < x < 1.$$

Доказательство. Воспользуемся асимптотической формулой Стирлинга

$$n! = \sqrt{2\pi n} n^n e^{-n} (1 + O(n^{-1}))$$

для факториалов $n!$, $k!$ и $(n - k)!$ в биномиальном коэффициенте C_n^k и представим функцию плотности биномиального распределения в асимптотическом виде:

$$\begin{aligned} f(k | n, p) &= \frac{n!}{k!(n - k)!} p^k (1 - p)^{n - k} = \\ &= \frac{\sqrt{2\pi n} n^n e^{-n} p^k (1 - p)^{n - k}}{\sqrt{2\pi k} k^k e^{-k} \sqrt{2\pi(n - k)} (n - k)^{n - k} e^{-n + k}} \left(1 + O\left(\frac{1}{n}\right)\right) = \\ &= \frac{\exp\{n \ln n - k \ln k - (n - k) \ln(n - k) + k \ln p + (n - k) \ln(1 - p)\}}{\sqrt{2\pi n \frac{k}{n} \left(1 - \frac{k}{n}\right)}} \\ &= (1 + O(n^{-1})). \end{aligned}$$

Доказательство завершается очевидными преобразованиями выражения, стоящего в фигурных скобках под экспонентой, к виду $\{-nH(\hat{p})\}$.

Лекция 12

Теорема 7.2. (Локальная предельная теорема Муавра–Лапласа). Пусть при $n \rightarrow \infty$ целое $k = np + O(\sqrt{n})$. Тогда

$$f(k | n, p) = \frac{1}{\sqrt{2\pi np(1 - p)}} \exp\left\{-\frac{(k - np)^2}{2np(1 - p)}\right\} \left(1 + O\left(n^{-1/2}\right)\right).$$

Доказательство. Так как по условию теоремы $\hat{p} = k/n = p + O(n^{-1/2})$, то естественно воспользоваться асимптотической формулой леммы 7.1, разлагая функции $(\hat{p}(1 - \hat{p}))^{-1/2}$ и $H(\hat{p})$ в ряд Тейлора по степеням $\hat{p} - p = O(n^{-1/2})$.

Имеем

$$\begin{aligned} (\hat{p}(1 - \hat{p}))^{-1/2} &= \left((p + O(n^{-1/2}))(1 - p + O(n^{-1/2})) \right)^{-1/2} = \\ &= (p(1 - p))^{-1/2} (1 + O(n^{-1/2})), \end{aligned}$$

и для доказательства теоремы остается показать, что

$$n H(\hat{p}) = \frac{(k - np)^2}{2np(1 - p)} + O\left(\frac{1}{\sqrt{n}}\right). \quad (1)$$

Разложим

$$H(\hat{p}) = \hat{p} \ln \frac{\hat{p}}{p} + (1 - \hat{p}) \ln \frac{1 - \hat{p}}{1 - p}$$

в ряд Тейлора в окрестности точки $\hat{p} = p$:

$$H(\hat{p}) = H(p) + (\hat{p} - p)H'(p) + \frac{(\hat{p} - p)^2}{2!}H''(p) + \frac{(\hat{p} - p)^3}{3!}H'''(p + \lambda(\hat{p} - p)),$$

где, как и в любом разложении Тейлора, $0 < \lambda < 1$.

Имеем $H(p) = 0$, и так как

$$H'(x) = \ln \frac{x}{p} - \ln \frac{1 - x}{1 - p},$$

то $H'(p) = 0$. Далее,

$$H''(x) = \frac{1}{x} + \frac{1}{1 - x},$$

откуда $H''(p) = (p(1 - p))^{-1}$. Наконец,

$$H'''(x) = -\frac{1}{x^2} + \frac{1}{(1 - x)^2},$$

что влечет ограниченность $H'''(p + \lambda(\hat{p} - p))$ при больших n , поскольку p отграничено от 0 и 1. Таким образом,

$$H(\hat{p}) = \frac{(\hat{p} - p)^2}{2p(1 - p)} + O((\hat{p} - p)^3),$$

что, очевидно, эквивалентно (1).

Теорема 7.3 (Интегральная предельная теорема Муавра–Лапласа). Для любых постоянных a и b и случайной величины $X \sim B(n, p)$ справедливо асимптотическое представление

$$\lim_{n \rightarrow \infty} P \left(a \leq \frac{X - np}{\sqrt{np(1-p)}} < b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx. \quad (2)$$

Доказательство. Используя теорему 7.2, представим вероятность $P(a \leq Y_n < b)$, где

$$Y_n = (X - np) / \sqrt{np(1-p)}$$

в виде

$$P(a \leq Y_n < b) = \sum_{k \in A} \frac{1}{\sqrt{2\pi np(1-p)}} \exp \left\{ -\frac{(k - np)^2}{2np(1-p)} \right\} \left(1 + O \left(\frac{1}{\sqrt{n}} \right) \right), \quad (3)$$

где множество целых чисел

$$A = \left\{ k : a \leq \frac{k - np}{\sqrt{np(1-p)}} < b \right\}.$$

Применение локальной предельной теоремы в данном случае оправдано: если $k \in A$, то при $n \rightarrow \infty$ справедливо асимптотическое представление $k = np + O(\sqrt{n})$.

Покажем теперь, что правая часть (3) представляет собой сумму Дарбу для интеграла в правой части равенства (2). Для этого положим

$$x_k = \frac{k - np}{\sqrt{np(1-p)}}, \quad \Delta x_k = x_k - x_{k-1} = \frac{1}{\sqrt{np(1-p)}}, \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

и разобьем отрезок $[a; b]$ точками x_k , $k \in A$. Поскольку $\Delta x_k \rightarrow 0$ при $n \rightarrow \infty$, а суммарная длина отрезков разбиения

$$\sum_{k \in A} \Delta x_k \approx b - a,$$

то число отрезков разбиения растет с ростом n , в то время как их длина стремится к нулю. Следовательно,

$$\sum_{k \in A} \varphi(x_k) \Delta x_k \longrightarrow \int_a^b \varphi(x) dx.$$

Для завершения доказательства остается только заметить, что $0 < \varphi(x) < 1$, и поэтому при $n \rightarrow \infty$

$$0 \leq \sum_{k \in A} \varphi(x_k) \Delta x_k \cdot O\left(\frac{1}{\sqrt{n}}\right) \leq \frac{b-a}{\sqrt{n}} \rightarrow 0.$$

З а м е ч а н и е. Интегральная теорема Муавра–Лапласа иногда формулируется в терминах следующего приближенного равенства для распределения биномиальной случайной величины X :

$$P(a \leq X < b) \approx \frac{1}{\sqrt{2\pi}} \int_{\frac{a-np}{\sqrt{np(1-p)}}}^{\frac{b-np}{\sqrt{np(1-p)}}} \exp\left\{-\frac{x^2}{2}\right\} dx, \quad n \gg 1. \quad (4)$$

В такой записи теоремы знак \approx означает асимптотическую эквивалентность правой и левой частей (4) (их отношение стремится к единице при $n \rightarrow \infty$) лишь в случае незначительной удаленности a и b от центра np биномиального распределения. Для этого достаточно сравнить запись одного и того же утверждения с помощью формул (2) и (4), чтобы убедиться в справедливости формулы (4) лишь при значениях a и b порядка $np + O(\sqrt{n})$. В противном случае как левая, так и правая части (4) с ростом n стремятся к единице, но с разной скоростью. Асимптотический анализ биномиальных вероятностей в областях, удаленных от np на порядок больший, чем $O(\sqrt{n})$, составляет содержание *теорем о больших отклонениях биномиального распределения*, которые в нашем курсе теории вероятностей рассматриваться не будут.

Как известно из общего курса анализа, интеграл Эйлера–Пуассона

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx = 1,$$

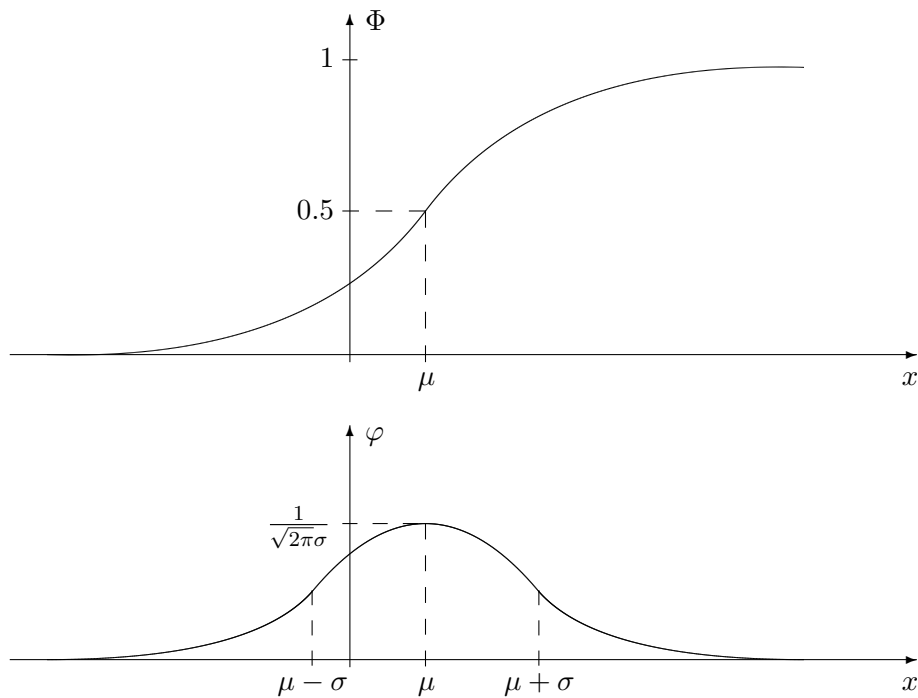
поэтому при любых $\mu \in \mathbb{R}$ и $\sigma \in \mathbb{R}_+$

$$\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} \exp\left\{-\frac{t^2}{2}\right\} dt = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^x \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} dt$$

есть функция распределения, а

$$\frac{1}{\sigma} \varphi \left(\frac{x - \mu}{\sigma} \right) = \frac{d}{dx} \Phi \left(\frac{x - \mu}{\sigma} \right) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} -$$

функция плотности. Эти функции определяют двухпараметрическое семейство *нормальных* или *гауссовских* распределений с носителем $\mathcal{X} = \bar{\mathbb{R}} = [-\infty, +\infty]$ и параметрическим пространством $\Theta = \mathbb{R} \times \mathbb{R}_+$. Мы будем обозначать это распределение $\mathcal{N}(\mu, \sigma^2)$.



Если $\mu = 0$, а $\sigma = 1$, то $\mathcal{N}(0, 1)$ называется *стандартным нормальным* распределением; ему соответствуют функция распределения $\Phi(x)$ и функция плотности $\varphi(x)$. Поскольку параметры нормального распределения являются параметрами сдвига (μ) и масштаба (σ), то семейство нормальных распределений замкнуто относительно линейных преобразований случайных величин: если $X \sim \mathcal{N}(0, 1)$, то $Y = \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

Так как $\exp\{-x^2/2\}$ – четная функция, то нормальное распределение симметрично относительно точки $x = \mu$, которая, как легко видеть, является модой распределения. Симметричность функции плотности влечет также очевидные равенства: $\Phi(-x) = 1 - \Phi(x)$ и $\Phi(0) = 1/2$. Графики функции распределения и функции плотности нормального закона $\mathcal{N}(\mu, \sigma^2)$ представлены на рисунке.

Так как среднее значение стандартного нормального распределения

$$\mathbf{E}X = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp \left\{ -\frac{x^2}{2} \right\} dx = 0,$$

(как интеграл от нечетной функции по всему \mathbb{R}), то $\sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$ имеет среднее значение μ . В силу той же нечетности подынтегральных функций все центральные моменты нечетного порядка

$$\mu_{2k+1} = \mathbf{E}(X - \mu)^{2k+1} = 0.$$

Четные моменты вычисляются с помощью гамма-функции Эйлера:

$$\begin{aligned} \mu_{2k} &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x - \mu)^{2k} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} dx = \frac{2\sigma^{2k}}{\sqrt{2\pi}} \int_0^{\infty} t^{2k} \exp \left\{ -\frac{t^2}{2} \right\} dt = \\ &= \frac{\sigma^{2k} 2^k \sqrt{2}}{\sqrt{2\pi}} \int_0^{\infty} x^{k-1/2} e^{-x} dx = \frac{\sigma^{2k} 2^k \sqrt{2}}{\sqrt{2\pi}} \Gamma \left(k + \frac{1}{2} \right) = \sigma^{2k} (2k - 1)!!. \end{aligned}$$

В частности, $\mathbf{D}X = \sigma^2$, что оправдывает обозначения параметров μ и σ^2 нормального распределения. Так как $\mu_4 = 3\sigma^4$, то коэффициент эксцесса $\gamma_2 = 0$. В силу этого пикообразность или сплюсченность вершины функции плотности любого распределения соотносится с кривой нормальной плотности, которая часто называется в честь Ф. Гаусса *гауссиадой*.

Итак, возвращаясь к нашим примерам с определениями видимой звездной величины и общего содержания серы в дизельном топливе, мы должны прийти к заключению о нормальности распределения наблюдаемой случайной величины (заметим, что это предположение блестяще подтверждается статистическим анализом реальных данных). В этом распределении μ играет роль параметра, неизвестное значение которого составляет предмет проводимого исследования (эксперимента), в то время как значение σ характеризует ошибку наблюдений.

§8. Векторные случайные величины. Независимость случайных величин

Лекция 13

При определении действительной случайной величины мы интерпретировали ее как некоторую числовую характеристику исследуемого объекта. Однако на практике чаще имеют дело с одновременным наблюдением нескольких числовых характеристик – случайным вектором, распределение которого так же, как и в одномерном случае, порождается распределением на измеримом пространстве (Ω, \mathcal{A}) элементарных исходов статистического эксперимента. Чтобы провести аналогию с определением скалярной случайной величины, мы должны вспомнить строение борелевских множеств в \mathbb{R}^n . Роль интервалов здесь играют *прямоугольники* – подмножества \mathbb{R}^n вида $B = B_1 \times \dots \times B_n$, где каждое B_k есть открытый (a_k, b_k) , полуоткрытый $(a_k, b_k]$ и $[a_k, b_k)$ или замкнутый $[a_k, b_k]$ интервал на действительной прямой \mathbb{R} . Конечные объединения непересекающихся прямоугольников образуют булеву алгебру подмножеств \mathbb{R}^n , а наименьшая σ -алгебра \mathcal{B}^n , содержащая эту булеву алгебру, образует класс измеримых подмножеств \mathbb{R}^n или *событий*. Таким образом мы получаем измеримое пространство $(\mathbb{R}^n, \mathcal{B}^n)$.

Определение 8.1. *Векторной случайной величиной* или *случайным вектором* называется измеримое отображение

$$X^{(n)} = X^{(n)}(\omega) = (X_1(\omega), \dots, X_n(\omega))$$

пространства элементарных исходов Ω , наделенного σ -алгеброй измеримых подмножеств \mathcal{A} , в n -мерное евклидово пространство \mathbb{R}^n с борелевской σ -алгеброй \mathcal{B}^n . Для любого $B \in \mathcal{B}^n$ справедливо включение

$$X^{(n) -1}(B) = \{\omega : X^{(n)}(\omega) \in B\} \in \mathcal{A}.$$

Теперь, по аналогии с одномерным случаем, зададим вероятность P_n на $(\mathbb{R}^n, \mathcal{B}^n)$, порожденную вероятностью P на (Ω, \mathcal{A}) , соотношением

$$P_n(B) = P\left(X^{(n) -1}(B)\right), \quad \forall B \in \mathcal{B}^n.$$

Как будет видно в дальнейшем, исходное вероятностное пространство (Ω, \mathcal{A}, P) играет более важную роль в характеристике распределения $X^{(n)}$,

если $n > 1$. Мы будем изучать вероятностные модели, которые можно записать в виде интеграла Лебега

$$P(X^{(n)} \in B) = \int_B f(x_1, \dots, x_n) d\mu_1(x_1) \cdots d\mu_n(x_n)$$

от неотрицательной функции $f(x_1, \dots, x_n)$ по мере $d\mu = d\mu_1 \cdots d\mu_n$, где каждая σ -конечная мера μ_i , $i = 1, \dots, n$ на борелевской прямой $(\mathbb{R}, \mathcal{B})$ является или считающей мерой, или мерой Лебега. В таком случае вычисление вероятности событий $B \in \mathcal{B}^n$ сводится или к суммированию вероятностей отдельных точек в \mathbb{R}^n , или к вычислению кратных интегралов Римана. Функция f в данном случае выступает в роли n -мерной *функции плотности*. Естественно, можно ввести также понятие n -мерной *функции распределения*

$$F(x_1, \dots, x_n) = P(X_1 < x_1, \dots, X_n < x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \dots, t_n) d\mu_1(t_1) \cdots d\mu_n(t_n),$$

однако при $n > 1$ с помощью этой функции можно выразить только вероятности “прямоугольников” в \mathbb{R}^n , в то время как вероятность попадания случайного вектора в подмножества более сложной конфигурации (например, эллипсоиды) приходится вычислять с помощью интеграла от функции плотности. Как и в одномерном случае, n -мерная *функция распределения однозначно определяет распределение вероятностей на $(\mathbb{R}^n, \mathcal{B}^n)$* , то есть имеет место n -мерный аналог теоремы 4.1.

Из определения функции распределения вытекает, что в случае непрерывного распределения ($\mu = \mu_1 \times \cdots \times \mu_n$ – мера Лебега) функция плотности f выражается через функцию распределения посредством дифференцирования

$$f(x_1, \dots, x_n) = \frac{\partial^n F(x_1, \dots, x_n)}{\partial x_1 \cdots \partial x_n},$$

а в дискретном случае (μ – считающая мера, приписывающая единицу каждой точке \mathbb{R}^n с целочисленными координатами)

$$f(x_1, \dots, x_n) = P(X^{(n)} = x^{(n)}) = P(X_1 = x_1, \dots, X_n = x_n).$$

Я полагаю, вы сами сможете записать аналогичные связи между F и f в “смешанном” дискретно-непрерывном случае, когда часть компонент случайного вектора имеет непрерывное распределение, а другая – дискретное.

Как вычислить совместное распределение отдельных компонент X_{i_1}, \dots, X_{i_k} случайного вектора $X^{(n)}$? Для этого достаточно в функции распределения $X^{(n)}$ устремить к $+\infty$ все переменные, отличные от x_{i_1}, \dots, x_{i_k} , или, что то же, проинтегрировать функцию плотности по каждой из переменных, отличных от x_{i_1}, \dots, x_{i_k} , в пределах $\pm\infty$.

Заметим, что в теории вероятностей принято называть распределения каждой компоненты (случайной величины) $X_i, i = 1, \dots, n$, — *маргинальными* или *частными* распределениями.

Пример 8.1 (равномерное распределение на круге.) В часть плоскости \mathbb{R}^2 , ограниченную окружностью $x^2 + y^2 = r^2$, наугад бросается точка, так что ее координаты (x, y) представляют реализацию случайного вектора (X, Y) . Как и в случае с бросанием точки на отрезок прямой, термин “наугад” понимается в смысле зависимости вероятности попадания точки в некоторую, измеримую по Лебегу часть B круга только от площади B . Те же рассуждения, что и при выводе равномерного распределения на отрезке, приводят нас к равномерному распределению (X, Y) с функцией плотности (по мере Лебега $d\mu = dxdy$) $f(x, y)$, равной постоянной $1/\pi r^2$, если $x^2 + y^2 \leq r^2$, и равной нулю вне этого круга.

Найдем функцию плотности $f^X(x)$ маргинального распределения X . Для этого мы должны проинтегрировать функцию $f(x, y)$ по переменной y в пределах $\pm\infty$ при каждом фиксированном значении $x \in \mathbb{R}$. Если x фиксировано, то $f(x, y)$ отлична от нуля и равна $1/\pi r^2$ только при значениях y , удовлетворяющих неравенству $-\sqrt{r^2 - x^2} \leq y \leq \sqrt{r^2 - x^2}$. Следовательно,

$$f^X(x) = \frac{1}{\pi r^2} \int_{-\sqrt{r^2 - x^2}}^{\sqrt{r^2 - x^2}} dy = \frac{2}{\pi r^2} \sqrt{r^2 - x^2},$$

если $|x| \leq r$, и $f^X(x) = 0$ в противном случае. Легко видеть, что маргинальное распределение второй компоненты Y случайного вектора имеет тот же вид. Таким образом, маргинальные распределения компонент отличны от равномерного и имеют четко выраженную моду, совпадающую с началом координат.

Маргинальные плотности компонент случайного вектора наиболее просто находятся в том случае, когда функция плотности $X^{(n)}$ распадается в произведение функций плотности отдельных компонент. Понятно, что хотя бы в дискретном случае это говорит о некоторой “независимости” компонент случайного вектора. Чтобы ввести строгое определение такой независи-

мости, мы должны обратиться к σ -подалгебрам алгебры \mathcal{A} , порожденным каждой компонентой X_i , $i = 1, \dots, n$, вектора $X^{(n)}$.

Пусть $X = X(\omega)$ – случайная величина на (Ω, \mathcal{A}) со значениями в измеримом пространстве $(\mathbb{R}, \mathcal{B})$. Рассмотрим класс $\mathcal{A}_X = \{X^{-1}(B), B \in \mathcal{B}\}$ всех прообразов элементов борелевского поля \mathcal{B} , полагая $X^{-1}(\mathbb{R}) = \Omega$. Имеет место

Предложение 8.1. *Класс \mathcal{A}_X подмножеств Ω является σ -алгеброй (подалгеброй \mathcal{A} .)*

Доказательство. Достаточно проверить аксиомы булевой σ -алгебры (см. определение 2.5).

(A1). По определению \mathcal{A}_X пространство элементарных исходов

$$\Omega = X^{-1}(\mathbb{R}) \in \mathcal{A}_X.$$

(A2). Покажем, что дополнение $(X^{-1}(B))^c \in \mathcal{A}_X$, каково бы ни было $B \in \mathcal{B}$. Действительно, событие, противоположное

$$X^{-1}(B) = \{\omega : X(\omega) \in B\},$$

означает, что $X(\omega)$ не принадлежит B , то есть $X(\omega) \in B^c$. Так как $B^c \in \mathcal{B}$, то

$$X^{-1}(B^c) = (X^{-1}(B))^c \in \mathcal{A}_X.$$

(A3)_S. Рассуждения, аналогичные предыдущему пункту, показывают, что

$$\bigcup_1^\infty X^{-1}(B_i) = X^{-1}\left(\bigcup_1^\infty B_i\right) \in \mathcal{A}_X.$$

Легко понять, что данное утверждение справедливо не только для скалярных случайных величин, но и случайных векторов. Теперь мы в состоянии ввести одно из фундаментальнейших понятий теории вероятностей и математической статистики.

Определение 8.2. Случайные величины (случайные векторы) X_1, \dots, X_n , заданные на одном и том же измеримом пространстве (Ω, \mathcal{A}) , называются *независимыми в совокупности* или *совместно независимыми*, если независимы σ -подалгебры $\mathcal{A}_{X_1}, \dots, \mathcal{A}_{X_n}$ σ -алгебры \mathcal{A} , порожденные соответствующими случайными величинами.

Таким образом, в соответствии с определением 3.4 независимости σ -алгебр, для любых элементов (событий) B_1, \dots, B_n борелевского поля \mathcal{B} справедливо равенство

$$P(X_1 \in B_1, \dots, X_n \in B_n) = \prod_1^n P(X_i \in B_i), \quad (1)$$

то есть совместное распределение независимых случайных величин распадается в произведение их маргинальных распределений. Оказывается, для независимости случайных величин достаточно потребовать выполнения более слабого условия, состоящего в возможности представления совместной функции распределения X_1, \dots, X_n в виде произведения маргинальных функций распределения.

Предложение 8.2. *(критерий независимости случайных величин).* Случайные величины X_1, \dots, X_n независимы в совокупности тогда и только тогда, когда их совместная функция распределения (функция плотности) распадается в произведение маргинальных функций распределения (маргинальных функций плотности):

$$F(x_1, \dots, x_n) = \prod_1^n F^{X_i}(x_i), \quad f(x_1, \dots, x_n) = \prod_1^n f^{X_i}(x_i).$$

Доказательство. Условимся обозначать полужирной буквой \mathbf{P} вероятность на исходном вероятностном пространстве (Ω, \mathcal{A}) , на котором определены случайные величины X_1, \dots, X_n , а обычной буквой P – вероятность на $(\mathbb{R}^n, \mathcal{B}^n)$, которая единственным образом определяется заданием функции распределения F . Тогда

$$\begin{aligned} F(x_1, \dots, x_n) &= P\left(\bigcap_{i=1}^n \{X_i \in (-\infty, x_i)\}\right) = \mathbf{P}\left(\bigcap_{i=1}^n X_i^{-1}((-\infty, x_i))\right) = \\ &= \prod_{i=1}^n \mathbf{P}(X_i^{-1}((-\infty, x_i))) = \prod_{i=1}^n P(X_i \in (-\infty, x_i)) = \prod_1^n F^{X_i}(x_i), \end{aligned}$$

то есть свойство мультипликативности совместной функции распределения есть частный случай равенства (1).

Для доказательства достаточности условия мультипликативности покажем, что для случайных величин X и Y равенство $F(x, y) = F^X(x)F^Y(y)$

при любых $x, y \in \mathbb{R}$ влечет

$$P(X \in B_1, Y \in B_2) = P(X \in B_1)P(Y \in B_2)$$

каковы бы ни были $B_1, B_2 \in \mathfrak{B}$ (общий случай, касающийся независимости $n > 2$ случайных величин, рассматривается с привлечением метода математической индукции). Отсюда будет следовать независимость порожденных сигма-алгебр. Действительно, любой элемент $A_1 \in \mathcal{A}_X$ имеет вид $X^{-1}(B_1)$ с некоторым $B_1 \in \mathfrak{B}$, и, аналогично, любой $A_2 \in \mathcal{A}_Y$ имеет вид $Y^{-1}(B_2)$ с $B_2 \in \mathfrak{B}$, так что при любых $A_1 \in \mathcal{A}_X$, $A_2 \in \mathcal{A}_Y$

$$\begin{aligned} \mathbf{P}(A_1 \cap A_2) &= \mathbf{P}(X^{-1}(B_1) \cap Y^{-1}(B_2)) = P(X \in B_1, Y \in B_2) = \\ &P(X \in B_1)P(Y \in B_2) = \mathbf{P}(X^{-1}(B_1)) \mathbf{P}(Y^{-1}(B_2)) = \mathbf{P}(A_1)\mathbf{P}(A_2), \end{aligned}$$

то есть сигма-алгебры \mathcal{A}_X и \mathcal{A}_Y независимы.

Перепишем условие независимости

$$\begin{aligned} P(X \in (-\infty, x))P(Y \in (-\infty, y)) &= \\ P(\{X \in (-\infty, x)\} \cap \{Y \in (-\infty, y)\}) & \end{aligned}$$

в виде

$$\begin{aligned} F^X(x) = P(X \in (-\infty, x)) &= \\ \frac{P(\{X \in (-\infty, x)\} \cap \{Y \in (-\infty, y)\})}{F^Y(y)}. & \end{aligned} \quad (2)$$

Поскольку функция распределения $F^X(x)$ однозначно определяет маргинальное распределение $P(X \in B_1)$ случайной величины X (теорема 4.1), то равенство (2) влечет

$$P(X \in B_1) = \frac{P(\{X \in B_1\} \cap \{Y \in (-\infty, y)\})}{F^Y(y)}$$

или, что то же,

$$F^Y(y) = P(Y \in (-\infty, y)) = \frac{P(\{X \in B_1\} \cap \{Y \in (-\infty, y)\})}{P(X \in B_1)}$$

для любых $B_1 \in \mathfrak{B}$. Используя снова теорему 4.1 об однозначном определении распределения вероятностей случайной величины Y посредством ее функции распределения $F^Y(y)$, получаем требуемое определение независимости:

$$P(X \in B_1, Y \in B_2) = P(X \in B_1)P(Y \in B_2),$$

каковы бы ни были $B_1, B_2 \in \mathfrak{B}$.

Утверждение теоремы, касающееся функции плотности, следует немедленно из соотношения между функцией распределения и функцией плотности:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) d\mu_1(u) d\mu_2(v).$$

Следующее утверждение, относящееся к функциям от независимых случайных величин, позволяет вычислять моментные характеристики некоторых распределений значительно проще, чем это делалось в §6.

Предложение 8.3. *Если X_1, \dots, X_n независимы в совокупности, то*

1⁰. *независимы в совокупности случайные величины*

$$Y_1 = g_1(X_1), \dots, Y_n = g_n(X_n),$$

где $g_i, i = 1, \dots, n$ – измеримые функции;

$$2^0. \mathbf{E} \prod_1^n X_i = \prod_1^n \mathbf{E} X_i;$$

$$3^0. \mathbf{D} \sum_1^n X_i = \sum_1^n \mathbf{D} X_i.$$

Доказательство. 1⁰. Поскольку σ -алгебры, порожденные случайными величинами Y_1, \dots, Y_n , являются подалгебрами соответствующих σ -алгебр, порожденных X_1, \dots, X_n , а последние независимы (см. определение 3.4), то данное утверждение следует непосредственно из определения 8.2 независимости случайных величин.

2⁰. Пусть $f_i(\cdot)$ – функция плотности X_i по мере $\mu_i, i = 1, \dots, n$. Тогда, в силу предложения 8.2, совместная функция плотности

$$f(x_1, \dots, x_n) = \prod_1^n f_i(x_i),$$

так что

$$\mathbf{E} \prod_1^n X_i = \int_{\mathbf{R}} x_1 f_1(x_1) d\mu_1(x_1) \cdots \int_{\mathbf{R}} x_n f_n(x_n) d\mu_n(x_n) = \prod_1^n \mathbf{E} X_i.$$

3⁰. Используя только что доказанное утверждение (2) и свойство линейности математического ожидания, получаем

$$\begin{aligned} \mathbf{D} \sum_1^n X_i &= \mathbf{E} \left(\sum_1^n (X_i - \mathbf{E}X_i) \right)^2 = \\ &= \mathbf{E} \left[\sum_1^n (X_i - \mathbf{E}X_i)^2 + \sum_{i \neq j} (X_i - \mathbf{E}X_i)(X_j - \mathbf{E}X_j) \right] = \\ &= \sum_1^n \mathbf{E}(X_i - \mathbf{E}X_i)^2 + \sum_{i \neq j} \mathbf{E}(X_i - \mathbf{E}X_i) \cdot \mathbf{E}(X_j - \mathbf{E}X_j) = \\ &= \sum_1^n \mathbf{E}(X_i - \mathbf{E}X_i)^2 = \sum_1^n \mathbf{D}X_i. \end{aligned}$$

Как будет видно в дальнейшем, вывод ряда вероятностных моделей строится на стохастическом представлении наблюдаемой случайной величины в виде суммы независимых случайных величин: $X = X_1 + \dots + X_n$, и при этом распределение каждой X_i , $i = 1, \dots, n$ имеет достаточно простой вид, например, вычислить моменты X_i намного проще, чем моменты X . В таком случае формулы предложения 8.3 указывают прямой путь к вычислению моментов, а иногда и распределения, случайной величины X . В сущности, мы уже использовали технику таких представлений, когда выводили биномиальное распределение – распределение числа успехов в испытаниях Бернулли.

Лекция 14

Пример 8.2 (о некоторых свойствах биномиального распределения). Результат каждого i -го испытания в схеме Бернулли можно регистрировать как значение индикаторной функции успеха, обозначая цифрой 1 успех, а цифрой 0 неудачу. Таким образом, с i -м испытанием соотносится случайная величина X_i , принимающая значение 1 с вероятностью p и значение 0 с вероятностью $1 - p$. Последовательности из n независимых испытаний Бернулли ставится в соответствие случайный вектор $X^{(n)} = (X_1, \dots, X_n)$, состоящий из независимых, одинаково распределенных по закону $B(1, p)$

компонент (напомним, $B(1, p)$ есть частный случай биномиального распределения, которое мы назвали двухточечным распределением). В таких обозначениях случайная величина X , реализация которой равна числу успехов в n испытаниях (числу X_i , принявших значение 1), представима в виде

$$X = \sum_1^n X_i,$$

и в силу предложения 8.3

$$\mathbf{E}X = \sum_1^n \mathbf{E}X_i = n\mathbf{E}X_1, \quad \mathbf{D}X = \sum_1^n \mathbf{D}X_i = n\mathbf{D}X_1.$$

Имеем:

$$\begin{aligned} \mathbf{E}X_1 &= 1 \cdot p + 0 \cdot (1 - p) = p, \\ \mathbf{E}X_1^2 &= \mathbf{E}X_1 = p, \quad \mathbf{D}X_1 = p - p^2 = p(1 - p), \end{aligned}$$

откуда немедленно получаем известные нам и полученные в результате более сложных выкладок формулы моментов биномиального распределения: $\mathbf{E}X = np$, $\mathbf{D}X = np(1 - p)$.

Укажем еще на одно интересное применение стохастического представления биномиальной случайной величины X в виде суммы независимых случайных величин.

Предложение 8.4 (теорема сложения для биномиального распределения). Если X_1, \dots, X_m независимы в совокупности и

$$X_k \sim B(n_k, p), \quad k = 1, \dots, m,$$

то

$$X = \sum_1^m X_k \sim B(n, p),$$

где $n = n_1 + \dots + n_m$.

Доказательство. Каждое X_k есть сумма n_k независимых, одинаково распределенных по закону $B(1, p)$ случайных величин. Следовательно, X есть сумма n таких же величин, откуда $X \sim B(n, p)$.

Распределения, для которых справедливы теоремы сложения, составляют особый класс *устойчивых* законов распределений, и изучению свойств таких распределений посвящаются отдельные монографии. Вы, наверное, догадываетесь, что устойчивым является пуассоновское распределение, как

предел биномиального. В дальнейшем мы покажем, что это в действительности так, разработав более совершенный математический аппарат доказательств теорем сложения. А сейчас мы докажем устойчивость нормального закона, получив предварительно общую формулу для распределения суммы независимых случайных величин.

Предложение 8.5 (*формула свертки распределений*). Пусть X_1 и X_2 независимы и имеют непрерывные распределения с функциями плотности $f_1(x)$ и, соответственно, $f_2(x)$ по мере Лебега $d\mu = dx$. Тогда функция плотности $f(x)$ распределения случайной величины $X = X_1 + X_2$ есть свертка функций f_1 и f_2 :

$$f(x) = \int_{-\infty}^{\infty} f_1(t)f_2(x-t)dt = \int_{-\infty}^{\infty} f_2(t)f_1(x-t)dt.$$

Доказательство. Совместная функция плотности $f(x_1, x_2)$ независимых случайных величин X_1 и X_2 равна (см. предложение 8.2) произведению их функций плотности: $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. Используя известную нам формулу (см. формулы после определения 8.1)

$$P((X_1, X_2) \in B) = \int_B f_1(x_1)f_2(x_2)dx_1dx_2$$

для вычисления вероятностей попадания случайного вектора в любую измеримую область B на плоскости \mathbb{R}^2 , представим функцию распределения суммы случайных величин в виде

$$F(x) = P(X_1 + X_2 < x) = \int_{t+s < x} \int f_1(t)f_2(s)dtds = \int_{-\infty}^{\infty} f_1(t)dt \int_{-\infty}^{x-t} f_2(s)ds.$$

Дифференцируя правую часть последнего равенства по x , получаем искомую первую формулу для плотности $f(x)$. Вторая формула справедлива в силу симметрии вхождения функций f_1 и f_2 в интегральное представление $F(x)$.

Предложение 8.6 (*теорема сложения для нормального распределения*.) Если X_1, \dots, X_n независимы в совокупности и каждое $X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$, $k = 1, \dots, n$, то

$$X = \sum_1^n X_k \sim \mathcal{N}\left(\sum_1^n \mu_k, \sum_1^n \sigma_k^2\right).$$

Доказательство. Предложение достаточно доказать для случая $n = 2$, поскольку для произвольного числа слагаемых доказательство проводится методом индукции. При $n = 2$ формула свертки дает следующее выражение для функции плотности $f(x)$ случайной величины $X = X_1 + X_2$:

$$f(x) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} \left(\frac{t - \mu_1}{\sigma_1} \right)^2 - \frac{1}{2} \left(\frac{x - t - \mu_2}{\sigma_2} \right)^2 \right\} dt.$$

Приводя квадратическую форму под знаком экспоненты к виду

$$-\frac{1}{2} \left(\frac{t - a}{b} \right)^2 + h(a, b),$$

где a и b зависят от параметров μ_i и σ_i , $i = 1, 2$, и используя известную нам формулу

$$\frac{1}{\sqrt{2\pi}b} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} \left(\frac{t - a}{b} \right)^2 \right\} dt = 1,$$

находим искомую функцию плотности

$$f(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp \left\{ -\frac{(x - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}.$$

**§9. Моментные характеристики многомерных распределений.
Мультиномиальное и многомерное нормальное
распределения**

Для описания положения в пространстве, рассеяния и формы многомерных распределений обычно используются *смешанные* центральные моменты, вычисляемые как математические ожидания от произведения различных степеней центрированных средними значениями компонент случайного вектора:

$$\mathbf{E} [(X_1 - \mu_1)^{k_1} \cdots (X_n - \mu_n)^{k_n}],$$

где $\mu_i = \mathbf{E}X_i$, $i = 1, \dots, n$, – вектор средних значений компонент случайного вектора $X^{(n)}$. Мы будем иметь дело только с моментами второго порядка

$$\lambda_{ij} = \mathbf{E}(X_i - \mu_i)(X_j - \mu_j), \quad i, j = 1, \dots, n.$$

Матрица $\Lambda = \|\lambda_{ij}\|$ моментов второго порядка называется *ковариационной матрицей* или *матрицей ковариаций*

$$\text{cov}(X_i, X_j) = \lambda_{ij}.$$

Естественно, диагональ ковариационной матрицы Λ составляют дисперсии

$$\sigma_i^2 = \lambda_{ii} = \text{cov}(X_i, X_i)$$

соответствующих компонент X_i , $i = 1, \dots, n$ случайного вектора $X^{(n)}$, в то время как смешанные моменты λ_{ij} при $i \neq j$ характеризуют степень *линейной связности* компонент X_i и X_j . Этот термин требует специального обсуждения, ввиду его исключительной распространенности в приложениях многомерного статистического анализа.

Всё вертится около следующего лебеговского варианта известного неравенства Коши–Буняковского.

Неравенство Шварца. Пусть X и Y – случайные величины, а $g(X)$ и $h(Y)$ – измеримые функции от соответствующих величин, обладающие конечными вторыми моментами. Тогда

$$|\mathbf{E}g(X)h(Y)| \leq [\mathbf{E}g^2(X)\mathbf{E}h^2(Y)]^{1/2}$$

с равенством тогда и только тогда, когда функции g и h линейно связаны: существуют такие постоянные a и b , что

$$P(ag(X) + bh(Y) = 0) = 1.$$

Применим это неравенство к функциям

$$g(X) = X - \mu_X, \quad h(Y) = Y - \mu_Y,$$

где

$$\mu_X = \mathbf{E}X, \quad \mu_Y = \mathbf{E}Y.$$

Если случайные величины X и Y независимы, то, в силу предложения 8.3

$$\text{cov}(X, Y) = \mathbf{E}(X - \mu_X)(Y - \mu_Y) = \mathbf{E}(X - \mu_X)\mathbf{E}(Y - \mu_Y) = 0,$$

то есть независимые случайные величины имеют нулевую ковариацию. Если же X и Y линейно связаны:

$$Y - \mu_Y = a(X - \mu_X),$$

то в неравенстве Шварца достигается знак равенства, так что

$$\text{cov}(X, Y) = \lambda_{XY} = \mathbf{E}(X - \mu_X)(Y - \mu_Y) = \pm\sqrt{\mathbf{D}X \cdot \mathbf{D}Y} = \pm\sigma_X\sigma_Y$$

(естественно, мы предполагаем, что X и Y принимают по крайней мере два различных значения с ненулевой вероятностью). Эти два крайних значения в неравенстве Шварца оправдывают введение следующей меры линейной связности пары случайных величин.

Определение 9.1. Пусть X и Y – две случайные величины с конечными дисперсиями. Моментная характеристика

$$\rho = \rho_{XY} = \frac{\lambda_{XY}}{\sigma_X\sigma_Y}$$

называется *коэффициентом корреляции* между случайными величинами X и Y .

Итак, если X и Y независимы, то $\rho = 0$, если же $Y = a + bX$ при некоторых постоянных a и b , то $|\rho| = 1$, причем $\rho = -1$, если $b < 0$, и $\rho = +1$, если $b > 0$. Однако, равенство $\rho = 0$ не означает, что случайные величины X и Y независимы!

Пример 9.1 (*зависимых случайных величин с нулевым коэффициентом корреляции*). Покажем, что случайные величины X и Y , равномерно распределенные в круге радиуса r , зависимы, но $\rho_{XY} = 0$.

Действительно, совместная функция плотности $f(x, y)$ случайных величин X и Y (см. пример 8,1) отлична от нуля только в круге $x^2 + y^2 \leq r^2$

и принимает постоянное значение, равное $1/\pi r^2$, внутри этого круга. Маргинальные плотности

$$f^X(x) = \frac{2}{\pi r^2} \sqrt{r^2 - x^2}, \quad |x| \leq r; \quad f^Y(y) = \frac{2}{\pi r^2} \sqrt{r^2 - y^2}, \quad |y| \leq r,$$

и $f^X(x) = f^Y(y) = 0$ вне квадрата $|x| \leq r, |y| \leq r$.

Имеем:

$$f^X(x)f^Y(y) = 4\pi^{-2}r^{-4} [(r^2 - x^2)(r^2 - y^2)]^{1/2},$$

что, очевидно, не совпадает с $f(x, y) = 1/\pi r^2$ в области $x^2 + y^2 \leq r^2$. Таким образом, в силу предложения 8.2, случайные величины X и Y зависимы.

Покажем, что, тем не менее, $\rho_{XY} = 0$. Функция $f(x, y)$ центрально симметрична, и поэтому $\mu_X = \mu_Y = 0$. Далее,

$$\lambda_{XY} = \frac{1}{\pi r^2} \int_{-r}^r x dx \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} y dy = 0.$$

Но если $\lambda_{XY} = 0$, то и $\rho = 0$.

Для ковариации пары случайных величин справедливы формулы, аналогичные тем, что были получены для дисперсии в предложениях 6.1 и 8.3.

Предложение 9.1. Для любой пары случайных величин (X, Y) и независимых двумерных векторов $(X_1, Y_1), \dots, (X_n, Y_n)$, обладающих конечными вторыми моментами, справедливы равенства

$$(1) \quad \lambda_{XY} = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y,$$

$$(2) \quad \lambda_{S_X S_Y} = \sum_1^n \lambda_{X_i Y_i},$$

где

$$S_X = \sum_1^n X_i, \quad S_Y = \sum_1^n Y_i.$$

Доказательство. (1) Имеем:

$$\lambda_{XY} = \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)] =$$

$$\mathbf{E}(XY - Y\mathbf{E}X - X\mathbf{E}Y + \mathbf{E}X\mathbf{E}Y) = \mathbf{E}XY - \mathbf{E}X\mathbf{E}Y.$$

(2) Делаем столь же тривиальные выкладки, что и выше, и при этом не забываем, что среднее от произведения независимых случайных величин равно произведению средних:

$$\begin{aligned} \lambda_{S_X S_Y} &= \mathbf{E} \left[\sum_1^n (X_i - \mathbf{E}X_i) \sum_1^n (Y_i - \mathbf{E}Y_i) \right] = \\ &= \mathbf{E} \left[\sum_1^n (X_i - \mathbf{E}X_i)(Y_i - \mathbf{E}Y_i) + \sum_{i \neq j} (X_i - \mathbf{E}X_i)(Y_j - \mathbf{E}Y_j) \right] = \\ &= \sum_1^n \mathbf{E} [(X_i - \mathbf{E}X_i)(Y_i - \mathbf{E}Y_i)] + \sum_{i \neq j} \mathbf{E}(X_i - \mathbf{E}X_i) \cdot \mathbf{E}(Y_j - \mathbf{E}Y_j). \end{aligned}$$

Последнее слагаемое в правой части равно нулю, ибо

$$\mathbf{E}(X_i - \mathbf{E}X_i) = \mathbf{E}X_i - \mathbf{E}X_i = 0,$$

а первое слагаемое есть сумма ковариаций каждого вектора.

Изучим две наиболее распространенные многомерные вероятностные модели.

Лекция 15

Мультиномиальное распределение $\mathcal{M}(m, n, \mathbf{p})$. Рассматривается схема независимых испытаний, в каждом из которых может произойти одно из $m \geq 2$ событий A_1, \dots, A_m с вероятностями

$$p_1, \dots, p_m, \quad \sum_1^m p_j = 1.$$

Типичный пример таких испытаний – наблюдения энтомолога по оценке численности видов насекомых, населяющих некоторый, достаточно изолированный район нашей планеты. Всего проводится n независимых испытаний, и регистрируются значения x_1, \dots, x_m компонент случайного вектора

$$X^{(m)} = (X_1, \dots, X_m), \quad \sum_1^m X_j = n,$$

где x_j – количество испытаний, в которых произошло событие A_j , $j = 1, \dots, m$.

Легко видеть, что мы имеем дело с многомерным аналогом схемы Бернулли, и для вывода распределения $X^{(m)}$ естественно воспользоваться техникой стохастических представлений наблюдаемого случайного элемента в виде суммы индикаторов, то есть поступить по аналогии с примером 8.2. Свяжем с каждым i -м испытанием случайный вектор $Y_i = X_{1i}, \dots, X_{mi}$, каждая компонента X_{ji} которого принимает значение 1, если в i -ом испытании произошло событие A_j , и $X_{ji} = 0$ в противном случае. Таким образом, все компоненты Y_i равны нулю за исключением одной компоненты, равной единице, и номер этой компоненты совпадает с номером исхода (события A_j), которым завершилось i -е испытание, $i = 1, \dots, n$, $j = 1, \dots, m$. Постулируется, что случайные векторы Y_1, \dots, Y_n независимы в совокупности (следствие независимости проведения испытаний).

При таком соглашении каждая компонента X_j наблюдаемого вектора $X^{(m)}$ имеет стохастическое представление

$$X_j = \sum_{i=1}^n X_{ji}, \quad (1)$$

в котором X_{j1}, \dots, X_{jn} независимы и одинаково $B(1, p_j)$ распределены: принимают значение 1 с вероятностью p_j и значение 0 с вероятностью $1 - p_j$, $j = 1, \dots, m$. Из представления (1) следует, что вероятность любого события в n мультиномиальных испытаниях (значений, которые принимают векторы Y_1, \dots, Y_n) определяется только количествами x_1, \dots, x_m испытаний, которые завершились соответствующими исходами A_1, \dots, A_m . Легко видеть, что эта вероятность равна

$$p_1^{x_1} \cdots p_m^{x_m}, \quad \sum_1^m x_j = n.$$

Теперь для того чтобы вывести функцию плотности

$$f(x_1, \dots, x_m) = P(X_1 = x_1, \dots, X_m = x_m),$$

достаточно решить комбинаторную задачу, которую мы умеем решать в случае $m = 2$: сколькими способами можно получить x_1 исходов A_1 , x_2 исходов A_2 , \dots , x_m исходов A_m в

$$n = \sum_1^m x_i$$

испытаниях? Решение задачи дают *мультиномиальные коэффициенты*

$$C_n^{x_1 \dots x_m} = \frac{n!}{x_1! \cdots x_m!}$$

(сравните с биномиальными коэффициентами C_n^x). Итак, функция плотности *мультиномиального распределения* $M(m, n, \mathbf{p})$ по m -кратному произведению считающих мер равна

$$f(x_1, \dots, x_m) = \frac{n!}{x_1! \cdots x_m!} p_1^{x_1} \cdots p_m^{x_m},$$

в области

$$\sum_1^m x_j = n$$

и $f(x_1, \dots, x_m) = 0$ в случае целых x_1, \dots, x_m , не удовлетворяющих последнему равенству, а также в случае дробных x_j , $j = 1, \dots, m$.

Вычислим моментные характеристики мультиномиального распределения, используя стохастическое представление (1). Вектор средних

$$\mathbf{E}X_j = \sum_{i=1}^n \mathbf{E}X_{ji} = n(1 \cdot p_j + 0 \cdot (1 - p_j)) = np_j, \quad j = 1, \dots, m;$$

вектор дисперсий (см. предложение 9.1)

$$\sigma_j^2 = \mathbf{D} \sum_{i=1}^n X_{ji} = \sum_{i=1}^n \mathbf{D}X_{ji} = np_j(1 - p_j), \quad j = 1, \dots, m.$$

Вычислим ковариации X_j с X_l при $j \neq l$ (см. предложение 9.1):

$$\lambda_{jl} = \sum_{i=1}^n \text{cov}(X_{ji}, X_{li}) = \sum_{i=1}^n [\mathbf{E}(X_{ji}X_{li}) - \mathbf{E}X_{ji}\mathbf{E}X_{li}].$$

Поскольку при i -м испытании только одна из компонент X_{1i}, \dots, X_{mi} равна единице, а остальные равны нулю, то $X_{ji}X_{li} \equiv 0$, и

$$\lambda_{jl} = -np_j p_l.$$

Коэффициенты корреляции

$$\rho_{jl} = -\frac{np_j p_l}{n(p_j(1 - p_j)p_l(1 - p_l))^{1/2}} = -\sqrt{\frac{p_j p_l}{(1 - p_j)(1 - p_l)}}, \quad j \neq l.$$

Отрицательные значения коэффициентов корреляции есть следствие связей между компонентами наблюдаемого вектора:

$$\sum_1^m X_j = n.$$

Многомерное нормальное распределение $\mathcal{N}_m(\mu, \Lambda)$. Мы трактовали мультиномиальную схему испытаний как многомерный аналог схемы независимых испытаний Бернулли. В таком случае естественно рассмотреть многомерный аналог предельной теоремы Муавра–Лапласа. Применяя формулу Стирлинга, нетрудно убедиться, что имеет место

Теорема 9.1. (Интегральная предельная теорема для мультиномиального распределения). *Для любых постоянных*

$$(a_1, b_1), \dots, (a_m, b_m)$$

и $(m + 1)$ -мерного случайного вектора

$$X^{(m+1)} = (X_1, \dots, X_m, X_{m+1}) \sim \mathcal{M}(m + 1, n, \mathbf{p})$$

справедливо асимптотическое представление

$$\lim_{n \rightarrow \infty} P \left(a_1 \leq \frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} < b_1, \dots, a_m \leq \frac{X_m - np_m}{\sqrt{np_m(1-p_m)}} < b_m \right) =$$

$$\frac{1}{(2\pi)^{m/2} \sqrt{|\mathbf{P}|}} \int_{a_1}^{b_1} \dots \int_{a_m}^{b_m} \exp \left\{ -\frac{1}{2} \mathbf{x}' \mathbf{P}^{-1} \mathbf{x} \right\} dx_1 \dots dx_m, \quad (2)$$

где $\mathbf{x}' = (x_1, \dots, x_m)$, \mathbf{x} – аналогичный вектор столбец; $\mathbf{P} = \|\rho_{ij}\|$ – корреляционная матрица, в которой $\rho_{ii} = 1$ и

$$\rho_{ij} = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}},$$

если $i \neq j$, $i, j = 1, \dots, m$; \mathbf{P}^{-1} – матрица, обратная к \mathbf{P} , наконец, $|\mathbf{P}|$ – определитель матрицы \mathbf{P} .

Интеграл (2) определяет непрерывное распределение вероятностей на прямоугольниках (следовательно, и на борелевском поле) пространства \mathbb{R}^m , причем функция плотности этого распределения

$$f_m(\mathbf{x}' | \mathbf{P}) = \frac{1}{(2\pi)^{m/2} \sqrt{|\mathbf{P}|}} \exp \left\{ -\frac{1}{2} \mathbf{x}' \mathbf{P}^{-1} \mathbf{x} \right\}, \quad \mathbf{x}' \in \mathbb{R}^m.$$

Если теперь вместо корреляционной матрицы \mathbf{P} мультиномиального распределения рассмотреть произвольную, положительно определенную корреляционную матрицу

$$\mathbf{P} = \|\rho_{ij}\|, \quad \rho_{ii} = 1, \quad i = 1, \dots, m,$$

то $f_m(\mathbf{x}' | \mathbf{P})$ будет функцией плотности случайного вектора

$$X^{(m)} = (X_1, \dots, X_m),$$

имеющего *стандартное m -мерное нормальное распределение* $\mathcal{N}_m(0, \mathbf{P})$. Далее, если ввести вектор средних $\mu = (\mu_1, \dots, \mu_m)$ и вектор дисперсий $\sigma^2 = (\sigma_1^2, \dots, \sigma_m^2)$, то случайный вектор

$$\sigma_1 X_1 + \mu_1, \dots, \sigma_m X_m + \mu_m$$

будет иметь *многомерное нормальное распределение* $\mathcal{N}_m(\mu, \Lambda)$ с функцией плотности

$$\varphi_m(\mathbf{x}' | \mu, \Lambda) = \frac{1}{(2\pi)^{m/2} \sqrt{|\Lambda|}} \exp \left\{ -\frac{1}{2} (x - \mu)' \Lambda^{-1} (x - \mu) \right\} =$$

$$\frac{1}{(2\pi)^{m/2} \sigma_1 \cdots \sigma_m \sqrt{|\mathbf{P}|}} \exp \left\{ -\frac{1}{2} \sum_{i,j} \frac{\mathbf{P}_{ij}}{|\mathbf{P}|} \cdot \frac{(x_i - \mu_i)(x_j - \mu_j)}{\sigma_i \sigma_j} \right\}, \quad \mathbf{x}' \in \mathbb{R}^m,$$

где определитель ковариационной матрицы

$$|\Lambda| = \sigma_1^2 \cdots \sigma_m^2 \cdot |\mathbf{P}|,$$

а $\mathbf{P}_{ij}/|\mathbf{P}|$ – элементы матрицы \mathbf{P}^{-1} , обратной к \mathbf{P} .

Нетрудно видеть, что если коэффициенты корреляции $\rho_{ij} = 0$, когда $i \neq j$, то есть \mathbf{P} есть единичная матрица, а в матрице ковариаций Λ отличны от нуля только диагональные элементы $\sigma_1^2, \dots, \sigma_m^2$, то нормальная функция плотности распадается в произведение маргинальных нормальных $\mathcal{N}_1(\mu_i, \sigma_i^2)$, $i = 1, \dots, m$ функций плотности. Поэтому справедливо

Предложение 9.2. *В случае нормального распределения случайного вектора некоррелированность его компонент влечет их независимость.*

Следует обратить особое внимание на требование *положительной определенности* корреляционной матрицы \mathbf{P} или, что то же, ковариационной матрицы Λ . Если эти матрицы положительно полуопределены, то есть имеют ранг $r < m$, то мы получим *несобственное m -мерное нормальное распределение*, вся вероятностная масса которого будет сосредоточена на гиперплоскости \mathbb{R}^r , а между компонентами случайного вектора $X^{(m)}$ будет существовать линейная зависимость.

Указанные свойства многомерного нормального распределения наиболее наглядно прослеживаются в случае $m = 2$ – нормального распределения на плоскости \mathbf{R}^2 . Функция плотности распределения случайного вектора $(X, Y) \sim \mathcal{N}_2(\mu, \Lambda)$ при $\rho = \rho_{XY} \neq \pm 1$ равна

$$\varphi_2(x, y | \mu, \Lambda) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) \right\}.$$

Как отмечалось выше, $\mu_1 = \mathbf{E}X$, $\mu_2 = \mathbf{E}Y$ есть вектор средних значений; $\sigma_1^2 = \mathbf{D}X$, $\sigma_2^2 = \mathbf{D}Y$ – дисперсии соответствующих случайных величин; $\rho = \text{cov}(X, Y)/\sigma_1\sigma_2$, – коэффициент корреляции между X и Y . В том, что это действительно так, можно убедиться и непосредственным вычислением интегралов, определяющих соответствующие моментные характеристики. Маргинальные функции плотности $f^X(x)$ и $f^Y(y)$ находятся также непосредственным интегрированием совместной функции плотности $\varphi_2(x, y | \mu, \Lambda)$ по соответствующим переменным y и x :

$$f^X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\}, \quad f^Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp \left\{ -\frac{(y-\mu_2)^2}{2\sigma_2^2} \right\},$$

то есть

$$X \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad Y \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

Особо отметим, что *существуют многомерные распределения, отличные от нормального, но имеющие нормальные маргинальные распределения.*

Подобные эллипсы

$$\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) = c^2$$

играют роль *кривых равных вероятностей*: нетрудно вычислить, что вероятность попадания (X, Y) в область, ограниченную этим эллипсом, равна $1 - \exp\{-c^2\}$.

Форма эллипса равных вероятностей дает хорошее представление о виде поверхности $z = \varphi_2(x, y | \mu, \Lambda)$ нормальной плотности. При $\rho = 0$, $\sigma_1 = \sigma_2$ эллипсы превращаются в окружности. Когда ρ приближается к $+1$ или -1 , эллипсы становятся более тонкими и вытянутыми, что является показателем стремления вероятностной массы сосредотачиваться около общей большей оси этих эллипсов.

Особый интерес представляет эллипс с $c = 2$, который называется *эллипсом рассеяния*. Он обладает достаточно высокой вероятностью $1 - e^{-4} \approx 0.98$ попадания случайной точки (X, Y) внутрь эллипса и еще одним замечательным свойством: равномерное распределение по области, ограниченной эллипсом рассеяния, имеет те же моменты первого (μ_1, μ_2) и второго $(\sigma_1^2, \sigma_2^2, \rho\sigma_1\sigma_2)$ порядков, что и нормальное распределение.

В заключение отметим, что двумерное нормальное распределение играет важную роль в *теории стрельб*: распределение координат точек попадания при стрельбе из закрепленного ствола хорошо согласуется с нормальным законом.

§10. Условное распределение вероятностей.

Условное математическое ожидание

Лекция 16

Одной из типичных задач теории вероятностей является прогноз возможного значения случайной величины Y по результату наблюдения другой случайной величины X . Примеры таких задач – прогноз метеорологических показателей (температуры воздуха, атмосферного давления, количества осадков) по данным их замеров в прошлом; прогноз стоимости ценных бумаг или обменного курса валюты на ближайшее будущее, причем прогноз осуществляется не столько по значениям этих финансовых показателей на сегодняшний день, сколько по данным информационного центра о состоянии промышленности страны, различных показателях рынка ценных бумаг, их движения, спроса и пр. С аналогичными проблемами мы сталкиваемся в медицине, технике, управлении производством. Легко понять, что во всех этих примерах прогноз носит вероятностный характер, и должен отвечать на вопросы типа: если наблюдается значение x случайной величины X , то какова вероятность того, что Y будет лежать в определенных пределах, каково наиболее вероятное значение Y , или чему равно среднее значение Y , если X приняло значение x ?

На все эти вопросы легко ответить в случае дискретных распределений случайного вектора (X, Y) . Пусть $f^{X,Y}(x, y) = P(X = x, Y = y)$ – совместная функция плотности X и Y по считающей мере на борелевской плоскости \mathbb{R}^2 . Тогда решение задачи прогноза дает *условная функция плотности* – условная вероятность

$$f^{Y|X}(y|x) = P\{Y = y | X = x\} = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f^{X,Y}(x, y)}{f^X(x)},$$

где $f^X(x)$ – маргинальная функция плотности X . Действительно, вероятность того, что прогнозируемое значение Y будет лежать в пределах $(a; b)$ равно условной вероятности

$$P\{Y \in (a; b) | X = x\} = \sum_{y \in (a; b)} f^{Y|X}(y|x),$$

наиболее вероятное значение Y – это мода условного распределения

$$\text{mod}(Y | X) = \arg \sup_y f^{Y|X}(y|x),$$

ожидаемое (среднее) значение Y – условное математическое ожидание

$$\mathbf{E}\{Y | X = x\} = \sum_y y f^{Y|X}(y | x).$$

Но как осуществить аналогичные расчеты в случае непрерывного распределения X , когда $P(X = x) = 0$ и на ноль, как известно, делить нельзя? Конечно, и в этом случае можно предложить некоторое “суррогатное”, построенное на аналогиях с дискретным случаем, определение условного распределения, но тем не менее отсутствие строгого определения условной вероятности, пригодного для распределений любого типа, долго тормозило развитие теории вероятностей и ее становление как самостоятельной математической дисциплины. Особенно страдала от этого теория случайных процессов, до тех пор пока в 20-х годах прошлого столетия Андрей Николаевич Колмогоров ввел строгое определение условного математического ожидания и условного распределения вероятностей. Увы, это слишком сложная теория, которую мы не сможем постичь с нашими скудными познаниями в области теории меры. Замечу только, что в этом определении решающую роль играет теорема Радона–Никодима, о которой мы упоминали в §6, когда вводили функцию плотности распределения случайной величины. Кроме того, условное математическое ожидание и условное распределение трактуются как случайные величины, изменение значений которых на множествах нулевой вероятности не влияет на их распределение, и это обстоятельство позволяет исключить из рассмотрения значения x случайной величины X , в которых $f^X(x) = 0$. Сейчас мы расскажем, как это делается без привлечения теоремы Радона–Никодима, когда совместное распределение X и Y принадлежит непрерывному типу (существует плотность $f^{X,Y}(x, y)$ по мере Лебега на плоскости).

Предположим дополнительно, что совместная функция плотности $f^{X,Y}(x, y)$ непрерывна по обоим переменным, и изучим асимптотическое поведение условной вероятности $P\{Y < y | x \leq X < x + \Delta x\}$ при $\Delta x \rightarrow 0$, в тех точках переменных x , в некоторых окрестностях которых маргинальная функция плотности $f^X(x) > 0$. Понятно, что таким образом мы пытаемся ввести условную функцию плотности Y , при условии $X = x$. По формуле условной вероятности находим

$$P\{Y < y | x \leq X < x + \Delta x\} = \frac{P(Y < y, x \leq X < x + \Delta x)}{P(x \leq X < x + \Delta x)} =$$

$$\frac{\int_{-\infty}^y dt \int_x^{x+\Delta x} f^{X,Y}(s, t) ds}{\int_x^{x+\Delta x} f^X(s) ds}.$$

Применяя теорему о среднем к интегралам в правой части этого равенства, получаем

$$P\{Y < y | x \leq X < x + \Delta x\} = \frac{\int_{-\infty}^y f^{X,Y}(x + \lambda_1 \Delta x, t) dt \Delta x}{f^X(x + \lambda_2 \Delta x) \Delta x},$$

где $0 < \lambda_i < 1$, $i = 0, 1$.

Теперь, используя непрерывность функций $f^{X,Y}$ и f^X , мы можем определить условную функцию распределения Y относительно события $X = x$, если устремим в обеих частях последнего неравенства $\Delta x \rightarrow 0$:

$$F^{Y|X}(y | x) = \lim_{\Delta x \rightarrow 0} P\{Y < y | x \leq X < x + \Delta x\} = \frac{\int_{-\infty}^y f^{X,Y}(x, t) dt}{f^X(x)}.$$

Условная функция плотности получается дифференцированием условной функции распределения:

$$f^{Y|X}(y | x) = \frac{d}{dy} F^{Y|X}(y | x) = \frac{f^{X,Y}(x, y)}{f^X(x)}.$$

Удивительно, но мы пришли к той же формуле условной плотности, что и в дискретном случае!

Поскольку значения x , в которых $f^X(x) = 0$, можно вообще исключить из области возможных значений случайной величины X (эти значения в совокупности дают множество нулевой вероятности), то, подставляя в определении условной плотности вместо x случайную величину X , приходим к следующему определению условного распределения и условного математического ожидания для дискретных и непрерывных распределений случайного вектора (X, Y) .

Определение 10.1. Если случайный вектор (X, Y) имеет дискретную или непрерывную совместную функцию плотности $f^{X,Y}(x, y)$, то *условное распределение случайной величины Y относительно случайной величины*

X определяется функцией плотности

$$f^{Y|X}(y | X) = \frac{f^{X,Y}(y, X)}{f^X(X)}, \quad y \in \mathbb{R},$$

а условное математическое ожидание Y относительно случайной величины X вычисляется по формуле

$$\mathbf{E}\{Y | X\} = \int_{\mathbb{R}} y f^{Y|X}(y | X) d\mu(y).$$

Таким образом, *условное математическое ожидание и условное распределение являются случайными величинами* и определяющие их формулы справедливы для почти всех значений случайной величины X .

Естественно, мы можем ввести также определение условного математического ожидания измеримой функции $g(Y)$ относительно случайной величины X , полагая

$$\mathbf{E}\{g(Y) | X\} = \int_{\mathbb{R}} g(y) f^{Y|X}(y | X) d\mu(y).$$

Если положить $g(y) = \mathbf{I}_B(y)$ – индикаторной функции борелевского множества B , то условное математическое ожидание от $g(Y)$ относительно X совпадает с условной вероятностью события B , и, таким образом, условное распределение Y относительно X есть частный случай условного математического ожидания:

$$P\{Y \in B | X\} = \mathbf{E}\{\mathbf{I}_B(Y) | X\}.$$

Легко видеть, что условное математическое ожидание обладает всеми свойствами обычного среднего, но ему присущи и некоторые специфические черты.

Предложение 10.1. *Для условного математического ожидания $\mathbf{E}\{Y | X\}$ справедливы равенства*

- (1) $\mathbf{E}^X \mathbf{E}\{Y | X\} = \mathbf{E}Y$,
- (2) *если X и Y независимы, то $\mathbf{E}\{Y | X\} = \mathbf{E}Y$.*

Доказательство. Оба равенства немедленно следуют из определения условного математического ожидания.

(1) Имеем

$$\begin{aligned}\mathbf{E}^X \mathbf{E}\{Y | X\} &= \int_{\mathbf{R}} \mathbf{E}\{Y | x\} f^X(x) d\mu(x) = \\ &= \int_{\mathbf{R}} d\mu(x) \int_{\mathbf{R}} y \frac{f^{X,Y}(x, y)}{f^X(x)} f^X(x) d\mu(y) = \int_{\mathbf{R}} y f^Y(y) d\mu(y) = \mathbf{E}Y.\end{aligned}$$

(2) Если X и Y независимы, то $f^{X,Y}(x, y) = f^X(x)f^Y(y)$, откуда

$$\mathbf{E}\{Y | X\} = \int_{\mathbf{R}} y \frac{f^X(x)f^Y(y)}{f^X(x)} d\mu(y) = \int_{\mathbf{R}} y f^Y(y) d\mu(y) = \mathbf{E}Y.$$

Теперь мы располагаем некоторой техникой для решения простейших задач прогноза.

Наилучший в среднем квадратическом прогнозе ожидаемого значения Y по результату наблюдения X .

Проблема оптимального предсказания ожидаемого значения Y по результату x наблюдения случайной величины X состоит в отыскании функции $g(X)$, которая минимизирует некоторую меру уклонения $g(X)$ от Y . Мы будем решать задачу минимизации *среднего квадратического уклонения* $\mathbf{E}(Y - g(X))^2$, в котором математическое ожидание вычисляется по совместному распределению X и Y . Решение задачи дает

Предложение 10.2. *Минимум функционала $\mathbf{E}(Y - g(X))^2$ достигается на функции*

$$g^*(X) = \mathbf{E}\{Y | X\}.$$

Доказательство. Используя свойство (1) условного математического ожидания, установленное в предложении 10.1, представим среднее квадратическое уклонение в виде

$$\mathbf{E}(Y - g(X))^2 = \mathbf{E}^X \mathbf{E}\{(Y - g(X))^2 | X\}$$

и будем минимизировать момент второго порядка $\mathbf{E}\{(Y - g(X))^2 | X\}$ случайной величины Y относительно “точки” $g(X)$, когда распределение Y определяется условной функцией плотности $f^{Y|X}(y | X)$. В §6 (предложение 6.1, утверждение 4⁰) было показано, что минимум достигается на среднем

значении Y и равен $\mathbf{D}Y$. Поскольку в данном случае момент второго порядка вычислялся по условному распределению Y относительно X , то минимум достигается на функции (случайной величине) $g^*(X) = \mathbf{E}\{Y | X\}$.

Итак, наилучший в среднем квадратическом прогноз ожидаемого значения Y по результату x наблюдения случайной величины X доставляет реализация в точке $X = x$ условного математического ожидания $\mathbf{E}\{Y | X\}$. Допуская некоторую вольность, запишем эту реализацию в виде функции $y = g^*(x) = \mathbf{E}\{Y | X = x\}$. Функция $y = g^*(x)$ называется *кривой средней квадратической регрессии* Y на X , а минимальное среднее квадратическое уклонение $\sigma_{Y|X}^2 = \mathbf{E}(Y - g^*(X))^2$ — *остаточной дисперсией*.

Найдем кривую $g^*(x)$ и вычислим остаточную дисперсию в случае совместного нормального распределения X и Y .

Предложение 10.3. Пусть $(X, Y) \sim \mathcal{N}_2(\mu, \Lambda)$. Тогда кривая средней квадратической регрессии Y на X есть прямая

$$y = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1). \quad (1)$$

Остаточная дисперсия $\sigma_{Y|X}^2 = \sigma_2^2(1 - \rho^2)$.

Доказательство. В случае двумерного нормального распределения

$$f^{X,Y}(x, y) = f^{Y|X}(y | x) \cdot f^X(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \cdot \exp \left\{ -\frac{1}{2(1 - \rho^2)} \left(\frac{(x - \mu_1)^2}{\sigma_1^2} - \frac{2\rho(x - \mu_1)(y - \mu_2)}{\sigma_1\sigma_2} + \frac{(y - \mu_2)^2}{\sigma_2^2} \right) \right\},$$

и остается только убедиться, что

$$f^X(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left\{ -\frac{(x - \mu_1)^2}{2\sigma_1^2} \right\},$$

а

$$f^{Y|X}(y | x) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1 - \rho^2}} \exp \left\{ -\frac{(y - g(x))^2}{2\sigma_2^2(1 - \rho^2)} \right\}, \quad (2)$$

где

$$g(x) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1).$$

Маргинальную функцию плотности $f^X(x)$ компоненты X мы вычисляли в конце предыдущего параграфа. Перемножая $f^X(x)$ и

$f^{Y|X}(y|x)$ и производя несложные алгебраические преобразования, убеждаемся, что это произведение равно плотности $f^{X,Y}(x,y)$ двумерного нормального распределения.

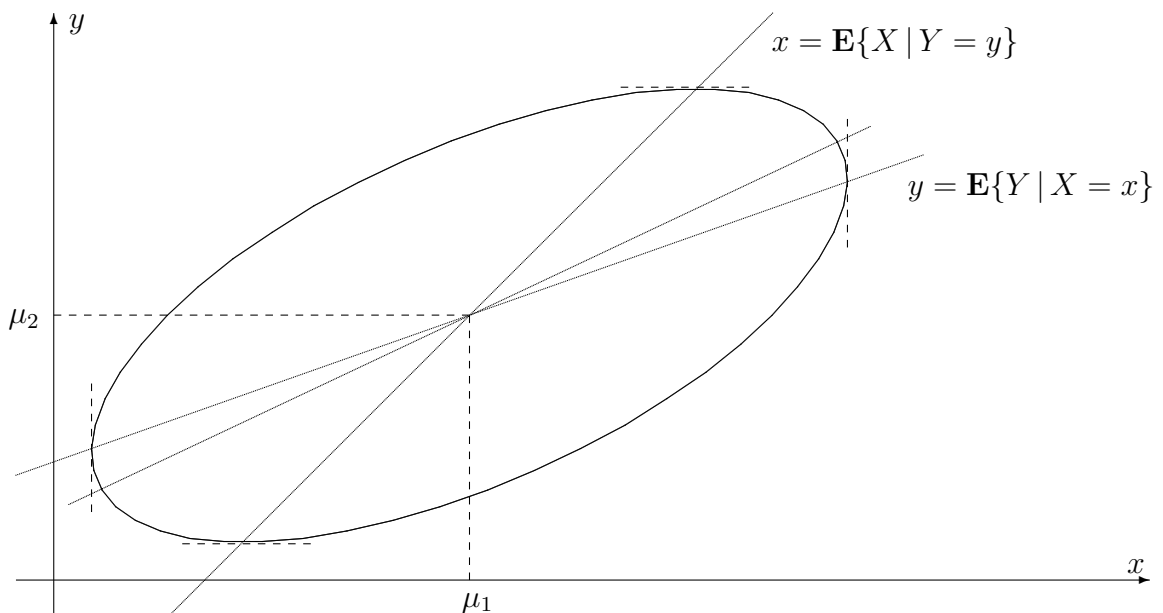
Из формулы (2) находим, что условная плотность имеет среднее значение $\mathbf{E}\{Y|X=x\} = g(x)$ и дисперсию $\sigma_2^2(1-\rho^2)$, которые совпадают с формулами, приведенными в формулировке предложения.

Легко понять, что если проблема состоит в предсказании значения x случайной величины X по наблюдению y случайной величины Y , то наилучший в среднем квадратическом прогноз x имеет вид

$$x = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2),$$

что не совпадает с функцией, обратной к (1). Это вполне естественно, поскольку мы минимизируем в каждом из этих случаев отклонения от прямой регрессии в различных (перпендикулярных друг к другу) направлениях: одно – вдоль оси OY , второе – вдоль оси OX .

Ниже приводится рисунок, иллюстрирующий расположение прямых регрессии Y на X и X на Y относительно главной оси эллипса рассеяния. Эта ось называется *прямой ортогональной регрессии* и дает наилучший в среднем квадратическом прогноз, когда ошибки измеряются не вдоль координатных осей, а по кратчайшему расстоянию к кривой прогноза $g(x)$.



§11. Сходимость случайных величин и функций распределений

Лекция 17

Для того чтобы продвинуться дальше в построении новых вероятностных моделей и возможно в большей степени дать математическое основание для применения методов математической статистики к идентификации вероятностных моделей, мы должны изучить проблемы сходимости последовательностей случайных величин и их функций распределения.

Пусть на вероятностном пространстве (Ω, \mathcal{A}, P) заданы две случайные величины $X = X(\omega)$ и $Y = Y(\omega)$, имеющие одно и то же распределение вероятностей. В таком случае, с точки зрения приложений теории вероятностей такие случайные величины следует считать эквивалентными (неразличимыми).

Определение 11.1. Если $P(X(\omega) = Y(\omega)) = 1$, то случайные величины X и Y называются *равными почти наверное* или *эквивалентными* и пишется $X \underset{\text{п.н.}}{=} Y$.

При исследовании предела последовательности случайных величин $\{X_n, n \geq 1\}$, заданных на одном и том же вероятностном пространстве (Ω, \mathcal{A}, P) , мы имеем дело, по существу, с проблемой сходимости последовательности функций $\{X_n(\omega), n \geq 1\}$, но при этом мы можем не обращать внимания на множество точек ω нулевой вероятности, в которых соответствующие числовые последовательности не имеют предела. Поэтому поточечная (в каждой точке $\omega \in \Omega$) сходимость функций претерпевает существенное изменение и превращается в следующее

Определение 11.2. Если для последовательности случайных величин $\{X_n(\omega), n \geq 1\}$, заданных на вероятностном пространстве (Ω, \mathcal{A}, P) , существует такая случайная величина $X = X(\omega)$, $\omega \in \Omega$, что

$$P(\{\lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}) = 1,$$

то X называется *пределом последовательности* $\{X_n, n \geq 1\}$ *почти наверное* и пишется $X_n \underset{\text{п.н.}}{\longrightarrow} X$.

В общей теории меры сходимость почти наверное называется сходимостью *почти всюду*, и это наиболее *сильная* из известных нам форма сходимости функций – случайных величин. Для нее, естественно, справедлив

критерий сходимости Коши: если $|X_n - X_m| \xrightarrow[\text{п.н.}]{} 0$ при $n, m \rightarrow \infty$, то последовательность $\{X_n, n \geq 1\}$ почти наверное сходится к некоторому пределу X . Существуют довольно сложные в приложениях достаточные признаки сходимости почти наверное, однако мы практически не будем в дальнейшем касаться этой формы сходимости, поскольку конкретные результаты (например, усиленный закон больших чисел) требуют для своего доказательства огромных временных затрат, – мы не можем позволить себе такой роскоши в рамках того скудного промежутка времени, который отведен нам учебным планом для изучения теории вероятностей и математической статистики. Мы будем, в основном, иметь дело с более “слабой” формой сходимости, которая в общей теории меры называется *сходимостью по мере*, а в теории вероятностей, как вы, наверное, уже догадываетесь, *сходимостью по вероятности*.

Определение 11.3. Последовательность $\{X_n, n \geq 1\}$ называется сходящейся к пределу X по вероятности, если для любого $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n(\omega) - X(\omega)| > \varepsilon) = 0.$$

Сходимость по вероятности обозначается $X_n \xrightarrow{P} X$.

Напомним известный вам из общей теории меры факт: сходимость почти наверное влечет сходимость по вероятности, но обратное, вообще говоря, не верно, и существуют примеры последовательностей, сходящихся по вероятности, но не имеющих предела почти наверное. Однако из всякой сходящейся по вероятности последовательности случайных величин можно извлечь подпоследовательность, сходящуюся к тому же пределу почти наверное.

Мы уже имели дело со сходимостью по вероятности, когда доказывали закон больших чисел Бернулли. Следующий результат обобщает этот закон на суммы независимых случайных величин с достаточно произвольным общим распределением.

Теорема 11.1 (закон больших чисел Чебышева). Пусть $\{X_n, n \geq 1\}$ – последовательность независимых одинаково распределенных случайных величин с конечным вторым моментом. Тогда

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{P} \mathbf{E}X_1.$$

Доказательство проводится столь же просто, как и в случае закона больших чисел Бернулли, когда $X_k \sim B(1, p)$, и также основано на использовании неравенства Чебышева

$$P(g(X) > c) \leq \mathbf{E}g(X)/c.$$

Положим

$$g(X) = (\bar{X}_n - \mathbf{E}X_1)^2$$

и $c = \varepsilon^2$, где ε – произвольное положительное число. Тогда

$$P(|\bar{X}_n - \mathbf{E}X_1| > \varepsilon) = P((\bar{X}_n - \mathbf{E}X_1)^2 > \varepsilon^2) \leq \frac{\mathbf{E}(\bar{X}_n - \mathbf{E}X_1)^2}{\varepsilon^2}.$$

Поскольку \bar{X}_n есть нормированная на n сумма независимых случайных величин с общим конечным средним $\mu = \mathbf{E}X_1$ и общей дисперсией $\sigma^2 = \mathbf{D}X_1$, то

$$\mathbf{E}(\bar{X}_n - \mathbf{E}X_1)^2 = \mathbf{E} \left[\frac{1}{n} \sum_1^n (X_k - \mu) \right]^2 = \frac{1}{n^2} \mathbf{D} \sum_1^n X_k = \frac{\sigma^2}{n}.$$

Следовательно,

$$P(|\bar{X}_n - \mathbf{E}X_1| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \rightarrow 0,$$

каково бы ни было $\varepsilon > 0$.

Мы покажем в дальнейшем, что в законе больших чисел можно убрать требование о существовании второго момента, – достаточно только существование среднего, но для этого мы должны будем разработать более совершенный математический аппарат анализа сходимости по вероятности последовательностей случайных величин. Отметим пока несомненную практическую ценность установленного закона: при независимых равно-точных наблюдениях некоторой постоянной μ , характеризующей состояние исследуемого объекта (например, общего содержания серы в партии дизельного топлива), арифметическое среднее \bar{X}_n результатов параллельных наблюдений, отягощенных случайной ошибкой с нулевым средним, является асимптотически точной оценкой μ в том смысле, что $\bar{X}_n \xrightarrow{P} \mu$.

Докажем еще два утверждения, касающиеся сходимости по вероятности, которые обычно называются в монографиях по теории вероятностей *теоремами типа Слуцкого*.

Предложение 11.1. Если последовательность случайных величин $\{X_n, n \geq 1\}$ сходится по вероятности к случайной величине X , а последовательность случайных величин $\{\xi_n, n \geq 1\}$ сходится по вероятности к нулю, тогда

$$(1) X_n + \xi_n \xrightarrow{P} X, \quad (2) X_n \xi_n \xrightarrow{P} 0.$$

Доказательство. (1) Требуется показать, что

$$P(|X_n + \xi_n - X| > \varepsilon) \rightarrow 0,$$

каково бы ни было $\varepsilon > 0$.

Используя неравенство $|a + b| \leq |a| + |b|$, получаем

$$\begin{aligned} 0 \leq P(|X_n + \xi_n - X| > \varepsilon) &\leq P(|X_n - X| + |\xi_n| > \varepsilon) \leq \\ P\left(\left\{|X_n - X| > \frac{\varepsilon}{2}\right\} \cup \left\{|\xi_n| > \frac{\varepsilon}{2}\right\}\right) &\leq P\left(|X_n - X| > \frac{\varepsilon}{2}\right) + \\ P\left(|\xi_n| > \frac{\varepsilon}{2}\right) &\rightarrow 0, \end{aligned}$$

так как по условию предложения $|X_n - X| \xrightarrow{P} 0$ и $|\xi_n| \xrightarrow{P} 0$.

(2) По аналогии с (1) сначала делаем оценку вероятности

$$\begin{aligned} P(|X_n \xi_n| > \varepsilon) &= P(|(X_n - X) + X| |\xi_n| > \varepsilon) \leq \\ P(|X_n - X| |\xi_n| + |X| |\xi_n| > \varepsilon) &\leq \\ P\left(|X_n - X| |\xi_n| > \frac{\varepsilon}{2}\right) + P\left(|X| |\xi_n| > \frac{\varepsilon}{2}\right). \end{aligned} \quad (1)$$

Покажем, что первое слагаемое в правой части (1) стремится к нулю, когда $n \rightarrow \infty$:

$$\begin{aligned} P(|X_n - X| |\xi_n| > \varepsilon/2) &\leq P\left(\left\{|X_n - X| > \sqrt{\varepsilon/2}\right\} \cup \left\{|\xi_n| > \sqrt{\varepsilon/2}\right\}\right) \leq \\ P\left(|X_n - X| > \sqrt{\varepsilon/2}\right) + P\left(|\xi_n| > \sqrt{\varepsilon/2}\right) &\rightarrow 0. \end{aligned}$$

Теперь покажем, что второе слагаемое в (1) можно сделать меньше любого $\delta > 0$, для всех $n \geq N = N(\delta)$, иными словами, покажем, что второе слагаемое также стремится к нулю при $n \rightarrow \infty$. С этой целью введем два противоположных события $\{|X| > A\}$ и $\{|X| \leq A\}$, в которых число A будет выбрано в дальнейшем по заданному δ , и представим второе слагаемое в виде

$$P\left(|X| |\xi_n| > \frac{\varepsilon}{2}\right) = P\left(\left\{|X| |\xi_n| > \frac{\varepsilon}{2}\right\} \cap \{|X| > A\}\right) + P\left(\left\{|X| |\xi_n| > \frac{\varepsilon}{2}\right\} \cap \{|X| \leq A\}\right). \quad (2)$$

Первое слагаемое в правой части этого представления не превосходит $P(|X| > A)$, и мы можем выбрать $A = A(\delta)$ по заданному δ настолько большим, что $P(|X| > A) < \delta/2$. Для выбранного таким образом A , которое не зависит от n , оценим второе слагаемое в представлении (2):

$$P\left(\left\{|X| \cdot |\xi_n| > \frac{\varepsilon}{2}\right\} \cap \{|X| \leq A\}\right) \leq P\left(A \cdot |\xi_n| > \frac{\varepsilon}{2}\right) = P\left(|\xi_n| > \frac{\varepsilon}{2A}\right).$$

Поскольку $\xi_n \xrightarrow{P} 0$, то существует такое $N = N(\delta)$, что для всех $n > N$ вероятность

$$P\left(|\xi_n| > \frac{\varepsilon}{2A}\right) \leq \frac{\delta}{2}.$$

Следует обратить особое внимание на то, как в изучаемых нами видах сходимости почти наверное и по вероятности играет существенную роль задание последовательностей случайных величин на едином вероятностном пространстве (Ω, \mathcal{A}, P) . По существу, близость членов X_n с большими значениями n к их пределу X зависит не столько от совпадения распределений X_n и X , сколько от близости функций $X_n(\omega)$ и $X(\omega)$. Сейчас мы введем еще один вид сходимости случайных величин, более слабый, чем сходимость по вероятности, и для этого вида близость распределений случайных величин становится доминирующей, по сравнению с близостью их как функций на едином пространстве Ω ; в этом виде сходимости случайные величины, как компоненты некоторой последовательности, могут быть определены даже на разных пространствах элементарных исходов.

Определение 11.4. Последовательность случайных величин $\{X_n, n \geq 1\}$ сходится к случайной величине X слабо или по распределению, если соответствующая последовательность их функций распределения $\{F_n(x), n \geq 1\}$ сходится к функции распределения $F(x)$ случайной величины X в каждой точке непрерывности функции $F(x)$.

Слабая сходимость обозначается двойной стрелкой: $X_n \Rightarrow X$, и поскольку речь идет не столько о сходимости случайных величин, сколько об их

распределениях, то аналогичное обозначение сохраняется и для последовательности функций распределения: $F_n \Rightarrow F$, и при этом говорят, что последовательность функций распределения $\{F_n, n \geq 1\}$ *слабо сходится* к функции $F(x)$.

Возникает естественный вопрос: почему в определении слабой сходимости ограничиваются только точками непрерывности предельной функции $F(x)$? Дело в том, что без этого условия последовательность $\{X_n, n \geq 1\}$ и, например, “сдвинутая” на бесконечно малую величину последовательность $\{X_n + 1/n, n \geq 1\}$ могут иметь различные “слабые” пределы, а это – нехорошо.

Приведем пример последовательности случайных величин, которая сходится по распределению, но не имеет предела по вероятности.

Пример 11.1. Пусть $\Omega = [0; 1]$, \mathcal{A} – борелевская σ -алгебра подмножеств этого отрезка и P – равномерное распределение $U(0, 1)$. На вероятностном пространстве (Ω, \mathcal{A}, P) введем последовательность случайных величин $\{X_n, n \geq 1\}$, полагая $X_n(\omega) = (-1)^n$, если $0 \leq \omega \leq 1/2$, и $X_n(\omega) = (-1)^{n-1}$, если $1/2 < \omega \leq 1$. Все случайные величины последовательности имеют одну и ту же функцию распределения $F(x)$, определяемую вероятностями $P(X_n = -1) = P(X_n = +1) = 1/2$, так что $\{X_n, n \geq 1\}$ сходится по распределению. Однако эта последовательность состоит из чередующейся пары различных случайных величин: все члены последовательности с четными номерами X_{2k} принимают значения $+1$ при $0 \leq \omega \leq 1/2$, в то время как соседний член X_{2k+1} с нечетным номером на этом отрезке равен -1 . Следовательно, не существует случайной величины $X = X(\omega)$, $\omega \in [0; 1]$ одинаково близкой при больших n ко всем X_n в смысле малости вероятности $P(|X_n - X| > \varepsilon)$.

Естественно, представляют несомненный интерес достаточные условия, при выполнении которых слабая сходимость влечет сходимость по вероятности. Одно из таких условий дает

Предложение 11.2. Если последовательность $X_n \Rightarrow C (\equiv \text{const})$, то $X_n \xrightarrow{P} C$.

Доказательство. Будем трактовать постоянную C как вырожденную случайную величину с функцией распределения $F(x) = 0$ при $x \leq C$ и $F(x) = 1$ при $x > C$. По условию предложения соответствующая последовательность функций распределения $F_n(x) \rightarrow F(x)$ при любом $x \neq C$, ибо C – единственная точка разрыва предельной функции распределения

$F(x)$. Требуется показать, что $P(|X_n - C| > \varepsilon) \rightarrow 0$, когда $n \rightarrow \infty$, каково бы ни было $\varepsilon > 0$.

Выразим вероятность $P(|X_n - C| > \varepsilon)$ через функцию распределения $F_n(x)$:

$$\begin{aligned} P(|X_n - C| > \varepsilon) &= P(X_n - C > \varepsilon) + P(X_n - C < -\varepsilon) = \\ &= P(X_n < C - \varepsilon) + 1 - P(X_n \leq C + \varepsilon) = \\ &= F_n(C - \varepsilon) + 1 - F_n(C + \varepsilon) - P(X_n = C + \varepsilon). \end{aligned}$$

Так как $C - \varepsilon$ и $C + \varepsilon$ — точки непрерывности функции $F(x)$, то $F_n(C - \varepsilon) \rightarrow 0$, а $F_n(C + \varepsilon) \rightarrow 1$. Остается показать, что $P(X_n = C + \varepsilon) \rightarrow 0$. Имеем

$$\begin{aligned} 0 \leq P(X_n = C + \varepsilon) &\leq P\left(C + \frac{\varepsilon}{2} \leq X_n < C + \frac{3\varepsilon}{2}\right) = \\ &= F_n\left(C + \frac{3\varepsilon}{2}\right) - F_n\left(C + \frac{\varepsilon}{2}\right) \rightarrow 0, \end{aligned}$$

поскольку $C + 3\varepsilon/2$ и $C + \varepsilon/2$ есть точки непрерывности предельной функции $F(x)$, в которых она принимает одно и то же значение 1.

Теперь мы приступим к построению критерия слабой сходимости, развивая попутно новую и очень сильную технику построения вероятностных моделей.

§12. Характеристические функции. Теоремы единственности и сложения

Лекция 18

Мы введем сейчас одну из интереснейших функциональных характеристик случайной величины X , которая единственным образом определяет распределение X . С ее помощью можно найти все моменты X без вычисления интегралов, вычисляя производные от этой характеристики. Наконец, она представляет универсальный инструмент для вывода распределений сумм независимых случайных величин, поэтому с ее помощью можно просто и без громоздких выкладок доказывать предельные теоремы типа тех, что мы называли интегральной предельной теоремой Муавра–Лапласа.

Определение 12.1. *Характеристической функцией $\varphi(t)$ случайной величины X с функцией плотности $f(x)$ по мере μ называется преобразование Фурье–Лебега $f(x)$:*

$$\varphi(t) = \mathbf{E}e^{itX} = \int_{\mathbf{R}} e^{itx} f(x) d\mu(x).$$

Напомним, что в преобразованиях Фурье \mathbf{i} – мнимая единица, так что $e^{itx} = \cos(tx) + \mathbf{i} \sin(tx)$, и интеграл, определяющий $\varphi(t)$ представляет собой интеграл от функции комплексного переменного (криволинейный интеграл второго рода) по действительной оси $\mathbf{R} = (-\infty, +\infty)$.

Характеристическая функция существует при любом распределении X , поскольку $|e^{itx}| = 1$, откуда

$$|\varphi(t)| \leq \int_{\mathbf{R}} |e^{itx}| f(x) d\mu(x) = \int_{\mathbf{R}} f(x) d\mu(x) = 1.$$

Если $d\mu(x) = dx$ – мера Лебега, то $\varphi(t)$ есть обычное преобразование Фурье

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx,$$

если же μ – считающая мера, то $\varphi(t)$ представляет дискретный аналог преобразования Фурье

$$\varphi(t) = \sum_x e^{itx} f(x).$$

Рассмотрим несколько интересных примеров по вычислению характеристических функций известных нам распределений.

Пример 12.1 (*биномиальное распределение* $B(n, p)$). Характеристическая функция вычисляется простым суммированием биномиального ряда:

$$\begin{aligned}\varphi(t) &= \mathbf{E}e^{itX} = \sum_{x=0}^n e^{itx} C_n^x p^x (1-p)^{n-x} = \\ &= \sum_{x=0}^n C_n^x (e^{it}p)^x (1-p)^{n-x} = (pe^{it} + (1-p))^n.\end{aligned}$$

Пример 12.2 (*распределение Пуассона* $P(\lambda)$). Используем известное разложение Маклорена для показательной функции:

$$\varphi(t) = \sum_{x=0}^{\infty} e^{itx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^{it})^x}{x!} = \exp\{\lambda(e^{it} - 1)\}.$$

Пример 12.3 (*равномерное распределение* $U(a, b)$). Характеристическая функция представляет собой криволинейный интеграл по отрезку $[a; b]$ действительной прямой:

$$\varphi(t) = \frac{1}{b-a} \int_a^b e^{itx} dx = \frac{1}{b-a} \int_{[a;b]} e^{itz} dz.$$

Поскольку e^{itz} есть аналитическая функция, то интеграл равен разности значений первообразной этой функции в конечных точках отрезка интегрирования:

$$\varphi(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}.$$

Пример 12.4 (*показательное распределение* $E(\theta)$). Характеристическая функция снова представляется криволинейным интегралом от аналитической функции, но на сей раз по бесконечному промежутку $[0; \infty)$ действительной прямой:

$$\varphi(t) = \frac{1}{\theta} \int_0^{\infty} \exp\{itx - x\theta^{-1}\} dx = \frac{1}{\theta} \lim_{A \rightarrow \infty} \int_{[0;A]} \exp\{z(it - \theta^{-1})\} dz =$$

$$\lim_{A \rightarrow \infty} \frac{1 - \exp\{A(\mathbf{i}t - \theta^{-1})\}}{1 - \mathbf{i}t\theta}.$$

Однако

$$|\exp\{A(\mathbf{i}t - \theta^{-1})\}| = \exp\{-A\theta^{-1}\} |\exp\{\mathbf{i}At\}| = \exp\{-A\theta^{-1}\},$$

и так как $\theta > 0$, то

$$\lim_{A \rightarrow \infty} \exp\{A(\mathbf{i}t - \theta^{-1})\} = 0.$$

Таким образом,

$$\varphi(t) = \frac{1}{1 - \mathbf{i}t\theta}.$$

Можно, конечно, обойтись и без этой комплексной зауми, а просто воспользоваться формулой Эйлера $e^{\mathbf{i}z} = \cos z + \mathbf{i} \sin z$ и каким-нибудь справочником по интегралам (например, родным “Демидовичем”, а лучше всего справочником по интегральным преобразованиям).

Пример 12.5 (*распределение Коши* $C(0, 1)$). Используя формулу Эйлера, найдем характеристическую функцию стандартного (параметр сдвига $a = 0$, параметр масштаба $b = 1$) распределения Коши:

$$\varphi(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\cos tx + \mathbf{i} \sin tx}{1 + x^2} dx.$$

Интеграл от синуса (так называемое *синус-преобразование Фурье*) равен нулю, как интеграл от нечетной функции по симметричному относительно начала координат промежутку, а для нахождения косинус-преобразования Фурье воспользуемся четностью функции $\cos x$ и ответом к примеру N 3825 задачника Демидовича:

$$\varphi(t) = \frac{2}{\pi} \int_0^{\infty} \frac{\cos tx}{1 + x^2} dx = e^{-|t|}.$$

Пример 12.6 (*стандартное нормальное распределение* $N(0, 1)$). Рассуждая так же, как и в случае распределения Коши, и используя ответ к примеру N 3809, получаем

$$\varphi(t) = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \cos tx \exp\left\{-\frac{x^2}{2}\right\} dx = \exp\left\{-\frac{t^2}{2}\right\}.$$

Характеристические функции распределений Коши и нормального при произвольных значениях параметров можно получить просто линейной заменой переменной интегрирования, но легко видеть, что справедлива общая формула для семейств распределений, зависящих от параметров сдвига и масштаба.

Предложение 12.1. *Характеристическая функция $\varphi_X(t)$ случайной величины X обладает следующими свойствами.*

$$1^0. \varphi(0) = 1, \quad |\varphi(t)| \leq 1.$$

$$2^0. \varphi_{bX+a} = e^{ita} \varphi_X(bt).$$

3⁰. Если X_1, \dots, X_n независимы в совокупности, то

$$\varphi_{\sum_1^n X_k}(t) = \prod_1^n \varphi_{X_k}(t);$$

в частности, если X_1, \dots, X_n независимы и одинаково распределены, то

$$\varphi_{\sum_1^n X_k}(t) = (\varphi_{X_1}(t))^n.$$

4⁰. Характеристическая функция $\varphi(t)$ равномерно непрерывна на всей действительной оси \mathbf{R} .

5⁰. Если случайная величина X обладает моментами $\alpha_k = \mathbf{E}X^k$, $k = 1, \dots, n$, то

$$\alpha_k = \mathbf{i}^{-k} \varphi^{(k)}(0)$$

и для характеристической функции справедливо разложение Тейлора

$$\varphi(t) = 1 + \sum_{k=1}^n \frac{(\mathbf{i}t)^k}{k!} \alpha_k + o(t^n), \quad t \rightarrow 0.$$

Доказательство. 1⁰. Это свойство, по существу, было установлено сразу же после определения характеристической функции, когда мы рассуждали о ее существовании.

2⁰. По определению характеристической функции

$$\varphi_{bX+a}(t) = \mathbf{E}e^{\mathbf{i}t(bX+a)} = e^{ita} \mathbf{E}e^{\mathbf{i}btX} = e^{ita} \varphi_X(bt).$$

3⁰. Опять работаем с определением характеристической функции:

$$\varphi_{\sum_1^n X_k}(t) = \mathbf{E} \exp \left\{ \mathbf{i}t \sum_1^n X_k \right\} = \mathbf{E} \prod_1^n e^{\mathbf{i}tX_k} = \prod_1^n \mathbf{E} e^{\mathbf{i}tX_k} = \prod_1^n \varphi_{X_k}(t).$$

Естественно, если X_1, \dots, X_n одинаково распределены, то произведение в правой части последнего равенства состоит из одинаковых сомножителей и мы получаем $\varphi_{X_1}^n(t)$.

4⁰. Требуется доказать, что

$$\sup_{t \in \mathbf{R}} |\varphi(t+h) - \varphi(t)| \rightarrow 0,$$

когда $h \rightarrow 0$. Оценим приращение характеристической функции:

$$\begin{aligned} |\varphi(t+h) - \varphi(t)| &= \\ \left| \int_{-\infty}^{\infty} (e^{\mathbf{i}(t+h)x} - e^{\mathbf{i}tx}) f(x) d\mu(x) \right| &\leq \int_{-\infty}^{\infty} |e^{\mathbf{i}tx}| |e^{\mathbf{i}hx} - 1| |f(x)| d\mu(x) = \\ \int_{-\infty}^{\infty} \sqrt{(\cos hx - 1)^2 + \sin^2 hx} |f(x)| d\mu(x) &= \int_{-\infty}^{\infty} \sqrt{2(1 - \cos hx)} |f(x)| d\mu(x). \end{aligned}$$

Так как $0 \leq 1 - \cos hx \leq 2$, то последний интеграл сходится равномерно (признак Вейерштрасса) и можно переходить к пределу при $h \rightarrow 0$ под знаком интеграла. Но

$$\lim_{h \rightarrow 0} (1 - \cos hx) = 0,$$

каково бы ни было $x \in \mathbf{R}$. Следовательно,

$$\lim_{h \rightarrow 0} \int_{-\infty}^{\infty} \sqrt{2(1 - \cos hx)} |f(x)| d\mu(x) = 0,$$

откуда

$$\lim_{h \rightarrow 0} \sup_{t \in \mathbf{R}} |\varphi(t+h) - \varphi(t)| \rightarrow 0.$$

5⁰. Формальное дифференцирование k раз под знаком интеграла в формуле, определяющей характеристическую функцию, приводит нас к соотношению

$$\varphi^{(k)}(t) = \mathbf{i}^k \int_{-\infty}^{\infty} x^k e^{\mathbf{i}tx} f(x) d\mu(x).$$

Если k -й момент α_k существует, то (напомним, $|e^{itx}| = 1$)

$$\left| \int_{-\infty}^{\infty} x^k e^{itx} f(x) d\mu(x) \right| \leq \int_{-\infty}^{\infty} |x|^k f(x) d\mu(x) < \infty,$$

поскольку существование интеграла Лебега от функции влечет его существование от модуля этой функции. Таким образом, в силу признака Вейерштрасса, интеграл сходится равномерно, формальное дифференцирование под знаком интеграла оправдано, и мы получаем искомую формулу для вычисления моментов случайной величины X , полагая $t = 0$: $\varphi^{(k)}(0) = i^k \alpha_k$.

Итак, существование момента k -го порядка влечет существование k -ой производной в точке $t = 0$ функции $\varphi(t)$. Если существует n моментов, то можно воспользоваться формулой Тейлора и получить асимптотическое ($t \rightarrow 0$) разложение характеристической функции:

$$\varphi(t) = \sum_{k=0}^n \frac{\varphi^{(k)}(0)}{k!} t^k + o(t^n) = 1 + \sum_{k=1}^n \frac{(it)^k}{k!} \alpha_k + o(t^n).$$

Используя свойство 2^0 , легко находим характеристическую функцию распределений $C(a, b)$ и $N(\mu, \sigma^2)$.

Если $X \sim C(0, 1)$, то $bX + a \sim C(a, b)$, и характеристическая функция распределения Коши с плотностью

$$f(x) = \frac{1}{\pi b [1 + ((x - a)/b)^2]}$$

равна (см. пример 12.5)

$$\varphi_{bX+a}(t) = \exp\{iat - b|t|\}.$$

Аналогично, если $X \sim N(0, 1)$, то $\sigma X + \mu \sim N(\mu, \sigma^2)$, и характеристическая функция нормального распределения с плотностью

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

равна (см. пример 12.6)

$$\varphi_{\sigma X + \mu}(t) = \exp\{i\mu t - \sigma^2 t^2 / 2\}.$$

Из всех свойств характеристической функции, установленных в предложении 12.1, наиболее привлекательным кажется свойство \mathfrak{Z}^0 , позволяющее находить характеристическую функцию суммы независимых случайных величин по характеристическим функциям слагаемых – открываются новые возможности в построении вероятностных моделей. Но при этом возникает естественный вопрос: существует ли взаимно однозначное соответствие между характеристическими функциями и функциями распределения (или плотности). Из курса математического анализа мы знаем, что на каждое преобразование Фурье

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx$$

существует обратное преобразование

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t) dt, \quad (1)$$

хотя, насколько мне известно, доказательства этой формулы обращения вам не давалось. Тем не менее, информацию о справедливости теоремы единственности вы получили, и мы теперь восполним пробел в вашем образовании, доказав аналогичную теорему для более общего преобразования Фурье–Лебега.

Теорема 12.1. (*формула обращения Леви*). Если $F(x)$ – функция распределения с.в. X , а $\varphi(t)$ – ее характеристическая функция, то для любых точек непрерывности x и y функции $F(x)$ имеет место формула обращения

$$F(y) - F(x) = \frac{1}{2\pi} \lim_{A \rightarrow \infty} \int_{-A}^A \frac{e^{-itx} - e^{-ity}}{it} \varphi(t) dt. \quad (2)$$

Доказательство. Заметим сначала, что правая часть формулы обращения (2) представляет собой несобственный интеграл в смысле главного значения, так как $\varphi(t)$ может оказаться неинтегрируемой функцией. Если существует $f(x) = dF(x)/dx$ и характеристическая функция $\varphi(t)$ интегрируема, то (2) нетрудно получить из формулы обращения преобразования Фурье (1), проинтегрировав обе части (1) в пределах от x до y .

Обратимся теперь непосредственно к доказательству формулы (2), для чего рассмотрим при $y > x$ интеграл

$$J_A = \frac{1}{2\pi} \int_{-A}^A \frac{e^{-itx} - e^{-ity}}{it} \varphi(t) dt = \frac{1}{2\pi} \int_{-A}^A dt \int_{-\infty}^{\infty} \frac{e^{it(u-x)} - e^{it(u-y)}}{it} f(u) d\mu(u),$$

в котором $\varphi(t)$ заменена на определяющий ее интеграл. Легко видеть, что при фиксированных x и y подынтегральная функция

$$\frac{e^{it(u-x)} - e^{it(u-y)}}{t}$$

в области $|u| < \infty$, $|t| < \infty$ непрерывна и ограничена, поэтому можно изменить порядок интегрирования:

$$J_A = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\int_{-A}^A \frac{e^{it(u-x)} - e^{it(u-y)}}{it} dt \right] f(u) d\mu(u).$$

Преобразуем внутренний интеграл I_A в пределах от $-A$ до A , для чего представим его в виде суммы интегралов по отрезкам $[-A, 0]$ и $[0, A]$ и в интеграле по отрезку $[-A, 0]$ сделаем замену t на $-t$. В результате получим

$$I_A = \int_0^A \left[\frac{e^{it(u-x)} - e^{-it(u-x)}}{it} - \frac{e^{it(u-y)} - e^{-it(u-y)}}{it} \right] dt =$$

$$2 \int_0^A \left[\frac{\sin(t(u-x))}{t} - \frac{\sin(t(u-y))}{t} \right] dt,$$

поскольку (формула Эйлера) $(e^{iz} - e^{-iz})/2i = \sin z$.

Вычисляя интеграл Дирихле

$$\int_0^{\infty} \frac{\sin(\alpha t)}{t} dt = \frac{\pi}{2} \operatorname{sgn} \alpha,$$

получаем следующее выражение для правой части (2):

$$\frac{1}{2} \int_{-\infty}^{\infty} [\operatorname{sgn}(u-x) - \operatorname{sgn}(u-y)] f(u) d\mu(u).$$

Представим последний интеграл в виде суммы трех интегралов по отрезкам $(-\infty, x]$, $[x, y]$ и $[y, \infty)$, на которых, соответственно,

$$\begin{aligned}\operatorname{sgn}(u - x) &= \operatorname{sgn}(u - y) = -1, \quad \operatorname{sgn}(u - x) = -\operatorname{sgn}(u - y) = +1, \\ \operatorname{sgn}(u - x) &= \operatorname{sgn}(u - y) = 1.\end{aligned}$$

Тогда этот интеграл, а следовательно, и правая часть (2), принимает окончательный вид

$$\int_x^y f(u) d\mu(u) = F(y) - F(x),$$

устанавливающий справедливость формулы обращения (2).

Итак, теперь можно не сомневаться, что, получив каким-либо способом характеристическую функцию наблюдаемой случайной величины X , мы, по сути дела, уже построили вероятностную модель, и остается только, используя формулу (2), найти функцию распределения X . Проиллюстрируем этот метод построения модели на одной из центральных задач *теории восстановления*, имеющей большие применения в практике и теории надежности систем, подвергаемых в процессе их эксплуатации ремонту (восстановлению), профилактике и резервированию компонент с высокой частотой отказа.

Гамма-распределение $G(\lambda, \theta)$. Рассматривается система, долговечность которой определяется моментом отказа X_1 ее отдельного элемента. Предположим, что $X_1 \sim E(\theta)$, то есть функционирование элемента протекает в рамках постулата “отсутствие последствия”. Система имеет резерв, состоящий из $n - 1$ таких же элементов, и при отказе работающего элемента мгновенно подключается запасной. Таким образом, общая долговечность системы определяется реализацией случайной величины $X = \sum_1^n X_i$, в которой слагаемые независимы и одинаково распределены по показательному закону $E(\theta)$ с характеристической функцией (см. пример 12.4) $\varphi_1(t) = (1 - i\theta t)^{-1}$.

В силу пункта 3⁰ предложения 12.1 характеристическая функция X равна $\varphi(t) = (1 - i\theta t)^{-n}$. Применяя обратное преобразование Фурье к $\varphi(t)$ (советую воспользоваться справочником – такие интегралы на нашем богоугодном факультете считать теперь не учат), получаем функцию плотности распределения долговечности

$$f(x) = f(x | n, \theta) = \frac{1}{(n - 1)! \theta^n} x^{n-1} \exp \left\{ -\frac{x}{\theta} \right\}, \quad x > 0,$$

(естественно, $f(x) = 0$ при $x \leq 0$).

Как будет видно в дальнейшем, полученное распределение долговечности с заменой целочисленного параметра n на произвольный положительный параметр λ описывает долговечность не только резервированных (или восстанавливаемых при отказе) систем, но и долговечность систем, подверженных износу, старению, накоплению усталости, в общем, всему тому, что постепенно накапливается, а потом приводит к “гибели”. В связи с этими замечаниями мы определяем *гамма-распределение* $G(\lambda, \theta)$ посредством функции плотности

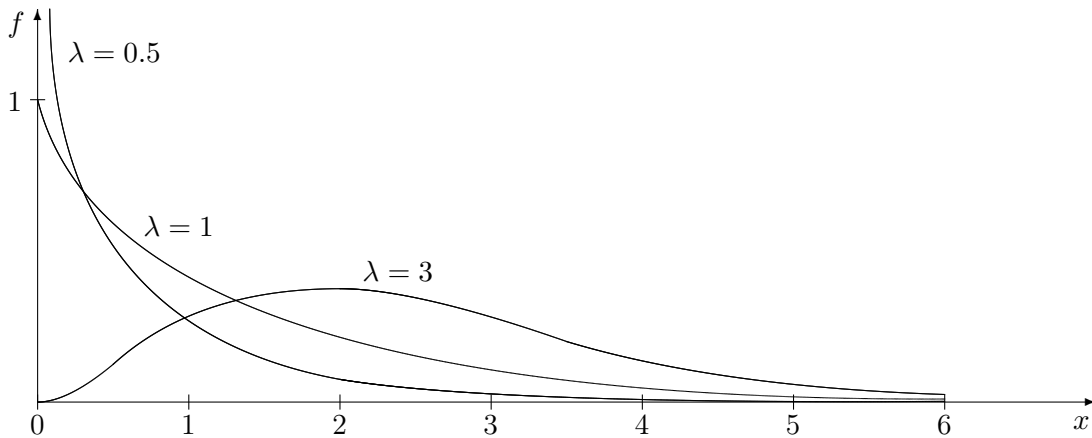
$$f(x | \lambda, \theta) = \frac{1}{\Gamma(\lambda)\theta^\lambda} x^{\lambda-1} \exp\left\{-\frac{x}{\theta}\right\}, \quad x > 0; \lambda > 0, \theta > 0,$$

где

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x} dx$$

– гамма-функция Эйлера.

Семейство гамма-распределений $\{G(\lambda, \theta), (\lambda, \theta) \in \mathbb{R}_+ \times \mathbb{R}_+\}$ содержит, как частный случай, показательное распределение ($\lambda = 1$). Гамма-распределение унимодально: если $\lambda \leq 1$, то $\text{mod}X = 0$, а при $\lambda > 1$ мода $\text{mod}X = \theta(\lambda - 1)$.



У гамма-распределения существуют моменты любого порядка:

$$\alpha_k = \mathbf{E}X^k = \frac{1}{\Gamma(\lambda)\theta^\lambda} \int_0^\infty x^{\lambda+k-1} \exp\left\{-\frac{x}{\theta}\right\} dx =$$

$$\frac{\Gamma(\lambda + k)\theta^{\lambda+k}}{\Gamma(\lambda)\theta^\lambda} = \lambda(\lambda + 1) \cdots (\lambda + k - 1)\theta^k.$$

В частности, $\mathbf{E}X = \lambda\theta$, $\mathbf{D}X = \lambda\theta^2$.

Теперь, используя аппарат характеристических функций, мы можем составить каталог изученных нами распределений, для которых справедлива теорема сложения.

Предложение 12.2. Пусть X_1, \dots, X_n независимы и $S_n = \sum_1^n X_k$. Тогда,

$$1^0 \text{ если } X_k \sim B(m_k, p), \quad k = 1, \dots, n, \quad \text{то } S_n \sim B\left(\sum_1^n m_k, p\right);$$

$$2^0 \text{ если } X_k \sim P(\lambda_k), \quad k = 1, \dots, n, \quad \text{то } S_n \sim P\left(\sum_1^n \lambda_k\right);$$

$$3^0 \text{ если } X_k \sim C(a_k, b_k), \quad k = 1, \dots, n, \quad \text{то } S_n \sim C\left(\sum_1^n a_k, \sum_1^n b_k\right);$$

$$4^0 \text{ если } X_k \sim N(\mu_k, \sigma_k^2), \quad k = 1, \dots, n, \quad \text{то } S_n \sim N\left(\sum_1^n \mu_k, \sum_1^n \sigma_k^2\right);$$

$$5^0 \text{ если } X_k \sim G(\lambda_k, \theta), \quad k = 1, \dots, n, \quad \text{то } S_n \sim G\left(\sum_1^n \lambda_k, \theta\right).$$

Доказательство. Следующая таблица характеристических функций отдельных слагаемых X_k и суммы S_n устанавливает справедливость всех утверждений предложения.

1^0 $B(m, p)$:

$$\varphi_{X_k}(t) = (pe^{it} + 1 - p)^{m_k}, \quad \varphi_{S_n}(t) = (pe^{it} + 1 - p)^{\sum_1^n m_k};$$

2^0 $P(\lambda)$:

$$\varphi_{X_k}(t) = \exp\{\lambda_k(e^{it} - 1)\}, \quad \varphi_{S_n}(t) = \exp\left\{\sum_1^n \lambda_k(e^{it} - 1)\right\};$$

3^0 $C(a, b)$:

$$\varphi_{X_k}(t) = \exp\{it a_k - |t| b_k\}, \quad \varphi_{S_n}(t) = \exp\left\{it \sum_1^n a_k - |t| \sum_1^n b_k\right\};$$

4^0 $N(\mu, \sigma^2)$:

$$\varphi_{X_k}(t) = \exp\left\{it\mu_k - \frac{t^2}{2}\sigma_k^2\right\}, \quad \varphi_{S_n}(t) = \exp\left\{it \sum_1^n \mu_k - \frac{t^2}{2} \sum_1^n \sigma_k^2\right\};$$

$$5^0 \text{ G}(\lambda, \theta) : \quad \varphi_{X_k}(t) = (1 - \mathbf{i}\theta t)^{-\lambda_k}, \quad \varphi_{S_n}(t) = (1 - \mathbf{i}\theta t)^{-\sum_1^n \lambda_k}.$$

Несколько слов о характеристической функции многомерного распределения. Если $X^{(n)} = (X_1, \dots, X_n)$ – случайный вектор с функцией плотности $f_n(x^{(n)}) = f_n(x_1, \dots, x_n)$ по мере

$$d\mu_n(x^{(n)}) = d\mu_1(x_1) \cdots d\mu_n(x_n),$$

то характеристическая функция определяется как n -мерное преобразование Фурье-Лебега

$$\varphi_n(t^{(n)}) = \mathbf{E} \exp \left\{ \mathbf{i} \left(t^{(n)}, X^{(n)} \right) \right\} = \int_{\mathbb{R}_n} \exp \left\{ \mathbf{i} \sum_1^n t_k x_k \right\} f_n(x^{(n)}) d\mu_n(x^{(n)}).$$

Очень просто, по прямой аналогии с биномиальным распределением, находится характеристическая функция мультиномиального распределения, и столь же просто, если воспользоваться ответом к задаче N 4220 из Демидовича, характеристическая функция n -мерного нормального распределения $\mathcal{N}(\mu, \Lambda)$:

$$\varphi_n(t^{(n)}) = \exp \left\{ \mathbf{i} \sum_1^n \mu_k t_k - \frac{1}{2} \sum_{1 \leq j, k \leq n} \lambda_{jk} t_j t_k \right\}.$$

Для многомерной характеристической функции также справедливы теоремы единственности и утверждения, аналогичные предложению 12.1. Используя аппарат многомерных характеристических функций, можно показать, что для мультиномиального и многомерного нормального распределений справедливы теоремы сложения, и доказать следующее удивительное свойство многомерного нормального распределения: *любое линейное преобразование $Y^{(m)} = \mathbf{A}X^{(n)}$ (с матрицей \mathbf{A} размерности $m \times n$) случайного вектора $X^{(n)} \sim \mathcal{N}(\mu, \Lambda)$ дает случайный вектор, имеющий m -мерное нормальное распределение со средним $\mu\mathbf{A}$ и ковариационной матрицей $\mathbf{A}\Lambda\mathbf{A}'$.*

§13. Характеристические функции. Критерий слабой сходимости

Лекция 20–21

Следующая теорема дает удобный критерий слабой сходимости распределений случайных величин.

Теорема 13.1. Пусть $\{\varphi_n, n \geq 1\}$ – последовательность характеристических функций и $\{F_n, n \geq 1\}$ – последовательность соответствующих функций распределений. Если при любом фиксированном $t \in \mathbf{R}$ последовательность характеристических функций сходится к некоторой непрерывной в точке $t = 0$ функции $\varphi(t)$, то $\varphi(\cdot)$ есть характеристическая функция некоторой случайной величины X с функцией распределения $F(\cdot)$ и $F_n \Rightarrow F$. Обратно, если $F_n \Rightarrow F$ и $F(\cdot)$ есть функция распределения, то $\varphi_n(t) \rightarrow \varphi(t)$ при любом $t \in \mathbf{R}$ и $\varphi(\cdot)$ – характеристическая функция случайной величины X с функцией распределения $F(\cdot)$.

Доказательство этой теоремы (в монографиях по теории вероятностей она обычно называется *теоремой непрерывности* для последовательностей характеристических функций) основано на ряде вспомогательных утверждений о слабой сходимости функций распределений.

Лемма 13.1. Всякая последовательность функций распределения $\{F_n, n \geq 1\}$ содержит подпоследовательность $\{F_{n_k}, k \geq 1\}$, слабо сходящуюся к некоторой ограниченной неубывающей и непрерывной слева функции $F(\cdot)$, т.е. $F_{n_k}(x) \rightarrow F(x)$ при $k \rightarrow \infty$ в любой точке x непрерывности функции $F(\cdot)$.

Замечание 13.1. Если последовательность $\{F_n(x), n \geq 1\}$ сходится в каждой точке x , то предельная функция $F(x)$, $x \in \mathbf{R}$ может и не быть функцией распределения, хотя, очевидно, $F(\cdot)$ не убывает и ее изменение на \mathbf{R} :

$$\text{var} F = \sup_x F(x) - \inf_x F(x) \leq 1,$$

ибо таковы функции распределения $F_n(\cdot)$, $n = 1, 2, \dots$. Пример такой последовательности дают функции $F_n(\cdot)$ равномерных распределений на отрезках $[n; n + 1]$, $n = 1, 2, \dots$. Поскольку $F_n(x) = 0$ при $x < n$, то для любого $x \in \mathbf{R}$ существует такое N (достаточно взять N больше x), что $F_n(x) = 0$ для всех $n \geq N$. Следовательно, $F_n(x) \rightarrow F(x) \equiv 0$ и $\text{var} F = 0$.

Доказательство леммы 13.1. Начнем с выбора подпоследовательности $\{F_{n_k}, k \geq 1\}$, которая сходится слабо к некоторому пределу F , обладающему указанными свойствами.

Пусть $\mathcal{D} = \{r_n, n \geq 1\}$ – счетное всюду плотное в \mathbb{R} множество, например, множество рациональных чисел. Числовая последовательность $\{F_n(r_1), n \geq 1\}$ ограничена, и поэтому содержит сходящуюся подпоследовательность $\{F_{1n}(r_1), n \geq 1\}$. Пусть $F_1(r_1)$ – предел этой подпоследовательности. Рассмотрим теперь последовательность чисел $\{F_{1n}(r_2), n \geq 1\}$; она также содержит сходящуюся подпоследовательность $\{F_{2n}(r_2), n \geq 1\}$ с некоторым пределом $F_2(r_2)$, причем

$$\lim_{n \rightarrow \infty} F_{2n}(r_1) = F_1(r_1),$$

ибо $\{F_{2n}(r_1), n \geq 1\}$ – подпоследовательность сходящейся к $F_1(r_1)$ последовательности $\{F_{1n}(r_1), n \geq 1\}$. Точно так же последовательность $\{F_{2n}(r_3), n \geq 1\}$ содержит подпоследовательность $\{F_{3n}(r_3), n \geq 1\}$ с пределом $F_3(r_3)$, причем

$$\lim_{n \rightarrow \infty} F_{3n}(r_2) = F_2(r_2), \quad \lim_{n \rightarrow \infty} F_{3n}(r_1) = F_1(r_1),$$

ибо

$$\{F_{3n}(r_1), n \geq 1\} \subseteq \{F_{2n}(r_1), n \geq 1\} \subseteq \{F_{1n}(r_1), n \geq 1\} –$$

индексы каждой последующей подпоследовательности выбирались из множества индексов предыдущей. Продолжая этот процесс, мы убеждаемся, что для любого $k \geq 1$ число $F_k(r_k)$ есть общий предел всех последовательностей

$$\{F_{jn}(r_k), n \geq 1\}, \quad j = k, k + 1, \dots,$$

причем каждая последующая последовательность есть подпоследовательность предыдущей.

Рассмотрим диагональную последовательность $\{F_{nn}(r_k), n \geq 1\}$. За исключением первых $k - 1$ членов ее последующие члены выбираются по одному из рассмотренных выше последовательностей, следовательно,

$$\lim_{n \rightarrow \infty} F_{nn}(r_k) = F_k(r_k).$$

Тем самым для всех $x \in \mathcal{D}$ определена неубывающая функция $F_0(x)$, равная $F_k(r_k)$, если $x = r_k$, и

$$\lim_{n \rightarrow \infty} F_{nn}(x) = F_0(x), \quad \forall x \in \mathcal{D}.$$

Функция $F_0(\cdot)$ ограничена и не убывает на \mathcal{D} , ибо этими свойствами обладает каждый член последовательности $\{F_{nn}, n \geq 1\}$. Теперь определим $F(x)$ при любом $x \in \mathbb{R}$, полагая

$$F(x) = \sup_{r < x, r \in \mathcal{D}} F_0(r).$$

Покажем, что $F(\cdot)$ – искомая функция, то есть она (1) не убывает, (2) непрерывна слева и (3) $F_{nn}(x) \rightarrow F(x)$ в каждой точке x непрерывности функции F .

(1) Монотонность F следует из аналогичного свойства F_0 : если $x \leq y$, то

$$F(x) = \sup_{r < x} F_0(r) \leq \sup_{r < y} F_0(r) = F(y).$$

(2) Непрерывность слева функции F в любой точке $x \in \mathbb{R}$ вытекает из определения точной верхней грани и монотонности функций F и F_0 . Требуется доказать, что для любых $\varepsilon > 0$ и $x \in \mathbb{R}$ существует такое $y_0 = y_0(\varepsilon, x) < x$, что $0 \leq F(x) - F(y) \leq \varepsilon$ при любом $y \in (y_0, x)$. По определению супремума существует такая возрастающая (супремальная) последовательность $\{r_k, k \geq 1\} \subset \mathcal{D}$, что $r_k < x$ при $\forall k = 1, 2, \dots$ и

$$\lim_k \uparrow F_0(r_k) = F(x).$$

Следовательно, существует такое $K = K(\varepsilon)$, что при $\forall k \geq K$ выполняется неравенство $0 \leq F(x) - F_0(r_k) < \varepsilon$. Но для любого $y \geq r_K$ имеет место неравенство

$$F_0(r_K) \leq \sup_{r < y} F_0(r) = F(y),$$

и поэтому $0 \leq F(x) - F(y) \leq \varepsilon$, каково бы ни было $y \geq r_K = y_0$. Итак, F непрерывна слева.

(3) Покажем теперь, что $F_{nn} \Rightarrow F$, то есть в любой фиксированной точке x непрерывности функции $F(\cdot)$, начиная с некоторого n , выполняется неравенство $|F_{nn}(x) - F(x)| < \varepsilon$, каково бы ни было наперед заданное число $\varepsilon > 0$.

Начнем с того, что в силу только что установленной непрерывности слева функции $F(\cdot)$ по заданному ε всегда можно подобрать такие $x', x'' \in \mathbb{R}$ и $r', r'' \in \mathcal{D}$, что $x' < r' < x < r'' < x''$, и при этом

$$0 < F(x'') - F(x) < \varepsilon/2, \quad 0 < F(x) - F(x') < \varepsilon/2.$$

Так как $F_{nn}(r) \rightarrow F_0(r)$ при $\forall r \in \mathbf{D}$, то, начиная с некоторого $n > N(\varepsilon)$, выполняется неравенство $|F_{nn}(r) - F_0(r)| < \varepsilon/2$, и поэтому

$$F_{nn}(x) - F(x) \leq F_{nn}(r'') - F(x) = \\ [F_{nn}(r'') - F_0(r'')] + [F_0(r'') - F(x)] \leq \varepsilon/2 + F_0(r'') - F(x) \quad ,$$

а также

$$F(x) - F_{nn}(x) \leq F(x) - F_{nn}(r') = \\ [F(x) - F_0(r')] + [F_0(r') - F_{nn}(r')] \leq F(x) - F_0(r') + \varepsilon/2.$$

Для доказательства сходимости $F_{nn}(x)$ к $F(x)$ достаточно показать, что

$$F_0(r'') \leq F(x''), \quad F_0(r') \geq F(x'),$$

а затем воспользоваться неравенством $F(x'') - F(x) < \varepsilon/2$. Но это почти очевидно, поскольку выполняются строгие неравенства $x' < r'$ и $r'' < x''$. Действительно,

$$F_0(r'') \leq \sup_{r < x''} F_0(r) = F(x'')$$

и, аналогично,

$$F_0(r') \geq \sup_{r < r'} F_0(r) = F(r') \geq F(x').$$

Следовательно,

$$F_0(r'') - F(x) \leq F(x'') - F(x) \leq \varepsilon/2, \quad F(x) - F_0(r') \leq F(x) - F(x') \leq \varepsilon/2,$$

откуда $-\varepsilon \leq F_{nn}(x) - F(x) \leq \varepsilon$.

Условимся, начиная с этого момента, записывать интеграл Лебега

$$\int_B g(x) dP(x), \quad B \in \mathfrak{B}$$

по вероятностной мере P на борелевской прямой $(\mathbf{R}, \mathfrak{B})$ как

$$\int_B g(x) dF(x),$$

используя тем самым вместо P функцию распределения F , которая, в силу теоремы 4.1, однозначно определяет распределение вероятностей P .

Лемма 13.2. *Для того чтобы последовательность функций распределений $\{F_n, n \geq 1\}$ слабо сходилась к некоторой функции распределения*

$F(\cdot)$, необходимо и достаточно, чтобы для любой непрерывной и ограниченной функции $g(x)$, $x \in \mathbb{R}$

$$\lim_{n \rightarrow \infty} \int_{\mathbb{R}} g(x) dF_n(x) = \int_{\mathbb{R}} g(x) dF(x). \quad (1)$$

Доказательство. *Необходимость.* Оценим разность

$$\Delta_n = \left| \int_{-\infty}^{\infty} g(x) dF(x) - \int_{-\infty}^{\infty} g(x) dF_n(x) \right|$$

и покажем, что Δ_n можно сделать сколь угодно малым, выбирая достаточно большое n , если $F_n \Rightarrow F$.

Зададимся некоторым $\varepsilon > 0$ и выберем на оси \mathbb{R} такие точки a и b , чтобы $F(x)$ была непрерывной в a и b и чтобы $F(a) < \varepsilon$ и $1 - F(b) < \varepsilon$. Поскольку $F_n \Rightarrow F$, то

$$\lim_{n \rightarrow \infty} F_n(a) = F(a), \quad \lim_{n \rightarrow \infty} F_n(b) = F(b)$$

и, следовательно,

$$F_n(a) \leq F(a) + \varepsilon, \quad F_n(b) \geq F(b) - \varepsilon, \quad (2)$$

начиная с некоторого $n > N(\varepsilon)$.

Разобьем каждый из интегралов, участвующих в определении Δ_n , на сумму трех интегралов по промежуткам $[-\infty; a]$, $[a; b]$, $[b; +\infty]$. Тогда $\Delta_n \leq \Delta_{1n} + \Delta_{2n} + \Delta_{3n}$, где

$$\Delta_{1n} = \left| \int_{-\infty}^a g dF - \int_{-\infty}^a g dF_n \right|, \quad \Delta_{2n} = \left| \int_a^b g dF - \int_a^b g dF_n \right|,$$

$$\Delta_{3n} = \left| \int_b^{\infty} g dF - \int_b^{\infty} g dF_n \right|.$$

Положим $M = \sup_x |g(x)| < \infty$ (напомним, функция g ограничена) и оценим Δ_{1n} и Δ_{3n} . Используя (2), получаем

$$\Delta_{1n} \leq \int_{-\infty}^a |g| dF + \int_{-\infty}^a |g| dF_n \leq M(F(a) + F_n(a)) \leq M(2F(a) + \varepsilon) \leq 3M\varepsilon,$$

$$\Delta_{3n} = \int_b^{\infty} |g| dF + \int_b^{\infty} |g| dF_n \leq M(1 - F(b) + 1 - F_n(b)) \leq 3M\varepsilon,$$

ибо $F(a) < \varepsilon$ и $1 - F(b) < \varepsilon$. Таким образом, Δ_{1n} и Δ_{3n} стремятся к нулю с ростом n . Покажем, что аналогичное заключение можно сделать относительно Δ_{2n} .

Разобьем отрезок $[a; b]$ на N частей точками x_1, \dots, x_{N-1} , выбрав их так, чтобы они оказались точками непрерывности $F(\cdot)$ (это возможно в силу известного свойства функции распределения: она имеет не более чем счетное множество скачков, и поэтому не может быть целого промежутка, состоящего из точек разрыва $F(\cdot)$). Итак, пусть

$$a = x_0 < x_1 < \dots < x_{N-1} < x_N = b.$$

Так как функция $g(\cdot)$ непрерывна на \mathbb{R} , то на конечном отрезке $[a; b]$ она равномерно непрерывна. Следовательно, при достаточно большом N разность $|g(x) - g(x_k)| < \varepsilon$ при $x_k \leq x < x_{k+1}$ и любом $k = 0, \dots, N$. Введем ступенчатую функцию $g_\varepsilon(x)$, положив ее равной $g(x_k)$, если $x \in [x_k; x_{k+1})$, $k = 0, \dots, N-1$, и обратимся к оценке Δ_{2n} . Имеем

$$\begin{aligned} \Delta_{2n} = & \left| \int_a^b (g - g_\varepsilon + g_\varepsilon) dF - \int_a^b (g - g_\varepsilon + g_\varepsilon) dF_n \right| \leq \\ & \left| \int_a^b (g - g_\varepsilon) dF \right| + \left| \int_a^b (g - g_\varepsilon) dF_n \right| + \left| \int_a^b g_\varepsilon dF - \int_a^b g_\varepsilon dF_n \right|. \end{aligned}$$

Каждое из первых двух слагаемых в правой части не превосходит ε , поскольку

$$|g - g_\varepsilon| < \varepsilon, \quad F(b) - F(a) \leq 1, \quad F_n(b) - F_n(a) \leq 1,$$

а для последнего слагаемого имеем оценку

$$\begin{aligned} \left| \int_a^b g_\varepsilon dF - \int_a^b g_\varepsilon dF_n \right| &= \left| \sum_{k=0}^{N-1} g(x_k) \int_{x_k}^{x_{k+1}} dF(x) - \sum_{k=0}^{N-1} g(x_k) \int_{x_k}^{x_{k+1}} dF_n(x) \right| = \\ & \left| \sum_{k=0}^{N-1} g(x_k) \{ (F(x_{k+1}) - F(x_k)) - (F_n(x_{k+1}) - F_n(x_k)) \} \right| \leq \\ & \sum_{k=0}^{N-1} |g(x_k)| \{ |F(x_{k+1}) - F_n(x_{k+1})| + |F(x_k) - F_n(x_k)| \}. \end{aligned}$$

Правая часть этого неравенства меньше наперед заданного $\varepsilon > 0$, поскольку N фиксировано, $|g(x)| \leq M$, а $F_n(x_k) \rightarrow F(x_k)$ при любом $k =$

$0, \dots, N$. Итак, Δ_{2n} сколь угодно мало и, следовательно, $\Delta_n \rightarrow 0$ при $n \rightarrow \infty$.

Достаточность. Пусть выполняется (1). Для любого $\varepsilon > 0$ и любой точки x непрерывности F рассмотрим непрерывную функцию $f_\varepsilon(t)$, принимающую значение 1 при $t < x$, значение 0, если $t > x + \varepsilon$, и меняющуюся линейно на $[x; x + \varepsilon]$. Так как

$$F_n(x) = \int_{-\infty}^x f_\varepsilon(t) dF_n(t) \leq \int_{-\infty}^{\infty} f_\varepsilon(t) dF_n(t),$$

то в силу (1)

$$\limsup_n F_n(x) \leq \int_{-\infty}^{\infty} f_\varepsilon(t) dF(t) \leq \int_{-\infty}^{x+\varepsilon} dF(t) = F(x + \varepsilon).$$

Аналогично, с помощью функции $f_\varepsilon^*(t) = f_\varepsilon(t + \varepsilon)$ получаем неравенство

$$F_n(x) \geq \int_{-\infty}^x f_\varepsilon^*(t) dF_n(t) = \int_{-\infty}^{\infty} f_\varepsilon^*(t) dF_n(t),$$

откуда

$$\liminf_n F_n(x) \geq \int_{-\infty}^{\infty} f_\varepsilon^*(t) dF(t) \geq F(x - \varepsilon).$$

Следовательно,

$$F(x - \varepsilon) \leq \liminf_n F_n(x) \leq \limsup_n F_n(x) \leq F(x + \varepsilon),$$

а так как x – точка непрерывности F , то в силу произвольности ε имеем равенство

$$\liminf_n F_n(x) = \limsup_n F_n(x) = \lim_n F_n(x) = F(x).$$

Замечание 13.2. В большинстве монографий по теории вероятностей слабая сходимость распределений определяется соотношением (1) – именно таким образом можно распространить понятие слабой сходимости на векторные случайные величины (или случайные величины с абстрактным пространством их значений). Слабая сходимость распределений обозначается тем же символом $P_n \Rightarrow P$.

Теперь мы имеем все необходимое, чтобы установить критерий слабой сходимости.

Доказательство теоремы непрерывности 13.1. Если $F_n \Rightarrow F$, то

$$\varphi_n(t) = \int_{-\infty}^{\infty} e^{itx} dF_n(x) \rightarrow \int_{-\infty}^{\infty} e^{itx} dF(x) = \varphi(t)$$

(достаточно применить лемму 13.2 к ограниченной непрерывной функции $g(x) = e^{itx}$).

Пусть теперь последовательность характеристических функций $\{\varphi_n, n \geq 1\}$ сходится к некоторой непрерывной в точке $t = 0$ функции $\varphi(t)$, и $\{F_n, n \geq 1\}$ – соответствующая последовательность функций распределения. Требуется доказать, что φ – характеристическая функция случайной величины с функцией распределения F и $F_n \Rightarrow F$.

В силу леммы 13.1 из последовательности $\{F_n, n \geq 1\}$ можно выбрать подпоследовательность $\{F_{n_k}, k \geq 1\}$, слабо сходящуюся к некоторой неубывающей, непрерывной слева функции F , причем $0 \leq F(x) \leq 1$. Если $\text{var}F = 1$, то есть F – функция распределения, то (см.(1)) $\varphi_{n_k}(t) \rightarrow \varphi_0(t)$, $k \rightarrow \infty$, при любом $t \in \mathbb{R}$, где $\varphi_0(\cdot)$ – характеристическая функция, соответствующая функции распределения $F(\cdot)$. Так как последовательность $\{\varphi_n(t), n \geq 1\}$ сходится, то все ее подпоследовательности имеют один и тот же предел $\varphi(t)$, откуда $\varphi_0(t) = \varphi(t)$ и $\varphi(t)$, $t \in \mathbb{R}$ – характеристическая функция. Наконец, в силу теоремы единственности 12.1 все подпоследовательности последовательности $\{F_n, n \geq 1\}$ имеют один и тот же слабый предел F , характеристическая функция которого есть φ , откуда $F_n \Rightarrow F$.

Итак, осталось показать, что $\text{var}F = 1$.

Допустим противное $\text{var}F = \delta < 1$. Так как $\varphi(\cdot)$ непрерывна в точке $t = 0$ и $\varphi(0) = 1$, ибо $\varphi_n(0) = 1$ при любом $n = 1, 2, \dots$, то для любого $\varepsilon \in (0; 1 - \delta)$ существует отрезок $[-\tau, \tau]$, на котором

$$|1 - \varphi(t)| < \varepsilon/2 = \varepsilon - \varepsilon/2 < 1 - \delta - \varepsilon/2.$$

Функция $\varphi(\cdot)$ интегрируема на любом отрезке $[-\tau, \tau]$, так как она есть предел интегрируемых на $[-\tau, \tau]$ и ограниченных функций $\varphi_n(\cdot)$ (см. начало доказательства формулы обращения). Следовательно, (напомним, $|a| - |b| \leq |a - b|$)

$$1 - \left| \frac{1}{2\tau} \int_{-\tau}^{\tau} \varphi(t) dt \right| \leq \frac{1}{2\tau} \left| \int_{-\tau}^{\tau} (1 - \varphi(t)) dt \right| \leq$$

$$\frac{1}{2\tau} \int_{-\tau}^{\tau} |1 - \varphi(t)| dt < 1 - \delta - \varepsilon/2,$$

откуда,

$$\left| \frac{1}{2\tau} \int_{-\tau}^{\tau} \varphi(t) dt \right| > \delta + \varepsilon/2. \quad (3)$$

Неравенство (3) получено нами только из предположения непрерывности функции $\varphi(\cdot)$ в точке $t = 0$. Покажем теперь, что из сделанного нами предположения $\text{var}F = \delta < 1$, вытекает неравенство, противоположное (3). Пусть $F_{n_k} \Rightarrow F$, а соответствующая последовательность характеристических функций $\varphi_{n_k}(t) \rightarrow \varphi(t)$ при $\forall t \in \mathbb{R}$. Имеем

$$\begin{aligned} \left| \int_{-\tau}^{\tau} \varphi_{n_k}(t) dt \right| &= \left| \int_{-\tau}^{\tau} \int_{-\infty}^{\infty} e^{itx} dF_{n_k}(x) dt \right| \leq \int_{-\infty}^{\infty} \left| \int_{-\tau}^{\tau} e^{itx} dt \right| dF_{n_k}(x) = \\ &= \int_{|x| > A} \left| \int_{-\tau}^{\tau} e^{itx} dt \right| dF_{n_k}(x) + \int_{|x| \leq A} \left| \int_{-\tau}^{\tau} e^{itx} dt \right| dF_{n_k}(x), \end{aligned}$$

где A – некоторое положительное число. Так как

$$F(A) - F(-A) \leq \text{var}F \leq \delta,$$

то

$$F_{n_k}(A) - F_{n_k}(-A) < \delta + \varepsilon/4,$$

начиная с некоторого k . Учитывая, что интеграл

$$\int_{-\tau}^{\tau} e^{itx} dt = \frac{e^{i\tau x} - e^{-i\tau x}}{ix} = \frac{2 \sin(\tau x)}{x}$$

и, следовательно, по модулю не превосходит 2τ (напомним, $|\sin x| \leq |x|$), получаем

$$\begin{aligned} \int_{|x| \leq A} \left| \int_{-\tau}^{\tau} e^{itx} dt \right| dF_{n_k}(x) &\leq 2\tau(\delta + \varepsilon/4), \\ \int_{|x| > A} \left| \int_{-\tau}^{\tau} e^{itx} dt \right| dF_{n_k}(x) &= 2 \int_{|x| > A} \left| \frac{\sin(\tau x)}{x} \right| dF_{n_k}(x) \leq \end{aligned}$$

$$\int_{|x|>A} \frac{2}{|x|} dF_{n_k}(x) \leq \frac{2}{A}.$$

Если выбрать $A = 4/\tau\varepsilon$, то

$$\frac{1}{2\tau} \left| \int_{-\tau}^{\tau} \varphi_{n_k}(t) dt \right| \leq \delta + \varepsilon/2,$$

что противоречит (3) при $k \rightarrow \infty$ и, следовательно, предположению $\delta = \text{var}F < 1$. Итак, F – функция распределения, φ – соответствующая ей характеристическая функция и $F_n \Rightarrow F$.

§14. Предельные теоремы теории вероятностей

Лекция 22

Характеристические функции являются весьма мощным инструментом для построения вероятностных моделей и позволяют без особых технических сложностей получить известные нам законы теории вероятностей, значительно расширяя их область действия. Сейчас мы получим более сильный, чем П.Л. Чебышева, закон больших чисел и обобщим предельную теорему Муавра–Лапласа на суммы независимых случайных величин с произвольным общим законом распределения.

Теорема 14.1 (закон больших чисел Хинчина). Пусть $\{X_n, n \geq 1\}$ – последовательность независимых одинаково распределенных случайных величин с конечным математическим ожиданием $\mu = \mathbf{E}X_1$. Тогда

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{P} \mu.$$

Доказательство. В силу конечности первого момента характеристическая функция каждого слагаемого допускает асимптотическое представление (предложение 12.1, п. 5⁰) $\varphi_{X_k}(t) = 1 + it\mu + o(t)$. Из свойств 2⁰ и 3⁰ предложения 12.1 следует, что

$$\varphi_{\bar{X}_n}(t) = \left(1 + it\frac{\mu}{n} + o\left(\frac{t}{n}\right) \right)^n.$$

Очевидно, $\varphi_{\bar{X}_n}(t) \rightarrow \varphi(t) = e^{it\mu}$, функция $\varphi(\cdot)$ непрерывна в точке $t = 0$ и соответствует характеристической функции константы μ – случайной величине, принимающей значение μ с вероятностью единица.

Таким образом, в силу теоремы непрерывности 13.1 $\bar{X}_n \Rightarrow \mu$, а поскольку слабая сходимость к постоянной влечет сходимость по вероятности (предложение 11.2), то $\bar{X}_n \xrightarrow{P} \mu$.

Наиболее сильный результат в законах больших чисел принадлежит А.Н.Колмогорову, который доказал, что при существовании математического ожидания $\bar{X}_n \xrightarrow{\text{п.н.}} \mu$. Конечно, нам, как всегда, не хватает времени доказать что-нибудь стоящее, и если в §11 я вводил понятие сходимости почти наверное, то это делалось только для того, чтобы сейчас хотя бы упомянуть об усиленном законе больших чисел А.Н.Колмогорова.

А что будет, если отказаться от условия конечности или существования среднего значения $\mathbf{E}X_1$? Следующий пример показывает, что сходимости к постоянной величине не будет.

Пример (нарушения закона больших чисел.) Пусть X_1, \dots, X_n независимы и одинаково распределены по закону Коши $C(0, 1)$. Характеристическая функция стандартного ($a = 0, b = 1$) распределения Коши $\varphi(t) = \exp\{-|t|\}$, характеристическая функция суммы $S_n = \sum_1^n X_k$ равна $\varphi^n(t) = \exp\{-n|t|\}$, наконец, характеристическая функция нормированной суммы $\bar{X}_n = S_n/n$ равна $\varphi^n(t/n) = \exp\{-|t|\}$, и мы снова получили то же самое стандартное распределение Коши! Конечно, внутри каждого из нас теплилась надежда, что \bar{X}_n будет сходиться при $n \rightarrow \infty$ к моде распределения Коши $\text{mod}(X_1) = 0$, но, увы, законы природы (математики) неумолимы и, вычисляя арифметическое среднее любого количества реализаций случайных величин с распределением Коши, мы также будем (в среднем) далеки от моды, как и на первом шаге нашего статистического эксперимента.

Изучим теперь более подробно асимптотическое ($n \rightarrow \infty$) распределение $\sum_1^n X_k$.

Теорема 14.2 (центральная предельная теорема). Пусть $\{X_n, n \geq 1\}$ – последовательность независимых, одинаково распределенных случайных величин с конечными математическим ожиданием $\mathbf{E}X_1 = \mu$ и дисперсией $\mathbf{D}X_1 = \sigma^2$. Тогда при любом $x \in \mathbf{R}$

$$\lim_{n \rightarrow \infty} P \left(\frac{\sum_1^n X_k - n\mu}{\sigma\sqrt{n}} < x \right) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt.$$

Доказательство. Рассуждения те же, что и при выводе закона больших чисел, но используется существование двух моментов у X_k . В силу п. 5⁰ предложения 12.1 характеристическая функция нормированной случайной величины $Y_k = (X_k - \mu)/\sigma$, у которой $\mathbf{E}Y_k = 0$ и $\mathbf{D}Y_k = 1$, допускает асимптотическое ($t \rightarrow 0$) представление

$$\varphi_{Y_k}(t) = \left(1 - \frac{t^2}{2} + o(t^2) \right).$$

Теперь, в силу п. 2⁰ предложения 12.1, характеристическая функция нор-

мированной суммы

$$\bar{S}_n = \frac{1}{\sqrt{n}} \sum_1^n Y_k = \frac{\sum_1^n X_k - n\mu}{\sigma\sqrt{n}},$$

имеет асимптотику

$$\varphi_{\bar{S}_n}(t) = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

Очевидно,

$$\varphi_{\bar{S}_n}(t) \rightarrow e^{-t^2/2},$$

если $n \rightarrow \infty$, а это, как нам известно из §12, есть характеристическая функция стандартного нормального распределения $\mathcal{N}(0, 1)$. Поскольку предельная нормальная функция $\Phi(x)$ непрерывна на всем \mathbb{R} , то функция распределения \bar{S}_n сходится к $\Phi(x)$ при любом $x \in \mathbb{R}$.

Существуют обширнейшие исследования по распределениям сумм случайных величин, в которых центральная предельная теорема обобщается на случай “слабо зависимых” или разно распределенных, но обладающих одинаковым порядком малости, случайных величин; рассматриваются суммы случайных векторов и суммы случайных элементов, принимающих значения в абстрактных пространствах, и т.д., и т.п., так что не перестаешь удивляться, как это можно что-то еще сделать в области того, где, кажется, всё уже сделано. Мы не будем углубляться в эту обширнейшую тематику и займемся более прикладными вопросами – продолжим построение вероятностных моделей, математической основой которых служат предельные теоремы теории вероятностей.

Вероятностные модели роста. Условимся употреблять терминологию, связанную с биологическими исследованиями; о приложениях к другим областям естествознания поговорим ниже, после вывода основного уравнения модели.

Предположим, что мы посадили с вами маленькое деревце (саженец) высоты x_0 , и во все последующие годы производим замеры x_1, x_2, \dots высоты растущего дерева. Нас интересуют прогноз высоты дерева по истечении n лет. Естественно, на ежегодный прирост высоты действует огромное количество природных факторов: температура, осадки, солнечное освещение, плодородие почвы и т.п., поэтому мы, очевидно, имеем дело со стохастическим прогнозом, который формулируется, примерно, как следующее заключение: “Через 60 лет с вероятностью 0,9 высота дерева будет не меньше

15 метров.” Конечно, такой прогноз, как и в случае однократного подбрасывания монеты, нельзя применить к одному посаженному дереву, но его можно использовать в прогнозе “зрелости” лесной посадки, состоящей из большого числа деревьев, и тогда наше заключение будет относиться приблизительно к 90% саженцев.

Итак, мы должны трактовать замеры x_1, x_2, \dots в терминах реализаций компонент последовательности случайных величин X_1, X_2, \dots и попытаться формализовать в математических терминах причину “разброса” в значениях ежегодных приращений $\Delta_k = X_k - X_{k-1}$ высоты дерева. Естественно предположить, что прирост Δ_k вызван суммарным действием всех тех причин роста, о которых мы говорили выше, то есть действием некоторого неотрицательного “импульса” $\xi_k (\geq 0)$. Между Δ_k и ξ_k существует приближенная линейная связь $\Delta_k = \alpha_k \xi_k$, где α_k зависит от высоты X_{k-1} дерева, которой оно достигло по истечении k лет. Положим $\alpha_k = g(X_{k-1})$ с естественным условием неотрицательности и непрерывности функции $g(\cdot)$. Таким образом, мы приходим к рекуррентным соотношениям, которые описывают ежегодный прирост высоты дерева,

$$X_k - X_{k-1} = \xi_k g(X_{k-1}), \quad k = 1, 2, \dots \quad (1)$$

Нам осталось только сделать некоторые предположения, касающиеся распределения случайных величин $\xi_k, k = 1, 2, \dots$. Будем считать, что эти случайные величины неотрицательны, независимы, одинаково распределены и обладают конечными моментами второго порядка: средним значением $a = \mathbf{E}\xi_k$ и дисперсией $b^2 = \mathbf{D}\xi_k$.

Напомним, что мы интересуемся распределением случайной величины X_n , реализация x_n которой указывает размер конкретного дерева по истечении n лет. Перепишем первые n рекуррентных соотношений (1) в виде

$$\xi_k = \frac{X_k - X_{k-1}}{g(X_{k-1})}, \quad k = 1, \dots, n$$

и просуммируем левые и правые части этих равенств. В результате получим

$$\sum_1^n \xi_k = \sum_1^n \frac{X_k - X_{k-1}}{g(X_{k-1})}.$$

Если каждый импульс вызывает незначительный прирост дерева, то есть все $\Delta_k = X_k - X_{k-1}$ малы, то, трактуя правую часть последнего ра-

венства как интегральную сумму, получаем приближенное равенство

$$\sum_1^n \xi_k = \int_{x_0}^X \frac{dt}{g(t)}, \quad (2)$$

где $X = X_n$ – окончательный размер дерева.

Так как функция $g(x)$ положительна, то интеграл в правой части (2) представляет собой некоторую монотонно возрастающую функцию $h(X)$. Применение центральной предельной теоремы 14.2 к левой части (2) приводит к утверждению: по истечении достаточно большого срока после посадки дерева ($n \gg 1$) распределение его высоты X определяется соотношением $h(X) \sim \mathcal{N}(\mu, \sigma^2)$, где $\mu = na$, $\sigma^2 = nb^2$. В силу монотонности функции $h(\cdot)$

$$F(x) = P(X < x) = P(h(X) < h(x)) = \Phi\left(\frac{h(x) - \mu}{\sigma}\right).$$

Осталось решить проблему с выбором функции $g(\cdot)$. Если постулировать, что прирост высоты дерева пропорционален достигнутой высоте, то есть положить $g(t) = t$, а именно такое предположение наиболее часто используется в моделях роста, то мы придем к следующему распределению случайной величины X .

Логарифмически-нормальное распределение $\mathcal{LN}(\mu, \sigma^2)$.

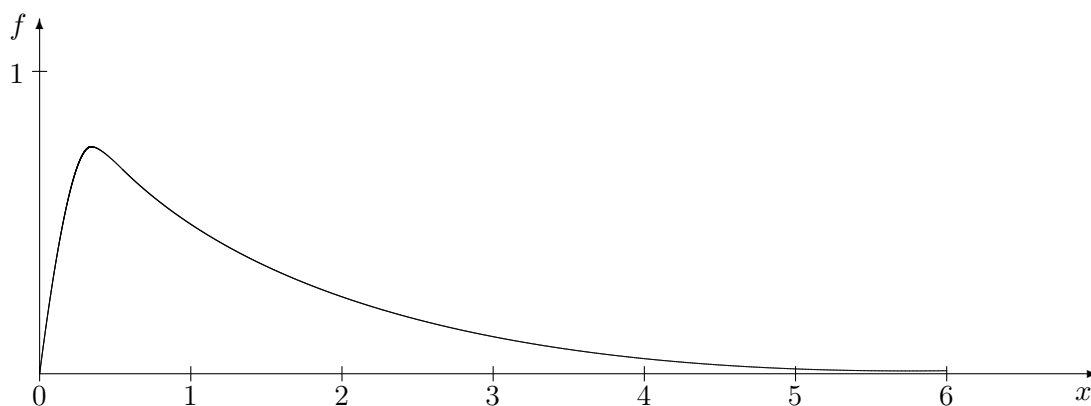
При $g(t) = t$ интеграл в правой части (2) с точностью до постоянного слагаемого $-\ln x_0$ равен $\ln X$, так что $\ln X \sim \mathcal{N}(\mu, \sigma^2)$, и функция распределения X

$$F(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \quad x > 0;$$

функция плотности

$$f(x | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi} x} \exp\left\{-\frac{(\ln x - \mu)^2}{2\sigma^2}\right\}.$$

Это унимодальное, резко асимметричное ($\gamma_1 > 0$) распределение, график плотности которого имеет следующий вид:



Конечно, в практических приложениях целесообразнее оперировать не с X , а с его натуральным логарифмом $Y = \ln X$, после чего производить все расчеты, используя модель нормального распределения.

Логарифмически-нормальный закон носит достаточно универсальный характер. Этому распределению подчиняется размер трещины в испытуемом образце материала, который подвергается циклическим нагрузениям “на изгиб” – вы можете сами без особых фантазийных усилий пересказать наши построения с высотой дерева в терминах размера трещины. Аналогичные рассуждения могут быть также применены в изучении роста доходов у отдельных лиц достаточно однородной человеческой популяции. Проводимые в этом направлении статистические исследования указывают на хорошее согласие с логарифмически-нормальным распределением достаточно низких доходов, в то время как для умеренных и высоких доходов более подходящим является распределение Парето.

В рамках построенной нами модели роста часто возникает задача, которую можно трактовать как некоторую альтернативу к проблеме вывода распределения размера, достигнутого к определенному сроку “растущим” объектом исследования. Пусть фиксирован некоторый уровень x размера (дерева, трещины, дохода) и нас интересует распределение момента времени (номера цикла), на котором этот размер будет достигнут. Удивительно, что в рамках нашей модели это распределение не зависит от выбора положительной функции $g(\cdot)$, и получить его можно путем следующих тривиальных рассуждений.

Распределение Бирнбаума–Саундерса $BS(\lambda, \theta)$. Пусть τ – случайная величина, реализующая момент достижения заданного размера x . Тогда событие $\tau > n$ эквивалентно событию $X_n < x$ (напомним, все $\xi_k \geq 0$)

– к моменту времени n высота дерева еще не достигла уровня x . Итак,

$$P(\tau > n) = P(X_n < x) = P(h(X_n) < h(x)) = \Phi\left(\frac{h(x) - na}{b\sqrt{n}}\right). \quad (3)$$

Заменим теперь n на “непрерывную” переменную t и введем новые параметры λ и θ , определив их уравнениями $\lambda\sqrt{\theta} = h(x)/b$, $\lambda/\sqrt{\theta} = a/b$. Цепочка равенств (3) позволяет нам записать распределение случайного момента времени τ , в который дерево (трещина, доход,) достигнет заданного уровня x :

$$F(t) = P(\tau < t) = 1 - \Phi\left(\lambda\left(\sqrt{\frac{\theta}{t}} - \sqrt{\frac{t}{\theta}}\right)\right), \quad t > 0.$$

Это унимодальное распределение, которое называется *распределением Бирнбаума–Саундерса*, и мы будем обозначать его $BS(\lambda, \theta)$. График плотности BS -распределения (я, надеюсь, вы достаточно образованы в области математического анализа, чтобы найти производную от $F(t)$) очень похож на функцию плотности гамма-распределения. BS -распределение играет большую роль при расчетах надежности объектов, долговечность которых определяется развитием трещин, приводящих к гибели объекта.

Рассмотрим еще одно распределение, часто используемое в практических расчетах надежности сложных систем.

Распределение Вейбулла $W(\lambda, \theta)$ (модель слабого звена). Имеется цепь, состоящая из большого числа n звеньев. Допустим, что прочностные x_1, \dots, x_n отдельных звеньев можно трактовать как реализации n независимых, одинаково распределенных случайных величин X_1, \dots, X_n . На оба конца цепи подается равномерно возрастающая нагрузка, и фиксируется напряжение, при котором происходит разрыв цепи. Очевидно это напряжение равно прочностной наислабейшего звена цепи, поэтому его можно трактовать как реализацию случайной величины

$$X = \min_{1 \leq k \leq n} X_k.$$

Если $F(x)$ – функция распределения каждого X_k , $k = 1, \dots, n$, то функция распределения X определяется посредством следующих расчетов, в которых существенно используется независимость X_1, \dots, X_n :

$$G_n(x) = P(X < x) = 1 - P(X \geq x) = 1 - P(X_1 \geq x, \dots, X_n \geq x) =$$

$$1 - \prod_{k=1}^n P(X_k \geq x) = 1 - (1 - F(x))^n.$$

При больших n естественно вместо $G_n(x)$ использовать ее асимптотику. Однако при каждом фиксированном $x (> 0)$ вероятность $G_n(x) \rightarrow 1$, если $n \rightarrow \infty$, и поэтому мы должны провести нормировку X по аналогии с тем, как это делалось в центральной предельной теореме, чтобы распределение не вырождалось, когда $n \rightarrow \infty$. Понятно также, что X по вероятности сходится к нулю, поэтому нормировку X следует производить домножением на некоторую растущую функцию от n . При этом нам не избежать условий на поведение функции распределения $F(x)$ при $x \rightarrow 0+$ – допустим, что $F(x) \sim ax^\lambda$, где a и λ – неотрицательные числа (удивительно, но все изученные нами распределения, сосредоточенные на положительной полуоси, удовлетворяют этому условию).

Функция распределения $W(x)$ нормированной случайной величины $Y = n^{1/\lambda}X$ не вырождается с ростом n , и предельное распределение находится с помощью следующих выкладок:

$$\begin{aligned} W(x) &= P(Y < x) = P\left(X < \frac{x}{n^{1/\lambda}}\right) = 1 - \left(1 - F\left(\frac{x}{n^{1/\lambda}}\right)\right)^n \sim \\ &1 - \left(1 - a\left(\frac{x}{n^{1/\lambda}}\right)^\lambda\right)^n = 1 - \left(1 - \frac{ax^\lambda}{n}\right)^n \sim 1 - e^{-ax^\lambda}. \end{aligned}$$

Заменяя параметр a на параметр θ , определяемый уравнением $a = \theta^{-\lambda}$, получаем *распределение Вейбулла* $W(\lambda, \theta)$ с функцией распределения

$$W(x | \lambda, \theta) = 1 - \exp\left\{-\left(\frac{x}{\theta}\right)^\lambda\right\}, \quad x > 0, \quad \lambda, \theta > 0.$$

Это также унимодальное распределение, график функции плотности которого “на глаз” не отличим от графика функции плотности гамма-распределения. Вейбулловскому распределению обычно следуют долговечности систем, состоящих из большого числа однотипных элементов (например, плата компьютера), отказ одного из которых (наислабейшего) приводит к отказу системы.

§15. Случайные процессы

Лекция 23

До сих пор мы изучали распределение конечного числа случайных величин $X_1(\omega), \dots, X_n(\omega)$, заданных на едином вероятностном пространстве (Ω, \mathcal{A}, P) , и вывод их совместного распределения сводился, по существу, к построению распределения вероятностей на произведении измеримых пространств значений этих величин (произведении борелевских прямых). Теперь мы приступаем к изучению распределений на бесконечном (возможно несчетном) произведении измеримых пространств. Допустим, что на пространстве элементарных исходов Ω задано семейство случайных величин $\{X_t, t \in T\}$, $X_t = X_t(\omega)$, индексированных параметром t , который пробегает множество значений T (например, векторная случайная величина имеет $T = \{1, \dots, n\}$). Пусть $(\mathcal{X}_t, \mathcal{B}_t)$, $t \in T$ – измеримые пространства значений X_t , соответствующие каждому $t \in T$. В дальнейшем будем рассматривать только случай $\mathcal{X}_t = \mathbb{R}$ с борелевской σ -алгеброй \mathcal{B}_t подмножеств \mathbb{R} , но для понимания конструкции распределений на бесконечномерных пространствах важно сохранить индекс t в обозначении пространств значений каждого представителя семейства $\{X_t, t \in T\}$. Это семейство называется *случайным процессом*.

Если зафиксировать некоторый элементарный исход ω_0 , то получим функцию $x(t) = X_t(\omega_0)$ на множестве T со значениями при каждом фиксированном t в \mathcal{X}_t . Эта функция называется *траекторией* или реализацией процесса $X_t, t \in T$. В связи с этим понятием следует трактовать случайный процесс как случайную функцию $X_t = X(t)$, помня при этом, что вся “случайность” состоит в зависимости $X(t)$ от $\omega \in \Omega$, в то время как траектория $x(t)$ есть “значение” случайного процесса $X(t)$ при фиксированном ω . Приведем несколько примеров случайных процессов и опишем вид их траекторий.

Пример 15.1 (точечные процессы). На телефонную станцию поступают заявки на междугородние разговоры, и при этом фиксируется время поступления заявки. В таких процессах с появлением определенных событий в случайные моменты времени обычно полагают $x(t)$ равной числу заявок, поступивших за промежуток времени $[0, t]$. Эти процессы служат хорошими математическими моделями при проектировании систем обслуживания (модели теории очередей), при анализе транспортных потоков на магистралях; они используются в ядерной физике, метеорной астрономии

и т.п. Множество T в данном случае – отрезок \mathbb{R}_+ вида $[0, T]$ с возможным бесконечным значением T . Пространство \mathcal{X}_t значений случайного процесса при любом $t \in T$ совпадает с множеством неотрицательных целых чисел. Траектория имеет вид ступенчатой функции, возрастающей скачками в случайные моменты времени, и величина каждого скачка равна единице.

Пример 15.2 (*ветвящиеся процессы*). Наблюдается некоторая биологическая популяция, состоящая из особей, способных размножаться и гибнуть. Такие данные, как число потомков в определенном колене отдельной особи, численность популяции к фиксированному моменту времени t , количество погибших и новорожденных особей и т.п., составляют особый интерес для популяционной генетики, и трудно переоценить роль вероятностных моделей в изучении динамики развития биологической популяции. Аналогичные модели используются в физике элементарных частиц, особенно при изучении ядерных реакций. Пространства T и \mathcal{X}_t те же, что и в первом примере, траектории также имеют вид ступенчатых функций, но величины скачков – произвольные целые числа.

Пример 15.3 (*броуновское движение в капилляре*). Длинный тонкий капилляр наполняется жидкостью, и в середину капилляра помещается частица, диаметр которой не намного меньше диаметра капилляра. Под действием молекул жидкости частица совершает хаотические движения, и для наблюдения за ними вводится система координат: капилляр рассматривается как действительная ось \mathbb{R} с нулем в середине капилляра. В каждый момент времени t (непрерывно) регистрируется расстояние $x(t)$ частицы от середины капилляра (естественно, $x(0) = 0$) с учетом знака (минус – слева от середины, плюс – справа). Если изобразить теперь траекторию движения частицы на плоскости в координатах $(t, x(t))$, то мы получим то, что физики называют траекторией одномерного броуновского движения. Вероятностные модели, определяющие распределения таких процессов, были предложены Винером, Эйнштейном и Смолуховским. В этом примере T – отрезок временной оси, $\mathcal{X}_t = \mathbb{R}$.

Пример 15.4 (*броуновское движение на плоскости*). В центр кювета, наполненного тонким слоем жидкости, помещается частица некоторого вещества, которая, как и в предыдущем примере совершает броуновское движение, но не на прямой \mathbb{R} , а на плоскости \mathbb{R}^2 (центр кювета служит началом декартовой системы координат (x, y)). Траектория броуновского движения представляет собой некоторую кривую на плоскости, определя-

емую параметрическими уравнениями $x = x(t)$, $y = y(t)$. Естественно, T – отрезок времени, а $\mathcal{X}_t = \mathbb{R}^2$.

Пример 15.5 (случайное поле.) Отшлифованная поверхность металла обычно подвергается проверке на “шероховатость”, для чего она помещается под микроскоп и замеряются некоторые характеристики отклонения различных точек поверхности металла от плоского уровня. Такая шероховатая поверхность $z = z(u, v)$, где (u, v) – фиксированная система декартовых координат, трактуется как реализация *случайного поля* $Z = Z(u, v)$, пространство T соответствует части плоскости $\mathbb{R}^2 = \{u, v\}$, занимаемой обрабатываемым объектом, $\mathcal{X}_t = \mathbb{R}$. Пример случайного поля, в котором кроме координат (u, v) пространство T включает временную ось \mathbb{R}_+ , – участок поверхности моря во время шторма.

Зададимся вопросом, какого рода события, связанные с рассмотренными случайными процессами $X(t)$, представляют наибольший интерес для их исследователей? В первую очередь следует обратить внимание на событие $\sup_{t \in T} X(t) \geq x_0$, а также на момент времени t , при котором процесс впервые достигнет уровня x_0 . Но для того чтобы вычислять вероятности таких событий, следует ввести понятие распределения вероятностей на измеримом пространстве траекторий процесса.

Пространство траекторий трактуется как прямое произведение

$$\mathcal{X} = \prod_{t \in T} \mathcal{X}_t$$

пространств значений процесса в каждой точке $t \in T$. Подмножества этого пространства, определяемые ограничениями вида

$$a_1 < X(t_1) < b_1, \dots, a_n < X(t_n) < b_n$$

при любом конечном n , называются *прямоугольниками*. Конечные объединения всевозможных непересекающихся прямоугольников (изменяются как значения n , так и наборы точек t_1, \dots, t_n из T) образуют, очевидно, булеву алгебру \mathcal{A} . Наименьшая σ -алгебра \mathcal{F} , содержащая \mathcal{A} , является искомой σ -алгеброй на пространстве траекторий \mathcal{X} .

Таким образом, мы имеем измеримое пространство $(\mathcal{X}, \mathcal{F})$, σ -алгебра \mathcal{F} которого порождается полуалгеброй прямоугольников, и естественно ожидать, что задание совместных функций распределения

$$F_{t_1, \dots, t_n}(x_1, \dots, x_n) = P(X(t_1) < x_1, \dots, X(t_n) < x_n)$$

случайных величин $X(t_1), \dots, X(t_n)$ при любых $n = 1, 2, \dots$ и любых наборах t_1, \dots, t_n однозначно определяет вероятность на σ -алгебре \mathcal{F} . То, что это действительно так, устанавливает знаменитая теорема А.Н. Колмогорова, положившая начало строгой математической теории случайных процессов. Заметим только, что в этой теореме накладывается естественное условие *согласованности* функций распределения: маргинальные функции распределения, соответствующие части $T_k = (t_{i_1}, \dots, t_{i_k})$, $k < n$, набора индексов t_1, \dots, t_n , должны совпадать с теми, что были построены для набора T_k . Впрочем, это условие соблюдается “автоматически,” поскольку построение функций распределения производится при произвольных значениях ее аргументов.

Следующие два примера, играющие важную роль в практических приложениях теории случайных процессов, иллюстрируют общую методологию и технические приемы, используемые при построении вероятностных моделей случайных процессов.

Пуассоновский процесс

На временной оси $T = \mathbb{R}_+$ в случайные моменты времени появляются некоторые события (см. пример 15.1), и наблюдается траектория $x(t)$ точечного случайного процесса $X(t)$, регистрирующая число событий, появившихся к моменту времени t . Следующие три постулата выделяют пуассоновский процесс из класса всевозможных точечных процессов.

(P1) *Стационарность*. Распределение числа событий, появившихся во временном промежутке $[t_1, t_2]$, зависит только от длины $t_2 - t_1$ этого промежутка, то есть

$$P(X(t_2) - X(t_1) = x) = p_x(t_2 - t_1).$$

(P2) *Независимость приращений*. Для любого упорядоченного набора моментов времени $0 = t_0 < t_1 < \dots < t_n$ случайные величины

$$X(t_k) - X(t_{k-1}), \quad k = 1, \dots, n,$$

где $X(t_0) = X(0) = 0$, независимы в совокупности.

(P3) *Ординарность или разреженность*. Вероятность

$$p_x(\Delta t) = P(X(t + \Delta t) - X(t) = x)$$

того, что за промежуток времени Δt произойдет ровно x ($= 0, 1, \dots$) событий допускает при $\Delta t \rightarrow 0$ асимптотическое представление

$$p_0(\Delta t) = 1 - \lambda \Delta t + o(\Delta t), \quad p_1(\Delta t) = \lambda \Delta t + o(\Delta t); \quad p_x(\Delta t) = o(\Delta t), \quad x \geq 2, .$$

В этом представлении $\lambda > 0$ – числовой параметр, называемый обычно *интенсивностью* пуассоновского потока событий (см. в связи с этим модель пуассоновского распределения в §5).

Используя постулаты (P1)–(P3), построим конечномерные распределения

$$f_{t_1, \dots, t_n}(x_1, \dots, x_n) = P(X(t_1) = x_1, \dots, X(t_n) = x_n)$$

пуассоновского процесса. Эти построения значительно облегчает

Лемма 15.1. *Функция $p_x(t) = P(X(t) = x)$, $t \geq 0$, $x = 0, 1, \dots$, однозначно определяет все конечномерные распределения пуассоновского процесса.*

Доказательство. Следующая цепочка равенств, в которой сначала используется постулат (P2), а потом – (P1), устанавливает соотношение между конечномерной плотностью процесса f_{t_1, \dots, t_n} и функцией $p_x(t)$:

$$\begin{aligned} f_{t_1, \dots, t_n}(x_1, \dots, x_n) &= P(X(t_1) = x_1, \\ X(t_2) - X(t_1) &= x_2 - x_1, \dots, X(t_n) - X(t_{n-1}) = x_n - x_{n-1}) = \\ &= \prod_{k=1}^n P(X(t_k) - X(t_{k-1}) = x_k - x_{k-1}) = \\ &= \prod_{k=1}^n P(X(t_k - t_{k-1}) = x_k - x_{k-1}) = \\ &= \prod_{k=1}^n p_{x_k - x_{k-1}}(t_k - t_{k-1}). \end{aligned}$$

Естественно, все эти выкладки имеют смысл лишь при

$$0 < t_1 < t_2 < \dots < t_n, \quad 0 \leq x_1 \leq \dots \leq x_n.$$

Вид функции $p_x(t)$, а вместе с ним и конечномерные распределения процесса Пуассона, устанавливает

Теорема 15.1. *Если справедливы постулаты (P1)–(P3), то*

$$p_x(t) = P(X(t) = x) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \quad t \geq 0, \quad x = 0, 1, \dots \quad (1)$$

Доказательство. Покажем сначала, что (1) выполняется в случае $x = 0$, для чего исследуем асимптотику при $\Delta t \rightarrow 0$ функции $p_0(t + \Delta t) = P(X(t + \Delta t) = 0)$.

Событие $X(t + \Delta t) = 0$ эквивалентно одновременному осуществлению двух независимых (в силу постулата (P2)) событий: $X(t) = 0$ и $X(t + \Delta t) - X(t) = 0$. Используя постулаты (P1) и (P3), находим, что

$$\begin{aligned} p_0(t + \Delta t) &= P(X(t) = 0) \cdot P(X(t + \Delta t) - X(t) = 0) = \\ &= p_0(t) \cdot p_0(\Delta t) = p_0(t)(1 - \lambda\Delta t + o(\Delta t)). \end{aligned}$$

Если полученное асимптотическое представление записать в виде

$$\frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + o(1)$$

и устремить Δt к нулю, то получим дифференциальное уравнение

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t)$$

с очевидным начальным условием $p_0(0) = 1$. Это уравнение с разделяющимися переменными, решение которого с учетом начальных условий

$$p_0(t) = e^{-\lambda t},$$

что совпадает с (1) при $x = 0$.

Проведем аналогичные построения для произвольного целого $x \geq 1$, для чего представим событие $X(t + \Delta t) = x$ в виде объединения $x + 1$ несовместных событий

$$\{X(t) = x - k\} \cap \{X(t + \Delta t) - X(t) = k\}, \quad k = 0, 1, \dots, x.$$

Используя, как и выше, постулаты (P1)–(P3), получаем

$$\begin{aligned} p_x(t + \Delta t) &= P(X(t + \Delta t) = x) = \\ &= \sum_{k=0}^x P(X(t) = x - k, X(t + \Delta t) - X(t) = k) = \\ &= \sum_{k=0}^x P(X(t) = x - k) \cdot P(X(t + \Delta t) - X(t) = k) = \end{aligned}$$

$$\sum_{k=0}^x p_{x-k}(t) \cdot p_k(\Delta t) = p_x(t)(1 - \lambda\Delta t) + p_{x-1}(t)\lambda\Delta t + o(\Delta t).$$

Если представить полученное соотношение в виде

$$\frac{p_x(t + \Delta t) - p_x(t)}{\Delta t} = -\lambda(p_x(t) - p_{x-1}(t)) + o(1)$$

и устремить Δt к нулю, то получим рекуррентную систему дифференциальных уравнений с начальными условиями:

$$\frac{dp_x(t)}{dt} = -\lambda(p_x(t) - p_{x-1}(t)), \quad p_x(0) = 0, \quad x = 1, 2, \dots$$

Поскольку выше мы определили $p_0(t) = e^{-\lambda t}$, то для $p_1(t)$ имеем линейное дифференциальное уравнение с постоянными коэффициентами

$$\frac{dp_1(t)}{dt} = -\lambda(p_1(t) - e^{-\lambda t}), \quad p_1(0) = 0,$$

решение которого стандартными методами дает

$$p_1(t) = \lambda t e^{-\lambda t},$$

что опять совпадает с (1) при $x = 1$.

Дальнейшее построение модели осуществляется по индукции. Предполагается, что (1) справедливо для некоторого $x \geq 2$, и решается линейное дифференциальное уравнение

$$\frac{dp_{x+1}(t)}{dt} = -\lambda \left(p_{x+1}(t) - \frac{(\lambda t)^x e^{-\lambda t}}{x!} \right), \quad p_{x+1}(0) = 0.$$

Нетрудно убедиться, что решение этого уравнения с учетом начального условия определяется формулой (1) с заменой x на $x + 1$. Таким образом, построение вероятностной модели пуассоновского процесса завершено.

Интересно заметить, что формула (1) при $t = 1$ дает функцию плотности распределения Пуассона $P(\lambda)$, так что (1) можно трактовать как обобщение теоремы сложения для распределения Пуассона на случай “дробного” числа слагаемых, по существу же происходит простое суммирование числа событий по всем t единицам времени.

Изучим некоторые свойства процесса Пуассона, которые вскрывают интересные связи распределения Пуассона $P(\lambda)$ с показательным, равномерным и гамма-распределениями. Начнем с выяснения вида распределения промежутков времени между появлениями событий в процессе Пуассона.

Предложение 15.1. *Случайные величины τ_1, \dots, τ_n , реализации которых указывают промежутки времени между появлениями событий в процессе Пуассона, независимы и одинаково распределены по показательному закону $E(\lambda^{-1})$.*

Доказательство. Требуется показать, что совместная функция плотности случайных величин τ_1, \dots, τ_n

$$f_n(t_1, \dots, t_n) = \lambda^n \exp \left\{ -\lambda \sum_1^n t_k \right\}, \quad (2)$$

в области $t_{[1]} = \min\{t_1, \dots, t_n\} > 0$.

Выберем $\Delta t < t_{[1]}$ и подсчитаем вероятность того, что в каждом из промежутков $[T_k, T_k + \Delta t)$, где $T_k = t_1 + \dots + t_k$, $k = 1, \dots, n$, произошло только по одному событию, в то время как в промежутках $[0, t_1)$ и $[T_k + \Delta t, T_{k+1})$, $k = 1, \dots, n-1$, событий не было. Очевидно, при $\Delta t \rightarrow 0$ асимптотика этой вероятности должна иметь вид $f_n(t_1, \dots, t_n)(\Delta t)^n$, и это обстоятельство позволит нам получить искомую функцию плотности f_n .

В силу постулата (P2) независимости приращений все из рассматриваемых $2n$ событий о появлении по одному или полному отсутствию инцидентов в указанных временных промежутках являются независимыми; вероятность появления ровно одного события в каждом из промежутков $[T_k, T_k + \Delta t)$, $k = 1, \dots, n$, равна (постулат (P1)) $p_1(\Delta t)$, а вероятности отсутствия событий в промежутках $[0, t_1)$ и $[T_k + \Delta t, T_{k+1})$ равны соответственно $p_0(t_1)$ и $p_0(t_{k+1} - \Delta t)$, $k = 1, \dots, n-1$. Таким образом, вероятность совместного осуществления всех $2n$ событий в терминах функции $p_x(t)$ равна

$$p_0(t_1)p_1^n(\Delta t) \prod_1^{n-1} p_0(t_{k+1} - \Delta t). \quad (3)$$

Если $\Delta t \rightarrow 0$, то применение формулы (1) дает

$$p_0(t_1) = \exp\{-\lambda t_1\}, \quad p_1^n(\Delta t) = \lambda^n \exp\{-n\lambda\Delta t\}(\Delta t)^n \sim \lambda^n(\Delta t)^n,$$

$$p_0(t_{k+1} - \Delta t) = \exp\{-\lambda t_{k+1} + \lambda \Delta t\} \sim \exp\{-\lambda t_{k+1}\}.$$

Подставляя полученные асимптотики в (3), получаем с точностью до множителя $(\Delta t)^n$ правую часть (2).

Доказанное предложение позволяет нам достаточно просто установить распределение случайной величины τ , реализация которой соответствует моменту первого достижения пуассоновским процессом заданного уровня h .

Следствие 15.1. *Случайная величина τ имеет гамма-распределение $G(m, \lambda^{-1})$, где параметр формы m принимает целочисленное значение, равное h , если h целое, и равное $[h] + 1$, если h дробное.*

Доказательство. Немедленно вытекает из результата предложения 15.1, поскольку

$$\tau = \sum_{k=1}^m \tau_k,$$

где τ_1, \dots, τ_m независимы и одинаково распределены в соответствии с показательным распределением $E(\lambda^{-1})$ (напомним, что именно таким образом вводилось гамма-распределение в §12).

Установленная связь гамма-распределения с пуассоновским потоком событий открывает новую область приложений этого распределения. Это – *вероятностные модели износа и старения*. Простейший пример построения такой модели дает исследование процесса износа протектора автомобильной шины. Резонно считать, что различного рода препятствия, возникающие на пути движения автомобиля и приводящие к резкому торможению, реализуют пуассоновский поток событий. Каждое резкое торможение приводит к уменьшению глубины r протектора на определенную (предположим, для простоты, – одинаковую) величину Δr . В таком случае “облысение” шин наступит после m торможений, где m в соответствии со следствием 15.1 определяется уровнем $h = r/\Delta r$.

Еще одно замечательное свойство пуассоновского процесса, характеризующее особого рода случайность в потоке событий, состоит в следующей специфике условного распределения моментов появления фиксированного числа n событий на фиксированном промежутке времени $[0, T]$. Точная формулировка этого свойства осуществляется в терминах специального случайного вектора, играющего важную роль в математической статистике.

Пусть X_1, \dots, X_n – случайный вектор, заданный на измеримом пространстве (Ω, \mathcal{A}) , с независимыми одинаково распределенными с плотностью $f(x)$ по мере Лебега компонентами. Вектор $X_{(1)}, \dots, X_{(n)}$, полученный из исходного вектора упорядочиванием его компонент при каждом фиксированном $\omega \in \Omega$, называется *вариационным рядом*. Таким образом, при каждом фиксированном $\omega \in \Omega$ компоненты вариационного ряда удовлетворяют неравенствам $X_{(1)}(\omega) \leq \dots \leq X_{(n)}(\omega)$, и если $x_1 = X_1(\omega), \dots, x_n = X_n(\omega)$, то $x_{(1)} = \min\{x_1, \dots, x_n\}$, $x_{(2)}$ равен второму по величине значению среди x_1, \dots, x_n , $x_{(3)}$ – третьему и т.д., так что реализация (при элементарном исходе ω) последней компоненты вариационного ряда $x_{(n)} = \max\{x_1, \dots, x_n\}$.

Функция плотности исходного вектора X_1, \dots, X_n с независимыми, одинаково непрерывно распределенными компонентами равна

$$f_n(x_1, \dots, x_n) = \prod_{k=1}^n f(x_k),$$

а функция плотности вариационного ряда отлична от нуля только в области $x_1 \leq x_2 \leq \dots \leq x_n$ и равна $g_n(x_1, \dots, x_n) = n! f_n(x_1, \dots, x_n)$. Для того чтобы убедиться в этом, достаточно применить метод, который использовался при доказательстве последнего предложения.

Для каждого фиксированного ряда $x_1 < x_2 < \dots < x_n$ аргументов функции плотности $g_n(\cdot)$ и $\Delta x < \min_{1 \leq k \leq n-1} (x_{k+1} - x_k)$ вычислим вероятность события A , состоящего в том, что одна из компонент исходного вектора X_1, \dots, X_n попадет в интервал $[x_1, x_1 + \Delta x)$ (событие A_1), другая, из оставшихся $n - 1$ компонент, в интервал $[x_2, x_2 + \Delta x)$ (событие A_2) и т.д., так что последняя из оставшихся компонент должна попасть в интервал $[x_n, x_n + \Delta x)$ (событие A_n). В силу независимости компонент события A_1, \dots, A_n независимы, и поэтому

$$P(A) = P\left(\bigcap_1^n A_k\right) = \prod_1^n P(A_k).$$

Если $F(x)$ – функция распределения, соответствующая плотности $f(x)$, то вероятность

$$P(A) = \prod_{k=1}^n [F(x_k + \Delta x) - F(x_k)].$$

Таким образом, все $n!$ элементарных исходов, связанных с выбором конкретной перестановки i_1, \dots, i_n индексов $1, \dots, n$ наблюдаемого вектора

X_1, \dots, X_n , имеют одну и ту же вероятность, и поэтому искомая вероятность равна

$$n! \prod_{k=1}^n [F(x_k + \Delta x) - F(x_k)].$$

Если $\Delta x \rightarrow 0$, то последнее выражение эквивалентно

$$n! \prod_{k=1}^n f(x_k) \cdot (\Delta x)^n,$$

так что множитель перед $(\Delta x)^n$ дает функцию плотности $g_n(x_1, \dots, x_n)$ вариационного ряда.

В частности, функция плотности вариационного ряда равномерного на интервале $[0, T]$ распределения

$$g_n(x_1, \dots, x_n) = n! T^{-n}, \quad 0 \leq x_1 \leq \dots \leq x_n \leq T. \quad (4)$$

Теперь сформулируем обещанное свойство пуассоновского процесса.

Предложение 15.2. Совместное распределение моментов τ_1, \dots, τ_n появления n событий на интервале $[0, T]$ пуассоновского процесса при условии, что в этом интервале появилось ровно n событий, совпадает с распределением вариационного ряда равномерного на интервале $[0, T]$ распределения.

Доказательство. Снова используем метод асимптотического представления функции плотности. Выберем на интервале $[0, T]$ упорядоченный ряд из n точек $0 < t_1 < \dots < t_n < T$, а также выберем Δt , меньшее любого из промежутков, ограниченных точками t_1, \dots, t_n . Пусть

$$A = B_0 \bigcap_{k=1}^n (A_k B_k) -$$

событие, состоящее в том, что в каждом из интервалов

$$[t_k, t_k + \Delta t), \quad k = 1, \dots, n,$$

появится ровно по одному пуассоновскому событию (эти пуассоновские события обозначаются A_k), а в интервалах

$$[0, t_1), [t_k + \Delta t, t_{k+1}), \quad k = 1, \dots, n-1, [t_n + \Delta t, T]$$

пуассоновских событий не было (эти “подсобытия” обозначаются B_k , $k = 0, \dots, n$). Событие, состоящее в том, что на интервале $[0, T]$ появилось ровно n пуассоновских событий (условие), обозначим B . В этих обозначениях доказательство предложения состоит в выводе следующей асимптотической формулы (см. формулу (4)):

$$P(A|B) = P(A \cap B)/P(B) = n!T^{-n}.$$

Повторяя рассуждения, которые мы проводили при доказательстве предложения 15.1 при $\Delta t \rightarrow 0$, получаем

$$\begin{aligned} P(A \cap B) &= \\ (\lambda \Delta t)^n \exp \left\{ -\lambda \left[n\Delta t + t_1 + \sum_1^{n-1} (t_{k+1} - t_k - \Delta t) + T - t_n - \Delta t \right] \right\} &\sim \\ \lambda^n \exp \left\{ -\lambda \left[t_1 + \sum_1^{n-1} (t_{k+1} - t_k) + T - t_n \right] \right\} (\Delta t)^n &= \lambda^n e^{-\lambda T} (\Delta t)^n. \end{aligned}$$

Поскольку вероятность появления ровно n событий в промежутке $[0, T]$ равна

$$P(B) = (\lambda T)^n e^{-\lambda T} / n!,$$

то

$$g_n(t_1, \dots, t_n) (\Delta t)^n \sim P(A \cap B) / P(B) \sim n! T^{-n} (\Delta t)^n.$$

Доказанное предложение проливает свет на феномен пуассоновости спорадического фона метеоров (см. пример 7 из §1). По-видимому, спорадические метеорные частицы равномерно заполняют пространство около орбиты Земли, и при ее движении мы наталкиваемся на отдельные частицы (пуассоновские события) так, что моменты этих столкновений выстраивают вариационный ряд равномерного распределения.

Винеровский процесс

Вернемся к примеру 15.4 и рассмотрим броуновское движение на плоскости. Частица вещества помещается в начало декартовой системы координат (x, y) на плоскости, и траектория ее движения описывается кривой с

параметрическим уравнением $x = x(t)$, $y = y(t)$. Нас интересуют конечномерные распределения двумерного процесса $Z(t) = (X(t), Y(t))$, для чего достаточно определить совместную функцию распределения

$$F(x_1, y_1, x_2, y_2, \dots, x_n, y_n) = P(X(t_1) < x_1, Y(t_1) < y_1, \dots, X(t_n) < x_n, Y(t_n) < y_n).$$

Мы начнем с очевидного условия независимости и одинаковой распределенности компонент $X(t)$ и $Y(t)$ процесса $Z(t)$. Хаотическое движение отдельных, не связанных друг с другом молекул толкает частицу в направлении оси OX вне зависимости от того, что делают другие молекулы, способствующие ее движению в направлении OY . Таким образом, броуновское движение на плоскости можно рассматривать как прямое произведение двух одномерных одинаково распределенных броуновских движений. Существует несколько моделей одномерного броуновского движения $X(t)$, $t \in \mathbb{R}_+$, из которых мы остановимся на простейшей, предложенной Н. Винером в начале XX века, и поэтому носящей название *винеровского процесса*.

Построение модели осуществляется по аналогии с выводом нормального распределения путем предельного перехода в биномиальном распределении при неограниченном возрастании числа испытаний Бернулли. Разобьем временную ось $T = \mathbb{R}_+$ на малые интервалы одинаковой длины $1/n$, введем “дискретное время” $t = k/n$, $k = 0, 1, \dots$, и будем предполагать, что частица движется “рывками” в эти моменты времени, передвигаясь с вероятностью $1/2$ вправо на некоторую величину α или, с той же вероятностью $1/2$, влево на такую же величину α , которая не зависит от времени t . Такой дискретный случайный процесс $X_n(t)$, $t = k/n$, $k = 0, 1, \dots$, траектория $x_n(t)$ которого определяет положения частицы в капилляре в моменты времени t , можно представить в виде суммы независимых одинаково распределенных случайных величин, принимающих всего два равных по модулю значения. Действительно, пусть X_1, X_2, \dots – бесконечная последовательность независимых случайных величин, каждая из которых принимает значения $+1$ или -1 с одинаковой вероятностью $1/2$. Тогда

$$X_n(t) = \alpha \sum_{i=0}^{tn} X_i,$$

при любых $t = k/n$, $k = 0, 1, \dots$

Так как случайные величины $X_n(t_1), \dots, X_n(t_m)$ однозначно определяются *приращениями*

$$X_n(t_i) - X_n(t_{i-1}), \quad i = 1, \dots, m, \quad t_0 = 0, \quad X_n(0) = 0$$

процесса $X_n(t)$ и эти приращения независимы в совокупности в силу независимости бинарных случайных величин X_1, X_2, \dots , а

$$X_n(t_i) - X_n(t_{i-1}) = \alpha \sum_{j=nt_{i-1}+1}^{nt_i} X_j,$$

то конечномерные распределения процесса с *независимыми приращениями* $\{X_n(t), t \geq 0\}$ однозначно определяются распределениями случайной величины $X_n(t)$ при каждом фиксированном значении t . Это есть следствие не только независимости бинарных случайных величин, но и того, что приращение $X_n(t_i) - X_n(t_{i-1})$ имеет то же распределение, что и $X_n(t_i - t_{i-1})$. Поэтому, если $f(x | t)$ – функция плотности $X_n(t)$, то функция плотности конечномерных распределений процесса равна

$$\prod_1^m f(x_i - x_{i-1} | t_i - t_{i-1}),$$

где, как и выше, $t_0 = 0, x_0 = 0$.

Не трудно понять, что мы затеяли всю эту игру с дискретным движением броуновской частицы только для того, чтобы потом перейти к пределу при $n \rightarrow \infty$, воспользовавшись центральной предельной теоремой. Но в таком случае необходимо нормировать $\sum_1^n X_i$. Так как $\mathbf{E}X_i = 0$, а $\mathbf{D}X_i = 1$, то условие невырождаемости процесса $X_n(t)$ при $n \rightarrow \infty$ состоит в выборе α пропорциональным $1/\sqrt{n}$. В связи с этим вводят параметр σ^2 , который называют *коэффициентом диффузии* (он характеризует скорость движения частицы), и полагают $\alpha = \sigma/\sqrt{n}$. При таком выборе α мы получаем дискретный случайный процесс

$$X_n(t) = \frac{\sigma}{\sqrt{n}} \sum_0^{tn} X_i.$$

В силу центральной предельной теоремы при каждом фиксированном значении t случайная величина $X_n(t)$ сходится слабо к случайной величине с нормальным $\mathcal{N}(0, \sigma^2 t)$ распределением. Используя теперь представление

конечномерных распределений через распределения приращений, мы можем дать следующее определение винеровского процесса.

Определение 15.1. Случайный процесс $\{X(t), t \geq 0\}$, у которого функция плотности конечномерного распределения определяется формулой

$$f_{t_1, \dots, t_n}(x_1, \dots, x_n) = (2\pi)^{-n/2} \sigma^{-n} \prod_1^n (t_i - t_{i-1})^{-1/2} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_1^n \frac{(x_i - x_{i-1})^2}{(t_i - t_{i-1})} \right\},$$

называется *винеровским случайным процессом*.

Лекция 25

Как и в случае пуассоновского процесса, для практических приложений несомненный интерес представляет распределение случайной величины $\tau = \inf\{t : X(t) \geq h\}$, реализация которой соответствует моменту первого достижения винеровским процессом уровня $h > 0$. К сожалению, для винеровского процесса техника вывода распределений функционалов от траекторий процесса достаточно сложна, и для овладения этой техникой требуется специальный аппарат, во многом выходящий за рамки общего курса теории вероятностей. Однако, что касается дискретного аналога винеровского процесса, который мы рассматривали до определения 15.1, то здесь распределение “первого перескока” можно получить, используя несложную технику комбинаторных выкладок.

Рассмотрим, как и выше, последовательность независимых случайных величин $\{X_i, i \geq 1\}$, принимающих всего два значения $+1$ и -1 с одинаковыми вероятностями $1/2$. Введем дискретный случайный процесс

$$S(t) = \sum_{i=0}^t X_i, \quad t = 0, 1, \dots; \quad S(0) = 0$$

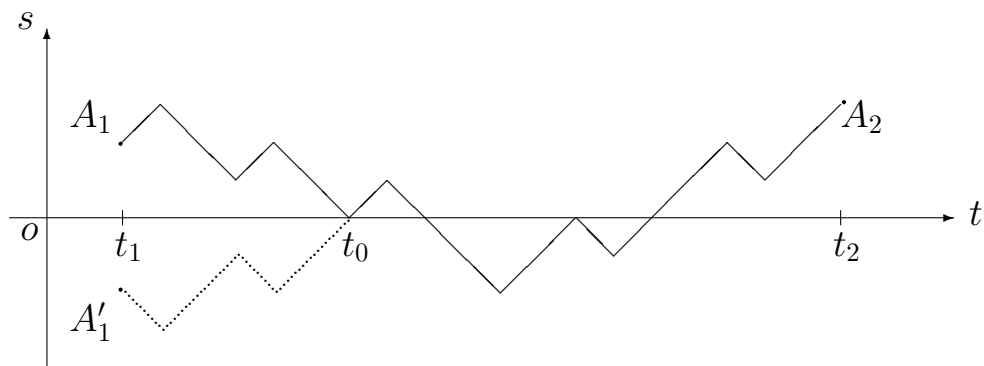
(для наглядности можно соединить последовательно точки $(t, s(t))$ траектории $s(t)$, $t = 0, 1, \dots$ процесса $S(t)$, представив траекторию в виде ломаной линии). Напомним, что дискретный аналог $X_n(t)$ винеровского процесса $X(t)$ получается из процесса $S(t)$ заменой t на tn с $t = k/n$ и последующим масштабированием его траектории: $X_n(t) = S(tn)\sigma/\sqrt{n}$.

Рассмотрим все траектории, проходящие через две заданные точки $A_1 = (t_1, s_1 = s(t_1))$ и $A_2 = (t_2, s_2 = s(t_2))$, $t_1 < t_2$, и назовем участок траектории между этими точками *путем из точки A_1 в точку A_2* . Эти пути обладают тем замечательным свойством, что у них число p слагаемых X_i , $i = t_1 + 1, \dots, t_2$, принявших значение $+1$, одинаково и равно $(t_2 - t_1 + s_2 - s_1)/2$, если, конечно, последнее число целое, – в противном случае не существует траектории, проходящей через эти точки. Действительно, если обозначить q число отрицательных (-1) слагаемых, то $p + q = t_2 - t_1$ и $p - q = s_2 - s_1$, что и дает указанную формулу для расчета p . Из этих же соотношений легко получить формулу для общего числа N путей, проходящих через точки A_1 и A_2 ; очевидно, $N = C_{p+q}^p = C_{p+q}^q$.

Следующие две леммы указывают простой метод для расчета числа путей из начала координат в точку (k, m) , которые расположены ниже уровня m .

Лемма 15.2 (*принцип отражения*). Число путей из точки $A_1 = (t_1, s_1)$, $s_1 > 0$, в точку $A_2 = (t_2, s_2)$, $s_2 > 0$, которые касаются или пересекают ось t хотя бы один раз, равно числу всевозможных путей из точки $A'_1 = (t_1, -s_1)$ в точку A_2 .

Доказательство. Между множеством путей из A_1 в A_2 , удовлетворяющих условию леммы, и множеством всевозможных путей из A'_1 в A_2 можно установить взаимно однозначное соответствие, используя следующий *принцип отражения* (см. рисунок).



Путь из A_1 в A_2 должен по крайней мере один раз коснуться оси времени t ; пусть $t_0 > t_1$ – абсцисса первого касания (напомним, $s_1 = s(t_1) > 0$). Такому пути с ординатами

$$s(t_1) > 0, s(t_1 + 1) > 0, \dots, s(t_0 - 1) > 0, s(t_0) = 0,$$

$$s(t_0 + 1), \dots, s(t_2)$$

сопоставим путь с ординатами

$$-s(t_1) < 0, -s(t_1 + 1) < 0, \dots, -s(t_0 - 1) < 0, s(t_0) = 0, \\ s(t_0 + 1), \dots, s(t_2),$$

который принадлежит второму множеству, то есть отразим участок пути из A_1 в A_2 на промежутке $[t_1, t_0]$ зеркально относительно оси t , а дальше оставим путь без изменения. Легко убедиться, что это взаимно однозначное соответствие – каждому пути второго множества отвечает такой же “зеркальный” образ из первого множества, ибо пути из второго множества обязательно пересекают ось t , так как $-s(t_1) < 0$, а $s(t_2) > 0$. Таким образом, оба множества содержат одинаковое число путей.

Рассмотрим теперь пути из начала координат $(0, 0)$ в точку (k, m) с $0 < m \leq k$. Общее число таких путей, как было показано выше, $N_{k,m} = C_k^p$, где $p = (m + k)/2$, если оно целое, в противном случае $N_{k,m} = 0$.

Лемма 15.3. Число путей из начала координат в точку (k, m) , $0 < m \leq k$, у которых $s(t) > 0$ при всех $t = 1, 2, \dots, k$, равно

$$N_{k-1,m-1} - N_{k-1,m+1} = \frac{m}{k} N_{k,m}.$$

Доказательство. Любой путь из $(0, 0)$ в (k, m) , удовлетворяющий условию леммы, проходит через точку $(1, 1)$. Следовательно, если вычесть из общего числа путей $N_{k-1,m-1}$ из точки $(1, 1)$ в точку (k, m) число M путей, которые соединяют эти точки, касаясь или пересекая ось t , то получим искомое число путей из $(0, 0)$ в (k, m) , лежащих в первом квадранте. В силу леммы 15.2 M равно общему числу путей из точки $(1, -1)$ в точку (k, m) , поэтому $M = N_{k-1,m+1}$.

Лемма 15.4. Число путей из начала координат в точку (k, m) , $0 < m \leq k$, у которых $s(t) < m$ при всех $t = 1, 2, \dots, k - 1$, равно

$$N_{k-1,m-1} - N_{k-1,m+1} = \frac{m}{k} N_{k,m}. \quad (5)$$

Доказательство. Достаточно поместить начало координат в точку (k, m) и трактовать уровень m как ось абсцисс. Используя формулу для расчета N , которая была получена в лемме 15.3, получаем (5).

Последняя лемма устанавливает распределение момента

$$\kappa = \min\{t : S(t) \geq m\}$$

первого выхода на уровень m дискретного процесса

$$S(t) = \sum_1^t X_i, \quad t = 0, 1, \dots$$

Действительно, формула (5) вычисляет количество траекторий определенного вида, связанного с их положением в момент $t = k$. Мы можем сгруппировать бесконечное множество траекторий процесса $S(t)$ в 2^k равновероятных класса в соответствии с различиями в путях, соединяющих начало координат с достижимыми точками, абсцисса которых равна k . Это равносильно к переходу к другому вероятностному пространству, где Ω состоит из 2^k равновероятных точек, и нас интересует вероятность события, состоящего из $mN_{k,m}/k$ элементарных исходов, так что справедлива

Лемма 15.5. *Вероятность того, что дискретный процесс $S(t)$ впервые достигнет уровня m в момент времени $t = k$, равна*

$$\frac{m}{2^k k} N_{k,m} = \frac{m}{2^k k} C_k^{(k+m)/2},$$

где k и m должны иметь одинаковую четность, $m \leq k$.

Теперь обратимся к дискретному аналогу $X_n(t)$ винеровского процесса $X(t)$ и моменту $\tau_n = \min\{t : X_n(t) \geq h\}$ первого выхода процесса $X_n(t)$ на уровень $h > 0$. Перепишем определение τ_n в терминах момента κ :

$$\frac{\kappa}{n} = \min \left\{ t : \frac{\sigma}{\sqrt{n}} S(nt) \geq \frac{m\sigma}{\sqrt{n}} \right\},$$

где $m = h\sqrt{n}/\sigma$. Из этой записи видно, что функция плотности случайной величины τ_n

$$g_n(t) = P(\tau_n = t) = \frac{m}{2^k k} C_k^{(k+m)/2},$$

где $k = nt$ с очевидными ограничениями на возможные значения переменной t и параметров h и σ .

Изучим асимптотическое поведение $g_n(t)$ при $n \rightarrow \infty$ и фиксированном h . Легко понять, что тем самым мы устанавливаем асимптотическое поведение вероятности

$$G(t + 1/n) - G(t) = P(t \leq \tau < t + 1/n) \sim g(t) \frac{1}{n}$$

при $n \rightarrow \infty$, и это позволит нам найти функцию плотности $g(t)$ момента τ первого достижения уровня h винеровским процессом $X(t)$.

Предложение 15.3. Если $n \rightarrow \infty$, то

$$g_n(t) \sim \frac{h}{\sqrt{2\pi\sigma t^{3/2}n}} \exp \left\{ -\frac{h^2}{2\sigma^2 t} \right\}.$$

Доказательство. Нам предстоит исследовать асимптотику выражения

$$\frac{m}{2^k k} \frac{k!}{\left(\frac{k+m}{2}\right)! \left(\frac{k-m}{2}\right)!},$$

в котором $k = nt$, $m = h\sqrt{n}/\sigma$ и $n \rightarrow \infty$.

Поскольку k , $k + m$ и $k - m$ с ростом n стремятся к бесконечности, то, как мы это делали раньше при доказательстве теоремы Муавра–Лапласа, воспользуемся формулой Стирлинга

$$n! = \sqrt{2\pi n} n^{n+1/2} e^{-n} (1 + O(1/n))$$

и представим функцию плотности в асимптотическом виде

$$g_n(t) \sim \frac{m}{2^k k \sqrt{2\pi}} \cdot \frac{k^{k+1/2} e^{-k} 2^{(k+m)/2+1/2+(k-m)/2+1/2}}{(k+m)^{(k+m)/2+1/2} (k-m)^{(k-m)/2+1/2} e^{-(k+m)/2} e^{-(k-m)/2}} = \frac{2m}{\sqrt{2\pi} k^{3/2}} \left(1 + \frac{m}{k}\right)^{-\frac{k+m}{2}-\frac{1}{2}} \left(1 - \frac{m}{k}\right)^{-\frac{k-m}{2}-\frac{1}{2}}$$

Поскольку $m/k \rightarrow 0$, то степени $1/2$ не влияют на асимптотику, и простые алгебраические преобразования дают

$$g_n(t) \sim \frac{2m}{\sqrt{2\pi} k^{3/2}} \left(1 - \frac{m^2}{k^2}\right)^{-\frac{k}{2}} \left(1 + \frac{m}{k}\right)^{-\frac{m}{2}} \left(1 - \frac{m}{k}\right)^{\frac{m}{2}}.$$

Если теперь подставить в правую часть $m = h\sqrt{n}/\sigma$, $k = nt$ и воспользоваться замечательным пределом, определяющим число e , то получим окончательный результат

$$g_n(t) \sim \frac{2h}{\sqrt{2\pi\sigma t^{3/2}n}} \exp \left\{ -\frac{h^2}{2t\sigma^2} \right\}.$$

Итак, мы установили, что $\tau_n \Rightarrow \tau$, функция плотности которого имеет вид

$$g(t) = \frac{2a}{\sqrt{2\pi}t^{3/2}} \exp\left\{-\frac{a^2}{2t}\right\}, \quad a = h/\sigma,$$

а функция распределения $G(t)$ выражается через функцию распределения $\Phi(\cdot)$ стандартного нормального закона $\mathcal{N}(0, 1)$ соотношением

$$G(t) = \frac{2a}{\sqrt{2\pi}} \int_0^t \frac{1}{x\sqrt{x}} \exp\left\{-\frac{a^2}{2x}\right\} dx =$$

$$\frac{2}{\sqrt{2\pi}} \int_{a/\sqrt{t}}^{\infty} \exp\left\{-\frac{u^2}{2}\right\} du = 2 \left[1 - \Phi\left(\frac{a}{\sqrt{t}}\right)\right].$$

Однако это совсем не означает, что мы получили функцию распределения момента первого перескока винеровским процессом заданного уровня. В нашем доказательстве имеется огромная “дыра” – мы не располагаем условиями, при которых сходимость последовательности конечномерных распределений процесса (слабая сходимость) влечет сходимость распределений функционалов от этого процесса. К счастью, в нашем случае с дискретным аналогом винеровского процесса все обстоит благополучно.

Следует отметить, что распределение первого перескока играет важную роль в моделях теории надежности, когда отказ системы вызывается усталостными разрушениями, вызванными хаотическими появлениями “пиковых” нагрузок, которые возникают во времени подобно локальным максимумам траектории винеровского процесса.

§1. Проблема статистического вывода

Лекция 1

Теория вероятностей создает базу для построения моделей реальных явлений, в основе которых лежат соотношения между частотами появления определенных событий. Располагая вероятностной моделью, мы можем рассчитать вероятности (относительные частоты) этих событий и тем самым оптимизировать свое поведение в условиях неопределенности. Математическая статистика строит модели индуктивного поведения в этих условиях на основе имеющихся вероятностных моделей. Основная проблема состоит в том, чтобы по наблюдениям элементарных исходов (обычно это – значения наблюдаемых случайных величин) дать метод выбора действий, при которых частота ошибок была бы наименьшей. Естественно, эта проблема сопряжена с решением сложных задач на экстремум, но даже в том случае, когда эти задачи не удается решить, теория вероятностей дает метод для расчета средней величины потерь, которые мы будем нести, используя конкретное, выбранное нами правило индуктивного поведения. Таким образом, *математическая статистика есть теория принятия оптимальных решений, когда последствия от действий, предпринимаемых на основе этих решений, носят случайный характер*. Математическая статистика использует методы теории вероятностей для расчета частоты “неправильных” решений или, более общо, для величины средних потерь, которые неизбежно возникают в условиях случайности, как бы мы ни пытались оптимизировать свое поведение в этих условиях.

Приведем два примера, иллюстрирующих задачи математической статистики и, отчасти, методы их решения, с тем чтобы в последующем формализовать общую проблему статистического вывода.

Пример 1.1. Определение общего содержания серы в дизельном топливе. Мы снова обращаемся к примеру 7.2 из курса теории вероятностей, где речь шла о важной в экологическом отношении характеристике дизельного топлива – процентном содержании элементарной серы, которая при сжигании и последующем соединении с водой дает серную кислоту. Необходимость использования методов теории вероятностей при аттестации дизельного топлива по этой характеристике была вызвана значительными расхождениями между результатами x_1, \dots, x_n параллельных и независимых испытаний n проб из партии дизельного топлива. Если даже исключить ошибки эксперимента, связанные с неправильным определением веса

пробы и титрованием, то все равно разброс в параллельных испытаниях будет значительным в силу случайного характера процесса сжигания пробы топлива и выпадения части элементарной серы в золу. Но в таком случае возникает естественный вопрос, что же мы измеряем и что же это за характеристика дизельного топлива, которую мы называли “общим содержанием серы”? В практике лабораторных испытаний обычно говорят о *среднем* значении этой характеристики, и дизельное топливо аттестуется величиной $\bar{x} = n^{-1} \sum_1^n x_k$ – арифметическим средним результатов параллельных испытаний. Это и есть то “индуктивное поведение” статистика в условиях случайности, о котором мы говорили в начале лекции, и оправдание разумности такого поведения естественно искать в рамках закона больших чисел.

Действительно, в примере 7.2 мы интерпретировали результат x определения общего содержания серы в одной пробе как результат наблюдения случайной величины X , распределенной по нормальному закону со средним μ и дисперсией σ^2 , причем значение (неизвестное экспериментатору) параметра μ являлось математическим выражением той, не совсем понятной для нас характеристики испытываемого топлива, которая называлась “общим содержанием серы”. В рамках этой вероятностной модели естественно трактовать результаты x_1, \dots, x_n параллельных испытаний n проб дизельного топлива как наблюдения n независимых копий X_1, \dots, X_n случайной величины X . Термин “копия” в данном случае употребляется для обозначения того факта, что каждая из наблюдаемых случайных величин имеет то же распределение, что и X . Таким образом, постулируется, что X_1, \dots, X_n независимы и одинаково распределены $\mathcal{N}(\mu, \sigma^2)$, так что в силу закона больших чисел при неограниченном возрастании объема испытаний n

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{P} \mu.$$

Итак, закон больших чисел гарантирует нам, что при достаточно большом объеме испытаний мы будем близки к истинному значению исследуемой характеристики топлива. Однако на практике в заводских лабораториях обычно сжигаются всего две пробы топлива, и только в исключительных случаях при проверке приборов или тестировании лаборантов делается четыре испытания. Естественно, при $n = 2$ говорить о законе “больших” чисел просто смешно, – следует искать некоторую количественную характеристику последствий от неточной аттестации партии дизель-

ного топлива. Легко понять, что в основу такой характеристики следует положить ошибку $|\bar{X} - \mu|$ в оценке параметра μ , но, к сожалению, значение μ нам неизвестно, а \bar{X} есть случайная величина, что окончательно делает проблему прогноза ожидаемых ошибок при аттестации конкретной партии топлива неразрешимой. Здесь наблюдается та же ситуация, что и при попытке предсказать сторону монеты, которая выпадет при ее подбрасывании. Точный прогноз невозможен, но методы теории вероятностей позволяют нам рассчитать, как часто мы будем ошибаться в прогнозе при достаточно длительной игре в орлянку. Следовательно, мы должны решить задачу о вычислении вероятности того, что ошибка в оценке μ будет слишком большой – превосходить некоторую предписанную величину Δ . Эта вероятность $P(|\bar{X} - \mu| > \Delta)$ обычно называется *риском* оценки \bar{X} , а вероятность $P(|\bar{X} - \mu| \leq \Delta)$ противоположного события – *надежностью* этой оценки.

Таким образом, риск оценки указывает частоту тех партий дизельного топлива, в паспорте которых общее содержание серы указано с недопустимо большой ошибкой. Зная риск оценки, мы можем вычислить средние затраты на выплату рекламаций по искам потребителей дизельного топлива. Вывести формулу для вычисления риска не представляет особого труда, если обратиться к теореме сложения для нормального распределения (предложение 12.2 курса ТВ). Выборочное среднее \bar{X} есть нормированная на n сумма независимых одинаково распределенных $\mathcal{N}(\mu, \sigma^2)$ случайных величин. В силу теоремы сложения эта сумма имеет также нормальное распределение, среднее значение которого равно сумме средних $n\mu$, а дисперсия равна сумме дисперсий $n\sigma^2$. При умножении на $1/n$ среднее умножается на ту же величину, а дисперсия умножается на ее квадрат. Таким образом, $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, надежность оценки

$$P(-\Delta \leq \bar{X} - \mu \leq \Delta) = \Phi(\Delta\sqrt{n}/\sigma) - \Phi(-\Delta\sqrt{n}/\sigma) = 2\Phi(\Delta\sqrt{n}/\sigma) - 1$$

(напомним, $\Phi(-x) = 1 - \Phi(x)$), а ее риск

$$P(|\bar{X} - \mu| > \Delta) = 2(1 - \Phi(\Delta\sqrt{n}/\sigma)).$$

При вычислении риска оценки необходимо знать величину стандартного отклонения σ . Но значение σ , очевидно, остается постоянным при аттестации различных партий – это параметр, характеризующий точность метода химического анализа топлива, и не имеет отношения к его химическому составу. Естественно, за достаточно короткий срок в лабораториях накапливается большой архивный материал данных испытаний различных партий

топлива, что позволяет оценить значение σ с достаточно высокой точностью. С тем, как это делается, мы познакомимся в одной из ближайших лекций.

Используя формулу риска, мы можем определить минимальный объем испытаний n , гарантирующий предписанную, достаточно малую величину риска. Действительно, если α – заданное ограничение на риск оценки, то разрешая неравенство $2(1 - \Phi(\Delta\sqrt{n}/\sigma)) \leq \alpha$ относительно переменной n , получаем, что требуемый объем испытаний определяется неравенством

$$n \geq \left(\frac{\Phi^{-1}(1 - \alpha/2)\sigma}{\Delta} \right)^2.$$

Пример 1.2. Выявление эффекта лечения. Группа пациентов в количестве 10 человек, обладающих схожими антропометрическими и антропологическими характеристиками, подвергается лечению по некоторой новой методике, подтверждение или опровержение эффективности которой составляет предмет статистического исследования. После лечения дается только качественное заключение о состоянии здоровья каждого пациента, так что результат испытания новой методики можно представить в виде последовательности x_1, \dots, x_{10} , компоненты которой принимают значения 1 (положительный исход лечения) или 0 (отрицательный исход).

Предлагается следующее статистическое правило: новая методика объявляется эффективной, если $x_i = 1$ для всех $i = 1, \dots, 10$, то есть все пациенты выздоровели. Если же лечение хотя бы одного пациента не привело к положительному исходу, новая методика не рекомендуется к дальнейшему клиническому использованию. Что можно сказать о надежности или, как говорят медики, “достоверности” такого правила индуктивного поведения?

Чтобы ответить на этот вопрос, мы должны построить вероятностную модель проводимых наблюдений. Естественно предполагать, что в силу “однородности” группы пациентов они обладают одинаковой вероятностью p положительного исхода лечения, и если в процессе лечения они не имели возможности излишне тесного общения, то исходы лечений можно представить в виде реализации десяти независимых бинарных случайных величин X_1, \dots, X_{10} , каждая из которых принимает значение 1 с вероятностью p и значение 0 с вероятностью $1-p$. Таким образом, мы пришли к модели испытаний в схеме Бернулли с вероятностью p успешного исхода. Вероятность того, что все 10 исходов были успешными равна p^{10} , и задавая различ-

ные значения p мы можем судить о том, как часто возможны различные результаты апробации нового метода лечения.

Предположим сначала, что новая методика неэффективна. При таком предположении значение p не должно превосходить величины $1/2$, и максимальное значение вероятности события $X_1 = 1, \dots, X_{10} = 1$ равно $2^{-10} = 1/1024 < 0,001$. Это очень редкое событие, и поэтому предположение о неэффективности новой методики должно быть отвергнуто. При этом вероятность 2^{-10} можно интерпретировать как риск внедрения в медицинскую практику неэффективного метода лечения: *используя предложенное правило выбора между двумя действиями (внедрение или отклонение методики) при испытаниях последующих методик, мы рискуем в среднем не более чем один раз из тысячи внедрить неэффективный метод лечения.*

Интересно заметить, что в предположении “нейтральности” нового метода ($p = 1/2$) вероятность любого исхода $X_1 = x_1, \dots, X_{10} = x_{10}$ одинакова и равна 2^{-10} , но исход $X_1 = 1, \dots, X_{10} = 1$ обладает наибольшей вероятностью принятия действительно эффективной методики, ибо

$$p \sum_1^{10} x_k (1-p)^{n-\sum_1^{10} x_k} \leq p^{10},$$

если $p > 1/2$. Столь же просто проверить, что результаты испытаний, в которых лечение только одного пациента окончилось неудачей, имеют вероятность $p^9(1-p)$, и такие 10 результатов x_1, \dots, x_{10} с одним $x_i = 0$ и другими $x_j = 1$ обладают большей вероятностью, чем исходы с двумя и более количеством неудач, если в действительности $p > 1/2$. Это замечание позволяет нам определить статистическое правило, обладающее наибольшей вероятностью принятия в действительности эффективной методики, но не с таким малым риском, как 2^{-10} .

Дело в том, что в медицинской практике установилась определенная граница риска, равная 0.05, и все события, обладающие меньшей вероятностью, объявляются “редкими” – ими можно пренебречь. В связи с этим позволим себе включить в область принятия новой методики дополнительные исходы с ровно одним неуспехом, и вычислим риск такого статистического правила при $p = 1/2$. Используя известные нам формулы биномиальных вероятностей, находим, что

$$P \left(\sum_1^{10} X_k \geq 9 \right) = p^{10} + C_{10}^1 p^9 (1-p),$$

и при $p = 1/2$ эта вероятность равна $2^{-10}(1 + 10) = 11/1024 \approx 0,01$, что по-прежнему достаточно мало по сравнению с 0.05. Следовательно, мы можем включить в область принятия новой методики еще C_{10}^2 результатов испытаний, в которых присутствуют ровно две неудачи. Риск такого статистического правила становится равным

$$P\left(\sum_{k=1}^{10} X_k \geq 8\right) = p^{10} + C_{10}^1 p^9(1-p) + C_{10}^2 p^8(1-p)^2,$$

и при $p = 1/2$ эта вероятность равна $2^{-10}(1 + 10 + 45) = 56/1024 \approx 0.05$.

Это как раз соответствует принятой в медицине норме риска статистического правила. Итак, мы рекомендуем новую методику к дальнейшему использованию в клинике, если лечение не более чем двух пациентов из десяти оказалось неудачным, и применение такого правила в испытаниях дальнейших методик может привести к принятию неэффективного метода лечения в среднем в пяти случаях из 100.

Мы рассмотрели две типичных задачи математической статистики – оценка параметров и проверка гипотез. Естественно, круг проблем математической статистики намного шире, но при надлежащей трактовке проблем большинство из них сводится или к задаче оценки параметров, или к задаче выбора одного из нескольких альтернативных высказываний об исследуемом объекте. Опираясь на рассмотренные примеры, мы можем теперь представить достаточно общую схему статистического вывода.

Лекция 2

Любое статистическое исследование, проводимое в рамках математической статистики, начинается с описания объекта исследования и формализации *пространства* \mathcal{D} решений d , одно из которых статистик принимает на основе наблюдений независимых копий случайной, возможно векторной, величины X , характеризующей состояние объекта в момент проведения наблюдений. Так, в примере с аттестацией партии дизельного топлива (объект исследования) \mathcal{D} есть интервал $(0; 100)$ (напомним, общее содержание серы измеряется в процентах к весу пробы), а в примере с определением эффективности нового метода лечения (объект исследования) пространство \mathcal{D} состоит из двух точек: d_0 – решение о неэффективности метода (принятие “нулевой” гипотезы) и d_1 – решение о внедрении нового метода в лечебную практику (принятие альтернативной гипотезы).

Наиболее важной и, по-видимому, наиболее сложной частью статистического исследования является этап построения *вероятностной модели*, который состоит в спецификации семейства $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ возможных распределений наблюдаемой случайной величины X . Этот этап связан с достаточно глубоким проникновением в природу исследуемого объекта и метода наблюдений X , – одной математикой здесь, как правило, не обойдешься. Семейство \mathcal{P} индексируется абстрактным *параметром* θ , совокупность значений которого Θ называется *параметрическим пространством*.

В первом примере мы выяснили, что семейство возможных распределений X есть семейство нормальных распределений $\mathcal{N}(\mu, \sigma^2)$ с двумерным параметром $\theta = (\mu, \sigma)$ и параметрическим пространством $\Theta = \mathbb{R} \times \mathbb{R}^+$. В дальнейшем мы предположили, что значение σ известно, и свели наше параметрическое пространство к евклидовой прямой: $\Theta = \mathbb{R}$ с $\theta = \mu$. Наконец, поскольку общее содержание серы измеряется в процентах, мы должны окончательно положить $\Theta = (0; 100.)$.

Во втором примере мы имели дело с бинарной случайной величиной X , принимающей значение 1 с вероятностью p и значение 0 с вероятностью $1 - p$. Таким образом, вероятностная модель представлялась семейством двухточечных распределений $B(1, p)$ с $\theta = p$ и параметрическим пространством $\Theta = (0; 1)$.

Следующий этап статистического исследования состоит в интерпретации решений d в терминах высказываний о соответствующих этому решению значениях параметра θ . Это необходимо сделать, если мы поставили себе задачу количественного измерения последствий от принятия неверных решений, – в наших примерах риск используемых правил представлял собой функцию от θ . Нетрудно понять, что в первом примере $\mathcal{D} = \Theta$, а во втором примере решению d_0 о неэффективности метода соответствует подмножество параметрического пространства $(0; 1/2]$, а альтернативному решению d_1 об использовании новой методики соответствует интервал $(1/2; 1)$ возможных значений параметра $\theta = p$. Именно таким образом мы сводим конкретные задачи по аттестации партии дизельного топлива и выявлению эффективности нового метода лечения к абстрактным задачам математической статистики – оценке параметра (среднего значения) θ нормального (θ, σ^2) распределения и, соответственно, различению двух гипотез $H_0 : \theta \in (0; 1/2]$ и $H_1 : \theta \in (1/2; 1)$ о величине вероятности θ успешного испытания в схеме Бернулли.

Параметрическая интерпретация решений позволяет статистику задать

потери $L(\theta, d)$, которые он несет от принятия решения d , когда θ представляет истинное значение параметра. Среднее значение этих потерь в длинном ряду однотипных статистических исследований с одним и тем же правилом принятия решения определяет величину риска, связанную с принятием неправильных решений. Так, в наших примерах риск определялся вероятностью принятия решения, отстоящего достаточно далеко от того решения, которое соответствовало истинному значению параметра, и, следовательно, *функция потерь* определялась индикатором некоторого подмножества в $\Theta \times \mathcal{D}$. Это так называемые функции потерь типа 0–1. В первом примере $L(\theta, d) = 1$, если $|d - \theta| > \Delta$, и $L(\theta, d) = 0$ в противном случае. Во втором примере $L(\theta, d) = 1$, если принималось решение d_1 , а $\theta \in (0; 1/2]$, или принималось d_0 , а $\theta \in (1/2; 1)$, в остальных точках произведения пространств $\Theta \times \mathcal{D}$ потери $L(\theta, d)$ полагались равными нулю. Отметим, что в задаче оценки параметров довольно часто используется квадратичная функция потерь $L(\theta, d) = |d - \theta|^2$.

Каждое из решений d статистик принимает на основе результата $x^{(n)} = x_1, \dots, x_n$ наблюдений над независимыми копиями $X^{(n)} = (X_1, \dots, X_n)$ случайной величины X . Строится измеримое отображение $\delta = \delta(\cdot)$ пространства возможных значений $X^{(n)}$ в пространство решений \mathcal{D} , с помощью которого принимается решение $d = \delta(x^{(n)})$. Это отображение называется *решающей функцией* или *статистическим правилом*. Так, в первом примере $\delta(X^{(n)}) = \bar{X}$, а во втором

$$\delta(X^{(n)}) = \begin{cases} d_0, & \text{если } \sum_1^n X_k < 8, \\ d_1, & \text{если } \sum_1^n X_k \geq 8. \end{cases}$$

Последствия от использования конкретной решающей функции в длинном ряду однотипных статистических исследований определяются величиной средних потерь $R(\theta; \delta) = \mathbf{E}_\theta L(\theta, \delta(X^{(n)}))$, которая зависит от θ ; функция $R(\theta, \delta)$, $\theta \in \Theta$, называется *функцией риска*.

Основная проблема математической статистики состоит в построении решающих функций δ , минимизирующих равномерно по $\theta \in \Theta$ функцию риска $R(\theta; \delta)$. Мы будем решать эту проблему для задач оценки параметров и проверки гипотез. Естественно, будут также изучаться традиционные, возможно не обладающие оптимальными свойствами, статистические правила, и в этом случае нашей основной задачей будет вычисление их функций риска.

Представленная выше схема статистического вывода весьма далека от общности. Большинство статистических задач имеет дело с наблюдениями одновременно за несколькими объектами, например, новый метод лечения применяется к одной группе пациентов, в то время как другая подвергается лечению традиционным методом, и по данным наблюдений копий двух случайных величин делается вывод о предпочтительности нового метода. Если мы хотим сократить число наблюдений, необходимое для достижения заданной (малой) величины риска, то целесообразно не фиксировать заранее n , а планировать прекращение испытаний после наблюдения каждой копии в зависимости от полученных ранее результатов. Существует большой класс задач управления наблюдениями – оптимального выбора случайной величины, наблюдаемой на каждом шаге статистического эксперимента, а также правила прекращения наблюдений. Все это далеко выходит за рамки тех “кратких начатков” теории статистических выводов, которые будут представлены в нашем семестровом курсе.

Мы завершим этот параграф набором простейших определений и понятий, которые постоянно используются в математической статистике.

Итак, с исследуемым объектом, относительно которого мы должны принять некоторое решение $d \in \mathcal{D}$, соотносится наблюдаемая случайная величина X , распределение которой P_θ известно с точностью до значения параметра θ . Семейство распределений $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, как обычно, называется *вероятностной моделью*. Пусть (X, \mathcal{A}) – измеримое пространство значений X . В дальнейшем будет всегда предполагаться, что на сигма-алгебре \mathcal{A} существует такая сигма-конечная мера μ , что при любом $\theta \in \Theta$ распределение X можно представить в виде интеграла

$$P_\theta(A) = \mathbf{P}(X \in A) = \int_A f(x | \theta) d\mu(x), \quad A \in \mathcal{A},$$

от плотности $f(x | \theta)$ распределения X по мере μ . В таком случае распределение независимых копий $X^{(n)} = (X_1, \dots, X_n)$ случайной величины X на произведении $(\mathcal{X}^n, \mathcal{A}^n)$ измеримых пространств (X, \mathcal{A}) определяется функцией плотности

$$f_n(x^{(n)} | \theta) = \prod_{k=1}^n f(x_k | \theta)$$

по мере $\mu_n = \underbrace{\mu \times \cdots \times \mu}_n$, то есть

$$P_{\theta,n}(A_n) = \mathbf{P}(X^{(n)} \in A_n) = \int_{A_n} f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}), \quad A_n \in \mathcal{A}^n.$$

Определение 1.1. Вектор $X^{(n)} = (X_1, \dots, X_n)$ независимых, одинаково распределенных по тому же закону, что и наблюдаемая случайная величина X , случайных величин называется *случайной выборкой объема n* . Измеримое пространство $(\mathcal{X}^n, \mathcal{A}^n)$ значений $X^{(n)}$ называется *выборочным пространством*, а семейство распределений $\mathcal{P}_n = \{P_{\theta,n}, \theta \in \Theta\}$ на этом пространстве – *статистической структурой* или *статистическим экспериментом*. Вектор $x^{(n)} = (x_1, \dots, x_n)$ результатов наблюдения случайной выборки $X^{(n)}$ называется *вектором (или совокупностью) выборочных данных*.

Зная распределение выборки, мы можем вычислять риск любого статистического правила δ с помощью n -кратного интеграла

$$R(\theta; \delta) = \int_{\mathcal{X}} \dots \int_{\mathcal{X}} L(\theta, \delta(x^{(n)})) f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}).$$

Конечно, если удастся найти распределение G_θ решающей функции δ на измеримом пространстве решений $(\mathcal{D}, \mathcal{D})$, то вычисление риска упрощается:

$$R(\theta; \delta) = \int_{\mathcal{D}} L(\theta, a) dG(a).$$

Так, в первом примере с выборкой из нормального (μ, σ^2) распределения решающей функцией служило выборочное среднее \bar{X} . Было показано, что \bar{X} имеет нормальное $(\mu, \sigma^2/n)$ распределение, и именно это обстоятельство позволило нам найти простое выражение риска статистического правила через функцию распределения стандартного нормального закона. Точно так же во втором примере с выбором из двухточечного распределения $B(1, p)$ решающая функция была основана на случайной величине $\sum_1^n X_k$, которая имеет распределение Бернулли $B(n, p)$. Риск нашего решающего правила по выявлению эффективности метода лечения выражался через функцию распределения $B(n, p)$.

Заметим, что функции от выборочного вектора $X^{(n)}$ играют важную, можно даже сказать самостоятельную, роль в математической статистике.

Определение 1.2. Любое измеримое отображение $T = T(X^{(n)})$ выборочного пространства $(\mathcal{X}^n, \mathcal{A}^n)$ в некоторое измеримое пространство $(\mathcal{T}, \mathcal{B})$ называется *статистикой*.

Существует довольно устоявшийся универсальный набор статистик, постоянно используемых в теории и практике статистического вывода; распределения этих статистик интенсивно изучались на протяжении последних двух столетий. В следующем параграфе мы познакомимся с набором статистик, которые являются выборочными аналогами стандартных характеристик распределения наблюдаемой случайной величины, а также рассмотрим статистики, редуцирующие размерность выборочного вектора до размерности параметрического пространства без потери информации.

§2. Выборочные характеристики. Достаточные статистики

Лекция 3

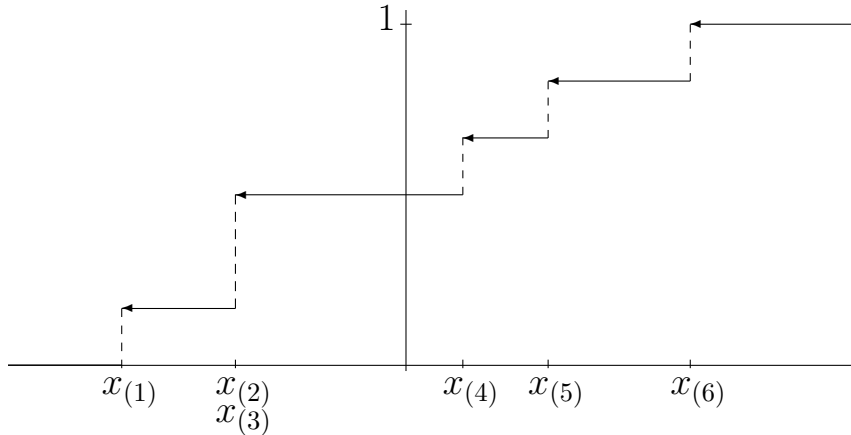
Построение вероятностных моделей в курсе теории вероятностей осуществлялось посредством спецификации функции распределения или функции плотности наблюдаемой случайной величины X . Любая из этих функций однозначно определяет распределение X на сигма-алгебре \mathcal{A} борелевских множеств, порожденной интервалами в пространстве $\mathcal{X} = \mathbb{R}$ возможных значений X , и с их помощью вычислялись такие характеристики распределения, как среднее, дисперсия, коэффициенты асимметрии и эксцесса, квантили, мода и пр. В прикладной статистике существует традиция, или, можно сказать, обязательное правило, представлять полученные экспериментальные данные с помощью статистик – выборочных аналогов этих функций и характеристик распределения X . Выборочные характеристики являются оценками истинных значений своих прообразов и позволяют судить в общих чертах о характере распределения наблюдаемой случайной величины.

Такая “описательная” статистика обычно начинается с построения *вариационного ряда*: выборочные данные x_1, \dots, x_n упорядочиваются по возрастанию их значений $x_{(1)} \leq \dots \leq x_{(n)}$, и полученный таким образом вектор с неубывающими компонентами служит реализацией случайного вектора $X_{(1)}, \dots, X_{(n)}$, который, собственно, и следует называть *вариационным рядом*. Компоненты вариационного ряда называются *порядковыми статистиками*, а $X_{(1)}$ и $X_{(n)}$ – *крайними членами* вариационного ряда. Мы уже сталкивались с порядковыми статистиками, когда изучали структуру пуассоновского процесса и строили вероятностную модель “слабого звена” (распределение Вейбулла).

Упорядоченные данные наносятся на ось абсцисс, и строится ступенчатая функция, возрастающая скачками величины $1/n$ в каждой точке $x_{(1)}, \dots, x_{(n)}$. Построенная таким образом дискретная функция распределения является реализацией случайной функции

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n \mathbf{I}(X_k < x)$$

($\mathbf{I}(A)$, как обычно, индикатор события A) и называется *эмпирической функцией распределения*.



Таким образом, дискретное эмпирическое распределение приписывает равные вероятности $1/n$ каждой из n компонент выборочного вектора, и при каждом фиксированном $x \in \mathbb{R}$ случайная величина $nF_n(x)$ подчиняется биномиальному распределению $B(n, F(x))$:

$$P(F_n(x) = k/n) = C_n^k F^k(x)(1 - F(x))^{n-k}, \quad k = 0, 1, \dots, n.$$

В силу закона больших чисел Бернулли $F_n \xrightarrow{P} F(x)$ при любом $x \in \mathbb{R}$. Более того, теорема Гливленко–Кантелли, утверждение которой

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{P} 0$$

мы приводим без доказательства, указывает на равномерность этой сходимости на всей числовой оси \mathbb{R} .

Мы закончим обсуждение свойств эмпирической функции распределения формулировкой широко известного результата А.Н. Колмогорова:

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n < x) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-k^2 x^2}.$$

Полученная им формула для асимптотического ($n \rightarrow \infty$) распределения статистики $\sqrt{n}D_n$, характеризующей величину расхождения между теоретическим F и эмпирическим F_n распределениями, используется для построения *критерия согласия* выборочных данных с предположением, что F является истинной функцией распределения, из которого извлекается выборка (гипотезой о том, что F есть функция распределения наблюдаемой случайной величины X).

Итак, мы установили, что эмпирическое распределение сходится по вероятности к истинному (или, как обычно говорят прикладники, теоретическому) распределению, и теперь можем обратиться к вычислению моментных

и квантильных характеристик распределения F_n . Его нецентральные

$$a_k = \int_{\mathbb{R}} x^k dF_n(x) = \frac{1}{n} \sum_{i=1}^n X_i^k$$

и центральные

$$m_k = \int_{\mathbb{R}} (x - a_1)^k dF_n(x) = \frac{1}{n} \sum_{i=1}^n (X_i - a_1)^k$$

моменты служат выборочными аналогами соответствующих теоретических моментов α_k , $k = 1, 2, \dots$, и μ_k , $k = 2, 3, \dots$, и называются *выборочными моментами*.

Если теоретические моменты существуют, то в силу закона больших чисел выборочные моменты сходятся по вероятности к своим теоретическим прообразам. Среди выборочных моментов особое место занимают моменты первого и второго порядков. Выборочный момент a_1 называется *выборочным средним* и имеет специальное обозначение \bar{X} ; *выборочная дисперсия* $m_2 = a_2 - \bar{X}^2$ обычно обозначается S^2 . Соответствующим образом определяются *выборочный коэффициент асимметрии* $g_1 = m_3/S^3$ и *выборочный коэффициент эксцесса* $g_2 = m_4/S^4 - 3$.

При выборе из m -мерного, $m > 1$, распределения эмпирическое распределение также приписывает массу n^{-1} каждому выборочному (векторному) значению $X_i = (X_{1i}, \dots, X_{mi})$, $i = 1, \dots, n$. В соответствии с этим мы можем определить вектор выборочных средних $\bar{X} = (\bar{X}_1, \dots, \bar{X}_m)$ с компонентами

$$\bar{X}_k = \frac{1}{n} \sum_{i=1}^n X_{ki}, \quad k = 1, \dots, m,$$

выборочную ковариационную матрицу $S = \|S_{kj}\|$ с элементами

$$S_{kj} = \frac{1}{n} \sum_{i=1}^n (X_{ki} - \bar{X}_k)(X_{ji} - \bar{X}_j) = \frac{1}{n} \sum_{i=1}^n X_{ki}X_{ji} - \bar{X}_k\bar{X}_j, \quad k, j = 1, \dots, m,$$

и матрицу *выборочных коэффициентов корреляции* $R = \|r_{kj}\|$ с элементами

$$r_{kj} = S_{kj} / \sqrt{S_{kk}S_{jj}}, \quad k, j = 1, \dots, m.$$

Смешанные моменты более высоких порядков в многомерном случае обычно не вычисляются.

Если выбор происходит из распределения, для которого справедлива теорема сложения (предложение 12.2 курса ТВ), то распределение выборочного среднего устанавливается достаточно просто. В общем же случае можно только утверждать об асимптотической ($n \rightarrow \infty$) нормальности этой статистики при условии существования второго момента у теоретического распределения. Аналогичное утверждение справедливо и для моментов любого k -го порядка, если у $F(x)$ существует момент порядка $2k$.

Обратимся теперь к выборочным аналогам квантилей распределения F наблюдаемой случайной величины X . Напомним, что для непрерывного распределения квантиль порядка p определялась как решение x_p уравнения $F(x) = p$, а в случае дискретного распределения – как наибольшее $x = x_p$ из носителя распределения, при котором $F(x_p) \leq p$. Поскольку эмпирическое распределение дискретно, и его функция распределения $F_n(\cdot)$ возрастает скачками в точках, соответствующих компонентам вариационного ряда, то *выборочная квантиль* порядка p полагается равной порядковой статистике $X_{([np])}$, где $[x]$, как обычно, означает целую часть x . Естественно, для повышения точности оценки истинной квантили x_p можно проводить интерполяцию между статистиками $X_{([np])}$ и $X_{([np]+1)}$. Так, *выборочная медиана*, будучи квантилью порядка $p = 0.5$, обычно определяется как $(X_{([n/2])} + X_{([n/2]+1)})/2$. Что же касается оценки моды распределения – точки наибольшего сгущения выборочных данных, то здесь нам придется обратиться к выборочным аналогам функции плотности.

При больших объемах наблюдений выборочные данные обычно подвергаются группировке, при этом индивидуальные выборочные значения не приводятся, а указываются лишь количества наблюдений, попавших в интервалы некоторого разбиения множества X значений наблюдаемой случайной величины. Поясним процедуру группировки на примере выборки из непрерывного одномерного распределения, когда $X = \mathbb{R}$.

В декартовой системе координат ось абсцисс разбивается на $r \geq 2$ интервалов

$$(-\infty, a_1], (a_1, a_2], \dots, (a_{r-2}, a_{r-1}], (a_{r-1}, +\infty),$$

причем внутренние интервалы выбираются, как правило, одинаковой длины: $a_i - a_{i-1} = \Delta$, $i = 2, \dots, r - 1$. Выборочные данные сортируются по интервалам разбиения и подсчитываются частоты n_i , $i = 1, \dots, r$ попадания данных в каждый интервал. Над каждым внутренним интервалом рисуется прямоугольник высоты $n_i/n\Delta$, так что площадь n_i/n каждого

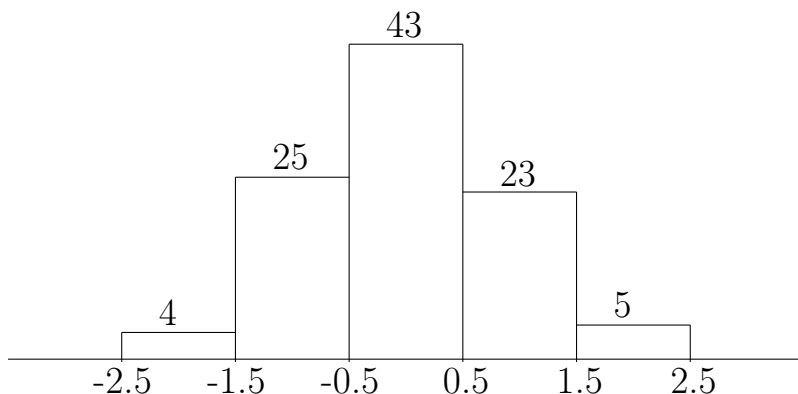
прямоугольника с номером $i = 2, \dots, r - 1$ служит реализацией частотной оценки ν_i/n вероятности попадания наблюдаемой случайной величины X в соответствующий интервал. Здесь ν_i – статистика, которую можно записать с помощью индикаторов событий

$$A_{ij} = \{X_j \in (a_{i-1}, a_i]\}, \quad i = 1, \dots, r, \quad a_0 = -\infty, \quad a_r = +\infty, \quad j = 1, \dots, n,$$

а именно

$$\nu_i = \sum_{j=1}^n I(A_{ij}).$$

Полученная таким образом случайная ступенчатая функция, принимающая нулевые значения на крайних интервалах $(-\infty, a_1]$, $(a_{r-1}, +\infty)$ и равная $\nu_i/n\Delta$ на внутренних интервалах с номерами $i = 2, \dots, r - 1$, называется *гистограммной оценкой* f_n функции плотности $f(x)$, $x \in \mathbb{R}$ распределения X , а ее реализация (ν_i заменяются на наблюдаемые частоты n_i , $i = 1, \dots, r$) – *гистограммой* выборки $x^{(n)}$.



В математической статистике существует ряд теорем, устанавливающих, что при определенных условиях на плотность f гистограммная оценка $f_n(x) \xrightarrow{P} f(x)$ при любом $x \in \mathbb{R}$, если $n \rightarrow \infty$ и одновременно $r \rightarrow \infty$, а $\Delta \rightarrow 0$ со скоростью, зависящей определенным образом от n и r .

В случае гистограммной оценки функции плотности естественно считать выборочным аналогом (оценкой) моды распределения X середину интервала разбиения, в котором гистограмма принимает наибольшее значение.

Заметим также, что вектор частот (ν_1, \dots, ν_r) имеет мультиномиальное распределение $\mathcal{M}(r, n, \mathbf{p})$ с вероятностями исходов $p_i = F(a_i) - F(a_{i-1})$, $i = 1, \dots, r$, что позволяет найти распределение оценки $f_n(x)$ при любом $x \in \mathbb{R}$ и построить критерий согласия выборочных данных с гипотезой о виде распределения наблюдаемой случайной величины. Это широко используемый

на практике *критерий хи-квадрат*, основанный на статистике (сравните с критерием Колмогорова D_n)

$$X^2 = \sum_1^r \frac{(\nu_i - np_i)^2}{np_i}.$$

Асимптотическое распределение этой статистики мы изучим в параграфе, посвященном статистической проверке гипотез.

Итак, мы рассмотрели основные выборочные аналоги распределения наблюдаемой случайной величины и его основных характеристик. Мы высказали также ряд утверждений о распределении этих статистик, что позволит нам в последующем вычислять последствия от их использования в качестве решающих функций. Для того чтобы уяснить, насколько важно знать хотя бы среднее значение статистики, претендующей на роль решающей функции, обратимся снова к примеру 1.1 по аттестации партии дизельного топлива, где обсуждалась сопутствующая проблема оценки дисперсии σ^2 наблюдаемой случайной величины $X \sim \mathcal{N}(\mu, \sigma^2)$.

Предлагалось оценивать σ^2 по накопленному в лаборатории архиву испытаний аттестуемых партий дизельного топлива, то есть по данным большого числа N выборок $X_1^{(n)}, \dots, X_N^{(n)}$ малого объема n . Каждая i -я выборка извлекается из нормального (μ_i, σ^2) распределения, причем средние μ_i могут быть различными для разных выборок, $i = 1, \dots, N$, но дисперсия σ^2 у всех выборок одна и та же. Предлагается следующая оценка σ^2 . В каждой выборке вычисляется выборочная дисперсия S_i^2 , $i = 1, \dots, N$, и затем берется их арифметическое среднее:

$$\hat{\sigma}_N^2 = (1/N) \sum_1^N S_i^2.$$

Распределение каждой S_i^2 не зависит от μ_i , $i = 1, \dots, N$, поскольку выборочная дисперсия инвариантна относительно сдвигов $X_k \rightarrow X_k + a$. Следовательно, предлагаемая оценка есть нормированная на N сумма независимых, одинаково распределенных случайных величин – копий статистики

$$S^2 = (1/n) \sum_1^n (X_k - \bar{X})^2,$$

и в силу закона больших чисел

$$\hat{\sigma}_N^2 \xrightarrow{P} \mathbf{E}S^2$$

при неограниченном возрастании объема N архивных данных. Вычислим это математическое ожидание:

$$\mathbf{E}S^2 = \mathbf{E} \left(\frac{1}{n} \sum_1^n X_k^2 - \bar{X}^2 \right) = \mathbf{E}X^2 - \mathbf{E}\bar{X}^2 = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2,$$

поскольку

$$\alpha_2 = \mathbf{E}X^2 = \mathbf{D}X + \mathbf{E}^2X, \quad \bar{X} \sim \mathcal{N}(\mu, \sigma^2/n).$$

Таким образом, предлагаемая оценка обладает значительным смещением при малом объеме n испытаний каждой партии дизельного топлива. Например, в случае $n = 2$ мы занижаем дисперсию в два раза, поскольку $\hat{\sigma}_N^2 \xrightarrow{P} \sigma^2/2$. Естественно, этот дефект легко устраним – достаточно использовать исправленную на смещение оценку $\tilde{\sigma}_N^2 = (n/(n-1))\hat{\sigma}_N^2$.

Лекция 4

В завершении этого параграфа мы изучим еще один класс замечательных статистик, используя которые можно редуцировать выборочные данные только к их значениям без потери информации. К сожалению, не все статистические структуры обладают такими статистиками, но, по существу, только в тех структурах, где имеются достаточные статистики, возможно построение оптимального статистического правила, на котором достигается минимум риска.

Идея, состоящая в том, что в определенных случаях для принятия решения без увеличения риска *достаточно* знать только значения некоторых статистик, а не все выборочные данные, не требует введения специальных мер информации, содержащейся в выборочных данных и статистиках, – все становится ясным при рассмотрении следующей простейшей задачи, с которой мы имели дело в самом начале курса теории вероятностей.

Предположим, что мы хотим узнать вероятность наследования доминантного признака в опытах Менделя и располагаем результатами x_1, \dots, x_n скрещиваний n пар, где, как обычно, каждое x_i есть индикатор наследования признака, $i = 1, \dots, n$, а совокупность выборочных данных представляет реализацию случайной выборки X_1, \dots, X_n из двухточечного распределения с функцией плотности

$$f(x | \theta) = P_\theta(X = x) = \theta^x(1 - \theta)^{1-x},$$

отличной от нуля только в точках $x = 0$ и 1 . Частотная оценка $\hat{\theta}_n = T/n$ вероятности θ наследования признака определяется статистикой $T = \sum_1^n X_k$, выборочное значение $t = \sum_1^n x_k$ которой соответствует числу потомков в эксперименте, наследовавших доминантный признак. Естественно, возникает вопрос, а нельзя ли извлечь дополнительную информацию о величине параметра θ из номеров k_1, \dots, k_t выборочных данных, принявших значение 1 ? Нетрудно понять, что это возможно только в том случае, если *распределение выборочного вектора $X^{(n)}$ при условии, что статистика T приняла фиксированное значение t , зависит от параметра θ* . Действительно, если мы будем наблюдать случайную величину, которая не имеет никакого отношения к интересующему нас параметру, то откуда этой информации взяться? Итак, найдем условное распределение $X^{(n)}$ относительно T .

Используя формулу условной вероятности, получаем, что

$$P_{\theta} \left(X^{(n)} = x^{(n)} \mid T = t \right) = \frac{P_{\theta} \left(\{X^{(n)} = x^{(n)}\} \cap \left\{ \sum_1^n X_k = t \right\} \right)}{P_{\theta} \left(\sum_1^n X_k = t \right)}.$$

Если значения компонент вектора $x^{(n)}$ таковы, что $\sum_1^n x_k \neq t$, то события $X^{(n)} = x^{(n)}$ и $\sum_1^n X_k = t$, очевидно, несовместны, и поэтому в этом случае условная вероятность равна нулю (не зависит от θ). Если же $\sum_1^n x_k = t$, то событие $X^{(n)} = x^{(n)}$ влечет событие $\sum_1^n X_k = t$, и формула для вычисления условной вероятности упрощается:

$$P_{\theta} \left(X^{(n)} = x^{(n)} \mid T = t \right) = \frac{P_{\theta} \left(X^{(n)} = x^{(n)} \right)}{P_{\theta} \left(\sum_1^n X_k = t \right)}.$$

Так как

$$P_{\theta} \left(X^{(n)} = x^{(n)} \right) = f_n \left(X^{(n)} \mid \theta \right) = \theta^{\sum_1^n x_k} (1 - \theta)^{n - \sum_1^n x_k},$$

$$P_{\theta} \left(\sum_1^n X_k = t \right) = C_n^t \theta^t (1 - \theta)^{n-t},$$

то в случае $\sum_1^n x_k = t$ условное распределение выборочного вектора $X^{(n)}$ относительно статистики T имеет вид

$$P_{\theta} \left(X^{(n)} = x^{(n)} \mid T = t \right) = \frac{1}{C_n^t}$$

и также не зависит от θ .

Итак, наши выкладки показывают, что распределение выборочного вектора на “плоскости” $\sum_1^n X_k = t$ не зависит от θ , и поэтому расположение значений $x_k = 1$ в последовательности x_1, \dots, x_n при фиксированном количестве таких значений не несет информации о параметре θ .

Определение 2.1. Статистика $T = T(X^{(n)})$ называется *достаточной* для статистической структуры $\mathcal{P}_n = \{P_{\theta,n}, \theta \in \Theta\}$, если условное распределение выборочного вектора $X^{(n)}$ относительно статистики T не зависит от θ .

В общей теории статистического вывода в рамках более общего определения статистического правила устанавливается замечательный факт: *если статистическая структура обладает достаточной статистикой T , то, каково бы ни было статистическое правило $\delta = \delta(X^{(n)})$, всегда существует правило $\delta^* = \delta^*(T)$, основанное только на T , риск которого совпадает с риском правила δ* . Таким образом, построение оптимальных статистических правил следует начинать с поиска достаточных статистик. Следующая теорема дает критерий существования у статистических структур достаточных статистик и, одновременно, указывает простой способ их нахождения.

Теорема 2.1. *Для того чтобы $T = T(X^{(n)})$ была достаточной статистикой для статистической структуры, определяемой функцией плотности $f_n(x^{(n)} | \theta)$, необходимо и достаточно, чтобы эта функция допускала представление*

$$f_n(x^{(n)} | \theta) = g_\theta \left(T(x^{(n)}) \right) h(x^{(n)}), \quad (1)$$

где функция h не зависит от параметра θ , а функция g зависит от θ и аргумента $x^{(n)}$ только через значения $T(x^{(n)})$ статистики $T = T(X^{(n)})$,

Доказательство теоремы мы проведем только для дискретного распределения наблюдаемой случайной величины, когда функция плотности выборки $f_n(x^{(n)} | \theta) = P_\theta (X^{(n)} = x^{(n)})$. В случае непрерывного распределения схема доказательства та же, но придется делать замену в n -кратном интеграле.

Достаточность. Пусть выполняется факторизационное представление (1); требуется показать, что условное распределение $X^{(n)}$ относительно T не зависит от θ . Как и в только что рассмотренном примере с двухточечным

распределением, воспользуемся формулой условной вероятности для вычисления условной плотности $X^{(n)}$ относительно T :

$$P_{\theta} \left(X^{(n)} = x^{(n)} \mid T(X^{(n)}) = t \right) = \frac{P_{\theta} \left(\{X^{(n)} = x^{(n)}\} \cap \{T(X^{(n)}) = t\} \right)}{P_{\theta}(T(X^{(n)}) = t)}.$$

События, стоящие в числителе, будут несовместными, если $T(x^{(n)}) \neq t$, и в этом случае условная вероятность равна нулю (не зависит от θ). Если же $T(x^{(n)}) = t$, то первое по порядку событие в числителе влечет второе, и поэтому формула для вычисления условной вероятности упрощается:

$$P_{\theta} \left(X^{(n)} = x^{(n)} \mid T(X^{(n)}) = t \right) = \frac{P_{\theta} \left(X^{(n)} = x^{(n)} \right)}{P_{\theta}(T(X^{(n)}) = t)}.$$

Так как

$$P_{\theta}(X^{(n)} = x^{(n)}) = f_n(x^{(n)} \mid \theta),$$

то используя представление (1), получаем, что (напомним, $T(x^{(n)}) = t$)

$$P_{\theta} \left(X^{(n)} = x^{(n)} \mid T(X^{(n)}) = t \right) = \frac{g_{\theta}(T(x^{(n)}))h(x^{(n)})}{\sum_{y^{(n)}: T(y^{(n)})=t} g_{\theta}(T(y^{(n)}))h(y^{(n)})} = \frac{h(x^{(n)})}{\sum_{y^{(n)}: T(y^{(n)})=t} h(y^{(n)})}.$$

Таким образом, условное распределение не зависит от θ , и поэтому статистика T достаточна для \mathcal{P}_n .

Необходимость. Пусть T – достаточная статистика, так что условное распределение

$$P_{\theta} \left(X^{(n)} = x^{(n)} \mid T(X^{(n)}) = t \right) = K(x^{(n)}, t),$$

где функция K не зависит от θ . Требуется показать, что в этом случае для функции плотности выборки справедливо представление (1).

Имеем

$$f_n(x^{(n)} \mid \theta) = P_{\theta} \left(X^{(n)} = x^{(n)} \right) = P_{\theta} \left(\{X^{(n)} = x^{(n)}\} \cap \{T(X^{(n)}) = T(x^{(n)})\} \right) =$$

$$P_\theta(T(X^{(n)}) = T(x^{(n)})) \cdot P_\theta\left(X^{(n)} = x^{(n)} \mid T(X^{(n)}) = T(x^{(n)})\right).$$

Мы получили представление (1) с

$$g_\theta(T(x^{(n)})) = P_\theta(T(X^{(n)}) = T(x^{(n)})), \quad h(x^{(n)}) = K(x^{(n)}, T(x^{(n)})).$$

Теорема доказана.

Рассмотрим несколько примеров на применения полученного критерия достаточности к статистическим структурам, соответствующим вероятностным моделям из нашего курса теории вероятностей. Начнем с двухточечного распределения (выбор в схеме Бернулли), где мы непосредственными вычислениями условного распределения убедились в достаточности статистики, реализующей число успешных испытаний, – посмотрим, как это делается с помощью представления (1).

1⁰. Двухточечное распределение $B(1, \theta)$ имеет функцию плотности

$$f(x \mid \theta) = \theta^x (1 - \theta)^{1-x},$$

отличную от нуля только в точках $x = 0$ и 1 . Параметрическое пространство этого распределения $\Theta = (0; 1)$, а функция плотности случайной выборки

$$f_n(x^{(n)} \mid \theta) = \theta^{\sum_1^n x_k} (1 - \theta)^{n - \sum_1^n x_k}.$$

Представление (1) выполняется с $h(x^{(n)}) \equiv 1$ и $T(x^{(n)}) = \sum_1^n x_k$. Следовательно, $T = \sum_1^n X_k$ – достаточная статистика.

2⁰. Распределение Пуассона $P(\theta)$, для которого

$$f(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!}, \quad x = 0, 1, \dots, \quad \Theta = \mathbb{R}_+,$$

функция плотности выборки

$$f_n(x^{(n)} \mid \theta) = \theta^{\sum_1^n x_k} e^{-n\theta} / \prod_1^n x_k!.$$

Следовательно, в представлении (1)

$$h(x^{(n)}) = \left[\prod_1^n x_k! \right]^{-1},$$

и $T = \sum_1^n X_k$ – достаточная статистика.

3⁰. Показательное распределение $E(\theta)$ с

$$f(x | \theta) = \frac{1}{\theta} \exp \left\{ -\frac{x}{\theta} \right\}, \quad x \geq 0, \quad \Theta = \mathbb{R}_+,$$

и

$$f_n(x^{(n)} | \theta) = \frac{1}{\theta^n} \exp \left\{ -\frac{1}{\theta} \sum_1^n x_k \right\}$$

также обладает достаточной статистикой $T = \sum_1^n X_k$.

4⁰. Равномерное распределение $U(a, b)$, функция плотности которого

$$f(x | \theta) = \frac{I(a \leq x) I(b \geq x)}{b - a}$$

отлична от нуля и постоянна на отрезке $[a; b]$, на что указывают стоящие в числителе индикаторные функции отрезка $[a; b]$. В этом распределении $\theta = (a, b)$ – двумерный параметр и параметрическое пространство

$$\Theta = \{(a, b) : (a, b) \in \mathbb{R}^2, a < b\}.$$

Статистическая структура определяется функцией плотности

$$f_n(x^{(n)} | \theta) = \frac{\prod_1^n I(a \leq x_k) I(b \geq x_k)}{(b - a)^n} = \frac{I(a \leq x_{(1)}) I(b \geq x_{(n)})}{(b - a)^n},$$

и, следовательно, вектор $T = (X_{(1)}, X_{(n)})$ крайних членов вариационного ряда является достаточной статистикой.

5⁰. Нормальное распределение $\mathcal{N}(\mu, \sigma^2)$. Это распределение обладает двумерным параметром $\theta = (\mu, \sigma)$ с областью значений (параметрическим пространством) $\Theta = \mathbb{R} \times \mathbb{R}_+$. Функции плотности наблюдаемой случайной величины X и случайной выборки $X^{(n)}$ определяются соответственно как

$$f(x | \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}$$

и

$$f_n(x^{(n)} | \theta) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_1^n (x_k - \mu)^2 \right\} =$$

$$\frac{1}{(2\pi)^{n/2}\sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_1^n x_k^2 - 2\mu \sum_1^n x_k + n\mu^2 \right) \right\}.$$

Последнее выражение для плотности $X^{(n)}$ показывает, что двумерная статистика $T = (T_1, T_2)$ с

$$T_1 = \sum_1^n X_k \quad T_2 = \sum_1^n X_k^2$$

достаточна для статистической структуры нормального распределения. Кроме того, поскольку $T_1 = n\bar{X}$ и $T_2 = n(S^2 + \bar{X}^2)$, то факторизационное равенство (1) указывает на достаточность статистик \bar{X} и S^2 , которые имеют конкретную статистическую интерпретацию и поэтому более удобны для практического использования. Понятно, что это замечание носит общий характер: *любые взаимно однозначные преобразования достаточной статистики наследуют свойство достаточности.*

Отметим также, что в случае известного (фиксированного) σ статистическая структура имеет параметрическое пространство, совпадающее с областью значений параметра μ , и достаточной статистикой будет выборочное среднее \bar{X} . Аналогичное утверждение имеет место для статистики S^2 при фиксированном μ .

6⁰. Гамма-распределение $G(\lambda, a)$ имеет функцию плотности

$$f(x | \theta) = \frac{1}{a^\lambda \Gamma(\lambda)} x^{\lambda-1} \exp \left\{ -\frac{x}{a} \right\}, \quad x > 0, \quad \theta = (a, \lambda), \quad a > 0, \quad \lambda > 0,$$

так что функция плотности выборочного вектора

$$f_n(x^{(n)} | \theta) = \frac{1}{a^{n\lambda} \Gamma^n(\lambda)} \left[\prod_1^n x_k \right]^{\lambda-1} \exp \left\{ -\frac{1}{a} \sum_1^n x_k \right\}.$$

Тождество (1) указывает, что достаточной является двумерная статистика

$$\left(\sum_1^n X_k, \prod_1^n X_k \right)$$

или более удобная в вычислительном отношении статистика

$$\left(\sum_1^n X_k, \sum_1^n \ln X_k \right).$$

Для этого распределения можно сделать то же замечание, что и для нормального: первая компонента достаточной статистики “отвечает” за масштабный параметр a , в то время как вторая соответствует параметру формы λ .

7⁰. *Биномиальное распределение* $V(m, p)$. Это дискретное распределение, сосредоточенное в точках $x = 0, 1, \dots, m$, с функцией плотности

$$f(x | \theta) = C_m^x p^x (1 - p)^{m-x},$$

зависящей от двумерного параметра $\theta = (m, p)$, первая компонента m которого может принимать только значения из множества $\mathbb{N} = \{1, 2, \dots\}$, а вторая компонента $p \in (0; 1)$. Функция плотности выборочного вектора

$$f_n(x^{(n)} | \theta) = \prod_{k=1}^n C_m^{x_k} \cdot p^{\sum_1^n x_k} (1 - p)^{nm - \sum_1^n x_k}.$$

Применение критерия (1) показывает, что для статистической структуры с параметрическим пространством $\Theta = \mathbb{N} \times (0; 1)$ достаточной статистикой может быть только весь выборочный вектор $X^{(n)}$, но если $\Theta = (0; 1)$ (значение параметра m известно), то $\sum_1^n X_k$ – достаточная статистика.

8⁰. *Распределение Коши* $C(a, b)$ имеет функцию плотности выборочного вектора

$$f_n(x^{(n)} | \theta) = \pi^{-n} b^{-n} \prod_{k=1}^n \left(1 + \left(\frac{x_k - a}{b} \right)^2 \right)^{-1},$$

и в силу критерия (1) его статистическая структура обладает только *тривиальной* достаточной статистикой $T = X^{(n)}$.

Мы не будем выписывать статистические структуры многомерных распределений в силу их чрезвычайной громоздкости, но нетрудно установить по аналогии с рассмотренными примерами, что у структуры мультиномиального распределения $\mathcal{M}(m, 1, \mathbf{p})$ с $m \geq 2$ исходами и вектором $\mathbf{p} = (p_1, \dots, p_m)$ вероятностей соответствующих исходов достаточным будет вектор, состоящий из частот этих исходов в мультиномиальной схеме испытаний, а у структуры многомерного нормального распределения $\mathcal{N}_m(\mu, \Lambda)$ достаточную статистику образуют вектор выборочных средних и выборочная ковариационная матрица.

На этом завершается вводная часть нашего курса математической статистики. Мы сделали постановку проблемы статистического вывода, провели классификацию основных статистических структур и теперь мы готовы к решению конкретных статистических проблем по оценке параметров распределения наблюдаемой случайной величины и проверке гипотез, касающихся структуры параметрического пространства этого распределения.

§3. Оценка параметров. Метод моментов

Лекция 5

Мы приступаем к решению статистической проблемы оценки неизвестного значения параметра θ , индексирующего семейство $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ возможных распределений наблюдаемой случайной величины X . Будут рассматриваться только конечномерные параметрические пространства $\Theta = \mathbf{R}^k$, $k \geq 1$. Информация о значении θ поступает к нам в виде выборочных данных $x^{(n)} = (x_1, \dots, x_n)$ – результатов наблюдений n независимых копий $X^{(n)} = (X_1, \dots, X_n)$ случайной величины X . Напомним, семейство \mathcal{P} мы назвали вероятностной моделью, а случайный вектор $X^{(n)}$ – случайной выборкой объема n .

В этой проблеме, о которой мы несколько раз упоминали в предыдущем параграфе, пространство решений \mathcal{D} совпадает с параметрическим пространством Θ , решающая функция $\delta = \delta(X^{(n)})$ – статистика с областью значений $\mathcal{T} = \Theta$ – называется *оценкой параметра θ* и обычно обозначается θ_n , $\hat{\theta}_n$, θ_n^* и тому подобное. Функции потерь $L(\theta, d)$ в проблеме оценивания обычно выбираются в виде неубывающей функции расстояния $|d - \theta|$ (в евклидовой метрике) между значением оценки $d = \hat{\theta}_n(x^{(n)})$ и истинным значением θ оцениваемого параметра.

Основная задача статистической теории оценивания состоит в построении оценки $\theta_n^ = \theta_n^*(X^{(n)})$, минимизирующей равномерно по $\theta \in \Theta$ функцию риска*

$$R(\theta; \hat{\theta}_n) = \mathbf{E}_\theta L(\theta, \hat{\theta}_n(X^{(n)})).$$

Таким образом, какова бы ни была статистическая оценка $\hat{\theta}_n$, для оценки θ_n^* с равномерно минимальным риском при любом $\theta \in \Theta$ справедливо неравенство $R(\theta; \theta_n^*) \leq R(\theta; \hat{\theta}_n)$.

Мы рассмотрим одно из решений этой задачи в случае оценки скалярного параметра ($\Theta = \mathbf{R}$) при квадратичной функции потерь $L(\theta, d) = (d - \theta)^2$, но сначала познакомимся с традиционно используемыми в статистической практике методами оценки параметров и изучим распределение этих оценок с целью вычисления их функции риска.

Конечно, далеко не все используемые на практике методы приводят к оптимальным оценкам, иногда бывает трудно найти оценку, обладающую хоть какими-нибудь привлекательными свойствами. Понятно, что считать оценкой любое измеримое отображение выборочного пространства \mathcal{X}^n в параметрическое пространство Θ не совсем разумно, и поэтому мы введем

некоторые условия, которым должна удовлетворять статистика $\hat{\theta}_n$, чтобы претендовать на роль *оценки*. Разрабатывая в дальнейшем методы оценивания и предлагая конкретные оценки, мы всегда будем проверять выполнимость этих условий.

Определение 3.1. Оценка $\hat{\theta}_n$ параметра θ называется *состоятельной*, если

$$\hat{\theta}_n(X^{(n)}) \xrightarrow{P} \theta$$

при любом $\theta \in \Theta$, когда объем выборки $n \rightarrow \infty$. Оценка $\hat{\theta}_n$ называется *несмещенной в среднем*, если

$$\mathbf{E}_\theta \hat{\theta}_n(X^{(n)}) = \theta,$$

каково бы ни было значение $\theta \in \Theta$.

Напомним, что $\hat{\theta}_n(X^{(n)}) \xrightarrow{P} \theta$ означает, что для любого $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P_\theta \left(\left| \hat{\theta}_n(X^{(n)}) - \theta \right| > \varepsilon \right) = 0,$$

или, что то же,

$$\lim_{n \rightarrow \infty} P_\theta \left(\left| \hat{\theta}_n(X^{(n)}) - \theta \right| \leq \varepsilon \right) = 1. \quad (1)$$

Здесь, как обычно, в случае векторного параметра θ запись $|\theta_1 - \theta_2|$ означает расстояние между точками θ_1 и θ_2 эвклидова пространства Θ .

В предыдущем параграфе мы показали, что выборочные моменты

$$a_i = \frac{1}{n} \sum_1^n X_j^i$$

являются состоятельными оценками соответствующих “теоретических” моментов $\alpha_i = \mathbf{E}_\theta X^i$, которые являются функциями оцениваемого параметра: $\alpha_i = \alpha_i(\theta)$, $i = 1, 2, \dots$. Этот результат указывает нам довольно простой метод построения состоятельных оценок в случае существования у распределения P_θ наблюдаемой случайной величины X момента порядка k , где k – число компонент $\theta_1, \dots, \theta_k$ оцениваемого параметрического вектора θ .

Приравняем теоретические моменты выборочным и разрешим полученную таким образом систему уравнений

$$\alpha_i(\theta_1, \dots, \theta_k) = a_i, \quad i = 1, \dots, k$$

относительно переменных $\theta_1, \dots, \theta_k$. Любое решение

$$\hat{\theta}_n(\mathbf{a}) = \left(\hat{\theta}_{1n}(\mathbf{a}), \dots, \hat{\theta}_{kn}(\mathbf{a}) \right), \quad \mathbf{a} = (a_1, \dots, a_k),$$

этой системы называется *оценкой θ по методу моментов*, и прежде чем исследовать свойства таких оценок, рассмотрим несколько примеров на применения *метода моментов*.

В курсе теории вероятностей, изучая новые вероятностные модели, мы всегда вычисляли их моментные характеристики. Например, мы знаем, что средние значения двухточечного распределения $B(1, \theta)$, распределения Пуассона $P(\theta)$ и показательного распределения $E(\theta)$ равны θ . Следовательно, выборочное среднее \bar{X} есть оценка по методу моментов параметра θ любого из этих распределений. Легко видеть, что эта оценка состоятельна и несмещена. Точно так же у нормального распределения $\mathcal{N}(\mu, \sigma^2)$ параметр μ означает среднее значение, а σ^2 – дисперсию этого распределения. Следовательно, выборочное среднее \bar{X} и выборочная дисперсия S^2 есть состоятельные оценки соответствующих компонент μ и σ^2 параметрического вектора $\theta = (\mu, \sigma^2)$. Исправляя смещение оценки S^2 компоненты σ^2 , получаем несмещенную оценку θ . Замечательно то, что все оценки являются достаточными статистиками, и это обстоятельство, как будет видно в дальнейшем, предопределяет их оптимальные свойства. Распределение оценки \bar{X} легко получить, используя теоремы сложения для распределений B , P , E и \mathcal{N} , распределение же S^2 при выборе из нормального распределения мы найдем несколько позже.

Рассмотрим теперь примеры, в которых приходится решать систему уравнений, и найденные оценки по методу моментов не являются функциями достаточных статистик.

Пример 3.1. *Оценка параметров биномиального распределения $B(m, p)$.* Проблема состоит в оценке обеих компонент m и p двумерного параметра $\theta = (m, p)$. Из курса теории вероятностей нам известно, что среднее значение биномиального распределения равно mp , а дисперсия – $mp(1 - p)$. Приравнивая эти теоретические моменты их выборочным аналогам, получаем систему для определения оценок по методу моментов:

$$mp = \bar{X}, \quad mp(1 - p) = S^2.$$

Разделив второе уравнение на первое, находим оценку

$$\hat{p}_n = \frac{\bar{X} - S^2}{\bar{X}}$$

параметра p , после чего, обращаясь к первому уравнению, находим оценку

$$\hat{m}_n = \frac{\bar{X}^2}{\bar{X} - S^2}$$

параметра m . Легко показать, что эти оценки обладают свойством состоятельности (общий метод доказательства таких утверждений смотрите в приведенной ниже теореме 3.1), но при малых n велика вероятность получить отрицательные значения оценок, оценка параметра m , как правило, не будет целым числом, наконец, можно показать, что оценка $p_n^* = \bar{X}$ параметра p будет обладать меньшим квадратичным риском, чем оценка \hat{p}_n . Все это, конечно, печально, однако другие методы, приводящие к более точным оценкам, обладают значительными вычислительными трудностями.

Пример 3.2. Оценка параметров гамма-распределения $G(\lambda, a)$. У этого двухпараметрического распределения среднее равно λa , а дисперсия – λa^2 . Решение системы уравнений

$$\lambda a = \bar{X}, \quad \lambda a^2 = S^2$$

дает оценки

$$\hat{a}_n = \frac{S^2}{\bar{X}}, \quad \hat{\lambda}_n = \frac{\bar{X}^2}{S^2},$$

которые, как и в предыдущем примере, не являются функциями достаточной статистики

$$T(X^{(n)}) = \left(\sum_1^n X_k, \prod_1^n X_k \right),$$

и как показывают не совсем простые вычисления, их риски далеки от возможного минимума. Тем не менее, очевидная вычислительная простота оценок параметров гамма-распределения по методу моментов обеспечивает их популярность в практических применениях.

Изучим теперь асимптотические свойства оценок по методу моментов – установим условия их состоятельности и исследуем поведение их распределений при больших объемах выборок. Для простоты мы ограничимся случаем одномерного параметра θ , оценка которого определяется решением уравнения $\mu(\theta) = \mathbf{E}_\theta X = \bar{X}$, и предположим, что это уравнение имеет единственное решение $\hat{\theta}_n = h(\bar{X})$. Понятно, что $h(\cdot) = \mu^{-1}(\cdot)$, так что $h(\mu(\theta)) \equiv \theta$. О возможности распространения наших результатов на случай векторного θ мы поговорим отдельно.

Теорема 3.1. Если наблюдаемая случайная величина X имеет конечное среднее значение $\mu = \mu(\theta)$ и функция $h(\cdot)$ непрерывна в области значений выборочного среднего \bar{X} , то $\hat{\theta}_n = h(\bar{X})$ является состоятельной оценкой параметра θ по методу моментов.

Доказательство. Мы воспользуемся формулой (1) в определении состоятельности оценки и покажем, что для любых $\varepsilon > 0$ и $\alpha > 0$ существует такое $N(\varepsilon, \alpha)$, что для всех $n > N(\varepsilon, \alpha)$ вероятность

$$P_{\theta} (|h(\bar{X}) - \theta| \leq \varepsilon) > 1 - \alpha. \quad (2)$$

Поскольку $h(\mu) = h(\mu(\theta)) = \theta$ и $\bar{X} \xrightarrow{P} \mu$, то нам достаточно показать, что свойство (или определение) непрерывности функции: $h(x) \rightarrow h(\mu)$ при $x \rightarrow \mu$, остается справедливым при замене обычной сходимости " \rightarrow " на сходимость по вероятности " \xrightarrow{P} ", то есть $\bar{X} \xrightarrow{P} \mu$ влечет $h(\bar{X}) \xrightarrow{P} h(\mu)$. Это почти очевидно, поскольку событие, состоящее в попадании в окрестность нуля случайной величины $|\bar{X} - \mu|$, влечет аналогичное событие для случайной величины $|h(\bar{X}) - h(\mu)|$, но все же проведем строгое доказательство на языке " $\varepsilon - \delta$ ".

Так как $\bar{X} \xrightarrow{P} \mu(\theta)$, а $h(\cdot)$ – непрерывная функция, то найдутся такие $\delta = \delta(\varepsilon, \alpha)$ и $N = N(\varepsilon, \alpha)$, что

$$P (|\bar{X} - \mu(\theta)| < \delta) > 1 - \alpha \quad (3)$$

для всех $n > N$ и событие $|\bar{X} - \mu(\theta)| < \delta$ повлечет событие

$$|h(\bar{X}) - h(\mu(\theta))| = |h(\bar{X}) - \theta| \leq \varepsilon.$$

В силу этого неравенство (2) становится следствием неравенства (3). Теорема доказана.

Анализ доказательства показывает, что теорема состоятельности остается справедливой в случае векторного параметра θ , если воспользоваться определением непрерывности векторной функции от векторного аргумента, связав его с расстояниями в евклидовых пространствах значений функции и ее аргумента.

Обратимся теперь к асимптотическому анализу распределения оценки $\hat{\theta}_n = h(\bar{X})$ при $n \rightarrow \infty$. Понятно, что в данном случае употребление термина "оценка" применительно к функции $h(\bar{X})$ ничего особенно не добавляет – речь идет просто об асимптотическом распределении статистики, имеющей вид функции от выборочного среднего.

Теорема 3.2. Если $X^{(n)}$ – случайная выборка из распределения с конечными средним значением μ и дисперсией σ^2 , функция $h(x)$ обладает

ограниченной второй производной $h''(x)$ в некоторой окрестности точки $x = \mu$ и $h'(\mu) \neq 0$, то

$$\lim_{n \rightarrow \infty} P(\sqrt{n}(h(\bar{X}) - h(\mu)) < x) = \Phi\left(\frac{x}{\sigma|h'(\mu)|}\right). \quad (4)$$

где $\Phi(\cdot)$ – функция распределения стандартного нормального закона.

Доказательство. Понятно, что мы должны воспользоваться центральной предельной теоремой (§14 курса ТВ) применительно к статистике $\bar{X} = \frac{1}{n} \sum_1^n X_k$:

$$\lim_{n \rightarrow \infty} P(\sqrt{n}(\bar{X} - \mu) < x) = \Phi\left(\frac{x}{\sigma}\right). \quad (5)$$

Стандартный прием использования этой теоремы при асимптотическом анализе функций от сумм независимых, одинаково распределенных случайных величин состоит в “линеризации” таких функций с помощью формулы Тейлора. В нашем случае мы разлагаем функцию $h(\cdot)$ в окрестности точки $\bar{X} = \mu$, используя только два члена разложения:

$$h(\bar{X}) = h(\mu) + (\bar{X} - \mu)h'(\mu) + \frac{(\bar{X} - \mu)^2}{2!}h''(\mu + \lambda(\bar{X} - \mu)),$$

где $0 \leq \lambda \leq 1$.

Перепишем это разложение в виде

$$\sqrt{n}(h(\bar{X}) - h(\mu)) = \sqrt{n}(\bar{X} - \mu)h'(\mu) + \frac{\sqrt{n}(\bar{X} - \mu)^2}{2!}h''(\mu + \lambda(\bar{X} - \mu)),$$

представив тем самым случайную величину $\sqrt{n}(h(\bar{X}) - h(\mu))$ (см. формулу (4)) в виде суммы двух случайных величин, первая из которых в силу формулы (5) имеет предельное нормальное распределение, указанное в правой части (4), а вторая сходится по вероятности к нулю. Действительно, $h''(\mu + \lambda(\bar{X} - \mu))$ по условию теоремы ограничено с вероятностью сколь угодно близкой к единице, начиная с некоторого n . В силу неравенства Чебышева (предложение 6.2 курса ТВ) для любого $\varepsilon > 0$ вероятность

$$P(\sqrt{n}(\bar{X} - \mu)^2 > \varepsilon) \leq \frac{\mathbf{E}\sqrt{n}(\bar{X} - \mu)^2}{\varepsilon} = \frac{\sqrt{n}\mathbf{D}\bar{X}}{\varepsilon} = \frac{\sigma^2}{\varepsilon\sqrt{n}} \rightarrow 0,$$

и поэтому стоящий перед $h''/2!$ множитель, а с ним и все второе слагаемое сходятся по вероятности к нулю. Доказательство теоремы завершается

ссылкой на предложение 11.1 курса ТВ: если одна последовательность случайных величин имеет невырожденное предельное распределение F , а вторая сходится по вероятности к нулю, то предельное распределение суммы этих последовательностей совпадает с F .

Итак, если $h(\bar{X})$ – оценка $\hat{\theta}_n$ параметра θ по методу моментов, то формула (4) теоремы 3.2 принимает вид

$$\lim_{n \rightarrow \infty} P \left(\sqrt{n} \left(\hat{\theta}_n - \theta \right) < x \right) = \Phi \left(\frac{x}{\sigma |h'(\mu)|} \right).$$

Однако асимптотическая формула (4) может применяться не только к вычислению надежности оценок по методу моментов, но и к аппроксимации распределений функций от выборочного среднего, представляющих некоторые оценки функций от “теоретического” среднего.

Пример 3.3. Оценка вероятности значений пуассоновской случайной величины. По выборке объема n из распределения Пуассона $P(\lambda)$ требуется оценить вероятность того, что за единицу времени произойдет не более одного события. Эта вероятность

$$h(\lambda) = P(X \leq 1) = (\lambda + 1) e^{-\lambda}.$$

Поскольку в распределении Пуассона $\lambda = \mathbf{E}X$, то естественно за оценку $h(\lambda)$ принять статистику $h(\bar{X})$ – функцию от выборочного среднего, удовлетворяющую всем условиям теоремы 3.2. Воспользуемся утверждением этой теоремы для приближенного вычисления надежности данной оценки при заданной точности Δ .

Надежность оценки $h(\bar{X})$

$$H(\lambda; h(\bar{X})) = P \left(|h(\lambda) - h(\bar{X})| \leq \Delta \right).$$

Так как

$$h'(\lambda) = -\lambda e^{-\lambda},$$

и для распределения Пуассона $\sigma^2 = \lambda$, то применение формулы (4) в обозначениях $\Delta = \varepsilon / \sqrt{n}$ приводит к следующей аппроксимации надежности оценки:

$$H(\lambda; h(\bar{X})) \approx 2\Phi \left(\varepsilon \lambda^{-3/2} e^{\lambda} \right) - 1.$$

Полученная аппроксимация указывает на хорошую надежность оценки при очень малых и достаточно больших значениях параметра λ ; минимум надежности достигается при $\lambda = 2/3$.

В том случае, когда оценка имеет вид функции от двух и более выборочных моментов, метод асимптотического анализа ее распределения тот же. Например, пусть $\hat{\theta}_n = h(a_1, a_2)$, и функция h удовлетворяет условиям, аналогичным требованиям к h в теореме 3.2. Используя формулу Тейлора, представим h в окрестности точки (α_1, α_2) в следующем виде:

$$h(a_1, a_2) = h(\alpha_1, \alpha_2) + (a_1 - \alpha_1)h'_1(\alpha_1, \alpha_2) + \\ (a_2 - \alpha_2)h'_2(\alpha_1, \alpha_2) + O_p(|a - \alpha|^2),$$

где $a = (a_1, a_2)$, $\alpha = (\alpha_1, \alpha_2)$. Легко проверяется, что $\sqrt{n}|a - \alpha|^2 \xrightarrow{P} 0$, так что случайная величина $\sqrt{n}(h(a_1, a_2) - h(\alpha_1, \alpha_2))$ асимптотически нормальна с параметрами, которые выражаются через первые четыре момента наблюдаемой случайной величины X .

§4. Оценка параметров. Метод максимального правдоподобия

Лекция 6

Мы приступаем к изучению более точного метода оценки неизвестного значения параметра. Он превосходит метод моментов и при наличии достаточных статистик дает оптимальные оценки с точки зрения квадратичного риска. Более того, при выполнении определенных условий регулярности этот метод приводит к асимптотически ($n \rightarrow \infty$) оптимальным оценкам для широкого класса вероятностных моделей и практически при любых функциях потерь.

Идея метода состоит в математической формализации “разумного” поведения человека в условиях неопределенности. Представим себе ситуацию, что мы ожидаем появления одного из нескольких событий, вероятности которых нам неизвестны и нас интересуют не столько значения этих вероятностей, сколько то событие, которое происходит наиболее часто. Ситуация осложняется тем, что мы располагаем всего одним испытанием, в результате которого произошло некоторое событие A . Конечно, мы примем решение, что A обладает наибольшей вероятностью, и вряд ли можно предложить нечто более разумное, чем такое правило принятия решения.

В этом и состоит *принцип максимального правдоподобия*, который буквально пронизывает всю теорию оптимального статистического вывода. Применение этого принципа к проблеме оценки параметров приводит к следующему статистическому правилу: *если $x^{(n)}$ – результат наблюдения случайной выборки $X^{(n)}$, то за оценку параметра следует брать то его значение, при котором результат $x^{(n)}$ обладает наибольшим правдоподобием.*

Вы спросите, что такое “правдоподобие” результата $x^{(n)}$? Давайте формализуем это понятие.

Если наблюдается дискретная случайная величина, то естественно называть правдоподобием результата $x^{(n)}$ при фиксированном значении параметра θ вероятность его наблюдения в статистическом эксперименте. Но в дискретном случае эта вероятность совпадает со значением функции плотности в точке $x^{(n)}$: $P_\theta(X^{(n)} = x^{(n)}) = f_n(x^{(n)} | \theta)$. Следовательно, оценка по методу максимального правдоподобия определяется точкой достижения максимума у функции плотности случайной выборки, то есть

$$\hat{\theta}_n(X^{(n)}) = \arg \max_{\theta \in \Theta} f_n(X^{(n)} | \theta). \quad (1)$$

Рассмотрим сразу же простой пример. Пусть $X^{(n)}$ – выборка в схеме Бернулли, и мы оцениваем вероятность θ успешного исхода. В этой модели

$$f(X^{(n)} | \theta) = \theta \sum_1^n X_k (1 - \theta)^{n - \sum_1^n X_k}.$$

Дифференцируя эту функцию по θ и приравнявая производную нулю, находим оценку максимального правдоподобия

$$\hat{\theta}_n = (1/n) \sum_1^n X_k.$$

Это – давно знакомая нам оценка вероятности успеха в испытаниях Бернулли, которую мы получили с помощью моментов и постоянно использовали при иллюстрации закона больших чисел.

Теперь определим правдоподобие в случае выбора из непрерывного распределения с функцией плотности (по мере Лебега)

$$f_n(x^{(n)} | \theta), \quad x^{(n)} \in \mathbb{R}^n, \quad \theta \in \Theta.$$

Пусть $x^{(n)}$ – совокупность выборочных данных, то есть точка в n -мерном выборочном пространстве \mathbb{R}^n . Окружим эту точку прямоугольным параллелепипедом малого размера, скажем,

$$V_\varepsilon = \prod_1^n [x_k - \varepsilon/2; x_k + \varepsilon/2].$$

В силу теоремы о среднем для кратного интеграла вероятность того, что выборочный вектор попадет в этот параллелепипед

$$P\left(X^{(n)} \in V_\varepsilon\right) \sim f_n(x^{(n)} | \theta) \cdot \varepsilon^n,$$

когда $\varepsilon \rightarrow 0$. Если трактовать эту вероятность, как правдоподобие результата $x^{(n)}$, которое, конечно, зависит от выбора малого ε , мы видим, что проблема максимизации правдоподобия сводится к проблеме отыскания точки достижения максимума по всем $\theta \in \Theta$ у функции плотности f_n . Таким образом, и в случае непрерывного распределения разумно назвать правдоподобием результата $x^{(n)}$ при фиксированном значении параметра θ опять-таки величину функции плотности выборки, то есть $f_n(x^{(n)} | \theta)$, и определить оценку максимального правдоподобия той же формулой (1).

Рассмотрим пример на построение такой оценки в случае выбора из непрерывного распределения. Пусть наблюдается случайная величина $X \sim \mathcal{N}(\mu, \sigma^2)$, так что функция плотности выборки

$$f_n(x^{(n)} | \theta) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_1^n (x_k - \mu)^2 \right\},$$

где $\theta = (\mu, \sigma)$ – двумерный параметр, значение которого нам неизвестно. В соответствии с формулой (1) необходимо отыскать точку достижения максимума функции $f_n(X^{(n)} | \mu, \sigma)$ по переменным $\mu \in \mathbb{R}$ и $\sigma \in \mathbb{R}_+$. Естественно, логарифм этой функции имеет те же точки экстремума, что и сама функция, но логарифмирование упрощает выкладки, поэтому ищем максимум функции

$$\mathcal{L}(\theta | X^{(n)}) = \ln f_n(X^{(n)} | \theta) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_1^n (X_k - \mu)^2.$$

Составляем уравнения, определяющие точки экстремума:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu} &= \frac{1}{\sigma^2} \sum_1^n (X_k - \mu) = 0, \\ \frac{\partial \mathcal{L}}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_1^n (X_k - \mu)^2 = 0. \end{aligned}$$

Из первого уравнения сразу находим оценку параметра μ : $\hat{\mu}_n = \bar{X}$. Подставляя \bar{X} вместо μ во второе уравнение, находим оценку σ : $\hat{\sigma}_n = S$ (выборочное стандартное отклонение). Очевидно, (\bar{X}, S) – точка максимума. Таким образом, метод максимального правдоподобия приводит к тем же оценкам \bar{X} и S^2 параметров μ и σ^2 , что и метод моментов.

Теперь дадим строгое определение правдоподобия и рассмотрим еще несколько примеров, в которых метод максимального правдоподобия дает оценки, отличные от метода моментов.

Определение 4.1. Случайная функция

$$L(\theta | X^{(n)}) = \prod_{i=1}^n f(X_i | \theta)$$

на параметрическом пространстве Θ называется *функцией правдоподобия*, а значение ее реализации $L(\theta_0 | x^{(n)})$ при результате наблюдения $X^{(n)} = x^{(n)}$ и фиксированном $\theta = \theta_0$ – *правдоподобием значения θ_0 при результате $x^{(n)}$* . Любая точка $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ (статистика) параметрического пространства Θ , доставляющая абсолютный максимум функции правдоподобия, называется *оценкой максимального правдоподобия* параметра θ .

Поскольку функция правдоподобия представляет собой произведение функций от θ , то при отыскании ее максимума методами дифференциаль-

ного исчисления удобнее иметь дело с логарифмом этой функции. Естественно, точки экстремума у *функции логарифмического правдоподобия*

$$\mathcal{L}(\theta | X^{(n)}) = \sum_{i=1}^n \ln f(X_i | \theta)$$

те же, что и у функции L , но если функция $L(\cdot | x^{(n)})$ имеет непрерывные частные производные по компонентам $\theta_1, \dots, \theta_k$ параметрического вектора θ , то проще дифференцировать \mathcal{L} чем L . В этом случае система уравнений

$$\frac{\partial \mathcal{L}(\theta | X^{(n)})}{\partial \theta_i} = 0, \quad i = 1, \dots, k \quad (2)$$

называется *уравнениями правдоподобия*. Это еще одна разновидность так называемых *оценочных уравнений*, – в предыдущем параграфе мы имели дело с уравнениями метода моментов.

Любое решение системы уравнений (2), доставляющее максимум функции $\mathcal{L}(\cdot | X^{(n)})$, может рассматриваться как оценка θ по методу максимального правдоподобия. Мы не будем изучать случаи, когда система (2) имеет несколько решений с возможно одинаковыми значениями функции правдоподобия в этих точках, так что требуются дополнительные априорные знания относительно вероятностной модели, позволяющие выбрать одно из этих решений. Во всех рассмотренных ниже примерах оценка максимального правдоподобия единственна.

Пример 4.1. Оценка параметра положения равномерного распределения $U(0, \theta)$. Равномерное на отрезке $[0; \theta]$ распределение имеет функцию плотности $f(x | \theta) = \theta^{-1}$, если $0 \leq x \leq \theta$, и $f(x | \theta) = 0$ вне этого отрезка. Следовательно, функция $L(\theta | X^{(n)})$ отлична от нуля и равна θ^{-n} только в области

$$\theta \geq X_{(n)} = \max_{1 \leq k \leq n} X_k.$$

Ее максимум по θ достигается в граничной точке $\theta = X_{(n)}$, так что наибольшее значение $X_{(n)}$ выборки $X^{(n)}$ есть оценка максимального правдоподобия параметра θ .

Легко видеть, что оценка θ по методу моментов равна $2\bar{X}$. Эта оценка на порядок хуже оценки максимального правдоподобия с точки зрения квадратичного риска

$$R(\theta; \hat{\theta}_n) = \mathbf{E}_\theta \left(\hat{\theta}_n(X^{(n)}) - \theta \right)^2.$$

Простые вычисления соответствующих математических ожиданий показывают, что $R(\theta; 2\bar{X}) = O(n^{-1})$, в то время как $R(\theta; X_{(n)}) = O(n^{-2})$. Данный пример интересен тем, что здесь функция правдоподобия не имеет гладкого максимума, и именно это обстоятельство, как будет видно в дальнейшем, обеспечивает такое различное поведение риска рассматриваемых оценок.

Пример 4.2. Оценка параметров гамма-распределения $G(a, \lambda)$. У этого распределения функция плотности

$$f(x | \theta) = \frac{1}{a^\lambda \Gamma(\lambda)} x^{\lambda-1} \exp\left\{-\frac{x}{a}\right\}, \quad x > 0, \quad \theta = (a, \lambda),$$

отлична от нуля только на положительной полуоси, и логарифмическое правдоподобие

$$\mathcal{L}(a, \lambda | X^{(n)}) = -n\lambda \ln a - n \ln \Gamma(\lambda) + (\lambda - 1) \sum_1^n \ln X_k - \frac{1}{a} \sum_1^n X_k.$$

Составляем уравнения правдоподобия:

$$\frac{\partial \mathcal{L}}{\partial a} = -\frac{n\lambda}{a} + \frac{1}{a^2} \sum_1^n X_k = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -n \ln a - n\psi(\lambda) + \sum_1^n \ln X_k = 0,$$

где $\psi(\lambda) = d \ln \Gamma(\lambda) / d\lambda$ – так называемая пси-функция Эйлера. Исключая из первого уравнения параметр a и подставляя полученный результат во второе, получаем трансцендентное уравнение

$$\ln \lambda - \psi(\lambda) = \ln \bar{X} - \frac{1}{n} \sum_1^n \ln X_k,$$

которое в силу свойства монотонности функции $\ln \lambda - \psi(\lambda)$ имеет единственное решение. При численном решении этого уравнения может оказаться полезной асимптотическая ($\lambda \rightarrow \infty$) формула

$$\ln \lambda - \psi(\lambda) = \frac{1}{2\lambda} + \frac{1}{12\lambda^2} + O\left(\frac{1}{\lambda^4}\right).$$

Пример 4.3. Оценка параметров структурированного среднего при нормальном распределении отклика. Данная задача весьма часто возникает при калибровке шкалы прибора. Две переменные x и y связаны линейным соотношением $y = a + bx$, и для градуировки значений y на шкале прибора необходимо знать значения параметров a и b этой зависимости. Однако для каждого стандартного фиксированного значения x прибор замеряет значение y с ошибкой, так что замеры происходят в рамках вероятностной модели $Y = a + bx + \xi$, где ошибка измерения (случайная величина) ξ имеет нормальное распределение с нулевым средним и некоторой дисперсией σ^2 , значение которой, как правило, также неизвестно. Случайная величина Y обычно называется *откликом* на значение *регрессора* x ; ее распределение при фиксированном x очевидно нормально ($a + bx, \sigma^2$).

Для оценки a и b производится n измерений отклика y_1, \dots, y_n при некоторых фиксированных значениях x_1, \dots, x_n регрессора x , оптимальный выбор которых, обеспечивающий наибольшую точность и надежность калибровки, составляет самостоятельную задачу особой области математической статистики – *планирование регрессионных экспериментов*. Мы будем предполагать, что значения x_1, \dots, x_n априори фиксированы. В таком случае значения y_1, \dots, y_n представляют реализации n независимых случайных величин Y_1, \dots, Y_n , и $Y_k \sim \mathcal{N}(a + bx_k, \sigma^2)$, $k = 1, \dots, n$. Совместная функция плотности Y_1, \dots, Y_n равна

$$f_n(y^{(n)} | a, b, \sigma) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_1^n (y_k - a - bx_k)^2 \right\},$$

так что логарифмическая функция правдоподобия, необходимая для оценки параметров a , b и σ методом максимального правдоподобия имеет вид

$$\mathcal{L}(a, b, \sigma | Y^{(n)}) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_1^n (Y_k - a - bx_k)^2.$$

Вычисляя производные этой функции по переменным a , b и σ , получаем уравнения правдоподобия

$$\sum_1^n (Y_k - a - bx_k) = 0,$$

$$\sum_1^n x_k (Y_k - a - bx_k) = 0,$$

$$n\sigma^2 - \sum_1^n (Y_k - a - bx_k)^2 = 0.$$

Конечно, это очень простая система уравнений, решение которой не может вызывать каких-либо затруднений, и мы сразу пишем оценки максимального правдоподобия

$$\hat{a}_n = \bar{Y} - \frac{m_{xY}}{s_x^2} \bar{x}, \quad \hat{b}_n = \frac{m_{xY}}{s_x^2}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_1^n (Y_k - \hat{a}_n - \hat{b}_n x_k)^2,$$

где

$$\bar{x} = \frac{1}{n} \sum_1^n x_k, \quad \bar{Y} = \frac{1}{n} \sum_1^n Y_k, \quad s_x = \frac{1}{n} \sum_1^n (x_k - \bar{x})^2, \quad S_Y = \frac{1}{n} \sum_1^n (Y_k - \bar{Y})^2,$$

$$m_{xY} = \frac{1}{n} \sum_1^n (x_k - \bar{x})(Y_k - \bar{Y}).$$

Легко видеть, что оценки по методу максимального правдоподобия параметров a и b совпадают с их оценками по *методу наименьших квадратов*. В этом методе “выравнивания” экспериментальных данных оценки ищутся из условия минимизации суммы квадратов *невязок*: $\sum_1^n (Y_k - a - bx_k)^2$, причем под невязкой понимается разность между откликом Y и его “теоретическим” средним значением $a + bx$.

Пример 4.4. *Оценка параметров двумерного нормального распределения: задачи регрессии и прогноза.* Оценка по методу максимального правдоподобия пяти параметров $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$ двумерного нормального распределения с функцией плотности

$$f(x, y | \theta) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right) \right\}$$

не представляет особой технической сложности. Эти оценки совпадают с оценками по методу моментов и, таким образом, равны выборочным аналогам тех характеристик двумерного нормального распределения, которые соответствуют указанным пяти параметрам:

$$\hat{\mu}_{1,n} = \bar{X}, \quad \hat{\mu}_{2,n} = \bar{Y}, \quad \hat{\sigma}_{1,n}^2 = S_X^2, \quad \hat{\sigma}_{2,n}^2 = S_Y^2, \quad \hat{\rho}_n = r.$$

Формулы для вычисления выборочных средних \bar{X} и \bar{Y} , выборочных дисперсий S_1^2 и S_2^2 , а также выборочного коэффициента корреляции r приведены в §2.

Полученные оценки часто используются для оценки параметров линейного прогноза $Y = a + bX$ значений случайной величины Y по результатам наблюдений X . В случае нормального распределения линейный прогноз обладает свойством оптимальности с точки зрения малости средней квадратичной ошибки и совпадает с кривой средней квадратичной регрессии (см. предложение 10.3 курса ТВ)

$$y = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x - \mu_1).$$

Однако формальная подгонка прогностической кривой с помощью прямой линии используется и вне рамок нормальной модели, и в этом случае оценки

$$\hat{a}_n = \bar{Y} - r \frac{S_2}{S_1} \bar{X}, \quad \hat{b}_n = r \frac{S_2}{S_1}$$

совпадают с оценками по *методу наименьших квадратов*: минимизируется, как и в примере 4.3, сумма квадратов невязок

$$\sum_1^n (Y_k - a - bX_k)^2.$$

Хотя оценки в обоих примерах имеют одинаковый вид, но решаемые в них статистические проблемы весьма различны: в примере 4.3 оценивались параметры некоторой функциональной зависимости с ошибками в наблюдениях отклика, в то время как в примере 4.4 решается задача выявления корреляционной связи и использования этой связи для прогноза.

Лекция 7

Исследуем теперь асимптотические свойства оценок по методу максимального правдоподобия.

Начнем с выяснения достаточных условий состоятельности этих оценок. Такие ограничения на вероятностную модель обычно называются *условиями регулярности*, и в данном случае они имеют следующий вид.

(R1) Параметрическое пространство Θ есть открытый интервал на прямой \mathbb{R} .

- (R2) Носитель \mathcal{X} распределения P_θ наблюдаемой случайной величины X не зависит от θ , то есть все множества $\mathcal{X} = \{x : f(x|\theta) > 0\}$ можно считать одинаковыми, каково бы ни было $\theta \in \Theta$.
- (R3) Распределения P_θ различны при разных θ , то есть при любых $\theta_1 \neq \theta_2$, $\theta_1, \theta_2 \in \Theta$, имеет место тождество

$$\mu\{x : x \in \mathcal{X}, f(x|\theta_1) = f(x|\theta_2)\} = 0,$$

где μ – мера, по которой вычисляется плотность $f(x|\theta)$ распределения P_θ .

Доказательство состоятельности оценок максимального правдоподобия, как и оценок по методу моментов, опирается на закон больших чисел, но при этом используется следующее достаточно простое, но играющее большую роль в теории вероятностей, неравенство.

Лемма 4.1. (неравенство Йенсена) Пусть X – случайная величина с конечным математическим ожиданием. Если функция $g(\cdot)$ дважды дифференцируема и выпукла ($g'' > 0$) на некотором интервале, содержащем носитель распределения X , и математическое ожидание $\mathbf{E}g(X)$ существует, то справедливо неравенство $\mathbf{E}g(X) \geq g(\mathbf{E}X)$, причем знак равенства достигается тогда и только тогда, когда распределение X сосредоточено в одной точке ($X = \text{const.}$).

Доказательство. Так как функция g дважды дифференцируема, то справедливо следующее представление Тейлора в окрестности точки $\mu = \mathbf{E}X$:

$$g(X) = g(\mu) + (X - \mu)g'(\mu) + (X - \mu)^2g''(\mu + \lambda(X - \mu))/2, \quad 0 < \lambda < 1.$$

Вычисляя математическое ожидание от обеих частей этого равенства, получаем

$$\mathbf{E}g(X) = g(\mathbf{E}X) + \mathbf{E}(X - \mu)^2g''(\mu + \lambda(X - \mu))/2 \geq g(\mathbf{E}X).$$

Знак равенства возможен только в случае

$$\mathbf{E}(X - \mu)^2g''(\mu + \lambda(X - \mu)) = 0.$$

Но поскольку $g'' > 0$, то последнее равенство с необходимостью влечет $(X - \mu)^2 = 0$, то есть $X = \text{const.}$

Покажем теперь, что справедлива

Теорема 4.1 (состоятельность). *Если функция логарифмического правдоподобия*

$$\mathcal{L}(\theta | X^{(n)}) = \sum_{k=1}^n \ln f(X_k | \theta) \quad (3)$$

имеет единственный максимум, то при выполнении условий регулярности (R1)–(R3) точка $\hat{\theta}_n$ достижения максимума этой функции (оценка максимального правдоподобия) является состоятельной оценкой параметра θ .

Доказательство. Покажем, что для любого фиксированного $\theta_0 \in \Theta$ и любого $\varepsilon > 0$ вероятность $P_{\theta_0}(|\hat{\theta}_n - \theta_0| < \varepsilon) \rightarrow 1$.

Если θ_0 – истинное значение параметра θ , то в силу условия (R1) θ_0 – внутренняя точка Θ . Тогда сформулированная выше задача состоит в доказательстве следующего утверждения: в некоторой ε -окрестности $(\theta_0 - \varepsilon; \theta_0 + \varepsilon)$ функция $\mathcal{L}(\cdot | X^{(n)})$ обладает локальным максимумом с вероятностью, стремящейся к единице при $n \rightarrow \infty$.

Но если происходит событие

$$A_n = \left\{ \mathcal{L}(\theta_0 | X^{(n)}) > \mathcal{L}(\theta_0 \pm \varepsilon | X^{(n)}) \right\},$$

то внутри этой окрестности имеется точка максимума, и нам остается только показать, что $P_{\theta_0}(A_n) \rightarrow 1$, ибо

$$P_{\theta_0}(|\hat{\theta}_n - \theta_0| < \varepsilon) \geq P_{\theta_0}(A_n).$$

Используя условие (R2) и вид функции \mathcal{L} (см. (3)), представим неравенство, определяющее событие A_n , в виде

$$\frac{1}{n} \sum_{k=1}^n \ln \frac{f(X_k | \theta_0 \pm \varepsilon)}{f(X_k | \theta_0)} < 0.$$

В силу закона больших чисел Хинчина левая часть этого неравенства сходится по вероятности к

$$\mathbf{E}_{\theta_0} \ln \frac{f(X_k | \theta_0 \pm \varepsilon)}{f(X_k | \theta_0)}, \quad (4)$$

и для доказательства утверждения достаточно показать, что это математическое ожидание строго меньше нуля (кстати, докажите сами, что при

справедливости условий теоремы математическое ожидание (4) всегда существует, в противном случае закон больших чисел Хинчина не применим).

Так как $g(x) = -\ln x$ – выпуклая функция, то в силу неравенства Йенсена

$$\begin{aligned} \mathbf{E}_{\theta_0} \ln \frac{f(X | \theta_0 \pm \varepsilon)}{f(X | \theta_0)} &\leq \ln \mathbf{E}_{\theta_0} \frac{f(X | \theta_0 \pm \varepsilon)}{f(X | \theta_0)} = \\ \ln \int_{\mathcal{X}} \frac{f(x | \theta_0 \pm \varepsilon)}{f(x | \theta_0)} \cdot f(x | \theta_0) d\mu(x) &= \ln 1 = 0, \end{aligned}$$

причем равенство нулю первого члена в этой цепочке неравенств возможно лишь в случае

$$\frac{f(X | \theta_0 \pm \varepsilon)}{f(X | \theta_0)} = \text{const.},$$

то есть, поскольку интеграл от плотности равен 1, лишь в случае

$$f(X | \theta_0 \pm \varepsilon) = f(X | \theta_0),$$

что невозможно в силу условия (R3). Таким образом, математическое ожидание (4) строго меньше нуля, и состоятельность оценки максимального правдоподобия доказана.

Изучим теперь асимптотическое распределение оценки максимального правдоподобия. Для этого нам потребуется ввести дополнительные условия регулярности.

(R4) Для каждой точки θ_0 параметрического пространства Θ существует некоторая ее окрестность, в которой функция плотности $f(x | \theta)$ трижды дифференцируема по параметру θ и

$$\left| \frac{\partial f(x | \theta)}{\partial \theta} \right| \leq H_1(x), \quad (5)$$

$$\left| \frac{\partial^2 f(x | \theta)}{\partial \theta^2} \right| \leq H_2(x), \quad (6)$$

$$\left| \frac{\partial^3 \ln f(x | \theta)}{\partial \theta^3} \right| \leq H_3(x),$$

причем функции H_1 и H_2 интегрируемы по мере μ на носителе \mathcal{X} распределения X и $\mathbf{E}_{\theta} H_3(X) < \infty$ в некоторой окрестности каждой точки θ параметрического пространства Θ .

(R5) Функция

$$I(\theta) = \mathbf{E}_\theta \left(\frac{\partial \ln f(X | \theta)}{\partial \theta} \right)^2 = \int_{\mathcal{X}} \left(\frac{\partial \ln f(x | \theta)}{\partial \theta} \right)^2 f(x | \theta) d\mu(x) > 0,$$

каково бы ни было $\theta \in \Theta$.

Естественно, столь громоздкие и, на первый взгляд, странные условия требуют некоторого комментария.

Условие (R4) означает, что соответствующие производные функции плотности равномерно интегрируемы на \mathcal{X} , и поэтому можно выносить производную по θ за знак интеграла.

Условие (R5) требует положительности очень важной, с точки зрения состоятельности статистического вывода, характеристики вероятностной модели: $I(\theta)$ называется *информацией по Фишеру* в точке θ , содержащейся в наблюдении случайной величины X . Если $I(\theta) = 0$, то возникают непреодолимые трудности с принятием корректного решения, соответствующего этой параметрической точке θ . Понятно, что аналогичным образом можно определить и информацию по Фишеру, содержащуюся в случайной выборке $X^{(n)}$:

$$I_n(\theta) = \mathbf{E}_\theta \left(\frac{\partial \ln f_n(X^{(n)} | \theta)}{\partial \theta} \right)^2.$$

Приведем несколько утверждений, касающихся свойств информации по Фишеру.

Лемма 4.2. 1^0 . При выполнении условия (R4) в части (6) для вычисления информации по Фишеру можно использовать формулу

$$I(\theta) = -\mathbf{E}_\theta \frac{\partial^2 \ln f(X | \theta)}{\partial \theta^2}.$$

2^0 . При выполнении условия (R4) в части (5) информация по Фишеру обладает свойством аддитивности: $I_n(\theta) = nI(\theta)$ – информация, содержащаяся в выборке, равна сумме информаций, содержащихся в наблюдении каждой ее компоненты.

Доказательство. 1^0 . Условие (R4) в части (6) обеспечивает возможность смены порядка дифференцирования и интегрирования функции плотности, поэтому

$$\mathbf{E}_\theta \frac{\partial^2 \ln f(X | \theta)}{\partial \theta^2} = \mathbf{E}_\theta \left(\frac{f''_{\theta\theta}(X | \theta)}{f(X | \theta)} - \left(\frac{f'_\theta(X | \theta)}{f(X | \theta)} \right)^2 \right) =$$

$$\int_{\mathcal{X}} \frac{f''_{\theta\theta}(x|\theta)}{f(x|\theta)} \cdot f(x|\theta) d\mu(x) - I(\theta) = \frac{d^2}{d\theta^2} \int_{\mathcal{X}} f(x|\theta) d\mu(x) - I(\theta) = -I(\theta).$$

2⁰. Используя независимость и одинаковую распределенность компонент случайной выборки, получаем, что

$$\begin{aligned} I_n(\theta) &= \mathbf{E}_\theta \left(\frac{\partial \sum_{k=1}^n \ln f(X_k|\theta)}{\partial \theta} \right)^2 = \\ &= \mathbf{E}_\theta \left(\sum_{k=1}^n \left(\frac{\partial \ln f(X_k|\theta)}{\partial \theta} \right)^2 + \sum_{i \neq j} \frac{\partial \ln f(X_i|\theta)}{\partial \theta} \cdot \frac{\partial \ln f(X_j|\theta)}{\partial \theta} \right) = \\ &= \sum_{k=1}^n \mathbf{E}_\theta \left(\frac{\partial \ln f(X_k|\theta)}{\partial \theta} \right)^2 + \sum_{i \neq j} \mathbf{E}_\theta \frac{\partial \ln f(X_i|\theta)}{\partial \theta} \cdot \mathbf{E}_\theta \frac{\partial \ln f(X_j|\theta)}{\partial \theta} = nI(\theta), \end{aligned}$$

поскольку, в силу неравенства (5) в условии (R4), математическое ожидание

$$\mathbf{E}_\theta \frac{\partial \ln f(X|\theta)}{\partial \theta} = \int_{\mathcal{X}} \frac{f'_\theta(x|\theta)}{f(x|\theta)} \cdot f(x|\theta) d\mu(x) = \frac{d}{d\theta} \int_{\mathcal{X}} f(x|\theta) d\mu(x) = 0.$$

Теперь приступим к выводу асимптотического распределения оценки максимального правдоподобия скалярного параметра θ .

Теорема 4.2 (асимптотическая нормальность). *При выполнении условий (R1)–(R5) и наличии единственного локального максимума у функции правдоподобия корень $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ уравнения правдоподобия*

$$\partial \mathcal{L}(\theta | X^{(n)}) / \partial \theta = 0$$

асимптотически ($n \rightarrow \infty$) нормален со средним θ и дисперсией $(nI(\theta))^{-1}$, то есть

$$\lim_{n \rightarrow \infty} P_\theta \left((\hat{\theta}_n - \theta) \sqrt{nI(\theta)} < x \right) = \Phi(x).$$

Доказательство. Если $\hat{\theta}_n$ – оценка по методу максимального правдоподобия (корень уравнения правдоподобия), то имеет место тождество

$\partial \mathcal{L}(\hat{\theta}_n | X^{(n)}) / \partial \theta = 0$. Используя условие (R4), разложим его левую часть по формуле Тейлора в окрестности истинного значения θ_0 параметра θ :

$$\begin{aligned} \partial \mathcal{L}(\hat{\theta}_n | X^{(n)}) / \partial \theta &= \mathcal{L}'(\theta_0 | X^{(n)}) + \\ &(\hat{\theta}_n - \theta_0) \mathcal{L}''(\theta_0 | X^{(n)}) + (\hat{\theta}_n - \theta_0)^2 \mathcal{L}'''(\theta_1 | X^{(n)}) / 2 = 0, \end{aligned}$$

где производные от функции правдоподобия \mathcal{L} вычисляются по параметру θ , а $\theta_1 = \theta_0 + \lambda(\hat{\theta}_n - \theta_0)$, $0 < \lambda < 1$.

Разрешим полученное уравнение относительно величины $\sqrt{n}(\hat{\theta}_n - \theta_0)$, которая, согласно утверждению теоремы, должна иметь в пределе при $n \rightarrow \infty$ нормальное распределение со средним 0 и дисперсией $[I(\theta_0)]^{-1}$:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\mathcal{L}'(\theta_0 | X^{(n)}) / \sqrt{n}}{-\mathcal{L}''(\theta_0 | X^{(n)}) / n - (\hat{\theta}_n - \theta_0) \mathcal{L}'''(\theta_1 | X^{(n)}) / 2n}. \quad (7)$$

Числитель правой части этого представления

$$\frac{1}{\sqrt{n}} \mathcal{L}'(\theta_0 | X^{(n)}) = \frac{1}{\sqrt{n}} \sum_1^n \frac{\partial \ln f(X_k | \theta)}{\partial \theta}$$

есть нормированная на \sqrt{n} сумма независимых, одинаково распределенных случайных величин с нулевыми средними и дисперсиями $I(\theta_0) > 0$ (см. доказательство пункта 2⁰ леммы 4.2). Таким образом, в силу центральной предельной теоремы числитель правой части (7) асимптотически нормален с этими параметрами, и для завершения доказательства теоремы достаточно показать, что знаменатель (7) сходится по вероятности к постоянной $I(\theta_0)$, и сослаться на пункт (2) предложения 11.1 (теорема типа Слуцкого) курса ТВ.

В силу закона больших чисел и утверждения 1⁰ леммы 4.2 первое слагаемое в знаменателе (7)

$$-\frac{1}{n} \mathcal{L}''(\theta_0 | X^{(n)}) = -\frac{1}{n} \sum_1^n \frac{\partial^2 \ln f(X_k | \theta_0)}{\partial \theta^2} \xrightarrow{P} -\mathbf{E}_{\theta_0} \frac{\partial^2 \ln f(X | \theta_0)}{\partial \theta^2} = I(\theta_0),$$

так что остается показать, что и второе слагаемое сходится по вероятности к нулю.

Так как при выполнении условий (R1)–(R3) оценка максимального правдоподобия состоятельна, то $\hat{\theta}_n - \theta_0 \xrightarrow{P} 0$. Множитель при этой разности

$$\frac{1}{n} \mathcal{L}'''(\theta_1 | X^{(n)}) = \frac{1}{n} \sum_1^n \frac{\partial^3 \ln f(X_k | \theta_1)}{\partial \theta^3}$$

в силу условия (R4), начиная с некоторого n по абсолютной величине не превосходит $(1/n) \sum_1^n H_3(X_k)$ (это то n , при котором θ_1 попадает в окрестность точки θ_0 с любой наперед заданной вероятностью $1 - \varepsilon$). Применяя к этой сумме закон больших чисел, получаем, что она сходится по вероятности к

$$\mathbf{E}_{\theta_0} H_3(X) < \infty,$$

и поэтому указанный выше сомножитель ограничен с вероятностью единица, а все второе слагаемое в знаменателе правой части (7) сходится по вероятности к нулю.

Доказанная теорема, как будет видно из основного результата следующего параграфа, устанавливает асимптотическую оптимальность оценок максимального правдоподобия с точки зрения квадратичного риска.

§5. Эффективность оценок

Лекция 8

Обсуждая в начале нашего курса общую проблему статистического вывода, мы говорили о главной задаче математической статистики – построении решающих правил $\delta_n = \delta_n(X^{(n)})$, минимизирующих равномерно по всем $\theta \in \Theta$ функцию риска $R(\theta; \delta_n)$. К сожалению, без дополнительных ограничений на класс решающих функций эта задача не разрешима. Действительно, рассмотрим проблему оценки параметра θ , в которой пространство решений \mathcal{D} совпадает с параметрическим пространством Θ , а решающая функция $\delta_n = \hat{\theta}_n$ – оценка θ . Возьмем в качестве оценки некоторую фиксированную точку $\theta_0 \in \Theta$, то есть при любом результате $x^{(n)}$ статистического эксперимента будем принимать одно и то же решение $d = \theta_0$. Если функция потерь обладает тем естественным свойством, что $L(\theta, \theta) = 0$, каково бы ни было значение $\theta \in \Theta$, то риск такой оценки $R(\theta; \theta_0) = L(\theta, \theta_0)$ при $\theta = \theta_0$ равен нулю. Таким образом, если мы хотим построить оценку с равномерно минимальным риском в классе всевозможных оценок θ , то мы должны найти оценку θ_n^* с функцией риска $R(\theta, \theta_n^*) \equiv 0$, и понятно, что такой оценки не существует. Поэтому мы будем всегда при поиске оптимальных решений указывать класс оценок, в которых ищется оптимальное решение.

Определение 5.1. Оценка $\theta_n^* = \theta_n^*(X^{(n)})$ называется *оптимальной* или *оценкой с равномерно минимальным риском* в классе \mathcal{K} оценок параметра θ , если для любой оценки $\hat{\theta}_n \in \mathcal{K}$ и каждого $\theta \in \Theta$ имеет место неравенство $R(\theta; \theta_n^*) \leq R(\theta; \hat{\theta}_n)$.

Ниже предлагается метод нахождения оптимальных оценок скалярного параметра θ при квадратичной функции потерь в классе несмещенных оценок: $\mathbf{E}_\theta \hat{\theta}_n(X^{(n)}) = \theta$ при любом $\theta \in \Theta \subseteq \mathbb{R}$, но при дополнительных ограничениях на вероятностную модель и соответствующее семейство распределений оценки. Эти ограничения аналогичны тем условиям регулярности, которые мы накладывали на вероятностную модель при изучении асимптотических свойств оценок максимального правдоподобия. Мы покажем, что квадратичный риск любой несмещенной оценки, удовлетворяющей этим условиям, не может быть меньше $[nI(\theta)]^{-1}$ – асимптотической дисперсии оценки максимального правдоподобия (см. теорему 4.2). Следовательно, метод максимального правдоподобия доставляет асимптотическое решение

проблемы оптимальной оценки. Более того, мы покажем, что при наличии достаточных статистик метод максимального правдоподобия может привести и к точному решению проблемы равномерной минимизации функции риска.

Сформулируем условия регулярности, при выполнении которых будет находиться нижняя (достижимая!) граница квадратичного риска оценки.

(B1) Носитель \mathcal{X} распределения P_θ наблюдаемой случайной величины X не зависит от $\theta \in \Theta$ (условие, совпадающее с (R2) в §4).

(B2) Информация по Фишеру $I(\theta)$ строго положительна при любом $\theta \in \Theta$ (условие, совпадающее с (R5) в §4).

(B3) Равенство

$$\int_{\mathcal{X}^n} f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}) = 1$$

можно дифференцировать по θ под знаком интеграла, то есть

$$\int_{\mathcal{X}^n} f'_n(x^{(n)} | \theta) d\mu_n(x^{(n)}) = 0.$$

По аналогии с (R4) в части (5) для этого достаточно потребовать существование такой интегрируемой по мере μ функции $H(x)$, что в некоторой окрестности любой точки $\theta \in \Theta$ выполняется неравенство $|\partial f(x | \theta) / \partial \theta| \leq H(x)$, $x \in \mathcal{X}$.

(B4) Оценка $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ должна принадлежать классу оценок \mathcal{K}' , среднее значение которых

$$\mathbf{E}_\theta \hat{\theta}_n(X^{(n)}) = \int_{\mathcal{X}^n} \hat{\theta}_n(x^{(n)}) f_n(x^{(n)} | \theta) d\mu_n(x^{(n)})$$

можно дифференцировать по $\theta \in \Theta$ под знаком интеграла.

Конечно, условие (B4) требует комментария. В “высокой” теории статистического вывода приводятся достаточные условия на семейство распределений $\{P_\theta, \theta \in \Theta\}$ наблюдаемой случайной величины X , которые обеспечивают выполнение условия (B4), но формулировка этих условий и, в особенности, доказательство того, что они влекут (B4), настолько технически и концептуально сложны, что могут составить предмет специального курса. Однако все изучаемые нами в курсе ТВ вероятностные модели, за

исключением равномерного распределения, удовлетворяют этим условиям, и поэтому любая оценка их параметров принадлежит классу \mathcal{K}' .

Прежде чем получить основной “технический” результат этого параграфа, вспомним одно замечательное неравенство из курса математического анализа. Это – неравенство Коши–Буняковского, которое в случае интегралов Лебега по вероятностной мере P называется неравенством Шварца. Пусть Y – случайная величина с распределением P и g, h – две интегрируемые с квадратом по мере P функции на области \mathcal{Y} значений Y . Для этих функций имеет место неравенство

$$(\mathbf{E} g(Y)h(Y))^2 \leq \mathbf{E} g^2(Y) \cdot \mathbf{E} h^2(Y)$$

или, что то же,

$$\left(\int_{\mathcal{Y}} g(y)h(y) dP(y) \right)^2 \leq \int_{\mathcal{Y}} g^2(y) dP(y) \cdot \int_{\mathcal{Y}} h^2(y) dP(y),$$

причем знак равенства достигается тогда и только тогда, когда функции g и h линейно зависимы: существуют такие постоянные a и b , что $ag(y) + bh(y) = 0$ для почти всех $y \in \mathcal{Y}$ по мере P .

Теорема 5.1 (неравенство Рао–Крамера) *При выполнении условий (B1)–(B4) для квадратичного риска любой оценки $\hat{\theta}_n \in \mathcal{K}'$ справедливо неравенство*

$$\mathbf{E}_{\theta} \left(\hat{\theta}_n(X^{(n)}) - \theta \right)^2 \geq \mathbf{D}_{\theta} \hat{\theta}_n(X^{(n)}) \geq \frac{[d\gamma(\theta)/d\theta]^2}{nI(\theta)}, \quad (1)$$

где $\gamma(\theta) = \mathbf{E}_{\theta} \hat{\theta}_n(X^{(n)})$, причем знак равенства между риском и дисперсией оценки $\hat{\theta}_n$ достигается на несмещенных оценках: $\gamma(\theta) = \theta$, а знак равенства во втором неравенстве (1) имеет место тогда и только тогда, когда существует такая параметрическая функция $C(\theta)$, $\theta \in \Theta$, что

$$\hat{\theta}_n(X^{(n)}) - \gamma(\theta) = C(\theta) \frac{\partial \mathcal{L}(\theta | X^{(n)})}{\partial \theta} \quad (2)$$

почти наверное по мере P_{θ} .

Доказательство. Продифференцируем обе части равенств

$$\int_{\mathcal{X}^n} f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}) = 1,$$

$$\int_{\mathcal{X}^n} \hat{\theta}_n(x^{(n)}) f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}) = \gamma(\theta)$$

по параметру θ , занося производные в левых частях под знаки интегралов, что можно сделать благодаря условиям (В3) и (В4). Полученный результат, используя условие (В1), представим в виде

$$\int_{\mathcal{X}^n} \frac{\partial \mathcal{L}(\theta | x^{(n)})}{\partial \theta} f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}) = 0,$$

$$\int_{\mathcal{X}^n} \hat{\theta}_n(x^{(n)}) \frac{\partial \mathcal{L}(\theta | x^{(n)})}{\partial \theta} f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}) = \gamma'(\theta).$$

Вычтем из второго равенства первое, умножив его предварительно на $\gamma(\theta)$:

$$\int_{\mathcal{X}^n} \left(\hat{\theta}_n(x^{(n)}) - \gamma(\theta) \right) \frac{\partial \mathcal{L}(\theta | x^{(n)})}{\partial \theta} f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}) = \gamma'(\theta).$$

Применим к левой части полученного равенства неравенство Шварца, полагая

$$y = x^{(n)}, \quad \mathcal{Y} = \mathcal{X}^n, \quad g(x^{(n)}) = \hat{\theta}_n(x^{(n)}) - \gamma(\theta),$$

$$h(x^{(n)}) = \partial \mathcal{L}(\theta | x^{(n)}) / \partial \theta, \quad dP(y) = f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}).$$

В результате получим неравенство

$$(\gamma'(\theta))^2 \leq \int_{\mathcal{X}^n} \left(\hat{\theta}_n(x^{(n)}) - \gamma(\theta) \right)^2 f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}).$$

$$\int_{\mathcal{X}^n} \left(\frac{\partial \mathcal{L}(\theta | x^{(n)})}{\partial \theta} \right)^2 f_n(x^{(n)} | \theta) d\mu_n(x^{(n)}), \quad (3)$$

в котором знак равенства достигается тогда и только тогда, когда выполняется соотношение (2).

Мы получили неравенства (1), поскольку первое из них очевидно (на дисперсии достигается минимум всевозможных средних квадратичных отклонений случайной величины от постоянной). Второе неравенство в (1) есть следствие неравенства (3), ибо первый интеграл в правой части (3) равен $\mathbf{D}_\theta \hat{\theta}_n$, а второй интеграл определяет фишеровскую информацию $I_n(\theta)$, содержащуюся в выборке. Наконец, из пункта 2⁰ леммы 4.2 следует, что $I_n(\theta) = nI(\theta)$.

Следствие 5.1. Если $\hat{\theta}_n$ принадлежит подклассу $\mathcal{K} \subseteq \mathcal{K}'$ несмещенных оценок класса \mathcal{K}' , то ее квадратичный риск

$$R(\theta; \hat{\theta}_n) = \mathbf{D}_\theta \hat{\theta}_n \geq [nI(\theta)]^{-1}, \quad (4)$$

причем знак равенства тогда и только тогда, когда выполняется равенство (2) с $\gamma(\theta) = \theta$.

Понятно, что это следствие есть частный случай доказанной теоремы. Оно указывает неконструктивный путь к построению несмещенных оценок с равномерно минимальным риском. Достаточно вычислить производную в правой части равенства (2) и затем подбирать статистику $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ и параметрическую функцию $C(\theta)$, для которых имеет место равенство

$$\hat{\theta}_n(X^{(n)}) - \theta = C(\theta) \frac{\partial \mathcal{L}(\theta | X^{(n)})}{\partial \theta}.$$

Обычно это можно сделать в случае статистических структур, обладающих достаточными статистиками, где, в силу теоремы факторизации (теорема 2.1 из §2), функция правдоподобия

$$L(\theta | X^{(n)}) = g_\theta(T(X^{(n)}))h(X^{(n)}),$$

и последнее равенство имеет вид

$$\hat{\theta}_n(X^{(n)}) - \theta = C(\theta) \frac{\partial \ln g_\theta(T(X^{(n)}))}{\partial \theta}. \quad (5)$$

Например, для показательного распределения с функцией плотности

$$f(x | \theta) = \theta^{-1} \exp\{-x/\theta\}, \quad x > 0,$$

функция

$$\ln g_\theta(X^{(n)}) = -n \ln \theta - \theta^{-1} \sum_1^n X,$$

ее производная

$$\partial \ln g_\theta(T(X^{(n)}))/\partial \theta = -n/\theta + \sum_1^n X/\theta^2,$$

и равенство (5) выполняется при $C(\theta) = \theta^2/n$ и $\hat{\theta}_n = \bar{X}$. Таким образом, выборочное среднее \bar{X} есть несмещенная оценка с равномерно минимальным риском для параметра θ показательного распределения. Напомним, что \bar{X} оценка θ как по методу моментов, так и по методу максимального правдоподобия.

Легко понять, что если в (4) достигается знак равенства, то $\hat{\theta}_n$ – оптимальная оценка в классе \mathcal{K} , но обратное, вообще говоря, может и не выполняться – мы не располагаем утверждением, что любая оптимальная

оценка имеет квадратичный риск, равный $[n I(\theta)]^{-1}$. Чтобы подчеркнуть это различие и указать в дальнейшем более конструктивный метод построения оптимальных оценок, введем еще одно определение, рассмотрев более общую задачу несмещенной оценки некоторой параметрической функции $\gamma(\theta)$.

Определение 5.2. Несмещенная оценка $\hat{\gamma}_n = \hat{\gamma}_n(X^{(n)})$ параметрической функции $\gamma(\theta)$ называется *эффективной* в классе \mathcal{K}' , если ее квадратичный риск

$$R(\gamma; \hat{\gamma}_n) = \mathbf{E}_\theta \left(\hat{\gamma}_n(X^{(n)}) - \gamma(\theta) \right)^2 = \mathbf{D}_\theta \hat{\gamma}_n(X^{(n)}) = [\gamma'(\theta)]^2/n I(\theta),$$

то есть (см. теорему 5.1 с $\hat{\theta}_n = \hat{\gamma}_n$) выполняется равенство

$$\hat{\gamma}_n(X^{(n)}) - \gamma(\theta) = C(\theta) \frac{\partial \mathcal{L}(\theta | X^{(n)})}{\partial \theta}. \quad (6)$$

Оценка $\hat{\gamma}_n$ называется *асимптотически эффективной* в классе \mathcal{K}' , если

$$\mathbf{E}_\theta \hat{\gamma}_n(X^{(n)}) \sim \gamma(\theta), \quad \mathbf{D}_\theta \hat{\gamma}_n(X^{(n)}) \sim [\gamma'(\theta)]^2/n I(\theta),$$

когда $n \rightarrow \infty$.

В силу теоремы 4.2 оценка по методу максимального правдоподобия скалярного параметра θ (в данном случае $\gamma(\theta) = \theta$) является асимптотически эффективной оценкой в классе \mathcal{K}' . Покажем, что она дает решение проблемы построения эффективной оценки в классе \mathcal{K} .

Пусть $\hat{\theta}_n$ – оценка максимального правдоподобия параметра θ . Определим оценку $\gamma(\hat{\theta}_n)$ параметрической функции $\gamma(\theta)$ с помощью подстановки вместо θ ее оценки $\hat{\theta}_n$.

Теорема 5.2. Если $\gamma(\hat{\theta}_n)$ есть несмещенная оценка параметрической функции $\gamma(\theta)$ и эффективная в классе \mathcal{K}' оценка γ_n^* параметрической функции $\gamma(\theta)$ существует, то при выполнении условий регулярности (R1)–(R5) и (B3)–(B4) почти наверное $\gamma_n^*(X^{(n)}) = \gamma(\hat{\theta}_n(X^{(n)}))$.

Доказательство. Если γ_n^* – эффективная оценка $\gamma(\theta)$, то она удовлетворяет равенству (6):

$$\gamma_n^*(X^{(n)}) - \gamma(\theta) = C(\theta) \partial \mathcal{L}(\theta | X^{(n)}) / \partial \theta, \quad (7)$$

каково бы ни было $\theta \in \Theta$. Но если $\hat{\theta}_n$ – оценка по методу максимального правдоподобия, то

$$\partial \mathcal{L}(\hat{\theta}_n | X^{(n)}) / \partial \theta = 0,$$

так что равенство (7) при $\theta = \hat{\theta}_n$ превращается в равенство

$$\gamma_n^*(X^{(n)}) - \gamma(\hat{\theta}_n) = 0$$

почти наверное по вероятности P_θ^n .

Из доказанной теоремы немедленно вытекает, что выборочное среднее \bar{X} есть эффективная (следовательно, и оптимальная) несмещенная оценка параметра θ таких распределений, как двухточечное, биномиальное при известном m , Пуассона, показательное; \bar{X} есть также несмещенная оценка с равномерно минимальным квадратичным риском среднего значения μ нормального (μ, σ^2) распределения.

§6. Доверительные интервалы

Лекция 9

Мы рассмотрели несколько методов построения *точечных* оценок для параметров, значения которых определяют распределение наблюдаемой случайной величины. Был получен ряд утверждений о распределении таких оценок, что позволяет судить о надежности оценки при заданной точности, то есть вычислять вероятности событий вида $|\hat{\theta}_n(X^{(n)}) - \theta| \leq \Delta$ при каждом фиксированном значении параметра θ . Поскольку именно значение θ нам неизвестно, то такого рода вычисления зачастую лишены практического смысла – слишком велик размах в надежности оценки $\hat{\theta}_n$ при различных θ , даже в случае когда мы располагаем некоторой априорной информацией о возможной области значений этого параметра. Поэтому в ряде практических ситуаций пытаются решать обратную задачу: для фиксированной надежности, скажем, $1 - \alpha$, где α мало, указать некоторую область значений θ , зависящую, естественно, от выборки $X^{(n)}$, которая с вероятностью, не меньшей $1 - \alpha$, накрывает истинное, неизвестное нам значение θ , причем такое надежностное утверждение должно выполняться при любых $\theta \in \Theta$. В таком случае по размерам области, которые определяются выборочными значениями $x^{(n)}$, можно судить о точности такой *интервальной* оценки.

Определение 6.1. Подмножество $\Delta_n = \Delta_n(X^{(n)})$ параметрического пространства Θ называется $(1 - \alpha)$ -*доверительной областью*, если

$$P_\theta \left(\Delta_n(X^{(n)}) \ni \theta \right) \geq 1 - \alpha, \quad (1)$$

каково бы ни было значение $\theta \in \Theta$. Заданное (фиксированное) значение $1 - \alpha$ называется *доверительным уровнем*, а наименьшее значение левой части неравенства (1) по всем $\theta \in \Theta$ – *доверительным коэффициентом*. В случае $\Theta \subseteq \mathbb{R}$ доверительная область вида $\Delta_n = (\underline{\theta}_n(X^{(n)}); \bar{\theta}_n(X^{(n)}))$ называется *доверительным интервалом*, в котором различаются *нижний* $\underline{\theta}_n$ и *верхний* $\bar{\theta}_n$ *доверительные пределы*. Доверительные интервалы вида $(\underline{\theta}_n; \infty)$ и $(-\infty; \bar{\theta}_n)$ называются соответственно *нижней* и *верхней доверительными границами*.

Естественно, конфигурация доверительной области выбирается статистиком, сообразуясь с ее геометрической наглядностью и, главное, возможностью гарантировать доверительную вероятность. В случае скалярного

параметра доверительная область обычно выбирается в виде интервала, причем в ряде случаев, например, при оценке надежности или вероятности нежелательного события, в виде одностороннего интервала. В случае многомерного параметра обычно строятся доверительные эллипсоиды или параллелепипеды.

Следует обратить особое внимание на правильную формулировку доверительного утверждения, которая подчеркивается в неравенстве (1) записью $\Delta_n(X^{(n)}) \ni \theta$ вместо обычного $\theta \in \Delta_n(X^{(n)})$. Говорить, что значение параметра θ с вероятностью, не меньшей $1 - \alpha$, принадлежит области Δ_n , значит сознательно вводить трудящихся на ниве прикладной статистики в заблуждение. Дело в том, что значение параметра θ в данной вероятностной модели не является случайной величиной, это постоянная, свойственная исследуемому объекту, а постоянная принадлежит какой-либо области только с вероятностью единица или ноль. Вся случайность заключена в самой доверительной области $\Delta_n(X^{(n)})$, и поэтому правильное доверительное утверждение гласит: *область $\Delta_n(X^{(n)})$ с вероятностью, не меньшей $1 - \alpha$, накрывает истинное (неизвестное) значение θ .*

В самом начале нашего курса математической статистики в примере 1.1 с определением содержания общей серы в дизельном топливе мы строили доверительный интервал фиксированной ширины для среднего значения нормального распределения, когда занимались планированием объема испытаний, необходимого для достижения заданной точности и надежности оценки. Рассмотрим еще раз этот пример в свете введенных понятий интервальной оценки параметра.

1⁰. ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ СРЕДНЕГО ЗНАЧЕНИЯ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПРИ ИЗВЕСТНОЙ ДИСПЕРСИИ. Итак, в примере 1.1 мы имели дело с выборкой $X^{(n)}$ из нормального (μ, σ^2) распределения, причем значение параметра σ нам было известно, так что в качестве неизвестного параметра θ выступало μ . Наша задача состоит в построении такого интервала $(\underline{\mu}_n(X^{(n)}), \bar{\mu}_n(X^{(n)}))$, что $P_\mu(\underline{\mu}_n \leq \mu \leq \bar{\mu}_n) \geq 1 - \alpha$, при любом значении $\mu \in \mathbb{R}$.

Напомним, что в этом примере оценкой μ служило выборочное среднее \bar{X} – несмещенная оценка μ с минимальным квадратичным риском. Эта линейная оценка обладает замечательным свойством инвариантности: распределение разности $\bar{X} - \mu$ не зависит от μ , и это обстоятельство подска-

зывает нам путь к построению доверительного интервала. Действительно,

$$P\left(\frac{|\bar{X} - \mu|}{\sigma}\sqrt{n} \leq \lambda\right) = 2\Phi(\lambda) - 1,$$

и если положить λ равным корню уравнения $2\Phi(\lambda) - 1 = 1 - \alpha$, то есть выбрать $\lambda = \lambda_\alpha = \Phi^{-1}(1 - \alpha/2)$, то интервал $(\bar{X} - \lambda_\alpha\sigma/\sqrt{n}, \bar{X} + \lambda_\alpha\sigma/\sqrt{n})$ будет $(1 - \alpha)$ -доверительным интервалом для среднего значения μ нормального (μ, σ^2) распределения при известной дисперсии σ^2 .

В этом простейшем примере на построение доверительного интервала ключевым моментом было использование *инвариантной* случайной функции $\hat{\theta}_n - \theta$ от оценки $\hat{\theta}_n = \bar{X}$ и параметра $\theta = \mu$. В принципе, именно на подобном выборе *опорной* функции $H(\hat{\theta}_n, \theta)$ с подходящей оценкой $\hat{\theta}_n$ параметра θ основаны исторически первые методы построения доверительных интервалов и множеств. Опорная функция $H(\cdot, \cdot)$ подбирается таким образом, чтобы она была монотонно возрастающей функцией второго аргумента θ , и при этом вероятность $P_\theta\left(H(\hat{\theta}_n(X^{(n)}), \theta) \leq \lambda\right)$ для некоторых значений λ должна оставаться достаточно высокой (близкой к единице), каково бы ни было значение $\theta \in \Theta$. Мы проиллюстрируем этот метод построения доверительных интервалов с помощью подбора инвариантных опорных функций на примере нормального (μ, σ^2) распределения, строя доверительные интервалы для каждого из параметров при известном и неизвестном значениях другого (“мешающего”) параметра.

2⁰. ВЕРХНЯЯ ДОВЕРИТЕЛЬНАЯ ГРАНИЦА ДЛЯ ДИСПЕРСИИ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПРИ ИЗВЕСТНОМ СРЕДНЕМ. При выборе из нормального (μ, σ^2) распределения с известным средним значением μ метод максимального правдоподобия приводит к несмещенной оценке

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \mu)^2$$

параметра σ^2 . Используя результаты предыдущего параграфа, нетрудно показать, что $\hat{\sigma}_n^2$ есть несмещенная оценка с равномерно минимальным риском.

Поскольку, в чем мы неоднократно убеждались,

$$Y_k = (X_k - \mu)/\sigma \sim \mathcal{N}(0, 1), \quad k = 1, \dots, n,$$

то естественно рассмотреть в качестве опорной функцию

$$H(\hat{\sigma}_n^2, \sigma^2) = \frac{1}{\sigma^2} \sum_1^n (X_k - \mu)^2.$$

Найдем ее распределение.

Лемма 6.1. *Если Y_1, \dots, Y_n независимы и одинаково распределены по стандартному нормальному закону $\mathcal{N}(0, 1)$, то $\sum_1^n Y_k^2$ имеет гамма-распределение $G(n/2, 2)$.*

Доказательство. Покажем, что Y^2 , где $Y \sim \mathcal{N}(0, 1)$, имеет гамма-распределение $G(1/2, 2)$, после чего просто воспользуемся теоремой сложения для гамма-распределения (см. предложение 12.2, пункт 5⁰ курса ТВ). Функция распределения Y^2 вычисляется по формуле

$$F(x) = P(Y^2 < x) = P(-\sqrt{x} < Y < \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = 2\Phi(\sqrt{x}) - 1,$$

так что ее функция плотности

$$f(x) = \frac{d}{dx} \left[\frac{2}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{x}} \exp\left\{-\frac{t^2}{2}\right\} dt - 1 \right] = \frac{1}{2^{1/2}\Gamma(1/2)} x^{1/2-1} e^{-x/2},$$

поскольку $\Gamma(1/2) = \sqrt{\pi}$. Мы видим, что это – функция плотности гамма-распределения $G(1/2, 2)$ с параметром формы $\lambda = 1/2$ и параметром масштаба $a = 2$, откуда, как было замечено выше, немедленно следует утверждение леммы.

Гамма-распределение $G(n/2, 2)$ очень часто используется в различных задачах математической статистики, и оно появилось раньше, чем гамма-распределение $G(\lambda, a)$ общего вида, под названием *хи-квадрат распределение с n степенями свободы*. Функция распределения хи-квадрат обычно обозначается $K_n(x)$, $x > 0$, а что касается термина “степени свободы”, то его смысл прояснится по мере других применений хи-квадрат распределения.

Теперь мы можем перейти к нашей основной задаче – построению доверительных границ для σ^2 . Если обратиться к практической стороне этой проблемы (см. в связи с этим пример 1.1), то легко понять, что статистика должна интересоваться только верхней (а не двусторонняя) граница σ^2 , на которую он будет ориентироваться, чтобы обезопасить себя от грубых ошибок при планировании статистического эксперимента. Таким образом, мы должны сформулировать доверительное утверждение в форме

$\sigma^2 \leq \bar{\sigma}_n^2$. Понятно, что нижняя доверительная граница и двусторонние границы (доверительный интервал), коль скоро они кому-то потребуются, строятся аналогичным образом.

В рамках такой формулировки задачи, мы должны рассмотреть событие

$$A_\lambda = \left\{ H(\hat{\sigma}_n^2, \sigma^2) = \frac{\sum_1^n (X_k - \mu)^2}{\sigma^2} \geq \lambda \right\},$$

выбирая λ из условия $P_{\mu, \sigma}(A_\lambda) = 1 - \alpha$. Как мы только что выяснили, эта вероятность не зависит от μ и σ , и в силу леммы 6.1 постоянная λ определяется квантилью хи-квадрат распределения с n степенями свободы – корнем уравнения $1 - K_n(\lambda) = 1 - \alpha$. Следовательно, верхняя $(1 - \alpha)$ -доверительная граница для σ^2 равна

$$\sum_1^n (X_k - \mu)^2 / K_n^{-1}(\alpha),$$

где, в соответствии с нашими стандартными обозначениями, $K_n^{-1}(\alpha)$ есть α -квантиль хи-квадрат распределения с n степенями свободы.

Используемые в рассмотренных примерах методы подбора опорных функций, основанные на принципе инвариантности статистик (оценок параметров μ и σ^2) относительно линейных преобразований, позволяют аналогичным образом подбирать такие функции и в случае неизвестных значений мешающего параметра. Так, если рассматривается задача построения доверительных границ для σ^2 при неизвестном μ , то естественно обратиться к оценке

$$S^2 = \frac{1}{n} \sum_1^n (X_k - \bar{X})^2$$

параметра σ^2 , замечая, что ее распределение не зависит от μ , поскольку каждая из разностей $X_k - \bar{X}$ инвариантна относительно сдвига, когда X_k заменяется на $X_k - \mu$, $k = 1, \dots, n$. Если разделить эти разности на σ , то мы получим случайные величины, распределение которых не зависит как от μ , так и от σ , и таким образом мы приходим к инвариантной опорной функции $H(S^2, \sigma^2) = S^2/\sigma^2$. Для вывода распределения этой функции можно обратиться к нормальным $(0, 1)$ случайным величинам $Y_k = (X_k - \mu)/\sigma$, $k = 1, \dots, n$, поскольку, в чем легко убедиться,

$$H(S^2, \sigma^2) = \frac{1}{n} \sum_1^n (Y_k - \bar{Y})^2.$$

Если обратиться к задаче доверительной интервальной оценки μ при неизвестном значении σ , то здесь инвариантную опорную функцию можно построить, комбинируя ее из опорных функций задач 1⁰ и 2⁰. Как мы видели при решении этих задач, распределения случайных величин $(\bar{X} - \mu)/\sigma$ и S/σ не зависят от μ и σ , и поэтому в качестве опорной функции при интервальной оценке μ можно использовать опорную функцию, определяемую отношением этих величин, то есть функцию $|\bar{X} - \mu|/S$.

Однако для построения доверительных интервалов на основе таких функций нам необходимо найти совместное распределение статистик \bar{X} и S^2 . Мы получим это распределение в следующей лекции, сформулировав его в виде утверждения, известного в математической статистике как *лемма Фишера*.

Лекция 10

Теорема 6.1. В случае выбора из нормального (μ, σ^2) распределения статистики \bar{X} и S^2 независимы, $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, а $nS^2/\sigma^2 \sim \chi_{n-1}^2$ (имеет хи-квадрат распределение с $n - 1$ степенью свободы).

Доказательство. Пусть Y_1, \dots, Y_n – случайная выборка из стандартного нормального распределения $\mathcal{N}(0, 1)$. Покажем, что статистики

$$\bar{Y} = \frac{1}{n} \sum_{k=1}^n Y_k \quad \text{и} \quad S_Y = \sum_{k=1}^n (Y_k - \bar{Y})^2$$

независимы, $\bar{Y} \sim \mathcal{N}(0, 1/n)$, а $S_Y \sim \chi_{n-1}^2$. Тогда утверждение теоремы будет следовать из того факта, что $\sigma Y_k + \mu$ имеют то же распределение, что и X_k , $k = 1, \dots, n$, и, следовательно, распределение \bar{X} совпадает с распределением $\sigma \bar{Y} + \mu$, а распределение S_Y – с распределением nS^2/σ^2 .

Введем случайные величины

$$Z_k = \sum_{i=1}^n c_{ki} Y_i, \quad k = 1, \dots, n,$$

которые определяются заданием матрицы $C = \|c_{ki}\|$ линейных преобразований случайных величин Y_1, \dots, Y_n . Пусть элементы первой строки этой матрицы $c_{11} = \dots = c_{1n} = 1/\sqrt{n}$, а остальные элементы матрицы C выберем так, чтобы произведение C на транспонированную матрицу C' было единичной матрицей: $CC' = I$. Как известно, такой выбор C возможен, и

полученная таким образом матрица называется ортонормированной. Случайные величины Z_1, \dots, Z_n распределены в соответствии с n -мерным нормальным законом, для спецификации которого достаточно найти вектор средних значений этих величин и матрицу их ковариаций.

Средние значения

$$m_k = \mathbf{E}Z_k = \mathbf{E} \sum_{i=1}^n c_{ki} Y_i = \sum_{i=1}^n c_{ki} \mathbf{E}Y_i = 0, \quad k = 1, \dots, n.$$

Далее, поскольку средние значения равны нулю, ковариации этих случайных величин

$$\begin{aligned} \text{cov}(Z_k, Z_j) &= \mathbf{E}(Z_k - m_k)(Z_j - m_j) = \mathbf{E}Z_k Z_j = \mathbf{E} \sum_{i=1}^n c_{ki} Y_i \cdot \sum_{i=1}^n c_{ji} Y_i = \\ &= \mathbf{E} \sum_{i=1}^n c_{ki} c_{ji} Y_i^2 + \mathbf{E} \sum_{i \neq l}^n c_{ki} c_{jl} Y_i Y_l. \end{aligned}$$

Если занести математические ожидания под знаки сумм и вспомнить, что Y_1, \dots, Y_n независимы, $\mathbf{E}Y_i = 0$, $\mathbf{E}Y_i^2 = \mathbf{D}Y_i = 1$, а при $i \neq l$ средние значения $\mathbf{E}Y_i Y_l = \mathbf{E}Y_i \mathbf{E}Y_l = 0$, то получим, что

$$\text{cov}(Z_k, Z_j) = \sum_{i=1}^n c_{ki} c_{ji}, \quad k, j = 1, \dots, n.$$

Поскольку для ортонормированной матрицы последняя сумма равна нулю, если $k \neq j$, и равна единице, если $k = j$, то мы приходим к заключению, что Z_1, \dots, Z_n независимы и одинаково распределены по стандартному нормальному закону $\mathcal{N}(0, 1)$. Таким образом, ортонормированные преобразования случайных величин Y_1, \dots, Y_n не изменили их совместное распределение.

Теперь представим наши статистики \bar{Y} и S_Y в терминах случайных величин Z_1, \dots, Z_n . Поскольку

$$Z_1 = \sum_{i=1}^n c_{1i} Y_i = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i,$$

то $\bar{Y} = Z_1/\sqrt{n}$. Далее, ортонормированное линейное преобразование сохраняет сумму квадратов компонент преобразуемого вектора, то есть

$$\sum_1^n Z_k^2 = \sum_1^n Y_k^2.$$

Следовательно, статистика S_Y в новых переменных приобретает вид

$$\frac{S_Y}{n} = \frac{1}{n} \sum_1^n Y_k^2 - \bar{Y}^2 = \frac{1}{n} \sum_1^n Z_k^2 - \frac{Z_1^2}{n} = \frac{1}{n} \sum_2^n Z_k^2.$$

Итак, распределение \bar{Y} совпадает с распределением Z_1/\sqrt{n} , а распределение S_Y – с распределением суммы квадратов $n - 1$ независимых в совокупности и независимых от Z_1 нормальных $(0, 1)$ случайных величин. Следовательно, \bar{Y} и S_Y независимы, $\bar{Y} \sim \mathcal{N}(0, 1/n)$, $S_Y \sim \chi_{n-1}^2$ (см. лемму 6.1), и “лемма Фишера” доказана.

Установив совместное распределение выборочного среднего и выборочной дисперсии в случае выбора из нормального распределения, мы можем приступить к построению доверительных интервалов для каждого из параметров μ и σ при неизвестном значении другого параметра.

3⁰. ВЕРХНЯЯ ДОВЕРИТЕЛЬНАЯ ГРАНИЦА ДЛЯ ДИСПЕРСИИ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПРИ НЕИЗВЕСТНОМ СРЕДНЕМ. Эта граница находится наиболее просто, поскольку распределение опорной функции

$$\frac{nS^2}{\sigma^2} = \frac{\sum_1^n (X_k - \bar{X})^2}{\sigma^2} = \sum_{k=1}^n \left(\frac{X_k - \mu}{\sigma} - \frac{1}{n} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \right)^2 = \sum_1^n (Y_k - \bar{Y})^2 \quad (2)$$

есть хи-квадрат распределение с $n - 1$ степенью свободы (см. теорему 6.1). Следовательно, верхняя $(1 - \alpha)$ -доверительная граница определяется квантилью $\lambda_\alpha = K_{n-1}^{-1}(\alpha)$ хи-квадрат распределения – корнем уравнения

$$P(nS^2/\sigma^2 \geq \lambda) = 1 - K_{n-1}(\lambda) = 1 - \alpha,$$

и доверительное утверждение $\sigma^2 \leq \bar{\sigma}_n^2 = nS^2/K_{n-1}^{-1}(\alpha)$ выполняется с заданной вероятностью $1 - \alpha$.

4⁰. ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ СРЕДНЕГО ЗНАЧЕНИЯ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПРИ НЕИЗВЕСТНОЙ ДИСПЕРСИИ. В этой задаче мы имеем дело с двусторонними доверительными границами (доверительным интервалом), и в соответствии с выбором опорной функции $|\bar{X} - \mu|/S$, о которой мы говорили перед доказательством теоремы 6.1, нам потребуется знание вероятности события вида $|\bar{X} - \mu|/S \leq \lambda$.

В начале XIX века английский математик В.Госсет, писавший под псевдонимом “Стьюдент” (Student), нашел распределение случайной величины

$T_\nu = \xi\sqrt{\nu}/\sqrt{\chi_\nu^2}$, где $\xi \sim \mathcal{N}(0, 1)$, а χ_ν^2 – случайная величина, не зависящая от ξ и распределенная по закону хи-квадрат с ν степенями свободы. Естественно, его исследования были связаны с проблемами статистического вывода о среднем значении μ нормального распределения при неизвестной дисперсии, и Стьюдент искал распределение опорной функции (см. (2)) в связи с переходом в записи опорной функции в терминах X_k к Y_k)

$$H = \frac{\bar{X} - \mu}{S} \sqrt{n-1} = \frac{\frac{1}{\sqrt{n}} \sum_1^n Y_k}{\sqrt{\sum_1^n (Y_k - \bar{Y})^2}} \sqrt{n-1},$$

$$Y_k = \frac{X_k - \mu}{\sigma} \sim \mathcal{N}(0, 1), \quad k = 1, \dots, n,$$

которая отличается от выбранной нами опорной функции только множителем $\sqrt{n-1}$, и поэтому также может быть использована в построении доверительного интервала для μ при неизвестном σ . То, что распределения T_{n-1} и H совпадают, следует из теоремы 6.1: случайная величина в числителе $\xi = \sum_1^n Y_k/\sqrt{n} \sim \mathcal{N}(0, 1)$ не зависит от $\sum_1^n (Y_k - \bar{Y})^2 \sim \chi_{n-1}^2$, разделив которую на значение степени свободы $n-1$, получаем $\frac{1}{\nu}\chi_\nu^2$ с $\nu = n-1$.

Найдем распределение случайной величины T_ν , которое называется *распределением Стьюдента с ν степенями свободы* или *t-распределением*. Совместная функция плотности независимых случайных величин ξ и $\eta = \chi_\nu^2$ равна

$$f(x, y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} \exp\left\{-\frac{y}{2}\right\},$$

так что функция распределения случайной величины T_ν

$$S_\nu(t) = P(\xi\sqrt{\nu/\eta} < t) = \int_{x\sqrt{\nu} < \sqrt{yt}} \int f(x, y) dx dy = \int_0^\infty dy \int_{-\infty}^{t\sqrt{y/\nu}} f(x, y) dx.$$

Дифференцируя это выражение по t , находим функцию плотности распределения Стьюдента

$$s_\nu(t) = \int_0^\infty \sqrt{y/\nu} f(t\sqrt{y/\nu}, y) dy =$$

$$\begin{aligned} & \frac{1}{\sqrt{\pi\nu}2^{(\nu+1)/2}\Gamma(\nu/2)} \int_0^\infty y^{\frac{\nu+1}{2}-1} \exp\left\{-\frac{y}{2}\left(1+\frac{t^2}{\nu}\right)\right\} dy = \\ & = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}. \end{aligned}$$

Вид полученной функции плотности говорит о том, что распределение Стьюдента можно трактовать как обобщение стандартного ($a = 0$, $b = 1$) распределения Коши $S(a, b)$, которое получается из распределения Стьюдента при числе степеней свободы $\nu = 1$. Это симметричное распределение, и поэтому $S_\nu(-t) = 1 - S_\nu(t)$, что позволяет нам довольно просто построить доверительный интервал для μ с помощью квантили распределения $S_{n-1}(\cdot)$:

$$P(|T_{n-1}| \leq t) = S_{n-1}(t) - S_{n-1}(-t) = 2S_{n-1}(t) - 1 = 1 - \alpha,$$

откуда $t_\alpha = S_{n-1}^{-1}(1 - \alpha/2)$, и $(1 - \alpha)$ -доверительный интервал для μ определяется пределами $\bar{X} \pm St_\alpha/\sqrt{n-1}$.

Итак, мы построили доверительные пределы для параметров μ и σ^2 нормального распределения. Таблицы нормального, хи-квадрат и стьюдентского распределений, а также квантилей этих распределений, необходимые для численной реализации доверительных оценок, смотрите в книге Большев Л.Н., Смирнов Н.В. Таблицы математической статистики, М.: Наука, 1983, которая в дальнейшем будет цитироваться как ТМС. Еще раз отметим, что возможность доверительной оценки этих параметров определялась, в основном, инвариантностью семейства нормальных распределений относительно линейной группы преобразований. Точно так же мы можем построить доверительные пределы для параметра θ показательного распределения или для параметра масштаба гамма распределения при известном параметре формы; мы вернемся к этим задачам позднее при обсуждении проблемы оптимизации доверительной оценки. Что же касается других распределений, то здесь проблема осложняется отсутствием инвариантных опорных функций и невозможностью получить распределение оценок параметра, для которого строятся доверительные пределы, в явном виде. Тем не менее существует достаточно общий подход к данной проблеме, основанный на асимптотической нормальности распределения оценок по методу моментов или методу максимального правдоподобия.

Пусть $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ – асимптотически нормальная со средним θ и дисперсией $\sigma^2(\theta)/n$ оценка параметра θ (например, при определенных условиях регулярности (см. теорему 4.2) оценка максимального правдоподобия асимптотически нормальна со средним θ и дисперсией $[nI(\theta)]^{-1}$). Тогда при $n \rightarrow \infty$ вероятность

$$P_\theta \left(\frac{|\hat{\theta}_n - \theta|}{\sigma(\theta)} \sqrt{n} \leq \lambda_\alpha \right) \rightarrow 1 - \alpha,$$

при любом $\theta \in \Theta$, если $\lambda_\alpha = \Phi^{-1}(1 - \alpha/2)$, и мы получаем *асимптотически* $(1 - \alpha)$ -доверительное множество

$$\Delta_n(X^{(n)}) = \left\{ \theta : \frac{|\hat{\theta}_n - \theta|}{\sigma(\theta)} \sqrt{n} \leq \lambda_\alpha \right\} \cap \Theta.$$

Если Δ_n есть интервал на прямой \mathbb{R} , то мы решили задачу интервальной оценки параметра θ . Если же это некоторое вычурное и непригодное к употреблению подмножество \mathbb{R} , то можно пойти на дальнейшие упрощения асимптотического утверждения, заменив в определении Δ_n параметрическую функцию $\sigma(\theta)$ на ее оценку $\sigma(\hat{\theta}_n)$. Достаточно потребовать непрерывность функции $\sigma(\theta)$, $\theta \in \Theta$, чтобы, ссылаясь на теорему Слуцкого (предложение 11.1 курса ТВ с $\xi_n \xrightarrow{P} \sigma(\theta)$), утверждать, что при $n \rightarrow \infty$

$$P_\theta \left(\hat{\theta}_n - \frac{\lambda_\alpha \sigma(\hat{\theta}_n)}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + \frac{\lambda_\alpha \sigma(\hat{\theta}_n)}{\sqrt{n}} \right) \rightarrow 1 - \alpha.$$

Проиллюстрируем “работу” этого метода на двух полезных в практическом отношении примерах.

5⁰. АСИМПТОТИЧЕСКИ ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ВЕРОЯТНОСТИ УСПЕХА В ИСПЫТАНИЯХ БЕРНУЛЛИ. В схеме испытаний Бернулли – выборе из распределения бинарной случайной величины X , принимающей значение 1 (“успех”) с вероятностью p и значение 0 (“неудача”) с вероятностью $1 - p$, оптимальной несмещенной оценкой p является выборочное среднее $\bar{X} = n^{-1} \sum_1^n X_k$ или, что то же, относительная частота успешных исходов в n испытаниях. Статистика $n\bar{X}$ имеет биномиальное распределение $B(n, p)$, и это позволяет насчитать таблицы доверительных пределов для p при различных значениях доверительного уровня $1 - \alpha$,

объема выборки n и числа успешных исходов $n\bar{x}$ (см., например, ТМС). Что же дает асимптотический подход к построению доверительных интервалов?

Выборочное среднее асимптотически нормально со средним p и дисперсией $p(1-p)/n$. Следовательно, $(1-\alpha)$ -доверительная область

$$\Delta_n = \left\{ p : 0 \leq p \leq 1, |\bar{X} - p| \leq \lambda_\alpha \sqrt{p(1-p)/n} \right\}.$$

Разрешая неравенства в фигурных скобках относительно p , получаем доверительный интервал

$$\frac{n}{n + \lambda_\alpha^2} \left(\bar{X} + \frac{\lambda_\alpha^2}{2n} \pm \lambda_\alpha \sqrt{\frac{\bar{X}(1-\bar{X})}{n} + \frac{\lambda_\alpha^2}{4n^2}} \right),$$

который при больших объемах испытаний n мало отличается от доверительного интервала

$$\bar{X} \pm \lambda_\alpha \sqrt{\frac{\bar{X}(1-\bar{X})}{n}},$$

полученного заменой $\sigma^2(p) = p(1-p)$ на ее оценку $\bar{X}(1-\bar{X})$:

6⁰. АСИМПТОТИЧЕСКИ ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ПАРАМЕТРА ИНТЕНСИВНОСТИ РАСПРЕДЕЛЕНИЯ ПУАССОНА. Распределение Пуассона $P(\theta)$ с функцией плотности (по считающей мере)

$$f(x|\theta) = P_\theta(X=x) = \theta^x e^{-\theta} / x!, \quad x = 0, 1, 2, \dots,$$

индексируется положительным параметром θ , оптимальная несмещенная оценка которого по выборке $X^{(n)}$ объема n , как и в предыдущем примере, определяется выборочным средним \bar{X} . Для распределения Пуассона также справедлива теорема сложения: $n\bar{X} \sim P(n\theta)$, и на основе этого можно построить точные доверительные пределы для θ , таблица которых имеется в упомянутом сборнике ТМС. Но оценка \bar{X} асимптотически нормальна $(\theta, \theta/n)$, что позволяет определить асимптотически доверительную область

$$\Delta_n = \left\{ \theta : \theta > 0, |\bar{X} - \theta| \leq \lambda_\alpha \sqrt{\theta/n} \right\}.$$

Решение неравенств в фигурных скобках относительно θ дает асимптотически доверительный интервал

$$\bar{X} + \frac{\lambda_\alpha^2}{2n} \pm \lambda_\alpha \sqrt{\frac{\bar{X}}{n} + \frac{\lambda_\alpha^2}{4n^2}}.$$

Наконец, заменяя $\sigma^2(\theta) = \theta$ ее оценкой \bar{X} , получаем также асимптотически доверительный, но, как показывают числовые расчеты, менее точный интервал $\bar{X} \pm \lambda_\alpha \sqrt{\bar{X}/n}$.

На этом мы заканчиваем изложение простейших методов построения доверительных и асимптотически доверительных интервалов на основе подбора опорных функций. О проблеме оптимального интервального оценивания мы поговорим позднее, изучив теорию оптимальной проверки гипотез – высказываний о возможных значениях параметра θ . Оставшиеся лекции будут посвящены именно этой теории.

§7. Статистическая проверка гипотез (критерии значимости)

Лекция 11

В приложениях математической статистики существует обширный класс задач, в которых требуется проверить истинность некоторого высказывания относительно исследуемого объекта или выбрать одно из альтернативных решений, которое определит дальнейшее поведение статистика по отношению к этому объекту. Например, при аттестации партии дизельного топлива по общему содержанию серы мы должны не только дать точечную оценку данной характеристики топлива, но и принять решение о качестве выпускаемого продукта, которое повлечет за собой одно из следующих действий – или отослать топливо потребителю, или произвести дополнительную очистку топлива от вредных примесей. Точно так же в примере 1.2 мы строили статистическое правило, позволяющее принять одно из двух решений относительно нового лечебного препарата – или признать его эффективным и внедрить в лечебную практику, или запретить его дальнейшее использование. В исследованиях, подобных опытам Менделя, часто надо проверить гипотезу относительно предполагаемого значения вероятности наследования доминантного признака. Селекционер, работающий над получением нового вида пшеницы, должен подкрепить свое заключение о превосходстве нового вида над тем, который уже используется в сельскохозяйственной практике, с помощью сопоставления данных об урожайности этих видов. И так далее, и тому подобное, – вы сами можете привести примеры таких задач по выбору одного из ряда альтернативных решений.

В нашем курсе математической статистики мы рассмотрим задачи, связанные только с выбором одного из двух решений. Пусть мы высказываем некоторое суждение (или предпринимаем действие) об исследуемом объекте, и пусть d_0 – решение об истинности этого суждения, в то время как d_1 – решение о его ложности. Таким образом, пространство решений \mathcal{D} в данной статистической проблеме состоит из точек: $\mathcal{D} = \{d_0, d_1\}$.

Для выбора одного из решений мы наблюдаем случайную выборку $X^{(n)}$ из некоторого распределения P_θ , значение параметра θ которого нам неизвестно. Пусть Θ – область возможных значений θ , которую мы назвали параметрическим пространством. В соответствии с принятой нами в §1 идеологией статистического вывода мы сопоставляем каждому решению $d \in \mathcal{D}$ определенное подмножество Θ_d пространства Θ , то есть интерпретируем каждое решение в терминах высказываний об истинном значении

параметра θ . В нашей статистической проблеме выбора одного из двух решений положим $\Theta_i = \Theta_{d_i}$, $i = 0, 1$, и введем ряд понятий и определений, используемых при решении этой проблемы.

Утверждение $H_0 : \theta \in \Theta_0$ называется *нулевой гипотезой*, а утверждение $H_1 : \theta \in \Theta_1$ – *альтернативной гипотезой* или (коротко) *альтернативой*. Гипотеза H_i называется *простой*, если соответствующее Θ_i состоит из одной точки параметрического пространства Θ ; в противном случае H_i называется *сложной* гипотезой; $i = 0, 1$. Так, в примере 1.2 с испытанием нового лечебного препарата параметр θ означал вероятность успешного лечения каждого пациента, и нулевая гипотеза $H_0 : \theta = 1/2$ о “нейтральности” препарата есть простая гипотеза, в то время как альтернативная гипотеза $H_1 : \theta > 1/2$ об его эффективности – сложная гипотеза.

Правило, по которому принимается или отвергается нулевая гипотеза H_0 , называется *критерием*. Иногда добавляется – *критерий согласия* (с нулевой гипотезой), особенно, когда альтернатива H_1 определена не совсем четко и под H_1 подразумевается “все остальное”. В случае полного равноправия гипотез говорят о критерии *различения гипотез*. Критерий определяется заданием особого подмножества S выборочного пространства X^n , которое называется *критической областью*: если выборочные данные $x^{(n)}$ попадают в эту область, то нулевая гипотеза H_0 отклоняется и принимается альтернативное решение – справедлива H_1 . Область $A = S^c = X^n \setminus S$ называется *областью принятия* нулевой гипотезы. Нам будет удобно проводить спецификацию критической области в виде ее индикаторной функции $\varphi = \varphi(X^{(n)})$, которая называется *критической функцией* или, поскольку она определяет статистическое правило проверки гипотезы, просто *критерием*. Итак, функция $\varphi(X^{(n)})$ есть бинарная случайная величина, принимающая значение 1, если произошло событие $X^{(n)} \in S$, и значение 0, если произошло противоположное событие $X^{(n)} \in A$. Понятно, что математическое ожидание $\mathbf{E}\varphi(X^{(n)})$ означает вероятность отклонения гипотезы H_0 .

В рассматриваемой статистической проблеме величина риска, связанная с отклонением верной гипотезы, обычно соотносится с функцией потерь типа 1 – 0: потери считаются равными 1, если принята гипотеза H_i , а в действительности $\theta \in \Theta_{1-i}$, $i = 0, 1$; если же принята H_i и $\theta \in \Theta_i$, $i = 0, 1$, то потери полагаются равными нулю. Легко видеть, что величина риска при любом значении параметра θ может быть определена с помощью функции $m(\theta) = \mathbf{E}_\theta\varphi(X^{(n)}) = P_\theta(X^{(n)} \in S)$, которая называется *функцией мощности* критерия φ . Эта функция указывает, как часто мы отклоняем нулевую

гипотезу, когда θ – истинное значение параметра, и хорошим следует считать тот критерий, у которого функция $m(\theta)$ принимает близкие к нулю значения в области Θ_0 и близкие к единице – в области Θ_1 . В связи с этим вводятся две компоненты функции риска: $\alpha(\theta) = m(\theta)$ при $\theta \in \Theta_0$ и $\beta(\theta) = 1 - m(\theta)$ при $\theta \in \Theta_1$. Функция $\alpha(\theta)$, $\theta \in \Theta_0$ называется *вероятностью ошибки первого рода* – она указывает относительную частоту отклонения гипотезы H_0 , когда она в действительности верна ($\theta \in \Theta_0$). Функция $\beta(\theta)$, $\theta \in \Theta_1$ называется *вероятностью ошибки второго рода* – она указывает относительную частоту принятия гипотезы H_0 , когда она ложна (верна альтернативная гипотеза $H_1 : \theta \in \Theta_1$). Заметим, что функция мощности $m(\theta)$ в области Θ_1 трактуется как вероятность отклонения гипотезы H_0 , когда в действительности выбор идет из распределения с альтернативным значением $\theta \in \Theta_1$, и поэтому часть $m(\theta)$ при $\theta \in \Theta_1$ называется *мощностью* критерия φ .

Легко понять, что при фиксированном объеме наблюдений n невозможно одновременно минимизировать вероятности обеих ошибок, – для уменьшения вероятности ошибки первого рода $\alpha(\theta) = P_\theta(X^{(n)} \in S)$, $\theta \in \Theta_0$, необходимо уменьшить критическую область S , что приведет к увеличению области A принятия нулевой гипотезы и, следовательно, к увеличению вероятности ошибки второго рода $\beta(u) = P_u(X^{(n)} \in A)$, $u \in \Theta_1$. Здесь возникает такая же ситуация, что и в проблеме построения оценки параметра θ с равномерно минимальным риском, – такие оценки существуют только в определенном классе статистических правил, например, в классе несмещенных оценок. Однако, даже и помимо задачи проверки гипотез с минимальной вероятностью ошибки, и намного раньше создания общей теории наиболее мощных критериев в статистической практике сложился следующий подход к управлению риском критерия.

Предположим, что отклонение гипотезы H_0 , когда она в действительности верна, приводит к более тяжким последствиям, чем ее принятие при справедливости альтернативы. В таком случае мы заинтересованы в первую очередь контролировать вероятность ошибки первого рода. С этой целью заранее фиксируется (выбирается) некоторый уровень α , выше которого вероятность ошибки первого рода не допустима, и критическая область S (критерий φ) определяется таким образом, что $\alpha(\theta) \leq \alpha$, каково бы ни было $\theta \in \Theta_0$. Это ограничение α на вероятность ошибки первого рода называется *уровнем значимости*, а сам критерий φ , для которого выполняется это ограничение, – *критерием уровня α* . Наибольшее значение

вероятности ошибки первого рода

$$\bar{\alpha} = \sup_{\theta \in \Theta_0} \alpha(\theta)$$

называется *размером* критерия φ , и если $\bar{\alpha} = \alpha$, то говорят о *критерии φ размера α* .

В этом выборе ограничения именно на вероятность ошибки первого, а не второго рода проявляется типичная асимметрия в практической ценности гипотезы и альтернативы. Например, если проверяется эффективность нового лекарственного препарата, то нулевой гипотезе должно соответствовать решение о его неэффективности, ибо, отклонив эту гипотезу, когда она верна, мы внедрим в лечебную практику бесполезное или вредное лекарство, что приведет к более тяжким последствиям, чем отклонение в действительности эффективного препарата. Но если мы ищем золото, анализируя состав кернов при бурении предполагаемого месторождения, то естественно принять за нулевую гипотезу утверждение о наличии золота, ибо отклонив ее, когда она верна, мы потеряем намного больше, чем стоимость нескольких дополнительных анализов, удостоверяющих, что золото в разбуренной местности отсутствует.

Следует также обратить особое внимание на общую методологию проверки гипотез, отражаемую в выборе малого значения уровня α . Если наши выборочные данные попадают в область S с исключительно малой вероятностью, то естественно предположить, что то утверждение, которое привело к этому маловероятному событию, не соответствует истине и отклонить его. Поступая таким образом, мы будем терять в действительности верную гипотезу H_0 крайне редко – не более, чем в $100\alpha\%$ случаев.

Простейший метод построения критериев значимости состоит в использовании состоятельных оценок тестируемого параметра θ . Рассмотрим простейший случай: θ – скалярный параметр, вероятностная модель не содержит других (мешающих) параметров и проверяется простая гипотеза $H_0 : \theta = \theta_0$ при альтернативе $H_1 : \theta \neq \theta_0$, где θ_0 – некоторое, априори фиксированное значение параметра θ (например, в опытах Менделя проверяется гипотеза: вероятность θ наследования доминантного признака равна $\theta_0 = 3/4$). Если $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ – состоятельная оценка θ , дисперсия которой стремится к нулю при $n \rightarrow \infty$ как $O(1/\sqrt{n})$, то естественно определить критическую область посредством неравенства $\sqrt{n}|\hat{\theta}_n(X^{(n)}) - \theta_0| > C$. Вероятность ошибки первого рода такого критерия

$$\alpha(\theta_0, C) = P_{\theta_0}(\sqrt{n}|\hat{\theta}_n(X^{(n)}) - \theta_0| > C),$$

и приравнивая эту вероятность заданному уровню значимости α , находим *критическую константу* $C = C(\alpha)$ как квантиль распределения случайной величины $\sqrt{n}|\hat{\theta}_n(X^{(n)}) - \theta_0|$; такой выбор C приводит к критерию уровня α . Если $\theta (\neq \theta_0)$ – некоторое альтернативное значение параметра, то, в силу состоятельности оценки, $\sqrt{n}|\hat{\theta}_n(X^{(n)}) - \theta_0| \xrightarrow{P} \infty$, и поэтому вероятность ошибки второго рода

$$\beta(\theta) = P_{\theta}(\sqrt{n}|\hat{\theta}_n(X^{(n)}) - \theta_0| \leq C(\alpha)) \rightarrow 0,$$

когда $n \rightarrow \infty$. Таким образом, мы получаем критерий заданного уровня α , обладающий к тому же свойством *состоятельности* – его вероятность ошибки второго рода стремится к нулю при неограниченном возрастании объема выборки n .

Сформулируем теперь основную задачу теории статистической проверки гипотез: *требуется найти такой критерий φ уровня α , который равномерно по всем $\theta \in \Theta_1$ максимизирует мощность $t(\theta)$ или, что то же, равномерно по $\theta \in \Theta_1$ минимизирует вероятность ошибки второго рода $\beta(\theta)$* . Мы укажем метод построения таких *равномерно наиболее мощных критериев* заданного уровня α в следующем параграфе, а пока обратимся к иллюстрациям введенных понятий и построению наиболее часто используемых на практике критериев, касающихся проверки гипотез о значениях параметров нормального распределения.

1⁰. ПРОВЕРКА ГИПОТЕЗЫ О ВЕЛИЧИНЕ СРЕДНЕГО ЗНАЧЕНИЯ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПРИ ИЗВЕСТНОЙ ДИСПЕРСИИ. Рассмотрим сначала наиболее часто встречающуюся в практических применениях математической статистики задачу проверки сложной гипотезы $H_0 : \mu \leq \mu_0$ при сложной альтернативе $H_1 : \mu > \mu_0$ о среднем значении μ нормального (μ, σ^2) распределения при известном значении дисперсии σ^2 . Выборочное среднее $\bar{X} = n^{-1} \sum_1^n X_k$ есть оптимальная оценка неизвестного значения μ , и поэтому, в соответствии с только что предложенным методом построения состоятельных критериев, рассмотрим критерий, отвергающий нулевую гипотезу H_0 , когда $\sqrt{n}(\bar{X} - \mu_0) > C$, или, что то же, $\bar{X} > C$, поскольку значения μ_0 и n фиксированы и известны. Постоянная C должна выбираться по заданному уровню значимости α , ограничивающему максимальное значение вероятности ошибки первого рода.

Так как при выборе из нормального (μ, σ^2) распределения статистика

$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, то функция мощности этого критерия

$$m(\mu) = P_{\mu}(\bar{X} > C) = 1 - \Phi\left(\frac{C - \mu}{\sigma}\sqrt{n}\right) = \Phi\left(\frac{\mu - C}{\sigma}\sqrt{n}\right).$$

Легко видеть, что $m(\mu)$ – строго возрастает с ростом μ , так что размер критерия

$$\bar{\alpha} = \max_{\mu \leq \mu_0} m(\mu) = m(\mu_0) = 1 - \Phi\left(\frac{C - \mu_0}{\sigma}\sqrt{n}\right).$$

Приравнивая размер критерия уровню значимости α , находим критическое значение $C(\alpha) = \mu_0 + \Phi^{-1}(1 - \alpha)\sigma/\sqrt{n}$.

Вероятность ошибки второго рода нашего критерия размера α

$$\begin{aligned} \beta(\mu) &= P_{\mu}(\bar{X} \leq C(\alpha)) = \Phi\left(\frac{C(\alpha) - \mu}{\sigma}\sqrt{n}\right) = \\ &\Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} + \Phi^{-1}(1 - \alpha)\right), \quad \mu > \mu_0, \end{aligned} \quad (1)$$

убывает с ростом μ по мере ее отхода от граничного значения μ_0 . Наибольшее значение $\beta(\mu)$ достигается в точке $\mu = \mu_0$ и равно $1 - \alpha$. Это значение не зависит от размера выборки n , и поэтому требуются дополнительные соображения при планировании объема наблюдений. Обычно используется метод введения так называемой *зоны безразличия* – интервала (μ_0, μ_1) , который выбирается из тех соображений, что при истинном значении $\mu \in (\mu_0, \mu_1)$ принятие нулевой гипотезы H_0 не приводит к слишком тяжелым последствиям. Однако при истинном $\mu \geq \mu_1$ вероятность принятия H_0 должна быть под контролем и не превосходить некоторого предписанного значения β . Это обстоятельство позволяет спланировать объем выборки n , определив его из неравенства $\beta(\mu_1) \leq \beta$. Используя формулу (1) для $\beta(\mu)$, находим, что *объем выборки* $n = n(\alpha, \beta, \mu_0, \mu_1)$, *необходимый для различения гипотез* $\mu \leq \mu_0$ и $\mu \geq \mu_1$ *с заданными ограничениями* α и β *на вероятности ошибок первого и второго рода*, равен наименьшему целому n , удовлетворяющему неравенству

$$n \geq \frac{[\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)]^2}{(\mu_1 - \mu_0)^2} \sigma^2.$$

Аналогичным методом строится критерий для проверки простой гипотезы $H_0 : \mu = \mu_0$ при сложной альтернативе $H_1 : \mu \neq \mu_0$. В этой задаче естественно определить критическую область посредством неравенства

$|\bar{X} - \mu_0| > C$. Функция мощности такого критерия

$$m(\mu) = 1 - \left[\Phi \left(\frac{C + \mu_0 - \mu}{\sigma} \sqrt{n} \right) - \Phi \left(\frac{-C + \mu_0 - \mu}{\sigma} \sqrt{n} \right) \right]$$

строго убывает при $\mu < \mu_0$, возрастает при $\mu > \mu_0$ и при $\mu = \mu_0$ равна вероятности ошибки первого рода. Таким образом, критическая константа $C = C(\alpha)$ определяется по заданному уровню значимости α из уравнения

$$m(\mu_0) = 1 - \left[\Phi \left(\frac{C}{\sigma} \sqrt{n} \right) - \Phi \left(\frac{-C}{\sigma} \sqrt{n} \right) \right] = 2 \left[1 - \Phi \left(\frac{C}{\sigma} \sqrt{n} \right) \right] = \alpha,$$

откуда $C(\alpha) = \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n}$.

Лекция 12

2⁰. ПРОВЕРКА ГИПОТЕЗЫ О ВЕЛИЧИНЕ ДИСПЕРСИИ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПРИ НЕИЗВЕСТНОМ СРЕДНЕМ ЗНАЧЕНИИ. Это типичная задача контроля за величиной случайной ошибки в параллельных наблюдениях некоторой характеристики исследуемого объекта. Так как превышение характеристики случайной погрешности σ над некоторым номиналом σ_0 в случае, когда мы утверждаем $\sigma \leq \sigma_0$, влечет более серьезные последствия, чем неоправданные претензии к слишком большому разбросу в данных, то следует принять за нулевую гипотезу $\sigma > \sigma_0$. Проверка этой гипотезы проводится при естественной альтернативе $H_1 : \sigma \leq \sigma_0$, причем мы не знаем значения мешающего параметра μ – среднего значения нормального распределения, из которого производится выбор.

Как нам известно, выборочная дисперсия $S^2 = n^{-1} \sum_1^n (X_k - \bar{X})^2$ есть состоятельная оценка σ^2 , ее распределение не зависит от μ , а случайная величина nS^2/σ^2 имеет хи-квадрат распределение с $n-1$ степенью свободы. Таким образом, разумно рассмотреть критерий с критической областью $nS^2 < C$. Функция мощности такого критерия

$$m(\sigma) = P_{\mu, \sigma} \left(\frac{nS^2}{\sigma^2} \leq \frac{C}{\sigma^2} \right) = K_{n-1} \left(\frac{C}{\sigma^2} \right)$$

монотонно убывает с ростом σ , поэтому наибольшее значение вероятности ошибки первого рода достигается при $\sigma = \sigma_0$, и критическое значение $C(\alpha)$

критерия требуемого размера α определяется из уравнения $K_{n-1}(C\sigma_0^{-2}) = \alpha$. Итак, $C(\alpha) = \sigma_0^2 K_{n-1}^{-1}(\alpha)$; вероятность ошибки второго рода

$$\beta(\sigma) = P_{\mu, \sigma}(nS^2 > C(\alpha)) = 1 - K_{n-1}\left(\frac{\sigma_0^2}{\sigma^2} K_{n-1}^{-1}(\alpha)\right), \quad \sigma \leq \sigma_0,$$

монотонно убывает по мере отхода истинного значения σ от номинала σ_0 .

3⁰. ПРОВЕРКА ГИПОТЕЗЫ О ВЕЛИЧИНЕ СРЕДНЕГО ЗНАЧЕНИЯ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПРИ НЕИЗВЕСТНОЙ ДИСПЕРСИИ (ОДНОВЫБОРОЧНЫЙ КРИТЕРИЙ СТЬЮДЕНТА). Вы, наверное, обратили внимание, что при построении критериев значимости мы по существу используем методы построения доверительных множеств? Это, действительно, так – между задачами доверительной оценки и проверки гипотез существует много общего, и, решив одну задачу, мы сразу же получаем решение другой. В конце этого параграфа мы формализуем этот параллелизм, а пока будем использовать его на интуитивном уровне: предлагается использовать для проверки гипотез о среднем значении нормального распределения статистику Стьюдента.

Рассмотрим сначала задачу проверки сложной гипотезы $H_0 : \mu \leq \mu_0$ при сложной альтернативе $H_1 : \mu > \mu_0$. Так как выборочное среднее \bar{X} есть состоятельная оценка значения μ , то статистика Стьюдента

$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n-1}$$

опосредственно, через выборочные данные, характеризует удаленность истинного среднего значения μ от границы μ_0 , разделяющей гипотезу и альтернативу. Поэтому предлагается отвергать нулевую гипотезу $\mu \leq \mu_0$, если $T > C$, выбирая C , как обычно, по заданному уровню значимости α . Для решения последней задачи необходимо исследовать поведение функции мощности $t(\mu; \sigma) = P_{\mu, \sigma}(T > C)$ критерия $T > C$. Если мы покажем, что $t(\mu; \sigma)$ есть монотонно возрастающая функция аргумента μ при любом фиксированном значении аргумента σ , то наибольшее значение вероятности ошибки первого рода $\alpha(\mu; \sigma) = t(\mu; \sigma)$, $\mu \leq \mu_0$ при каждом фиксированном σ будет достигаться в точке $\mu = \mu_0$. Следовательно, размер критерия в таком случае будет равен (см. пункт 4⁰ предыдущего параграфа)

$$\bar{\alpha} = t(\mu_0; \sigma) = P_{\mu_0, \sigma}(T > C) = 1 - S_{n-1}(C),$$

где $S_\nu(\cdot)$ – функция распределения Стьюдента с ν степенями свободы. Таким образом, мы получим свободный от неизвестного значения σ критерий $T > C(\alpha)$ требуемого размера α с критической константой $C(\alpha) = S_{n-1}^{-1}(1 - \alpha)$. Это и есть то статистическое правило, которое обычно называется *критерием Стьюдента* или *t-критерием*.

Покажем теперь, что вероятность (функция мощности) $P_{\mu,\sigma}(T > C)$ монотонно возрастает с ростом μ при любых фиксированных значениях σ и C . С этой целью представим статистику T в следующем виде:

$$T = \frac{\bar{X} - \mu}{S} \sqrt{n-1} + \frac{\mu - \mu_0}{\sigma} \cdot \frac{\sigma}{S} \sqrt{n-1}.$$

Если μ – среднее значение нормального распределения, из которого происходит выбор, то первое слагаемое в этом представлении есть стьюдентовская случайная величина с $n - 1$ степенью свободы. Второе слагаемое есть произведение параметрической функции $\Delta(\mu) = (\mu - \mu_0)\sqrt{n-1}/\sigma$ на положительную случайную величину σ/S , распределение которой не зависит от μ и σ . При фиксированном σ функция $\Delta(\mu)$ возрастает с ростом μ и при этом все второе слагаемое возрастает, что влечет увеличение вероятности события перескока статистикой T порога C , то есть вероятности события $T > C$.

Итак, мы построили критерий проверки *односторонней* гипотезы $\mu \leq \mu_0$ при односторонней альтернативе $\mu > \mu_0$. Функция мощности этого критерия зависит от μ и σ только через параметрическую функцию $\Delta = (\mu - \mu_0)\sqrt{n-1}/\sigma$, которая называется *параметром нецентральности*. Распределение статистики $T = (\bar{X} - \mu_0)\sqrt{n-1}/S$ при произвольных μ и σ , через которое выражается функция мощности критерия Стьюдента, называется *нецентральным* распределением Стьюдента с $n - 1$ степенью свободы; таблицы этого распределения, зависящего от параметра нецентральности Δ , можно найти в ТМС.

Понятно, что построение критерия проверки простой гипотезы $\mu = \mu_0$ при *двусторонней* (сложной) альтернативе $\mu \neq \mu_0$ не вызывает принципиальных затруднений. Это критерий с критической областью $|T| > C$, где критическая константа $C = C(\alpha) = S_{n-1}^{-1}(1 - \alpha/2)$.

4⁰. СРАВНЕНИЕ СРЕДНИХ ЗНАЧЕНИЙ ДВУХ НОРМАЛЬНЫХ РАСПРЕДЕЛЕНИЙ С ОБЩЕЙ НЕИЗВЕСТНОЙ ДИСПЕРСИЕЙ (ДВУХВЫБОРОЧНЫЙ КРИТЕРИЙ СТЬЮДЕНТА). Пусть X и Y – независимые случайные величины, причем $X \sim \mathcal{N}(\mu_1, \sigma^2)$, а $Y \sim \mathcal{N}(\mu_2, \sigma^2)$, так что $\mathbf{D}X = \mathbf{D}Y$. По

двум независимым выборкам $X^{(n)} = (X_1, \dots, X_n)$ и $Y^{(m)} = (Y_1, \dots, Y_m)$ (возможно, разного объема) требуется проверить гипотезу *однородности* $H_0 : \mu_1 = \mu_2$ при альтернативе $H_1 : \mu_1 > \mu_2$. Типичный пример такой задачи – выявление эффекта нового метода лечения на группе из n пациентов посредством сравнения с контрольной группой из m пациентов, лечение которых проводится по старой методике.

Эта задача является для нас несколько новой, поскольку до сих пор мы имели дело только с одной выборкой. Тем не менее, она сводится к той, что мы только что рассмотрели в \mathfrak{Z}^0 , с помощью следующих построений.

Рассмотрим сначала разность выборочных средних $\bar{X} - \bar{Y}$. Эта статистика имеет нормальное распределение со средним $\mathbf{E}(\bar{X} - \bar{Y}) = \mu_1 - \mu_2$ и дисперсией $\mathbf{D}(\bar{X} - \bar{Y}) = \mathbf{D}\bar{X} + \mathbf{D}\bar{Y} = \sigma^2(n^{-1} + m^{-1})$. Следовательно, при справедливости нулевой гипотезы $\mu_1 = \mu_2$ случайная величина

$$\xi = \frac{\bar{X} - \bar{Y}}{\sigma} \sqrt{\frac{nm}{n+m}}$$

имеет стандартное нормальное распределение $\mathcal{N}(0, 1)$. Далее, нормированные выборочные дисперсии nS_X^2/σ^2 и mS_Y^2/σ^2 независимы и распределены по закону хи-квадрат с $n - 1$ и $m - 1$ степенями свободы соответственно. Так как для хи-квадрат распределения, как частного случая гамма-распределения, имеет место теорема сложения, то случайная величина $\eta = (nS_X^2 + mS_Y^2)/\sigma^2$ имеет хи-квадрат распределение с $n + m - 2$ степенями свободы. Таким образом, мы приходим к двухвыборочной статистике Стьюдента

$$T_{n,m} = \frac{\xi}{\sqrt{\eta/(n+m-2)}} = \frac{\bar{X} - \bar{Y}}{\sqrt{nS_X^2 + mS_Y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

распределение которой при справедливости нулевой гипотезы есть распределение Стьюдента с $n + m - 2$ степенями свободы.

Как и в случае одновыборочного критерия Стьюдента в \mathfrak{Z}^0 нетрудно показать, что при любых фиксированных C и σ функция мощности двухвыборочного критерия Стьюдента $T_{n,m} > C$ есть монотонно возрастающая функция параметра нецентральности $\Delta = (\mu_1 - \mu_2)\sqrt{n+m-2}/\sigma$, так что критическая константа C определяется по заданному уровню значимости из уравнения

$$P(T_{n,m} > C) = 1 - S_{n+m-2}(C) = \alpha$$

и равна квантили распределения Стьюдента: $C(\alpha) = S_{n+m-2}^{-1}(1 - \alpha)$. Понятно, что при альтернативе $H_1 : \mu_1 \neq \mu_2$ критическая константа $C(\alpha) = S_{n+m-2}^{-1}(1 - \alpha/2)$.

При использовании этого критерия следует обратить особое внимание на предположение о равенстве дисперсий наблюдаемых случайных величин: $\sigma_X^2 = \sigma_Y^2$. Задача сравнения средних двух нормальных распределений с неравными дисперсиями и с гарантированным ограничением α на вероятность ошибки первого рода называется *проблемой Беренса–Фишера*. Известно лишь асимптотическое решение этой проблемы при больших n и m .

5⁰. СРАВНЕНИЕ ДИСПЕРСИЙ ДВУХ НОРМАЛЬНЫХ РАСПРЕДЕЛЕНИЙ ПРИ НЕИЗВЕСТНЫХ СРЕДНИХ (КРИТЕРИЙ ФИШЕРА). Независимые выборки $X^{(n)}$ и $Y^{(m)}$ берутся из соответствующих нормальных распределений $\mathcal{N}(\mu_1, \sigma_1^2)$ и $\mathcal{N}(\mu_2, \sigma_2^2)$, относительно параметров которого проверяется гипотеза $\sigma_1^2 = \sigma_2^2$ при альтернативе $\sigma_1^2 > \sigma_2^2$ с мешающими параметрами μ_1 и μ_2 .

В этой задаче естественно рассмотреть критерий, основанный на статистике $F = nS_X^2/mS_Y^2$, которая распределена как

$$\frac{\chi_{n-1}^2}{\chi_{m-1}^2} \cdot \frac{\sigma_1^2}{\sigma_2^2}.$$

Функция мощности критерия $F > C$ (который называется *критерием Фишера* или *F-критерием*)

$$m \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = P_{\mu_1, \mu_2, \sigma_1, \sigma_2}(F > C) = P \left(\frac{\chi_{n-1}^2}{\chi_{m-1}^2} > C \cdot \frac{\sigma_2^2}{\sigma_1^2} \right)$$

есть монотонно возрастающая функция отношения дисперсий σ_1^2/σ_2^2 . Для ее вычисления необходимо знать распределение отношения двух независимых случайных величин, распределенных по закону хи-квадрат с $n - 1$ и $m - 1$ степенями свободы. Это так называемое *распределение Фишера* $F_{n-1, m-1}$, плотность которого

$$f_{n-1, m-1}(x) = \frac{\Gamma\left(\frac{n+m-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right) \Gamma\left(\frac{m-1}{2}\right)} \cdot \frac{x^{\frac{n-1}{2}-1}}{(x+1)^{\frac{n+m-2}{2}}}, \quad x > 0,$$

вычисляется столь же просто, как это мы делали при выводе распределения Стьюдента. Таблицы распределения Фишера можно найти в ТМС.

Критическая константа C критерия Фишера заданного размера α определяется как квантиль этого распределения: $C(\alpha) = F_{n-1, m-1}^{-1}(1 - \alpha)$.

Мы завершим иллюстрацию методов построения критериев с помощью состоятельных оценок тестируемого параметра примером, в котором не всегда размер критерия совпадает с заданным уровнем значимости.

6⁰. ПРОВЕРКА ГИПОТЕЗЫ О ВЕРОЯТНОСТИ УСПЕХА В ИСПЫТАНИЯХ БЕРНУЛЛИ. Рассмотрим задачу проверки гипотезы $p = p_0$ против альтернативы $p < p_0$ о вероятности p успешного исхода в испытаниях Бернулли. Пример такой задачи – проверка гипотезы о вероятности наследования доминантного признака в опытах Менделя, когда альтернативная модель предписывает этой вероятности меньшее значение. Предлагаемый ниже метод решения позволяет строить критерии проверки такой гипотезы при альтернативах $p > p_0$ или $p \neq p_0$ посредством простой замены неравенства, определяющего критическую область, на обратное или двустороннее.

Итак, если мы располагаем выборкой $X^{(n)}$ из двухточечного распределения $B(1, p)$, то относительная частота успешных испытаний (выборочное среднее) \bar{X} является несмещенной оценкой p с минимальной дисперсией. В соответствии с предложенной выше идеологией проверки гипотез с помощью оценок тестируемого параметра мы должны отвергать гипотезу $p = p_0$ в пользу $p < p_0$, если $\bar{X} - p_0 < C$. Поскольку статистика $T = n\bar{X} = \sum_{k=1}^n X_k$ имеет биномиальное распределение $B(n, p)$, а значение p_0 задано, то для вычисления функции мощности удобнее записать критическую область в виде $T < C$. Но статистика T принимает только целочисленные значения $0, 1, \dots, n$, поэтому бессмысленно рассматривать дробные значения критических констант. Таким образом, мы приходим к наиболее удобной форме записи критической области в виде $T < C$, где C принимает значения $1, 2, \dots, n$

Функция мощности такого критерия

$$m(p) = P_p(T < C) = \sum_{k=0}^{C-1} C_n^k p^k (1-p)^{n-k},$$

и поскольку проверяется простая гипотеза, то критическая константа C должна определяться по заданному уровню значимости α из неравенства

$$m(p_0) = \sum_{k=0}^{C-1} C_n^k p_0^k (1-p_0)^{n-k} \leq \alpha. \quad (2)$$

Очевидно, что чем больше C , тем больше мощность критерия, и поэтому $C(\alpha)$ следует выбирать как наибольшее целое число, удовлетворяющее неравенству (2). Размер критерия с таким $C(\alpha)$ не обязательно равен α , так что мы можем получить критерий уровня α , но не размера α (в предыдущих примерах с тестовыми статистиками, имеющими распределение непрерывного типа, мы имели критерии размера α). Более того, если p_0 настолько мало, что $(1 - p_0)^n > \alpha$, то не существует таких C , при которых имеет место неравенство (2). В таком случае мы должны принимать нулевую гипотезу при любом результате статистического эксперимента, обеспечивая тем самым нулевой размер такого критерия “уровня α ”.

При больших объемах выборки n можно использовать нормальные аппроксимации биномиального распределения, получая таким образом критерий, размер которого асимптотически ($n \rightarrow \infty$) равен α . Статистика T асимптотически нормальна со средним np и дисперсией $np(1 - p)$, поэтому неравенство (2) для определения критической константы имеет асимптотический аналог

$$\Phi \left(\frac{C - np_0}{\sqrt{np_0(1 - p_0)}} \right) \leq \alpha,$$

откуда $C(\alpha) \approx np_0 - \Phi^{-1}(1 - \alpha)\sqrt{np_0(1 - p_0)}$. Легко понять, что такой метод построения критериев асимптотического уровня α применим для любой критической области, в задании которой используется асимптотически нормальная оценка тестируемого параметра (см. пояснения в предыдущем параграфе перед пунктом 5⁰).

Этот пример показывает, что в случае дискретных распределений задача построения равномерно наиболее мощных критериев значительно усложняется, поскольку один из двух критериев одного и того же уровня α может иметь большую мощность только потому, что он имеет больший размер. Мы столкнемся с этой проблемой в следующем параграфе, но следует заметить, что современная теория наиболее мощных критериев обходит этот неприятный момент за счет расширения понятия статистического правила, вводя так называемые *рандомизированные* критерии. К сожалению, я не располагаю временем познакомить вас с этим замечательным объектом теории статистического вывода.

Мы закончим этот параграф, как и было обещано, формулировкой *принципа двойственности* между задачами проверки гипотез и доверительного оценивания. Пусть $A(\theta_0) \subset \mathcal{X}^n$ – область принятия некоторого критерия

уровня α , тестирующего гипотезу $H_0 : \theta = \theta_0$, и пусть $A(\theta_0)$ определена при любом $\theta_0 \in \Theta$. Для каждого результата $x^{(n)}$ наблюдения случайной выборки $X^{(n)}$ введем подмножество $\Delta(x^{(n)})$ параметрического пространства Θ , положив

$$\Delta(x^{(n)}) = \{\theta : x^{(n)} \in A(\theta)\}.$$

Тогда $\Delta(X^{(n)})$ есть $(1 - \alpha)$ -доверительное множество для параметра θ , поскольку

$$P_\theta \left(\Delta(X^{(n)}) \ni \theta \right) = P_\theta \left(X^{(n)} \in A(\theta) \right) \geq 1 - \alpha.$$

Например, критерий Стьюдента проверки гипотезы $\mu = \mu_0$ о среднем значении нормального (μ, σ^2) распределения с неизвестной дисперсией σ^2 имеет область принятия (см. п. 3⁰ данного параграфа)

$$A(\mu_0) = \left\{ X^{(n)} : \frac{|\bar{X} - \mu_0|}{S} \sqrt{n-1} \leq S_{n-1}^{-1} (1 - \alpha/2) \right\}.$$

Подставим в это неравенство вместо фиксированного μ_0 параметр μ и разрешим неравенство относительно μ . В результате получим доверительное утверждение (см. п. 4⁰ предыдущего параграфа)

$$\bar{X} - St_\alpha / \sqrt{n-1} \leq \mu \leq \bar{X} + St_\alpha / \sqrt{n-1},$$

в котором $t_\alpha = S_{n-1}^{-1} (1 - \alpha/2)$.

Вы сами можете сопоставить доверительные интервалы, построенные в §6, с критериями из §7. При этом сопоставлении можно вывести полезное правило, касающееся доверительной оценки скалярного параметра θ . Если имеется состоятельный критерий проверки гипотезы $\theta = \theta_0$ при двусторонней альтернативе $\theta \neq \theta_0$, то его области принятия соответствует двусторонний доверительный интервал. Если же альтернативная гипотеза носит односторонний характер, то при альтернативе $\theta < \theta_0$ мы получаем верхнюю доверительную границу, а при $\theta > \theta_0$ – нижнюю.

Естественно, принцип двойственности применим и к доверительным интервалам, как статистическим правилам проверки гипотез: гипотеза $\theta \in \Theta_0$ отвергается тогда и только тогда, когда $(1 - \alpha)$ -доверительная область принадлежит подмножеству Θ_1 , и такое статистическое правило (критерий) гарантирует заданное ограничение α на вероятность ошибки первого рода.

§8. Равномерно наиболее мощные критерии

Лекция 13

Метод построения критериев заданного уровня α , который равномерно по всем альтернативным значениям параметра θ максимизирует мощность критерия, существенно опирается на следующее, почти очевидное утверждение, которое в теории проверки гипотез обычно называется *леммой Неймана–Пирсона*.

Рассмотрим вероятностную модель, состоящую всего из двух распределений P_0 и P_1 , с общим носителем X и функциями плотности $f_0(x)$ и $f_1(x)$, $x \in X$. По выборке $X^{(n)}$ проверяется простая гипотеза H_0 : выборка взята из распределения P_0 при простой альтернативе H_1 : выборке соответствует распределение P_1 . Определим критическую функцию $\varphi^*(X^{(n)})$ как индикаторную функцию критической области

$$L(X^{(n)}) = \prod_{k=1}^n \frac{f_1(X_k)}{f_0(X_k)} > C.$$

Статистика L называется *статистикой отношения правдоподобия*, а критерий φ^* – *критерием отношения правдоподобия* или *критерием Неймана–Пирсона*. Критерий φ^* отвергает нулевую гипотезу, если правдоподобие альтернативы

$$f_{1,n}(X^{(n)}) = \prod_1^n f_1(X_k)$$

в C раз превосходит правдоподобие нулевой гипотезы

$$f_{0,n}(X^{(n)}) = \prod_1^n f_0(X_k).$$

Этот критерий обладает следующим замечательным свойством.

Теорема 8.1. *Критерий отношения правдоподобия φ^* является наиболее мощным критерием в классе всех критериев проверки простой гипотезы при простой альтернативе, размер которых не превосходит размера критерия φ^* . Если критерий φ^* имеет размер α , то он обладает наибольшей мощностью в классе всех критериев уровня α .*

Доказательство. Пусть $\varphi = \varphi(X^{(n)})$ – любой другой критерий, размер которого

$$\mathbf{E}_0 \varphi(X^{(n)}) \leq \mathbf{E}_0 \varphi^*(X^{(n)}). \quad (1)$$

Требуется показать, что тогда критерий φ^* имеет большую мощность, чем критерий φ , то есть

$$\mathbf{E}_1 \varphi^*(X^{(n)}) \geq \mathbf{E}_1 \varphi(X^{(n)}).$$

Рассмотрим интеграл

$$\int_{\mathcal{X}^n} \left[\varphi^*(x^{(n)}) - \varphi(x^{(n)}) \right] \left[f_{1,n}(x^{(n)}) - C f_{0,n}(x^{(n)}) \right] d\mu_n(x^{(n)}) =$$

$$\mathbf{E}_1 \varphi^*(X^{(n)}) - \mathbf{E}_1 \varphi(X^{(n)}) - C \left[\mathbf{E}_0 \varphi^*(X^{(n)}) - \mathbf{E}_0 \varphi(X^{(n)}) \right].$$

Достаточно показать, что этот интеграл неотрицателен, и тогда первое утверждение теоремы будет следовать из неравенства:

$$\mathbf{E}_1 \varphi^*(X^{(n)}) - \mathbf{E}_1 \varphi(X^{(n)}) - C \left[\mathbf{E}_0 \varphi^*(X^{(n)}) - \mathbf{E}_0 \varphi(X^{(n)}) \right] \geq 0$$

которое влечет (см. (1))

$$\mathbf{E}_1 \varphi^*(X^{(n)}) - \mathbf{E}_1 \varphi(X^{(n)}) \geq C \left[\mathbf{E}_0 \varphi^*(X^{(n)}) - \mathbf{E}_0 \varphi(X^{(n)}) \right] \geq 0.$$

Покажем, что функции $\varphi^*(x^{(n)}) - \varphi(x^{(n)})$ и $f_{1,n}(x^{(n)}) - C f_{0,n}(x^{(n)})$, произведение которых интегрируется, одновременно положительны или отрицательны при любых $x^{(n)} \in \mathcal{X}^n$. Действительно, если $\varphi^*(x^{(n)}) - \varphi(x^{(n)}) > 0$, то это влечет $\varphi^*(x^{(n)}) = 1$, поскольку критическая функция равна единице, если она не равна нулю. Но, по определению критерия отношения правдоподобия, равенство $\varphi^*(x^{(n)}) = 1$ возможно лишь в случае

$$f_{1,n}(x^{(n)}) - C f_{0,n}(x^{(n)}) > 0.$$

Точно так же устанавливается, что неравенство $\varphi^*(x^{(n)}) - \varphi(x^{(n)}) < 0$ влечет

$$f_{1,n}(x^{(n)}) - C f_{0,n}(x^{(n)}) < 0.$$

Итак, критерий φ^* наиболее мощен в классе всех критериев, размер которых не превосходит размера φ^* . Если же $\mathbf{E}_0 \varphi^*(X^{(n)}) = \alpha$, то это утверждение, очевидно, влечет его наибольшую мощность в классе всех критериев уровня α .

Применение этой теоремы к построению равномерно наиболее мощных критериев мы проиллюстрируем на одном частном примере, из которого будет виден общий подход к данной задаче.

Пример 8.1. Проверка надежности при показательном распределении долговечности. В примере 3.3 мы рассматривали проблему оценки надежности изделия с показательным распределением долговечности. Напомним, случайная величина X , реализация x которой соответствует промежутку времени от начала работы до момента отказа некоторого изделия, называется долговечностью, и по функции распределения $F(x)$, $x \geq 0$ случайной величины X можно рассчитать надежность $H(t)$ изделий, соответствующую гарантийному времени t : $H(t) = P(X \geq t) = 1 - F(t)$.

Пусть долговечность X распределена по показательному закону с функцией распределения

$$F(x | \theta) = 1 - \exp\{-x/\theta\},$$

значение параметра θ которой не известно. Мы должны удостовериться, что надежность выпускаемых изделий достаточно высока: $H(t) \geq P_0$, где P_0 – наименьшая допустимая доля изделий, которые должны прослужить гарантийный срок t .

Это типичная задача проверки гипотез, решение которой начинается с определения нулевой гипотезы H_0 . При этом следует помнить, что в статистическом критерии контролируется вероятность отклонения H_0 , когда она в действительности верна. В нашей конкретной проблеме спецификация нулевой гипотезы во многом зависит от того, что повлечет за собой отказ изделия. Если мы выпускаем бытовые приборы, то отказ изделия до гарантийного срока t повлечет издержки на ремонт, которые могут быть незначительными по сравнению со стоимостью изделия. В таком случае естественно выбрать в качестве нулевой гипотезы утверждение о надежности изделий – отклонив эту гипотезу, когда она верна, мы потеряем дорогостоящую продукцию, ремонт которой нам обошелся бы значительно дешевле, чем ее уничтожение или продажа по бросовой цене. Если же отказ изделия приводит к катастрофическим последствиям, например, к гибели людей, то здесь рассуждать нечего, и за нулевую гипотезу следует брать утверждение о “ненадежности”. Отклонив такую гипотезу, когда она в действительности верна, мы столкнемся с неприемлемо большой долей отказов до истечения гарантийного срока, и поэтому риск от принятия “плохих” изделий должен быть контролируем. Остановимся на этом варианте и приступим к построению равномерно наиболее мощного критерия проверки гипотезы “ненадежности” $H_0 : H(t) < P_0$ при альтернативе $H_1 : H(t) \geq P_0$, когда $H(t) = \exp\{-t/\theta\}$.

В терминах значений параметра θ нулевая гипотеза принимает вид $H_0 :$

$\theta < \theta_0 = -t/\ln P_0$. Зафиксируем некоторое альтернативное значение $\theta_1 > \theta_0$, и рассмотрим задачу проверки простой гипотезы $H'_0 : \theta = \theta_0$ при простой альтернативе $H'_1 : \theta = \theta_1$. Наиболее мощный критерий проверки простой гипотезы при простой альтернативе имеет критическую область вида (см. теорему 8.1)

$$L(X^{(n)}) = \prod_{k=1}^n \frac{f_1(X_k)}{f_0(X_k)} = \frac{\theta_0}{\theta_1} \exp \left\{ \left(\frac{1}{\theta_0} - \frac{1}{\theta_1} \right) \sum_1^n X_k \right\} > C,$$

где критическая константа C определяется по заданному уровню значимости α из условия $P_{\theta_0}(L(X^{(n)}) > C) \leq \alpha$. Поскольку статистика

$$T_n = \sum_1^n X_k$$

имеет гамма-распределение $G(n, \theta_0)$, то для определения C в последнем неравенстве следует положить знак равенства. Кроме этого, статистика отношения правдоподобия $L(X^{(n)})$ есть монотонная функция статистики T_n , поэтому критическую область $L(X^{(n)}) > C$ можно записать в эквивалентной форме $T_n > C$ и находить новое C из равенства

$$P_{\theta_0}(T_n > C) = 1 - G_n(C/\theta_0) = \alpha$$

(собственно говоря, нам все равно, какое C определять, но на практике, вне сомнения, удобнее иметь дело с критической областью $T_n > C$).

Итак,

$$C(\alpha) = \theta_0 \cdot G_n^{-1}(1 - \alpha),$$

где $G_n^{-1}(\cdot)$ – квантиль стандартного гамма-распределения $G(n, 1)$, и критерий

$$\varphi^*(X^{(n)}) = I_{\{T_n > C(\alpha)\}}(X^{(n)})$$

заданного размера α является наиболее мощным в классе всех критериев уровня α , проверяющих гипотезу H'_0 при альтернативе H'_1 . Это означает, что для любого другого критерия φ с

$$\mathbf{E}_{\theta_0} \varphi(X^{(n)}) \leq \alpha$$

выполняется неравенство

$$\mathbf{E}_{\theta_1} \varphi(X^{(n)}) \leq \mathbf{E}_{\theta_1} \varphi^*(X^{(n)}). \quad (2)$$

Но критерий φ^* не зависит от выбора альтернативного значения θ_1 параметра θ – критическая константа $C(\alpha) = \theta_0 \cdot G_n^{-1}(1 - \alpha)$! Следовательно, неравенство (2) справедливо при любых $\theta_1 > \theta_0$, и мы приходим к заключению, что *критерий φ^* есть равномерно наиболее мощный критерий в классе всех критериев уровня α , проверяющих простую гипотезу $H_0' : \theta = \theta_0$ при сложной альтернативе $H_1 : \theta > \theta_0$.*

Далее, функция мощности критерия φ^* , как критерия различения исходных сложных гипотез $H_0 : \theta < \theta_0$ и $H_1 : \theta \geq \theta_0$, равна

$$m(\theta) = \mathbf{E}_\theta \varphi^*(X^{(n)}) = P_\theta(T_n > C(\alpha)) = 1 - G_n(G_n^{-1}(1 - \alpha)\theta_0/\theta), \quad \theta > 0.$$

Это – возрастающая функция θ , поэтому максимум вероятности ошибки первого рода (размер критерия) равен

$$m(\theta_0) = 1 - G_n(G_n^{-1}(1 - \alpha)) = \alpha.$$

Таким образом, *критерий φ^* есть критерий размера α проверки гипотезы H_0 при альтернативе H_1 , обладающий равномерно наибольшей мощностью в классе всех критериев φ с ограничением $\mathbf{E}_{\theta_0} \varphi(X^{(n)}) = \alpha$. Но в таком случае он будет равномерно наиболее мощным и в более узком классе критериев уровня α , то есть критериев φ , удовлетворяющих ограничению $\mathbf{E}_\theta \varphi(X^{(n)}) \leq \alpha$ при любом $\theta < \theta_0$.*

Более того, нетрудно убедиться, что критерий φ^* обладает минимальной вероятностью ошибки первого рода $\alpha(\theta) = m(\theta)$, $\theta \leq \theta_0$ в классе всех критериев уровня α . Для этого достаточно поменять местами нулевую гипотезу и альтернативу и выбрать уровень значимости, равный $1 - \alpha$.

В этом примере построение равномерно наиболее мощного критерия стало возможным благодаря особому свойству статистической структуры показательного распределения: *статистика $L(X^{(n)})$ отношения правдоподобия есть монотонная функция статистики $T_n = \sum_1^n X_k$. Это – частный случай статистических структур, обладающих достаточной статистикой T , ибо в силу теоремы факторизации у таких структур $L(X^{(n)}) = g_{\theta_1}(T)/g_{\theta_0}(T)$ зависит от $X^{(n)}$ только через значения $T(X^{(n)})$. Дополнительное свойство монотонности отношения правдоподобия относительно T обеспечивает существование и возможность конструктивного построения равномерно наиболее мощного критерия, причем критическая область такого критерия обязательно имеет вид $T > C$ или $T < C$. Например, критерий*

$\sum_1^n X_k > C$ при соответствующем выборе C по заданному уровню значимости α будет равномерно наиболее мощным критерием в классе всех критериев уровня α проверки гипотезы $\theta < \theta_0$ при альтернативе $\theta \geq \theta_0$, когда θ есть среднее значение нормального распределения (дисперсия предполагается известной) или параметр масштаба гамма-распределения (параметр формы известен). Но если θ – параметр таких распределений, как двухточечное или Пуассона, то критерий φ^* с критической областью $\sum_1^n X_k > C$ обладает равномерно наибольшей мощностью только в классе тех критериев, размер которых не больше размера φ^* .

Другие критерии, которые мы рассматривали в предыдущем параграфе, также обладают свойством равномерной наибольшей мощности, и при доказательстве этого также используется лемма Неймана–Пирсона, но методика доказательства совершенно другая и требует разработки методов построения критериев, обладающих свойством инвариантности – независимости от мешающих параметров. Но это уже совсем другая область теории проверки гипотез, поговорить о которой у нас не хватает времени. Я лучше расскажу вам о некоторых дополнительных ухищрениях в практических применениях статистических критериев, которые позволяют с большей степенью наглядности оценить степень согласия проверяемой гипотезы с выборочными данными.

Все рассматриваемые нами критерии заданного уровня α обладают тем свойством, что их критические области можно записать в виде $T(X^{(n)}) > C(\alpha)$, где T – некоторая статистика, характеризующая расхождение выборочных данных с предполагаемыми значениями параметра. Увеличение уровня значимости α приводит к уменьшению $C(\alpha)$, и мы получаем систему вложенных друг в друга критических областей. Это замечательное свойство наших критериев позволяет несколько изменить методологию их практического использования. До сих пор мы фиксировали уровень значимости α , находили по нему критическую константу $C(\alpha)$ и сравнивали ее с выборочным значением $t = T(x^{(n)})$ статистики $T = T(X^{(n)})$. Поступим теперь следующим образом. Получив выборочные данные $x^{(n)}$, вычислим значение $t = T(x^{(n)})$ и рассмотрим критерий $T(X^{(n)}) > t$. Размер такого критерия $\alpha_{кр.} = P_0(T(X^{(n)}) > t)$ называется *критическим уровнем значимости*, который трактуется как вероятность получить столь же большие расхождения между выборочными данными и нулевой гипотезой, как и для выборочных данных $x^{(n)}$.

Естественно, мы по-прежнему можем работать с заданным уровнем значимости α , отклоняя нулевую гипотезу, если $\alpha_{\text{кр.}} < \alpha$, и принимая ее в противном случае. Кстати, принимая гипотезу, не следует утверждать, что она верна. На этот счет существует более деликатное выражение: “выборочные данные согласуются с выдвинутой гипотезой,” ибо, как говорил один из создателей математической статистики сэра Д.Фишер, “гипотезы не проверяются, а разве лишь отвергаются”. Так вот, в свете этого высказывания более разумно просто сообщать полученный критический уровень значимости, сопровождая его следующим комментарием, который можно считать международным статистическим стандартом. Если $\alpha_{\text{кр.}} \leq 0.01$, то говорят, что расхождение между гипотезой и выборочными данными *высоко значимо*, если $0.01 < \alpha_{\text{кр.}} \leq 0.05$, то просто – *значимо*, если же $0.05 < \alpha_{\text{кр.}} \leq 0.10$ – *почти значимо*, и в случае $\alpha_{\text{кр.}} > 0.10$ – *не значимо*. Заметим также, что в некоторых применениях критериев значимости (особенно, в медицине) $\alpha_{\text{кр.}}$ называют *достоверностью*. Существуют и другие, совершенно фантастические названия $\alpha_{\text{кр.}}$, которые я не буду здесь приводить в силу их крайне неприличного звучания.

Поговорим теперь об оптимальных свойствах доверительных границ, соответствующих равномерно наиболее мощным критериям. Рассмотрим только случай верхней доверительной границы $\bar{\theta}_n = \bar{\theta}_n(X^{(n)})$.

Определение 8.1 Верхняя $(1 - \alpha)$ -доверительная граница $\bar{\theta}_n$ называется *равномерно наиболее точной*, если она равномерно по всем θ и θ' , удовлетворяющим неравенству $\theta' > \theta$, минимизирует вероятность

$$P_{\theta}(\bar{\theta}_n(X^{(n)}) \geq \theta').$$

Таким образом, в случае равномерно наиболее точной границы $\bar{\theta}_n$ интервал $(-\infty; \bar{\theta}_n]$ с заданной вероятностью $1 - \alpha$ накрывает истинное значение параметра θ , но он с минимальной вероятностью накрывает любые значения θ , лежащие правее истинного.

Если мы проверяем гипотезу $H : \theta = \theta_0$ при альтернативе $K(\theta_0) : \theta < \theta_0$, и область принятия $A(\theta_0)$ равномерно наиболее мощного критерия размера α обладает тем свойством, что подмножество

$$\Delta_n(x^{(n)}) = \{\theta : x^{(n)} \in A(\theta)\}$$

параметрического пространства $\Theta \subseteq \mathbb{R}$ есть интервал $(-\infty : \bar{\theta}_n(x^{(n)}))$, то $\bar{\theta}_n(X^{(n)})$ есть равномерно наиболее точная верхняя $(1 - \alpha)$ -доверительная

граница. Все объясняется довольно просто: вероятность

$$P_{\theta}(\bar{\theta}_n(X^{(n)}) \geq \theta') = P_{\theta}(X^{(n)} \in A(\theta'))$$

есть вероятность ошибки второго рода у критерия проверки гипотезы $H : \theta = \theta'$ при альтернативе $K(\theta') : \theta < \theta'$. Равномерно наиболее мощный критерий естественно обладает равномерно минимальной вероятностью ошибки второго рода. Все построенные нами в §6 доверительные границы обладают оптимальными свойствами с точки зрения малой вероятности накрытия тех значений параметра, которые не соответствуют истине.

§9. Проверка модельных предположений. Критерии согласия

Лекция 14

Рассмотренные нами методы построения оптимальных решающих функций в проблемах оценки параметров и проверки параметрических гипотез существенно опирались на такие особые свойства вероятностных моделей, как существование достаточных статистик, монотонность отношения правдоподобия относительно некоторой статистики, независимость выборок и прочее. Оценить же последствия от использования конкретных решающих функций (найти функцию риска статистического правила) вообще не представляется возможным без знания вероятностной модели. Отсюда возникает необходимость разработки общих методов тестирования (проверки) предлагаемой вероятностной модели $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ по данным случайной выборки, или нескольких выборок, которые предположительно извлекаются из некоторых распределений семейства \mathcal{P} . Значимые расхождения между модельными и эмпирическими распределениями вынуждают статистику пересмотреть посылки, положенные в основу построения вероятностной модели, и тем самым избежать больших потерь от использования заведомо плохих решающих правил (точнее правил, которые оптимальны не для той модели).

Понятно, что речь идет о проверке статистических гипотез без особой спецификации альтернатив к нулевой гипотезе. Статистические правила проверки модельных предположений обычно называются *критериями согласия*, и в математической статистике сложился некоторый традиционный набор таких критериев, обладающих большой универсальностью. Это критерии, с помощью которых можно не только проверять принадлежность распределения наблюдаемой случайной величины к определенному семейству, но и тестировать некоторые более “грубые” черты модели, как то: независимость компонент наблюдаемого случайного вектора (векторной случайной величины), возможность объединения нескольких выборок в одну (проверка гипотезы однородности выборок) и множество других предположений, касающихся структуры выборочных данных. Мы познакомимся в этом параграфе с набором универсальных статистических процедур, объединяемых общим названием *критерии хи-квадрат*. Об одном из них мы уже упоминали в §2 в связи с построением гистограммы выборки; это –

1⁰. КРИТЕРИЙ СОГЛАСИЯ ХИ-КВАДРАТ. Решается статистическая проблема проверки гипотезы о виде распределения наблюдаемой случайной величины X (возможно, векторной). Начнем с простейшего случая, когда построение вероятностной модели привело к полной спецификации распределения, то есть проблема состоит в проверке простой гипотезы H : распределение X на измеримом пространстве (X, \mathcal{A}) ее значений есть $P(A)$, $A \in \mathcal{A}$.

Построение критерия согласия выборочных данных с распределением P начинается с разбиения пространства X на $r \geq 2$ частей

$$A_1, \dots, A_r; \quad x = \sum_1^r A_i.$$

Рекомендации по выбору числа r и способу разбиения носят довольно расплывчатый характер, и если не уточнять возможные альтернативы к P , то, как вы сами понимаете, таких рекомендаций не может быть в принципе. Главное, разбиение не должно определяться выборочными значениями, надо стремиться к областям одинаковой конфигурации и размера, не следует делать слишком подробное разбиение. Например, если $X = \mathbb{R}$ (наблюдается действительная случайная величина), то прямая \mathbb{R} разбивается на r интервалов вида

$$(-\infty, a], (a, a + \Delta], (a + \Delta, a + 2\Delta], \dots, \\ (a + (r - 3)\Delta, a + (r - 2)\Delta], (a + (r - 2)\Delta, +\infty),$$

так что длина внутренних интервалов постоянна и равна Δ . Конечно, выбор r зависит от объема выборки n , но даже при исключительно больших n не делается более 15-20 разбиений; этого вполне достаточно, чтобы в гистограмме отразить всю специфику формы тестируемого распределения.

После разбиения X проводится сортировка выборочных данных по областям разбиений и подсчитываются количества

$$\nu_1, \dots, \nu_r, \quad \sum_1^r \nu_i = n,$$

данных, попавших в соответствующие области A_1, \dots, A_r . Вычисляются “теоретические” вероятности $p_i = P(A_i)$, $i = 1, \dots, r$ попадания выборочных данных в эти области и вычисляется значение x^2 тестовой статистики

$$X^2 = \sum_{i=1}^r \frac{(\nu_i - np_i)^2}{np_i}.$$

Гипотеза H отвергается, если $x^2 > C$, где критическая константа C выбирается по заданному уровню значимости α как наименьшее число, удовлетворяющее неравенству $P(X^2 > C) \leq \alpha$. Естественно, на практике используют критический уровень значимости $\alpha_{кр.} = P(X^2 > x^2)$, сопровождая его комментариями типа тех, которые были приведены в предыдущем параграфе после введения понятия критического уровня значимости. Однако точное распределение статистики X^2 найти в явном виде не представляется возможным; предельное распределение X^2 при $n \rightarrow \infty$ установил К.Пирсон в самом начале XX века.

Теорема 9.1. *Если число разбиений $r \geq 2$ фиксировано, а объем выборки $n \rightarrow \infty$, то распределение X^2 сходится к распределению хи-квадрат с $r - 1$ степенью свободы.*

Доказательство. Очевидно, для вывода предельного распределения X^2 следует в первую очередь обратиться к совместному распределению частот

$$\nu_1, \dots, \nu_r, \sum_1^r \nu_i = n.$$

Это мультиномиальное распределение $\mathcal{M}(r, n, p)$, (см. §9 курса ТВ) с функцией плотности

$$f(x_1, \dots, x_r) = P(\nu_1 = x_1, \dots, \nu_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r},$$

сосредоточенное на целочисленной решетке $\sum_1^r x_i = n$. Теорема 9.1 из курса ТВ утверждает, что совместное распределение первых $r - 1$ частот ν_1, \dots, ν_{r-1} аппроксимируется $r - 1$ -мерным нормальным распределением. Естественно, предельное распределение всего вектора частот ν_1, \dots, ν_r при соответствующей нормировке на их средние значения и стандартные отклонения будет вырожденным, ибо $\sum_1^r \nu_i = n$. Вырожденные распределения лучше всего исследовать с помощью характеристических функций, ибо такие распределения можно записать в явном виде, только переходя к системе координат на той гиперповерхности, где сосредоточено такое распределение, и это чрезвычайно усложняет технику асимптотического анализа распределений. Итак, найдем совместную характеристическую функцию ν_1, \dots, ν_r .

Вспомним схему мультиномиальных испытаний. Мы наблюдаем выборку Y_1, \dots, Y_n из распределения случайного вектора $Y = (X_1, \dots, X_r)$, все

компоненты которого, за исключением одной (скажем, X_j), могут принимать только нулевые значения, в то время как $X_j = 1$. Каждая компонента Y_i выборки $Y^{(n)} = (Y_1, \dots, Y_n)$ есть независимая копия Y , так что $Y_i = (X_{1i}, \dots, X_{ri})$ и X_{ji} – копия (в смысле одинаковости распределения) X_j , $j = 1, \dots, r$, $i = 1, \dots, n$. В таких обозначениях

$$\nu_j = \sum_{i=1}^n X_{ji}, \quad j = 1, \dots, r.$$

Если мы найдем характеристическую функцию $\varphi_Y(\mathbf{t})$, $\mathbf{t} = (t_1, \dots, t_r)$, наблюдаемого вектора Y , то характеристическая функция $\varphi_\nu(\mathbf{t})$ вектора частот $\nu = (\nu_1, \dots, \nu_r)$ будет вычисляться по формуле $\varphi_\nu(\mathbf{t}) = \varphi_Y^n(\mathbf{t})$, ибо характеристическая функция суммы независимых случайных величин равна произведению характеристических функций слагаемых (пункт 3⁰ теоремы 12.1 курса ТВ). Но характеристическая функция вектора Y (напомним, $\sum_1^r X_j = 1$)

$$\varphi_Y(\mathbf{t}) = \mathbf{E} \exp \left\{ \mathbf{i} \sum_1^r t_j X_j \right\} = \sum_1^r p_j e^{\mathbf{i} t_j},$$

и поэтому

$$\varphi_\nu(\mathbf{t}) = \left(\sum_1^r p_j e^{\mathbf{i} t_j} \right)^n.$$

Теперь перейдем к асимптотическому анализу характеристической функции вектора X нормированных частот

$$X_j = \frac{\nu_j - np_j}{\sqrt{np_j}}, \quad j = 1, \dots, r,$$

сумма квадратов компонент которого составляет тестовую статистику X^2 (извините, что использую букву X в новом смысле, но не хочется вводить для обозначения случайных величин новые символы). Характеристическая функция случайного вектора, компоненты которого подвергнуты линейному преобразованию, вычисляется по формуле, аналогичной пункту 2⁰ теоремы 12.1:

$$\varphi_X(\mathbf{t}) = \exp \left\{ -\mathbf{i} \sum_1^r t_j \sqrt{np_j} \right\} \left(\sum_1^r p_j \exp \left\{ \frac{\mathbf{i} t_j}{\sqrt{np_j}} \right\} \right)^n.$$

Разложим логарифм этой функции в ряд Маклорена по степеням t_1, \dots, t_r , как это делалось при доказательстве центральной предельной теоремы:

$$\begin{aligned} \ln \varphi_X(\mathbf{t}) &= -\mathbf{i} \sqrt{n} \sum_1^r t_j \sqrt{p_j} + \\ n \ln \left[1 + \frac{\mathbf{i}}{\sqrt{n}} \sum_1^r t_j \sqrt{p_j} - \frac{1}{2n} \sum_1^r t_j^2 + O(n^{-3/2}) \right] &= \\ = -\frac{1}{2} \sum_1^r t_j^2 + \frac{1}{2} \left(\sum_1^r t_j \sqrt{p_j} \right)^2 + O(n^{-1/2}). \end{aligned}$$

Таким образом, характеристическая функция предельного распределения вектора X нормированных частот есть

$$\lim_{n \rightarrow \infty} \varphi_X(\mathbf{t}) = \exp \left\{ -\frac{1}{2} \left[\sum_1^r t_j^2 - \left(\sum_1^r t_j \sqrt{p_j} \right)^2 \right] \right\}.$$

Это – характеристическая функция r -мерного нормального распределения с нулевыми средними и матрицей ковариаций $\Lambda = \mathbf{I} - \mathbf{p}\mathbf{p}'$, где \mathbf{I} – единичная матрица, а $\mathbf{p} = (\sqrt{p_1}, \dots, \sqrt{p_r})$ – вектор столбец.

Рассмотрим квадратичную форму

$$Q(\mathbf{t}) = \sum_1^r t_j^2 - \left(\sum_1^r t_j \sqrt{p_j} \right)^2,$$

коэффициенты которой определяют ковариации компонент вектора $Z = (Z_1, \dots, Z_r)$, распределенного по нормальному закону. Если произвести ортогональное преобразование \mathbf{A} вектора \mathbf{t} , полагая $\mathbf{u} = \mathbf{A}\mathbf{t}$ и фиксируя последнюю строку матрицы \mathbf{A} таким образом, чтобы в новом векторе $\mathbf{u} = (u_1, \dots, u_r)$ компонента $u_r = \sum_1^r t_j \sqrt{p_j}$, то мы получим квадратичную форму (вспомните аналогичные ортогональные преобразования нормального вектора при выводе распределения выборочной дисперсии в лемме Фишера)

$$Q(\mathbf{t}) = \sum_1^r t_j^2 - \left(\sum_1^r t_j \sqrt{p_j} \right)^2 = \sum_1^r u_j^2 - u_r^2 = \sum_1^{r-1} u_j^2.$$

Таким образом, существует ортогональное преобразование $Y = \mathbf{B}Z$ вектора Z , после которого Y_1, \dots, Y_{r-1} независимы и одинаково нормально распределены со средними, равными нулю, и единичными дисперсиями, а Y_r

имеет нулевое среднее и нулевую дисперсию, то есть $Y_r = 0$ почти наверное. Все это, конечно, следствие вырожденности нормального распределения вектора Z – оно сосредоточено на гиперплоскости $\sum_1^r Z_j \sqrt{p_j} = 0$.

Изучим теперь предельное распределение статистики $X^2 = \sum_1^r X_j^2$. Поскольку предельное распределение вектора X совпадает с распределением вектора Z , то предельное распределение статистики X^2 определяется распределением квадратичной формы $\sum_1^r Z_j^2$. Как известно, ортогональные преобразования не меняют суммы квадратов, поэтому

$$\sum_1^r Z_j^2 = \sum_1^r Y_j^2 = \sum_1^{r-1} Y_j^2.$$

Следовательно, предельное распределение статистики X^2 есть распределение суммы квадратов $r - 1$ независимых случайных величин, имеющих общее стандартное нормальное распределение. По определению это – хи-квадрат распределение с $r - 1$ степенями свободы. Теорема Пирсона доказана.

Лекция 15

Рассмотрим теперь более сложную статистическую проблему, в которой проверяется гипотеза о принадлежности распределения P наблюдаемой случайной величины некоторому параметрическому семейству $\mathcal{P} = \{P_\theta, \theta \in \Theta \subseteq \mathbb{R}^s\}$, индексированному s -мерным параметром $\theta = (\theta_1, \dots, \theta_s)$. В таком случае

$$X^2(\theta) = \sum_{i=1}^r \frac{(\nu_i - np_i(\theta))^2}{np_i(\theta)}$$

не может называться статистикой и ее нельзя использовать для проверки сложной гипотезы $H : P \in \mathcal{P}$. Естественно воспользоваться какой-либо оценкой $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ параметра θ и рассмотреть тестовую статистику

$$\hat{X}^2 = X^2(\hat{\theta}_n) = \sum_{i=1}^r \frac{(\nu_i - np_i(\hat{\theta}_n))^2}{np_i(\hat{\theta}_n)}.$$

Понятно, что распределение статистики \hat{X}^2 может зависеть от метода оценки параметра θ . Однако, если определить оценку $\hat{\theta}_n$ из условия минимума случайной функции $X^2(\theta)$, то, как показал Фишер, при определенных условиях регулярности, которым удовлетворяют все рассмотренные нами

в курсе ТВ вероятностные модели, *предельное распределение статистики* \hat{X}^2 *есть хи-квадрат распределение с* $r - s - 1$ *степенями свободы.* Если же $\hat{\theta}_n$ – оценка θ по методу максимального правдоподобия, то предельное распределение \hat{X}^2 , также при условиях регулярности типа тех, что обеспечивали асимптотическую нормальность $\hat{\theta}_n$, имеет функцию распределения $K(x)$, для которой справедлива двусторонняя оценка

$$K_{r-1}(x) \leq K(x) \leq K_{r-s-1}(x),$$

при любом $x > 0$.

Доказательство этих утверждений достаточно громоздко и мы не будем им заниматься из-за недостатка времени. Идейная сторона проблемы нам ясна, и коль скоро нам сообщили распределение тестовой статистики, то мы можем использовать его для расчета критического уровня значимости. В случае оценки максимального правдоподобия, когда мы располагаем двусторонней оценкой $\alpha_{кр.}$, рекомендуется при отклонении гипотезы ориентироваться на

$$\alpha_{кр.} = 1 - K_{r-1}(x^2) \quad (> 1 - K_{r-s-1}(x^2)),$$

а в случае ее принятия – на

$$\alpha_{кр.} = 1 - K_{r-s-1}(x^2) \quad (< 1 - K_{r-1}(x^2)),$$

чтобы уменьшить риск от принятия неправильного решения.

Критерий хи-квадрат является наиболее универсальным статистическим методом тестирования вероятностной модели, поскольку предельное распределение статистики не зависит от распределения наблюдаемой случайной величины даже в том случае, когда это распределение зависит от параметров, значение которых неизвестно. Критерий Колмогорова

$$\sqrt{n}D_n = \sqrt{n} \sup_x |F_n(x) - F(x)| > C,$$

о котором говорилось в начале §2, можно использовать только для проверки простой гипотезы $F(\cdot) = F_0(\cdot)$ о виде функции распределения. Если $F_0(x | \theta)$ зависит от параметра θ и в статистику $\sqrt{n}D_n$ вместо $F(x)$ подставляется $F_0(x | \hat{\theta}_n(X^{(n)}))$, то распределение модифицированной таким образом статистики $\sqrt{n}D_n$ зависит как от вида функции F_0 , так и от параметра θ . Существует, правда, несколько случаев особой связи между x и θ в записи функции F_0 , при наличии которой распределение тестовой статистики

не зависит от θ . Это, например, такие функции распределения с параметрами масштаба и сдвига, как нормальное и показательное. Для тестирования таких распределений составляются специальные таблицы критических констант и критических уровней значимости. Следует заметить, что прямое использование критерия Колмогорова с оценками неизвестных значений параметров является наиболее распространенной ошибкой в практических приложениях методов тестирования вероятностных моделей.

Обратимся теперь к проверке гипотез, касающихся не столько вида распределения наблюдаемых случайных величин, сколько их особых свойств, наличие которых позволяет значительно упростить вероятностную модель и добиться ее более четкой спецификации.

2⁰. КРИТЕРИЙ НЕЗАВИСИМОСТИ ХИ-КВАДРАТ (ТАБЛИЦЫ СОПРЯЖЕННОСТИ ПРИЗНАКОВ). Следующая задача выявления зависимости между определенными признаками наблюдаемых объектов часто возникает в практических приложениях математической статистики. Предположим, что мы случайно выбрали n особей из некоторой этнической популяции, и хотим выяснить, существует ли зависимость между цветом волос и цветом глаз. Мы различаем $s \geq 2$ уровней первого признака (например, блондин, брюнет, шатен и рыжий) и $r \geq 2$ уровней второго (например, карие, серые, голубые и зеленые). Все n особей разбиваются на sr групп в соответствии с наличием тех или иных уровней каждого признака, и составляется следующая таблица частот особей в каждой группе.

Признаки	1	2	...	s	Сумма
1	ν_{11}	ν_{12}	...	ν_{1s}	$\nu_{1\cdot}$
2	ν_{21}	ν_{22}	...	ν_{2s}	$\nu_{2\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
r	ν_{r1}	ν_{r2}	...	ν_{rs}	$\nu_{r\cdot}$
Сумма	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$...	$\nu_{\cdot s}$	n

Такие таблицы, в которых суммы

$$\nu_{i\cdot} = \sum_{j=1}^s \nu_{ij}, \quad \nu_{\cdot j} = \sum_{i=1}^r \nu_{ij},$$

называются *таблицами сопряженности признаков*. Требуется проверить нулевую гипотезу о том, что переменные признаки, по которым построена таблица, независимы. Построим вероятностную модель, соответствующую

такого рода табличным данным и составим статистику X^2 для проверки гипотезы независимости.

Пусть p_{ij} – вероятность того, что случайно отобранная особь имеет i -й уровень по первому признаку и j -й – по второму, $i = 1, \dots, r$, $j = 1, \dots, s$. Гипотеза независимости означает, что $p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$, где

$$p_{i\cdot} = \sum_{j=1}^s p_{ij}, \quad p_{\cdot j} = \sum_{i=1}^r p_{ij}$$

при любых $i = 1, \dots, r$ и $j = 1, \dots, s$. Для проверки гипотезы независимости предлагается использовать тестовую статистику

$$X^2 = \sum_{i,j} \frac{(\nu_{ij} - n p_{i\cdot} p_{\cdot j})^2}{n p_{i\cdot} p_{\cdot j}}, \quad (1)$$

в которой суммирование распространяется на все rs групп таблицы сопряженности признаков. Понятно, что X^2 является тестовой статистикой только в случае известных значений $r + s - 2$ параметров $p_{i\cdot}$ и $p_{\cdot j}$, $i = 1, \dots, r$, $j = 1, \dots, s$ (напомним,

$$\sum_1^r p_{i\cdot} = \sum_1^s p_{\cdot j} = 1,$$

так что с помощью этих соотношений два из $r + s$ параметров, например, $p_{r\cdot}$ и $p_{\cdot s}$, можно выразить через остальные $r + s - 2$ параметров). В этом случае X^2 имеет в пределе ($n \rightarrow \infty$) хи-квадрат распределение с $rs - 1$ степенями свободы.

Конечно, вся проблема состоит в том, что эти параметры неизвестны. Оказывается, оценки максимального правдоподобия

$$\hat{p}_{i\cdot} = \frac{\nu_{i\cdot}}{n}, \quad \hat{p}_{\cdot j} = \frac{\nu_{\cdot j}}{n}, \quad i = 1, \dots, r, \quad j = 1, \dots, s,$$

этих параметров асимптотически эквивалентны оценкам по методу минимума статистики X^2 , и поэтому подстановка в правую часть (1) этих оценок приводит к статистике

$$\hat{X}^2 = n \sum_{i,j} \frac{(\nu_{ij} - \nu_{i\cdot} \nu_{\cdot j} / n)^2}{\nu_{i\cdot} \nu_{\cdot j}} = n \left(\sum_{i,j} \frac{\nu_{ij}^2}{\nu_{i\cdot} \nu_{\cdot j}} - 1 \right),$$

предельное распределение которой есть хи-квадрат с $rs - (r + s - 2) - 1 = (r - 1)(s - 1)$ степенями свободы.

Естественно, статистику \hat{X}^2 можно использовать для проверки независимости компонент двумерного вектора (X, Y) , и при этом таблица сопряженности представляет частотные данные для построения гистограммы двумерной выборки $(X_1, Y_1), \dots, (X_n, Y_n)$. Соответствующим образом нормированная статистика X^2 может служить мерой зависимости признаков (или компонент X и Y случайного вектора).

3⁰. КРИТЕРИЙ ОДНОРОДНОСТИ ХИ-КВАДРАТ. Анализируются данные $s \geq 2$ независимых мультиномиальных схем испытаний с одинаковым числом $r \geq 2$ возможных исходов и соответствующими объемами n_1, \dots, n_s наблюдений в каждой схеме. Проверяется гипотеза *однородности*: все схемы испытаний имеют одинаковый вектор вероятностей

$$\mathbf{p} = (p_1, \dots, p_r), \quad \sum_1^r p_i = 1,$$

появления соответствующих исходов, причем значения компонент вектора \mathbf{p} не известны. Обозначая ν_{ij} частоту появления i -го исхода в j -м испытании, представим данные наблюдений в виде таблицы, аналогичной таблице сопряженности признаков

исх. \ схем.	1	2	...	s	Сумма
1	ν_{11}	ν_{12}	...	ν_{1s}	$\nu_{1.}$
2	ν_{21}	ν_{22}	...	ν_{2s}	$\nu_{2.}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
r	ν_{r1}	ν_{r2}	...	ν_{rs}	$\nu_{r.}$
Сумма	n_1	n_2	...	n_s	n

Составим сначала статистику хи-квадрат для случая известного вектора вероятностей \mathbf{p} :

$$X^2 = \sum_{j=1}^s \sum_{i=1}^r \frac{(\nu_{ij} - n_j p_i)^2}{n_j p_i}.$$

Внутренняя сумма

$$X_j^2 = \sum_{i=1}^r \frac{(\nu_{ij} - n_j p_i)^2}{n_j p_i}$$

представляет статистику хи-квадрат для j -ой схемы мультиномиальных испытаний, и поэтому имеет в пределе ($n_j \rightarrow \infty$) хи-квадрат распределение с $r - 1$ степенями свободы. Статистика X^2 есть сумма s независимых

статистик, каждая из которых имеет предельное хи-квадрат распределение, так что, в силу теоремы сложения, предельное распределение X^2 есть хи-квадрат распределение с $(r - 1)s$ степенями свободы.

В случае неизвестных значений вероятностей исходов, которые при справедливости нулевой гипотезы одинаковы для всех схем испытаний, используем их оценки $\hat{p}_i = \nu_{i\cdot}/n$, $i = 1, \dots, r$, (всего оценивается $r - 1$ параметр). Подстановка этих оценок в X^2 дает статистику

$$\hat{X}^2 = n \sum_{i,j} \frac{(\nu_{ij} - n_j \nu_{i\cdot}/n)^2}{n_j \nu_{i\cdot}} = n \left(\sum_{i,j} \frac{\nu_{ij}^2}{n_j \nu_{i\cdot}} - 1 \right),$$

предельное распределение которой есть хи-квадрат с $(r - 1)s - (r - 1) = (r - 1)(s - 1)$ степенями свободы. Замечателен тот факт, что мы получили тестовую статистику такого же вида и с тем же предельным распределением, что и при проверке гипотезы независимости признаков.

Естественно, построенный критерий можно использовать для проверки гипотезы однородности распределений, из которых извлекаются $s \geq 2$ выборок. Выборочные данные при этом подвергаются группировке в соответствии с одинаковым для всех выборок разбиением пространства X на $r \geq 2$ областей.