1 Структура математической статистики

По ряду причин в этом курсе для математической статистики и машинного обучения будут использоваться[1] англоязычные аббревиатуры MS и ML. Начнем с обсуждения того, какие задачи относятся к математической статистике и что следует понимать под их статистическим решением.

В настоящее время очень популярны ML-рецепты работы с данными и поэтому прежде всего надо бы пояснить связь и различия между методами ML и MS.

- ML нацелено на автоматическое обнаружение закономерностей в данных при помощи вычислительных алгоритмов и их дальнейшая структуризации в новые, но схожие данные. Т.е. основная задача ML — изучение и создание автоматических систем, которые способны делать предсказания с помощью анализа части данных. Основные понятия ML возникли во второй половине XX века, долгое время использовался термин Теория Распознавания Образов, причем ML создавалось преимущественно программистами и для программистов,а методы и подходы сильно зависели от вычислительной мощности компьютеров. ML, как правило, старается уменьшить число изначальных допущений и свободно использует эвристические рассуждения, не претендуя на формальные обоснования и строгий стиль математики. Предпочитаемый метод рассуждений в ML — индуктивный, от частного к общему.

  Таким образом, методы ML никогда не приводят к окончательному логическому доказательству соответствия произвольных данных установленным свойствам, вместо этого целеполаганием является возможность практического использования и потому запас методов анализа данных быстро растет. Хотя использование ML ориентировано на возможность математического описания строения исследуемых данных, но есть и понимание, что шансов на построение окончательной математической модели нет.

- MS возникла именно как часть математики гораздо раньше ML и в определенном смысле является разделом Теории Вероятностей. В своем самом общем виде задача Математической Статистики как науки обратна задаче Теории Вероятностей, а именно: в Теории Вероятностей при *заданной* случайной величине ξ изучались свойства ее значений $\xi(\omega)$, а в Математической Статистике исходя из свойств нескольких значений $x_k = \xi(\omega_k)$ (в совокупности обычно называемых *данными* или data) делают выводы о свойствах случайной величины. На первый взгляд, сделать это совершенно невозможно: действительно, значения случайных величин вообще говоря могут быть любыми и, вдобавок, в случае непрерывных распределений вероятность конкретного конечного множества значений должна быть нулевой! Математическая статистика выходит из этого логического затруднения, используя следующий прием: в выводах участвуют семейства (приближенных) ответов вместе с указанием вероятностной меры на этом семействе. Методы математической статистики в докомпьютерную эпоху создавались с возможностью их применить не прибегая к трудоемким вычислительным процедурам, поэтому во многих случаях предпочтение было отдано аналитическим формулам, которые заведомо лишь аппроксимируют точный ответ.

  Данные, которые изучаются в статистике ни в коем случае не произвольны, а возникают из (вполне идеалистической) математической модели независимых повторений фиксированного опыта, то есть значения – это независимые и одинаково распределенные компоненты случайного вектора.

  Стоит помнить следующее:

  1. математическая статистика ни в коем случае не сборник рецептов «как в произвольных наборах чисел обнаружить структуру и дать ее описание», но собрание методов *как в наборах чисел, заведомо отвечающих некоторой вероятностной модели (обычно независимых) повторений измерения сделать уточняющие выводы о вероятностных свойствах одного измерения*:
  2. статистические выводы *всегда имеют вероятностный характер*, иными словами, статистическое решение отличается от категоричных утверждений «да» и «нет» тем, что обязательно указывает на вероятность «да/нет».

  Последнее свойство чрезвычайно важно для понимания смысла бытовых рассуждений в духе «этот факт доказан статистикой», ведь на самом деле, доказательный смысл статистические решения приобретают лишь по принятию веры в следующий постулат: **маловероятные события на практике не наблюдаемы**.

  Таким образом, имеется предположение о законе распределения с.в. ξ, одновременное с предположением, что данные $x_k = \xi(\omega_k)$ отвечают повторным независимым испытаниям. На основании наблюдения ее

[1] в англоязычной литературе вместо «решение средствами математической статистики» говорят о *statistical inference*, или в случае компьютерных алгоритмов искуственного интеллекта употребляют термин *machine learning*

значений $x_1 = \xi(\omega_1), x_2 = \xi(\omega_2), \dots x_n = \xi(\omega_n)$ при предположительно независимом n- кратном повторении статистические методы никогда не приведут к окончательному логическому доказательству соответствия неизвестного распределения случайной величины ξ какой-то явной формуле. Вместо этого появляется мера истинности на возможных ответах, которые звучат следующим образом: «с вероятностью такой-то можно утверждать, что ... ». Принимая вдобавок эвристическое соображение, что *малая вероятность явления должна соответствовать его ненаблюдаемости*, можно сказать, что методы статистики направлены *на отбрасывание* указанных предположений для случайной величины ξ.

1.1 Соглашения о структуре данных

Параметрическая статистика (абстрактная формулировка)

Имеется наблюдение X – набор данных, их природа, вообще говоря, не важна: это может быть набор чисел; числовая последовательность; запись, сделанная самописцем, и т.п.), К имеющемуся наблюдению X мы примысливаем выборочное пространство - совокупность $\mathcal{X}$ таких исходов, которые могли бы появиться в нашем опыте вместо X. Мы предполагаем, что элемент X был выбран из $\mathcal{X}$ случайно (случайный выбор), согласно некоторому распределению вероятностей на $\mathcal{X}$. Это вероятностное распределение P, на множестве $\mathcal{X}$ нам, как правило, не известно. Исходя из условий опыта, мы можем указать лишь некоторые свойства P. Иначе говоря, мы можем указать совокупность вероятностных мер на $\mathcal{X}$, которой принадлежит распределение P. В этой схеме задачей математической статистики являются выводы о распределении P, которые можно получить на основании наблюдения X. Во многих практически важных случаях (не всех!) семейство вероятностных мер имеет естественную параметризацию такую, что множество параметров конечномерно. Параметрическая статистика интересуется выводами о значении параметра, которые можно получить на основании наблюдения X.

Статистическая модель данных

В статистике в основном используется модель бесконечного числа независимых повторений опыта (характеризующегося случайной величиной ξ), то есть все мыслимые последовательности значений $x_1, x_2, \dots$ (возникающие как $x_1 = \xi(\omega_1), x_2 = \xi(\omega_2), \dots$) лежат в $\mathbb{R}^\infty$, где возникает вероятностная мера (поскольку речь идет о случайном векторе). Соответствующее вероятностное пространство называется в русскоязычной литературе *генеральной совокупностью*, а конечномерные проекции $x_{i_1}, x_{i_2}, \dots x_{i_n}$ называются *выборками* размера n. Разумеется, соответствующая такой проекции вероятностная мера на $\mathbb{R}^n$ — это мера связанная с n-кратным независимым повторением ξ.

Модель независимых повторений удобно понимать и как последовательность независимых, одинаково распределенных случайных величин. Такие последовательности изучались в Теории Вероятностей, но здесь полезно напомнить явный пример (принадлежащий Радемахеру) вероятностного пространства и соответствующей последовательности. Вероятностным пространством служит $\{[0,1], \mathcal{B}([0,1]), P\}$ –единичный отрезок с обычной вероятностной мерой на нем.

Любое число t на отрезке $[0,1]$ (оно же случайный исход $\omega \in \Omega = [0,1]$) можно представить в виде

$$t = \frac{\epsilon_1}{2} + \frac{\epsilon_2}{2^2} + \frac{\epsilon_2}{2^2} + \dots \qquad \epsilon_k = 0, 1$$

Если дополнительно потребовать запрета на использование бесконечных хвостов с числителями из одних только единиц, то такое представление будет даже однозначным. Тем самым всякое число t окажется закодированным последовательностью $\epsilon_1(t), \epsilon_2(t), \epsilon_3(t), \dots$, где функции $\epsilon_k(t)$ однозначно определены. Тем самым, функции $\xi_k(\omega) = \epsilon_k(t)$ следует называть случайными величинами и возникает последовательность (ступенчатых) случайных величин $\{\xi_k\}$, у которых (в силу специфики выбора вероятностного пространства) можно даже нарисовать их графики!

- Построить функцию распределения $F_{\xi_k}(x)$.
- Проверить что случайные величины $\{\xi_k\}$ попарно независимы. На самом деле они даже независимы в совокупности, и это рассуждение стоит также воспроизвести самостоятельно.

Имеется некая связь этого примера с методом компьютерной генерации *псевдослучайных* чисел на отрезке $\Omega = [0,1]$: действительно, последовательность чисел $\{x_k\}$ из единичного отрезка можно соотнести с мерой, которая возникает для любого интервала $(a,b) \subset [0,1]$ в терминах предельной частоты попадания членов нашей последовательности в этот интервал:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{1}_{(a,b)}(x_k)$$

Если значение этой меры, совпадает с $b-a$, то последовательность псевдослучайных чисел называется **равномерной**. Однако будет ли последовательность случайной в том смысле, что значения x_k отвечают статистической модели повторных независимых повторений равномерно распределенной с.в. ξ? Достаточно ясно, как придумать «очевидно неслучайную последовательность», но вопрос не сводится только к детерминизму функциональной зависимости.

Достаточно простой способ генерации равномерных последовательностей получается из поворота окружности единичной длины на иррациональный (в градусах) угол, что соответствует сдвигу $R : \mathbb{T} \to \mathbb{T}$ $x \mapsto x + r$ в группе окружности $\mathbb{T}$ на иррациональный элемент. Очевидным образом сдвиг сохраняет метрику, также более-менее понятно, что траектория $R(x) = x, R^1(x) = x + r, R^2(x) = x + 2r, \ldots$ всюду плотна в $\mathbb{T}$ и можно доказать, что возникающая последовательность $R^k(x)$ действительно равномерна. Однако насколько хорош этот способ в смысле моделирования случайности? Для этого надо бы исследовать соответствующую меру на выборочных пространствах $[0,1]^n$: будет ли она произведением соответствующих одномерных мер?

Возвращаясь к общей структуре статистических методов заметим, что для разных показателей n придется рассматривать разные выборочные пространства X^n. Для изучения свойств выборок последовательности нам понадобятся отображения $T_n : X^n \to \mathbb{R}$, которые задают последовательнось случайных величин, называющихся **статистиками** (или статистическими функционалами).

Важным для выводов о свойствах с.в. ξ является поведение статистик при $n \to \infty$. Во многих случаях выборочную статистику удается связать с подходящим выражением вида $W(F_\xi)$, например, одним из вариантов выборочных статистик $T_n : X^n \to \mathbb{R}$ для исследования моментов случайной величины с.в. ξ является следующая конструкция[2]: в интегральной формуле, выражающей k-й момент, заменить интеграл на сумму:

$$E(\xi^k) = \int_{\mathbb{R}} x^k dF_\xi(x) \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

На самом деле здесь использована подстановка в формулу $W(F_\xi)$ вместо истинной функции распределения F_ξ, так называемой *эмпирической функции распределения* — функции распределения дискретной случайной величины, которая с равными вероятностями принимает значения из списка $x_1, x_2, \ldots x_n$, мотивировка этой подстановки составляет теорему Гливенко, она изложена далее в разделе 2.

В каждом конкретном контексте удобные формулы для статистик в значительной степени определяли успех Математической Статистики как прикладной науки: действительно, применение статистических методов рассчитано на людей, которые не обязаны быть специалистами в математике и потому ручное вычисление необходимых величин не должно было быть трудозатратным. В настоящее время эта проблема не так актуальна и за численными свойствами статистик подчас скрыты достаточно сложные алгоритмы.

Пока же необходимо кратко напомнить основные факты из Теории Вероятностей о сходимости последовательностей случайных величин.

Вспомогательные сведения из Теории Вероятностей (напоминание)

В теории вероятностей употребляются следующие основные виды сходимости последовательности случайных величин $\{\xi_n\}$ к случайной величине ξ: *сходимость по вероятности, сходимость почти всюду (иначе называется «сходимость с вероятностью единица»), сходимость в среднем порядка m, сходимость по распределению.*

Определение Последовательность называется сходящейся по вероятности (используется обозначение $\xi_n \xrightarrow{P} \xi$), если для любого $\varepsilon > 0$ $P\{|\xi_n - \xi| > \varepsilon\} \to 0$ при $n \to \infty$

Этот вид сходимости появлялся в связи с законом больших чисел в схеме Бернулли, а именно

$$P\left(\left|\frac{S_n}{n} - p\right| > \varepsilon\right) \to 0, \quad n \to \infty$$

Определение Последовательность называется сходящейся почти всюду (обозначение $\xi_n \xrightarrow{\text{п.в}} \xi$), если множество $W \subset \Omega$ тех исходов ω, в которых последовательность значений $\xi_n(\omega)$ стремится к $\xi(\omega)$, во-первых, измеримо, а во-вторых, имеет вероятность 1. Иначе говоря, множество исходов $\omega \in \Omega$, для которых последовательность значений $\xi_n(\omega)$ не сходится к пределу, равному $\xi(\omega)$, имеет нулевую вероятностную меру.

[2] так называемые plug-in формулы

Этот вид сходимости ослабляет обычное в математическом анализе понятие поточечной сходимости последовательности значений функций тем, что разрешает значениям на «маленьком» множестве аргументов вести себя как попало. В каждом из остальных же аргументов $\hat{\omega}$ числовая последовательность $\xi_n(\hat{\omega})$ должна сходиться к $\xi(\hat{\omega})$. Объединяя условие существования предела последовательности с условием малости множества, на котором такой предел не существует, получаем критерий сходимости почти всюду

Предложение 1.1 *Для выполнения свойства $\xi_n \overset{n.e}{\to} \xi$ необходимо и достаточно, чтобы* $\forall \varepsilon > 0$

$$P\left(\sup_{k \geqslant n} |\xi_k - \xi| \geqslant \varepsilon\right) \to 0, \quad n \to \infty$$

Отсюда, между прочим, получается, что из сходимости почти всюду $\xi_n \overset{\text{п.в}}{\to} \xi$ вытекает сходимость по вероятности $\xi_n \overset{P}{\to} \xi$ (но не наоборот!).

Определение Последовательность называется сходящейся в среднем порядка m (обозначение $\xi_n \overset{L^m}{\to} \xi$), если $E|\xi_n - \xi|^m \to 0$ при $n \to \infty$

Если выбрать $m = 2$ и вспомнить интегральную формулу для вычисления второго момента, то легко видеть, что определение сходимости в среднем в точности воспроизводит определения сходимости в известном из функционального анализа гильбертовом пространстве L^2. Действительно, «расстояние» между функциями $f, g \in L^2$ в гильбертовом пространстве вводилось через скалярное произведение функций, то есть через $\int fg$ — нетрудно видеть аналогию с рассмотрениеи случайных величин с конечным вторым моментом ξ, η как функций на Ω и скалярного произведения для них вида $E(\xi\eta)$. На самом деле несложно проверить, что из сходимости в среднем порядка $m > 0$ вытекает сходимость пометодов вероятности $\xi_n \overset{P}{\to} \xi$, в виду сказанного выше, эта проверка по крайней мере для случая $m = 2$— упражнение по функциональному анализу, поэтому мы ограничимся простой ссылкой на этот факт.

Таким образом, среди первых трех определений самым слабым требованием оказалось условие сходимости по вероятности. Тем не менее,

Определение Последовательность называется слабо сходящейся или сходящейся по распределению (обозначение $\xi_n \Rightarrow \xi$), если при $n \to \infty$ имеет место сходимость функций распределения $F_{\xi_n}(x)$ к $F_\xi(x)$ для всех тех аргументов x, где $F_\xi(x)$ непрерывна.

Заметьте, что последнее определение использует только закон распределения и даже не предполагает, что случайные величины ξ_n определены на одном и том же вероятностном пространстве Ω.

В подробных учебниках по функциональному анализу тщательно разобраны примеры и контрпримеры специальных видов последовательностей функций, которые бы удовлетворяли только одному типу сходимости. Это достаточно увлекательная тема, но она требует изрядного времени.

- Важное упражнение на понимание: о каком типе сходимости случайных величин идет речь в Законе Больших Чисел?.

Для практических приложений статистики, имеющей дело с конечными выборками, сходимость почти всюду $\xi_n \overset{\text{п.в}}{\to} \xi$ дает не больше, чем сходимость по вероятности: действительно, для данного фиксированного n и сходимость $\xi_n \overset{\text{п.в}}{\to} \xi$ и сходимость $\xi_n \overset{P}{\to} \xi$ означают, что с.в. ξ_n в подходящем смысле приближенно равна с.в. ξ , если n «достаточно велико». Поэтому обычно в статистике используют теоремы про сходимость по вероятности, даже если можно доказать более сильные теоремы.

2 Пример (непараметрического) статистического исследования. Теоремы Гливенко и Колмогорова

Здесь будет рассматриваться числовая с.в. ξ с функцией распределения F_ξ и выборки $x_1, x_2, \dots x_n$. Для действительного $y \in \mathbb{R}$ положим $F_n^*(y)$ равным $1/n$ умноженному на числу тех x_i в наборе $x_1, x_2, \dots x_n$, которые меньше либо равны y.

$F_n^*(y)$ – это так называемая *эмпирическая или выборочная функция распределения.* Заметим, что график $F_n^*(y)$ при $y \in \mathbb{R}$ совпадает с графиком функции распределения дискретной случайной величины, которая с равными вероятностями принимает значения из списка $x_1, x_2, \dots x_n$. Объектом какой природы является эмпирическая функция распределения? Проще всего выразить ее через случайные величины – индикаторы:

$$F_n^*(y) = \frac{1}{n} \sum_i^n I[y, i]$$

здесь для действительного $y \in \mathbb{R}$ и i-того повторения эксперимента ξ_i положим случайную величину $I[y, i] = \mathbf{1}_B$, где событие B это $\xi_i <= y$ и, как обычно, $\mathbf{1}_B$ обозначает индикатор события B,

$$\mathbf{1}_B(\omega) = \begin{cases} 1, & \omega \in B \\ 0, & \omega \notin B \end{cases}$$

Вот список простых задач (подробнее обсудим их на упражнениях), проясняющих суть дела:

- Объяснить, почему $F_n^*(y)$ является случайной величиной.
- Как связаны функции распределения F_ξ и $F_{I[y,i]}$?
- Объяснить, почему при $i \neq j$ случайные величины $I[y, i]$ $I[y, j]$ независимы.
- Найти математическое ожидание случайной величины $I[y, i]$. Доказать, что дисперсия случайной величины $I[y, i]$ существует и не превосходит единицы.

Теперь применим к последовательности случайных величин $I[x, i]$ Закон Больших Чисел, и получим

Теорема 2.1 *При $n \to \infty$ последовательность случайных величин $F_n^*(y)$ любом $y \in \mathbb{R}$ сходится к $F_\xi(y)$ по вероятности.*

Следующая теорема (Гливенко) утверждает, что при достаточно больших значениях n эмпирическая функция распределения с большой вероятностью *равномерно* похожа на истинную функцию распределения:

Теорема 2.2 *При $n \to \infty$ последовательность случайных величин $\sup_x |F_n^*(x) - F_\xi(x)|$ сходится по вероятности к 0.*

Более сильная формулировка этой теоремы также принадлежит Гливенко (он доказал утверждение для непрерывной случайной величины, а Кантелли — для общей) утверждает сходимость с вероятностью единица. (Теорема Гливенко-Кантелли и есть закон больших чисел в функциональном пространстве) .

Полное доказательство теоремы Гливенко достаточно громоздко, тем более, что пока вообще неясно как можно оценивать максимум расстояния между эмпирической функцией распределения и истинной $F_\xi(x)$ для совсем произвольной с.в. ξ. Однако идею доказательства можно коротко сформулировать, что и будет сейчас сделано.

Лемма 2.3 *Пусть $g : \mathbb{R} \to \mathbb{R}$ непрерывная и строго монотонная функция, а α, β случайные величины с функциями распределения соответственно F_α, F_β. Тогда*

$$\sup_x |F_\alpha(x) - F_\beta(x)| = \sup_x |F_{g(\alpha)}(x) - F_{g(\beta)}(x)|$$

Доказательство Там, где определена обратная функция g^{-1}

$$F_{g(\alpha)}(y) - F_{g(\beta)}(y) = P(g(\alpha) \leqslant y) - P(g(\beta) \leqslant y) = P(\alpha \leqslant g^{-1}(y)) - P(\beta \leqslant g^{-1}(y)) = F_\alpha(g^{-1}(y)) - F_\beta(g^{-1}(y))$$

Следовательно, если $\sup_y |F_\alpha(y) - F_\beta(y)|$ достигается на последовательности аргументов $\{y_k\}$, то такое же значение $|F_{g(\alpha)}(y) - F_{g(\beta)}(y)|$ достигается и на последовательности точек $\{g(y_k)\}$ ∎

Если функция распределения $F_\xi(x)$ строго монотонна вне тех аргументов, где она равна нулю и единице, то, имея в виду известную формулу $\eta = F_\xi(\xi)$ для одномерной копулы, после строго монотонной замены переменной можно считать, что с.в. ξ уже преобразована к равномерному на $[0,1]$ распределению. Чтобы применить лемму 2.3 осталось рассматривать эмпирическую функцию распределения как функцию распределения дискретной случайной величины, принимающей значения $\{x_1, x_2, \dots x_n\}$ с равными вероятностями. Но это и означает, что случай произвольной с.в. сведен к частному утверждению о поведении эмпирической функции распределения для равномерной на $[0,1]$ случайной величины, то есть закон распределения с.в. $\sup_x|F_n^*(x)-F_\xi(x)|$ зависит лишь от n, но не от вида $F_\xi(x)$. Осталось для равномерной случайной величины ξ, используя тот факт, что $F_\xi(x) = x$ $x \in [0,1]$ оценить вероятность значений $\sup_x|F_n^*(x) - x| > \text{const} > 0$ и убедиться в том, что эта вероятность с ростом n убывает; на качественном (не количественном) уровне это более-менее техническое вычисление и мы его опустим.

На количественном уровне вычисление этой вероятности далеко не тривиально, укажем здесь аналитическое DKW-неравенство (авторы Dvoretzky–Kiefer–Wolfowitz, идею доказательства смотрите в неравенстве Хефдинга 3.1):

$$P\left(\sup_x|F_n^*(x) - F_\xi(x)| > \varepsilon\right) \leqslant 2e^{-2n\varepsilon^2}$$

и доступную в Сети компьютерную программу на языке C, вычисляющую с разумной точностью вероятности $K(n,t) = P\left(\sup_x|F_n^*(x) - F_\xi(x)| \leqslant t\right)$ (George Marsaglia, Wai Wan Tsang, Jingbo Wang "Evaluating Kolmogorov's Distribution"). 3.1

А.Н. Колмогоров в 1933 году использовал идею ремасштабирования величин $\sup_x|F_n^*(x) - x|$ при разных n с использованием независимых приращений $\varepsilon_j = \pm\frac{1}{n}$ в точках $x_1, x_2, \dots x_n \in [0,1]$: действительно, приращения равновероятны и можно использовать свойства распределения Бернулли, чтобы оценить среднеквадратичное отклонение при данном n — это будет величина порядка $\frac{1}{\sqrt{n}}$. Отсюда понятно, что «физически правильно» для $t > 0$ рассматривать величины не $P\left(\sup_x|F_n^*(x) - x| \leqslant t\right)$, а $P\left(\sqrt{n}\sup_x|F_n^*(x) - x| \leqslant t\right)$ при больших n. Колмогоров показал, что соответствующие вероятности с ростом n сходятся к пределу и явно (при $t > 0$) вычислил[3] предельную функцию распределения $K(t)$ (при $t \leqslant 0$ следует положить $K(t) = 0$)

$$K(t) = \lim_{n\to\infty} P\left(\sqrt{n}\sup_x|F_n^*(x) - x| \leqslant t\right) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2t^2}$$

Имееся также иное выражение для $K(t)$

$$K(t) = \frac{\sqrt{2\pi}}{t}\sum_{k=1}^{\infty}\exp\left[-\frac{(2k-1)^2\pi^2}{8t^2}\right]$$

Поскольку $K(t)$ является функцией распределения некоторой неотрицательной с.в. θ, то можно говорить о математическом ожидании $E(\theta) = \sqrt{\frac{\pi}{2}}\ln 2 = 0.86873110\dots$ и дисперсии $D(\theta) = \pi^2/12 - [E(\theta)]$ — величине порядка $(0.260332872\dots)^2$. График соответствующей плотности распределения $f_\theta(t) = \frac{d}{dt}K(t)$ показан на Рис.1 слева. На практике общепринято использовать функцию $K(t)$ как приближение для вероятностей $P\left(\sqrt{n}\sup_x|F_n^*(x) - x| \leqslant t\right)$ и более того, бытует уверенность в быстрой сходимости при $n \to \infty$ вероятностей $P\left(\sqrt{n}\sup_x|F_n^*(x) - x| \leqslant t\right) = P\left(\sup_x|F_n^*(x) - x| \leqslant t/\sqrt{n}\right) = K(n, t/\sqrt{n})$ к предельной функции $K(t)$, но на самом деле разность $K(n, t/\sqrt{n}) - K(t)$ весьма медленно убывает с ростом n, см. Рис.1 справа.

2.0.1 Приложение: проверка случайности равномерной последовательности

Для заданной равномерной на $[0,1]$ последовательности можно использовать приведенные выше соображения для проверки ее «на случайность»: действительно, если верна гипотеза, что члены последовательности можно

[3] Уточнение этого результата для величин $\lim_{n\to\infty} P\left(\sqrt{n}\min_x\left(F_n^*(x) - x\right) \leqslant t\right)$ и $\lim_{n\to\infty} P\left(\sqrt{n}\max_x\left(F_n^*(x) - x\right) \leqslant t\right)$ принадлежит Смирнову. Подробное доказательство связано с аккуратным рассмотрением свойств модели случайного блуждания, которую традиционно рассматривают в курсе Теории Вероятностей (можно найти соответствующее рассуждение в книге В.Феллера "Введение в теорию вероятностей и ее приложения", том 2, Глава X, §5, пункт «г»), а здесь мы упомянем лишь общую идею.

Для n независимых одинаково распределенных случайных приращений $\varepsilon_j = \pm\frac{1}{n}$ с нулевым средним положим $T_m = \varepsilon_1 + \dots + \varepsilon_m$, нас интересует распределение максимума абсолютных значений $\max_m(|T_1|, |T_1|, \dots |T_m|)$. При больших значениях m Центральная Предельная Теорема делает правдоподобным заключение, что асимптотическое поведение этого максимума такое же, как если бы ε_j были бы нормально распределены, но тогда для вычисления закона распределения максимума абсолютных значений можно было бы использовать формулы гауссова распределения, отсюда собственно и возникает ответ (при $t > 0$) для выражения предельной функции $K(t)$ через сумму экспонент гауссова типа.

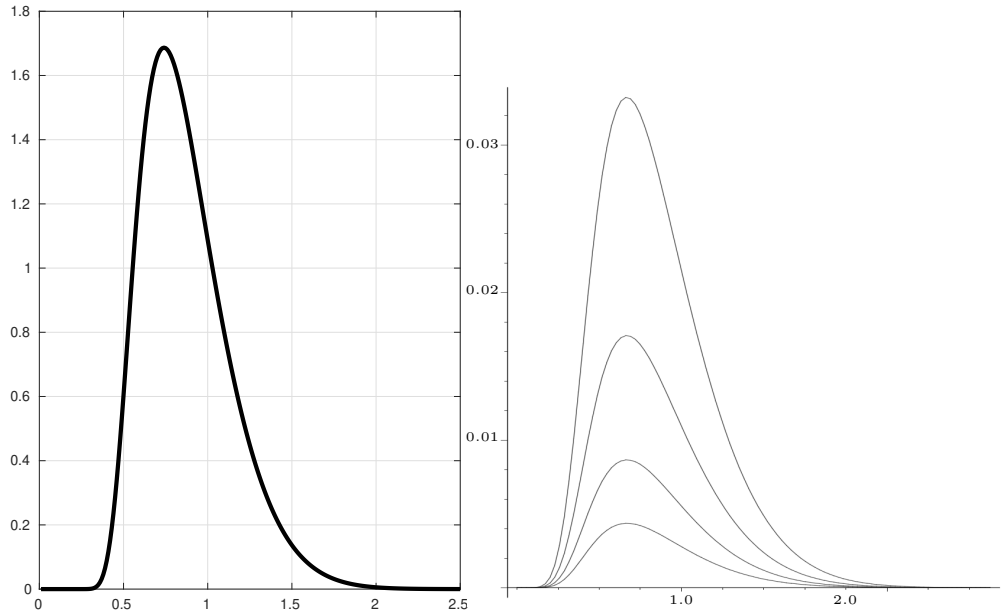

Рис. 1: Слева: график плотности $\frac{d}{dt}K(t)$. Справа: разности $K(n, t/\sqrt{n}) - K(t)$ при $n = 64, 256, 1024, 4096$.

рассматривать как результаты последовательных независимых случайных измерений, то, как и выше, рассмотрим отклонения эмпирических функций распределения от диагонали. При фиксированном n эти отклонения являются случайной величиной, для которой в принципе известен закон распределения, в свою очередь перенормированные законы распределения приближаются универсальным законом $K(t)$ для случайной величины θ. Вычисляя по конкретной равномерной последовательности величину отклонения, тем самым, вычисляем значение случайной величины θ. Пусть, например, эти значения окажутся не превосходящими какой-то величины, скажем 0.4, но какова вероятность для случайной величины θ выпадения чисел не превосходящих 0.4? Явное знание формулы для $K(t)$ показывает, что $K(0.4) < 0.003$, то есть предположив, что наша исходная равномерная последовательность может быть описана как случайные независимые повторения мы получим крайне маловероятные значения возможных отклонений. Общий принцип математической статистики гласит, что если мы обнаруживаем крайне маловероятное значение, то это повод отбросить исходное предположение, а оно в данном случае состояло в справедливости для равномерной последовательности модели независимых случайных повторений одинакового эксперимента.

Более общим образом, если мы будем получать значения θ сильно отличающиеся от $E(\theta) = 0.86873110\ldots$ то, например, по неравенству Чебышева мы опять увидим маловероятные события, указывающие на то, что либо последовательность не случайна, либо не является равномерной. Напротив, значения вне маргинальных областей (ниже мы сформулируем точнее, как именно принято в математической статистике выбирать маргинальные области) указывают на то, что нет оснований не доверять для равномерной последовательности модели независимых случайных повторений одинакового эксперимента.

- Задача. Сравнить куски $k = 1$ до $k = 100$ двух последовательностей дробных частей $a_k = \{a_{k-1} + \pi\}$ и $b_k = \{b_{k-1} \cdot \pi\}$, где $a_0 = b_0 = 1$ — какой из них лучше соответствует модели независимых случайных повторений равномерной на $0, 1$ случайной величины?

2.1 Общая статистическая терминология

Как уже было сказано, исследуют предположения о законе распределения с.в. ξ, очень часто это предположение состоит в принадлежности функции распределения $F_\xi(x)$ к некоторому классу функций, который будет обозначаться далее готической буквой, например $\mathfrak{F}$. Если класс распределений характеризуется одним или несколькими числовыми параметрами (например, гауссовский закон характеризуется двумя параметрами a и σ) то говорят о *пространстве параметров*. В большинстве ситуаций выясняют не информацию о функции распределения F_ξ, а значение какого-то параметра[4] (чаще всего момента), связанного с этим распределением. В этом смысле методы математической статистики очень похожи на процедуру измерения и во многих случаях говорят о *методах обработки данных* , *анализе сигналов* и т д.

[4] В этой версии MS называется параметрической статистикой

Много классических прикладных учебников по математической статистике с самого начала используют предположения о гауссовом распределении данных (или хотя бы об обязательном наличии конечномерного пространства параметров в статистической задаче), в этом курсе мы постараемся придерживаться другой линии и начнем с примеров из*непараметрической статистики*.

Утверждение относительно свойств случайной величины ξ называется **стастической гипотезой**. Следующая терминология является общепринятой:

- Статистическая гипотеза, однозначно определяющая функцию распределения F_ξ, называется *простой*. При простой гипотезе проверяют согласие наблюдаемой выборки с законом F_ξ.
- Статистическая гипотеза, утверждающая принадлежность функции распределения F_ξ к некоторому семейству распределений, называется *сложной*. Сложная проверяемая гипотеза имеет вид $F_\xi(x) \in \mathfrak{F} = \{F(x,\theta)\}$, где θ взято из множества обычно скалярных или векторных параметров. Если конкретное значение $\hat{\theta}$ сначала вычисляют по той же самой выборке и потом по ней же проверяют согласие с $F(x,\hat{\theta})$, то это сложная гипотеза. Если оценку параметра $\hat{\theta}$ вычисляют по другой выборке, то это ситуация простой гипотезы.

Если проверяют одну какую-то конкретную гипотезу, то ее принято называть *нулевой*, но чаще всего рассматривают пару или даже набор конкурирующих гипотез-альтернатив.

Например, в эксперименте с подбрасыванием монеты изначально нуль-гипотеза и альтернативная ей могут быть сформулированы в терминах симметричности герба и решки:

H_0 Монета симметрична, герб и решка равновероятны.

H_1 Монета несимметрична, вероятность герба больше.

H_2 Монета несимметрична, вероятность герба меньше.

.

Выше была указана разница между «принять» и «доказать», так что нужно говорить о правиле (включающем соглашение о вероятностном смысле утверждений), когда от гипотезы следует отказаться и когда для отказа не хватает аргументов. Правило, по которому принимается решение опровергнуть или сохранить гипотезу называется *критерием*.

Перед анализом выборки фиксируют некоторое малое значение $\epsilon > 0$, которое называется *уровнем значимости* и далее сравнивают значения квантиля q_ϵ (здесь в зависимости от формулировок альтернативных гипотез H_k возможны варианты: $q_{1-\epsilon}$, $q_{1-\epsilon/2}$, $q_{\epsilon/2}$) с наблюдаемым по выборке значением $T_n(x_1, x_2, \dots x_n)$.

2.1.1 Пример

Пусть у нас есть простая гипотеза H_0, что выборки получены из случайной величины ξ с непрерывным законом распределения $F_\xi(x)$. В процессе проверки согласия по выборке вычисляют значение $\hat{T} = T_n(x_1, x_2, \dots x_n)$ некоторой подходящей статистики T_n использующей абсолютную величину несовпадения экспериментальной функции распределения и теоретической. критерия. Затем для того, чтобы сделать вывод о принятии или отклонении гипотезы H_0 необходимо знать условное распределение $G_{T_n|H_0}$ статистики (то есть закон распределения статистики T_n при справедливости гипотезы H_0, при известном $F_\xi(x)$ определение этого распределения является более-менее стандартной задачей Теории Вероятностей) и задать *уровень значимости*, то есть некоторый параметр $0 < \alpha < 1$, который по смыслу будет вероятностью отклонить справедливую гипотезу H_0). Согласно общей для математической статистики концепции, что маловероятные эффекты не должны наблюдаться в типичном опыте, если значение по выборке попало в определяемую квантилем маловероятную (ее еще называют *критической*) область, то это следует считать опровержением нулевой гипотезы H_0. Таким образом, если вероятность

$$P(T_n > \hat{T}) = 1 - G_{T_n|H_0}(\hat{T}) > \alpha$$

то критерий утверждает, что нет оснований для отклонения гипотезы H_0.

Если статистические выводы вынуждают нас отвергнуть гипотезу на некотором уровне значимости, то в силу вероятностного характера используемого критерия может возникнуть ошибка в принятии или отбрасывании гипотезы H_0 на заданном уровне значимости. Различают *ошибку первого рода*, состоящую в том, что мы отклонили верную гипотезу в силу того, что маловероятное значение для $T_n(x_1, x_2, \dots x_n)$ все-таки возникло при рассмотрениях (такую ошибку можно сделать редкой изменяя уровень значимости в критерии) и *ошибку второго рода*, когда H_0 не отклонили, хотя в действительности имеет место альтернатива.

2.2 Непараметрические критерии согласия экспериментального и теоретического распределения

Рассмотрим два критерия, которые опираются на рассмотренные выше статистики.

2.2.1 Непараметрический критерий Колмогорова-Смирнова

Здесь в качестве статистики используют (нормализованную домножением на $\sqrt{n}$) обсуждавшуюся уже величину $\sup_x|F_n^*(x) - F_\xi(x)|$. В докомпьютерную эпоху для вычисления соответствующих вероятностей общепринятым было использование функции Колмогорова $K(t)$ для анализа выборок с $n \geqslant 20$, начиная рубежа XX-XXI веков в компьютерных реализациях использовались уже трудоемко вычислимые значения функций $K(n, t/sqrtn$, что уточняет ответ — см. правый график на Рис. 1.

2.2.2 Непараметрический критерий Андерсона-Дарлинга

Отклонение эмпирической функции распределения $F_n^*(t)$ от теоретической $F_\xi(t)$ можно попробовать измерять иначе, естественным кандидатом представляется (взвешенная при помощи неотрицательной функции $w\left(F_\xi(t)\right)$) квадратичная метрика, в которой квадрат расстояния между эмпирической и теоретической[5] функциями распределения вычисляется по формуле:

$$\int\limits_{-\infty}^{\infty} \left(F_n^*(t) - F_\xi(t)\right)^2 w(F_\xi(t)) dF_\xi(t)$$

При выборе $w\left(F_\xi(t)\right) \equiv 1$ получается так называемый «критерий Ω^2» Крамера-Мизеса, основанный на сравнении соответствующей статистики с предельным распределением, при выборе $w\left(F_\xi(t)\right) = [F_\xi(t)(1 - F_\xi(t))]^{-1}$ получается аналогичный по структуре «критерий AD» Андерсона-Дарлинга[6].

Первым этапом для критерия служит преобразование выборки к гипотетически равномерной и вычислениями расстояний в этом случае, далее, как было и при выводе предельной по n функции Колмогорова $K(t)$, при конкретном значении n квадраты расстояний в метрике Андерсона-Дарлинга нормализуют умножением на n и далее рассматривают с использованием ЦПТ предельный при $n \to \infty$ закон, который в этом случае уже не допускает такого относительно простого выражения как для $K(t)$. Выбор весовой функции Андерсона-Дарлинга позволяет более внимательно учитывать разницу функций распределения на хвостах распределения, в то время как критерий Колмогорова-Смирнова с очевидностью более чувствителен к отклонениям в центральной части. Следует также помнить, что под статистикой *muna* Андерсона-Дарлинга квадрата расстояния D^2 разные программные реализации понимают сходные по смыслу, но различающиеся выражения для самого расстояния:

$$D_1 = \sqrt{n \int\limits_{0<F_\xi(t)<1} \frac{\left(F_n^*(t) - F_\xi(t)\right)^2}{F_\xi(t)(1 - F_\xi(t))} dF_\xi(t)}$$

$$D_2 = \sqrt{n} \int\limits_{0<F_\xi(t)<1} \frac{|F_n^*(t) - F_\xi(t)|}{\sqrt{F_\xi(t)(1 - F_\xi(t))}} dF_\xi(t)$$

$$D_3 = \sqrt{n} \max_{0<F_\xi(t)<1} \frac{|F_n^*(t) - F_\xi(t)|}{\sqrt{F_\xi(t)(1 - F_\xi(t))}}$$

При конкретном значении n нахождение квантилей для распределения D^2 — трудная вычислительная задача, об аналитическом ответе здесь речь вообще не идет.

Принято считать, что пара критериев KS и AD вместе достаточно точно ловит отклонения как для сравнения с теоретическим распределением, так и в ситуации отклонений двух эмпирических функций распределения между собой. Важным обстоятельством в критериях KS и AD является учет необычно маленьких значений соответствующих статистик: как выше было показано в разделе 2.0.1 маленькие значения — это указание на проблему реальной независимости в исследуемой выборке. Как именно это обстоятельство было учтено (и учтено ли вообще!) в программных реализациях надо каждый раз разбираться отдельно.

[5] напомним, что речь идет об абсолютно непрерывной теоретической функции распределения

[6] Оба критерия реализованы в `python`-пакете SciPy, языке R и т п

Потеря критериями свойства «свободы от распределения»

При проверке сложных гипотез, когда по той же самой выборке оценивают и параметры наблюдаемого закона распределения вероятностей, непараметрические критерии согласия KS и AD теряют свойство инвариантности от распределения ξ. В этом случае предельные распределения статистик этих критериев будут явно зависеть от закона, которому подчинена наблюдаемая выборка и от используемого метода оценивания параметров и, разумеется, от объема выборки. Игнорирование того, что проверяют сложную гипотезу, неучет различия в сложных гипотезах приводят к некорректному применению непараметрических критериев согласия KS и AD и, как следствие, к неверным статистическим выводам, поскольку различия в предельных распределениях одних и тех же статистик при проверке простых и сложных гипотез могут быть весьма существенны.

2.3 Оптимальный критерий Неймана-Пирсона

При рассмотрении (сложных) гипотез встречаются версии, когда точное значение истинного параметра не является основным объектом исследования, а больший интерес связан с проверкой того, что истинное значение параметра принадлежит некоторому определенному возможных подмножеству значений параметра.

Для конкретизации: по выборке $\mathbf{x} = (x_1, x_2, \dots x_n)$ и надо решить лежит ли параметр θ распределения $F_\xi(x, \theta)$ в множестве Θ_0 (нулевая гипотеза H_0) или же (альтернативная гипотеза H_1) $\theta \in \Theta_1$, где $\Theta_0 \cap \Theta_0 = \emptyset$ и заведомо $\theta \in \Theta_0 \cup \Theta_0$.

2.3.1 Свойства

Начнем с примера. Известнейший вопрос в физике элементарных частиц за последнюю четверть века заключался в том, действительно ли известный бозон Хиггса существует или нет. Один из способов определить, действительно ли эта элементарная частица существует через его распад на два фотона. Используя Стандартную Модель физики элементарных частиц, мы можем вычислить, сколько таких двухфотонных событий было бы произведено *в среднем*, если бы не было бозона Хиггса, обозначим это число через b. Точно так же можно вычислить среднее число s дополнительных двухфотонных событий, если частица Хиггса действительно существовует. Cобытия возникновения хорошо задокументированы и наблюдаемым в эксперименте числам отвечает распределению Пуассона, скажем, с некоторым средним значением μ. Следовательно, $H_0 : \mu = b$ и $H_1 : \mu = b + s$.

Анализ ситуации определяет тестовая функция $\delta : \mathbf{x} \mapsto \{0, 1\}$, конструируемая обычно на основе подходящей статистики T (*тестовая статистика*) и заданной для нее критической области C

$$\delta(\mathbf{x}) = \mathbf{1}_{\{T(\mathbf{x}) \in C\}} = \begin{cases} 1, & T(\mathbf{x}) \in C \\ 0, & T(\mathbf{x}) \notin C \end{cases}$$

Таким образом, δ – с.в. Бернулли, ее значения указывают какую гипотезу надо принять: H_0 или H_1

При статистических исследованиях указанных гипотез возможные соотношения вывода и реальной принадлежности θ описываются следующей таблицей: Когда реальность отвечает нулевой гипотезе мы надеемся, что распределение тестовой функции $\delta()$ будет концентрироваться вокруг значения 0, наоборот, когда истинно H_1 надеемся, что распределение сосредоточится вокруг значения 1? то есть хорошее правило принятия решений должно концентрироваться вокруг значения i, когда H_i в реальности верна. Мы можем сравнивать решающие правила, глядя на что-то вроде их «среднеквадратичной ошибки» $\operatorname{mse}(\delta, H_i) = E\left[(\delta - i)^2\right]$, $i \in \{0, 1\}$

Определение Пусть $H_0 : \theta \in \Theta_0$ и $H_1 : \theta \in \Theta_1$ — пара конкурирующих гипотез. Два отображения

$$\begin{aligned} h : \Theta_0 \to \{0, 1\} \quad & h(\theta) = P_\theta(\delta = 1) \qquad \theta \in \Theta_0 \\ g : \Theta_1 \to \{0, 1\} \quad & g(\theta) = P_\theta(\delta = 0) \qquad \theta \in \Theta_1 \end{aligned}$$

называют соответственно вероятностями ошибок первого и второго рода[7].

[7] Функция на $\Theta_0 \bigcup \Theta_1$, значения которой равны $P_\theta(\delta = 1)$, называется функцией мощности

	$\theta \in \Theta_0$	$\theta \in \Theta_1$
$\delta = 0$	Нет ошибки	Ошибка II рода
$\delta = 1$	Ошибка I рода	Нет ошибки

Таблица 1:

Заметим, что $h(\theta) \neq 1 - g(\theta)$ хотя бы потому, что их области определения не совпадают.

Чтобы иметь хорошую тестовую функцию, мы должны попытаться выбрать тестовую статистику T и критическую область такими , чтобы вероятность ошибки типа I была мала для всех $\theta \in \Theta_0$ и в то же время вероятность ошибки типа II была мала для всех значений $\theta \in \Theta_1$. Выбор названий «тип I или II» связан с тем, что на практике гипотезы обычно не симметричны в смысле их важности: ошибка первого рода представляется более драматичной по последствиям, нулевой гипотезой считается та, ошибочное отрицание которой наносит теории больший вред.

Действительно, ошибка I рода опаснее, т.к. она заставляет нас отказаться от правильного предположения. В то же время ошибка II рода (не отвергнуть гипотезу, когда она не верна) не закрывает возможности все же отвергнуть ложную гипотезу в результате дальнейших ее проверок. Поэтому при проверке статистических гипотез возможность ошибки первого рода стараются уменьшить. Желательно, конечно, иметь такие статистические критерии, для которых малы (близки к нулю) вероятности обеих ошибок. Но это обычно невозможно и вот почему:

Пусть $\delta(\mathbf{x}) = \mathbf{1}_{\{T(\mathbf{x}) \in C\}}$ и мы желаем уменьшать для нее вероятность ошибки первого рода $h(\theta) = P_\theta(\delta = 1)$ $\forall \theta \in \Theta_0$, причем уменьшать для всех $\theta \in \Theta_0$. То есть критическую область C надо уменьшить до $C_* \subset C$ и перейти, тем самым, к другой тестовой функции δ_*. Однако при уменьшении критической области C дополнение к ней, очевидно увеличивается и потому $\forall \theta \in \Theta_1$ будет выполняться

$$P_\theta(\delta_* = 0) = P_\theta\left(T(\mathbf{x}) \notin C_*\right) \geqslant P_\theta\left(T(\mathbf{x}) \notin C\right) = P_\theta(\delta = 0)$$

Тем самым вероятность ошибки второго рода для этого критерия вообще говоря растет!

В рамках подхода Неймана-Пирсона, учитывающего неравноправность ошибок первого и второго рода, в указанной ситуации двух гипотез следует поступать иначе:

1. Зафиксировать уровень значимости α критерия (то есть вероятность ошибки первого рода)
2. Рассматривать лишь те тестовые функции $\delta : \mathbf{x} \to \{0, 1\}$ с данным или меньшим уровнем значимости, то есть с
$$\sup_{\theta \in \Theta_0} P_\theta(\delta = 1) \leqslant \alpha$$
Мы обозначим класс таких тестовых функций $\mathfrak{D}\left(\Theta_0, \alpha\right)$
3. Внутри класса $\mathfrak{D}\left(\Theta_0, \alpha\right)$ выбирать те тестовые функции, которые минимизируют вероятность ошибки второго рода, то есть $g(\theta) = P_\theta(\delta = 0)$ при $\theta \in \Theta_1$

Определение Тестовая функция δ дя проверки гипотезы H_0 относительно H_1 называется оптимальной для уровня значимости α при выполнении двух свойств: $\delta \in \mathfrak{D}\left(\Theta_0, \alpha\right)$ и для всех $\theta_1 \in \Theta_1$ при любой тестовой функции $\psi \in \mathfrak{D}\left(\Theta_0, \alpha\right)$ имеет место $P_{\theta_1}(\psi = 1) \leqslant P_{\theta_1}(\delta = 1)$

Оптимальный выбор критерия возможен лишь в немногих случаях. (Впрочем некоторые из них важны для статистической практики.) И там где он удается, всё основано на так называемой лемме Неймана-Пирсона. Она относится[8] к простейшей ситуации: и гипотеза H_0 и альтернатива H_1 — обе простые, то есть оба множества Θ_0, Θ_1- одноточечные; каждое из них задает единственное распределение вероятностей.

Оптимальный критерий для проверки простой гипотезы $H_0 : \theta = \theta_0$ против простой альтернативы гипотезы $H_1 : \theta = \theta_1$ мы построим в элементарной ситуации, когда зависящие от $\theta_0 \neq \theta_1$ два распределения либо оба дискретны, либо оба имеют плотности $f(\mathbf{x}, \theta_0)$, $f(\mathbf{x}, \theta_1)$. Доказательства для распределений, имеющих плотности и для дискретных распределений происходят одинаково - с той разницей, что интегралы заменяются суммами. Поэтому достаточно рассмотреть что-либо одно; для определённости - плотности.

Мера на выборках влечет, что функция плотности (вещественнозначная функция случайного векторного аргумента $\mathbf{x}$) может рассматриваться как случайная величина (называемая правдоподобием, ее распределение зависит от параметра), в частности отношение Λ плотностей —

$$\Lambda = \frac{f(\mathbf{x}, \theta_1)}{f(\mathbf{x}, \theta_0)}$$

как функция того же аргумента — также есть с.в. В частности можно вычислять вероятности $P_{\theta_0}\left(\Lambda > \text{const}\right)$. (Мы здесь пренебрегаем различиями в обозначениях выборки как случайного вектора и как векторного аргумента.)

[8] Для односторонних критических областей вида $C = \{T \geqslant c_0\}$ или вида $C = \{T \leqslant c_1\}$ и некоторых очень специальных распределений оптимальные критерии также возможны

Лемма 2.4 (Нейман-Пирсон) *Пусть мы находимся в указанной выше ситуации проверки гипотезы* $H_0 : \theta = \theta_0$ *против* $H_1 : \theta = \theta_1$ *на уровне* α $\theta_0 \neq \theta_1$, *распределение с.в.* Λ *непрерывно при справедливости нулевой гипотезы и положительная константа* q *такова, что* $P_{\theta_0}(\Lambda > q) = \alpha$. *Тогда тестовая функция*

$$\delta(\mathbf{x}) = \mathbf{1}_{\{\Lambda > q\}}$$

является оптимальной на уровне значимости α.

Доказательство Проверяем первое условие оптимальности:

$$P_{\theta_0}(\delta = 1) = P_{\theta_0}(\Lambda > q) = \alpha$$

то есть, действительно, $\delta \in \mathfrak{D}(\{\theta_0\}, \alpha)$.

Для проверки второго условия оптимальности заметим сначала, что

$$f(x, \theta_1) - q \cdot f(\mathbf{x}, \theta_0) > 0 \text{ при } \delta(\mathbf{x}) = 1 \text{ и } f(x, \theta_1) - q \cdot f(\mathbf{x}, \theta_0) \leqslant 0 \text{ при } \delta(\mathbf{x}) = 0$$

выбирая $\psi \in \mathfrak{D}(\{\theta_0\}, \alpha)$. имеем последовательность неравенств

$$\psi(\mathbf{x}) \cdot [f(\mathbf{x}, \theta_1) - q \cdot f(\mathbf{x}, \theta_0)] \leqslant \delta(\mathbf{x}) \cdot [f(\mathbf{x}, \theta_1) - q \cdot f(\mathbf{x}, \theta_0)]$$

$$\int\limits_{\mathbb{R}^n} \psi(\mathbf{x}) \cdot [f(\mathbf{x}, \theta_1) - q \cdot f(\mathbf{x}, \theta_0)]\, d\mathbf{x} \leqslant \int\limits_{\mathbb{R}^n} \delta(\mathbf{x}) \cdot [f(\mathbf{x}, \theta_1) - q \cdot f(\mathbf{x}, \theta_0)]\, d\mathbf{x}$$

$$\int\limits_{\mathbb{R}^n} [\psi(\mathbf{x}) - \delta(\mathbf{x})] \cdot f(\mathbf{x}, \theta_1) d\mathbf{x} \leqslant q \cdot \int\limits_{\mathbb{R}^n} [\psi(\mathbf{x}) - \delta(\mathbf{x})] \cdot f(\mathbf{x}, \theta_0) d\mathbf{x}$$

$$E_{\theta_1}(\psi - \delta) \leqslant q \cdot E_{\theta_0}(\psi - \delta)$$

$$P_{\theta_1}(\psi = 1) - P_{\theta_1}(\delta = 1) \leqslant q \cdot [P_{\theta_0}(\psi = 1) - P_{\theta_0}(\delta = 1)]$$

Принимая во внимание $\psi \in \mathfrak{D}(\{\theta_0\}, \alpha)$ получаем, что правая часть последнего неравенства меньше либо равна нулю, а значит, левая часть также не может быть строго положительной, что и обеспечивает нужное по определению оптимальности неравенство $P_{\theta_1}(\psi = 1) \leqslant P_{\theta_1}(\delta = 1)$ ∎

2.3.2 Вопросы для самоконтроля

1. Для всех ли уровней значимости α можно построить оптимальный критерий? В случае непрерывного распределения? В случае дискретного распределения?

2. Объясните подробно, как из третьей строчки неравенств возникает четвертая (это так называемое «Правило ленивого математика» для вычислений в теории вероятностей).

3 Сводка сведений: неравенства, распределения и статистики

Ранее было объяснено, что для проверки статистических гипотез используют статистические функционалы $S_n(x_1, x_2, \ldots x_n)$, то есть весьма специфические с.в. Мы рассмотрим несколько общих неравенств для с.в. и заодно явно докажем простую форму Закона Больших Чисел. На ЗБЧ опирается большинство статистических выводов: действительно, математическая статистика в основе своей использует соображение о связи вероятностей в одном повторении и частот появления событий при множественных независимых повторениях. Есть даже математическая теория, принадлежащая математику Рихарду фон Мизесу[9], в которой само понятие вероятности определено на основе частоты и чрезвычайно трудноформулируемого описания «типичных последовательностей независимых повторений» — это так называемая *фриквентистская версия теории вероятностей*, у нее есть стойкий круг приверженцев. Далее рассмотрим примеры распределений, отвечающих некоторым статистикам.

[9] не путать с его братом Людвигом фон Мизесом, приверженцем экстравагантной философии либертарианства!

3.1 Важные неравенства

Теорема 3.1 (Неравенство Маркова) *Пусть ξ - неотрицательная случайная величина, у которой есть математическое ожидание. Тогда $\forall \varepsilon > 0\ P\{\xi \geqslant \varepsilon\} \leqslant \frac{E(\xi)}{\varepsilon}$.*

Доказательство $E(\xi) \geqslant E\left(\xi \cdot I_{(\xi \geqslant \varepsilon)}\right) \geqslant \varepsilon E\left(I_{(\xi \geqslant \varepsilon)}\right) = \varepsilon P\{\xi \geqslant \varepsilon\}$ ∎

Простыми подстановками отсюда получаются еще два утверждения, они называются неравенствами Чебышёва:

Теорема 3.2 *Если соответствующие моменты с.в. определены, то $P\{\xi \geqslant \varepsilon\} \leqslant \frac{E(\xi^2)}{\varepsilon^2}$*

Теорема 3.3 *Если соответствующие моменты с.в. определены, то $P\{|\xi - E(\xi)| \geqslant \varepsilon\} \leqslant \frac{D(\xi)}{\varepsilon^2}$*

Теорема 3.4 (Закон больших чисел в форме Чебышёва) *Пусть случайные величины $\xi_1, \xi_2, \ldots \xi_n$ независимы и их дисперсии меньше некоторой константы C. Тогда для любого положительного ε при увеличении $n \to \infty$ вероятность*

$$P\left(\left|\frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} - \frac{E(\xi_1) + E(\xi_2) + \ldots + E(\xi_n)}{n}\right| \geqslant \varepsilon\right)$$

стремится к нулю.

Доказательство По неравенству Чебышёва достаточно установить, что $D\left((\xi_1 + \xi_2 + \ldots + \xi_n)/n\right) \to 0$. Но поскольку в данном случае дисперсия суммы равна сумме дисперсий имеем:

$$D\left(\frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n}\right) = \frac{1}{n^2} D\left(\xi_1 + \xi_2 + \ldots + \xi_n\right) = \frac{1}{n^2} \sum_i D(\xi_i) \leqslant \frac{nC}{n^2} \to 0$$ ∎

Эта (довольно абстрактная) формулировка в применении к схеме Бернулли независимых повторений опыта с двумя исходами соответствует базовому в статистике утверждению, что частота «успехов» стремится к вероятности «успеха» в одном испытании. Интересно, что прямое доказательство результата про «успехи» в повторениях достаточно громоздко, а общая теорема для более-менее произвольной с.в. гораздо проще.

В статистике важна версия ЗБЧ, в которой условие на существование вторых моментов вообще отсутствует:

Теорема 3.5 (ЗБЧ в версии Хинчина) *Для последовательности $\xi_1, \xi_2, \ldots$ независимых одинаково распределенных случайных величин с конечными математическими ожиданиями $E(\xi_k) = a$ имеет место сходимость по вероятности:*

$$\frac{S_n}{n} = \frac{\xi_1 + \xi_2 + \ldots + \xi_n}{n} \xrightarrow{P} a$$

Доказательство[10] короткое и основано на свойствах характеристических функций с.в..

Теорема 3.6 (Неравенство Чернова) *Если соответствующие моменты с.в. определены, то $P\{\xi \geqslant \varepsilon\} \leqslant \inf\limits_{t \geqslant 0} e^{-t\varepsilon} E\left(e^{t\xi}\right)$*

Доказательство Для $t > 0$ по монотонности экспоненты $P(\xi \geqslant \varepsilon) = P\left(e^{t\xi} \geqslant e^{t\varepsilon}\right)$, далее используем неравенство Маркова. ∎

Теорема 3.7 (Оценка хвостов гауссианы) *Пусть $\xi \sim \mathcal{N}(a, s^2)$, тогда*[11] *$\forall \varepsilon > 0\ E\left(e^{t\xi}\right) \leqslant \exp\left[\frac{1}{8} t^2 (b-a)^2\right]$,*

[10] Как обычно, вычитая из всех случайных величин константу a, сведем все к случаю $a = 0$. Теперь используем характеристические функции: пусть $\varphi_\xi(t) = E\left(\exp(it\xi)\right)$, поскольку первый момент наличествует и равен нулю, то производная хар.функции определена и $\varphi'_\xi(0) = ia = 0$.

Из независимости имеем

$$\varphi_{S_n/n}(t) = \left[\varphi_{\xi_k}\left(\frac{t}{n}\right)\right]^n = \left[1 + \varphi'_\xi(0)\frac{t}{n} + o\left(\frac{t}{n}\right)\right]^n = \left[1 + o\left(\frac{t}{n}\right)\right]^n \to 1$$

Таким образом хар.функция $\varphi_{S_n/n}(t)$ сходится к хар.функции случайной величины, принимающей значение ноль с вероятностью единица (и, тем самым, доказана сходимость по распределению). Но верна и сходимость по вероятности, действительно, для данного $\varepsilon > 0$ рассмотрим функцию-«шапочку» $\psi(x)$ равную единице в нуле и нулю при аргументах по модулю больших, чем ε. Тогда $E\left(\psi(S_n/n)\right) \to E\left(\psi(0)\right) = 1$. Но $E\left(\psi(S_n/n)\right) \leqslant 1 - P(|S_n/n| \geqslant \varepsilon)$ поэтому $P(|S_n/n| \geqslant \varepsilon) \to 0$ ∎

[11] Справедливо также усиление этого результата — так называемое неравенство Милля, в котором коэффициент в правой части не двойка, но обратно пропорционален ε

$$P\{|\xi - a| \geqslant \varepsilon\} \leqslant 2\exp\left[-\frac{\varepsilon^2}{2s^2}\right]$$

В частности, для среднего арифметического $\overline{\xi}$ m независимых одинаково распределенных как $\mathcal{N}(a, s^2)$ с.в. имеем

$$P\left\{|\overline{\xi} - a| \geqslant \varepsilon\right\} \leqslant 2\exp\left[-\frac{m\varepsilon^2}{2s^2}\right]$$

Доказательство Известное вычисление математематического ожидания логнормального закона дает $E\left(e^{t\xi}\right) = \exp\left[at + \frac{t^2 s^2}{2}\right]$. По неравенству Чернова

$$P\{\xi - a \geqslant \varepsilon\} \leqslant \inf_{t\geqslant 0} e^{-t\varepsilon} E\left(e^{t(\xi - a)}\right) = \inf_{t\geqslant 0} \exp\left[\frac{t^2 s^2}{2} - t\varepsilon\right]$$

Минимум правой части достигается при $t = \varepsilon s^{-2}$, и по симметрии гауссовского закона получаем нужную оценку для $P\{|\xi - a| \geqslant \varepsilon\}$ ∎

Похожие расуждения обеспечивают **неравенство Хёфдинга**.

Лемма 3.8 *Для с.в. ξ, такой что $a \leqslant \xi \leqslant b$ и $E(\xi) = 0$ и любого $t > 0$ выполнено $E\left(e^{t\xi}\right) \leqslant \exp\left[\frac{1}{8}t^2(b-a)^2\right]$*

Доказательство Выпуклость экспоненты дает неравенство $e^{t\xi} \leqslant \frac{b-\xi}{b-a}e^{ta} + \frac{\xi - a}{b-a}e^{tb}$. Применяя к обеим частям неравенства оператор математического ожидания с учетом $E(\xi) = 0$ имеем $E\left(e^{t\xi}\right) \leqslant \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} = e^{h(\frac{t}{b-a})}$, причем функцию h можно выразить явно

$$h(x) = \frac{a}{b-a}x + \ln\left[1 + \frac{a}{b-a}(1 - e^x)\right]$$

и потому в нуле $h'(0) = h(0) = 0$ и вдобавок $\forall x > 0 \;\; h''(x) \leqslant \frac{1}{4}$.

Разложение Тейлора с остаточным членом поэтому дает требуемое

$$h(x) = h(0) + h'(0)\cdot x + \frac{h''(\hat{x})x^2}{2} \leqslant \frac{x^2}{8}$$

что при подстановке $x = t(b-a)$ и дает $E\left(e^{t\xi}\right) \leqslant \exp\left[\frac{1}{8}t^2(b-a)^2\right]$ ∎

Теорема 3.9 (Неравенство Хёфдинга) *Пусть случайные величины $\xi_1, \xi_2, \ldots \xi_n$ независимы, $a_i \leqslant \xi_i \leqslant b_i$ и у них одинаковое математическое ожидание $E(\xi_i) = m$, тогда для $\varepsilon > 0$ и их усреднения $\overline{\xi}$ при $\forall t > 0$*

$$P\left\{|\overline{\xi} - m| \geqslant \varepsilon\right\} \leqslant 2e^{-nt\varepsilon}\prod_{i=1}^{n}\exp\left[\frac{t^2(b_i - a_i)^2}{8}\right]$$

В частности, для таких **и вдобавок одинаково распределенных** *с.в. минимум в правой части достигается при $t = \frac{4\varepsilon}{(b-a)^2}$ и потому*

$$P\left\{|\overline{\xi} - m| \geqslant \varepsilon\right\} \leqslant 2\exp\left[-\frac{2n\varepsilon^2}{(b-a)^2}\right]$$

Применительно к схеме Бернулли неравенство Хёфдинга означает, что наблюдаемая в эксперименте частота «успехов» распределена вокруг вероятности «успеха» в одном испытании с экспоненциально быстрым убыванием по n вероятностей возможных отклонений.

Доказательство Без ограничения общности положим $m = 0$, далее запишем неравенство Чернова для $t > 0$, учитывая свойство независимости:

$$P\left\{\overline{\xi} \geqslant \varepsilon\right\} = P\left\{\sum_i \xi_i \geqslant n\varepsilon\right\} \leqslant e^{-nt\varepsilon}E\left[\exp\left(t\sum_i \xi_i\right)\right] = e^{-nt\varepsilon}\prod_i E\left(e^{t\xi_i}\right)$$

По Лемме 3.8 $E\left(e^{t\xi_i}\right) \leqslant \exp\left[\frac{1}{8}t^2(b_i - a_i)^2\right]$, теперь, используя $P\left\{|\overline{\xi}| \geqslant \varepsilon\right\} = P\left\{\overline{\xi} \geqslant \varepsilon\right\} + P\left\{-\overline{\xi} \geqslant \varepsilon\right\}$, получаем требуемое. ∎

3.2 Важные распределения

Разнообразие возможных распределений в мат.статистике принято рассматривать, объединяя их в зависящие от параметра семейства. Иными словами, в выражении явной формулой функции (вариант плотности) распределения параметр, вариации этого параметра и задают семейство. Чтобы оставаться в рамках разумной простоты предполагаем, что при вариациях параметра распределение из семейства сохраняет свойство непрерывности или, напротив, дискретности.

3.2.1 Семейство гауссовых плотностей

В силу ЦПТ пределы сумм независимых случайных величин распределены по закону Гаусса $\mathcal{N}(a, \sigma^2)$. Для статистики важно следующее свойство гауссовых плотностей: сумма двух независимых гауссовых случайных величин с математическими ожиданиями a_1 и a_2, дисперсиями σ_1^2 и σ_2^2 снова гауссова случайная величина с математическим ожиданием $a_1 + a_2$ и дисперсией $\sigma_1^2 + \sigma_2^2$. Иными словами, *семейство гауссовых плотностей замкнуто относительно операции свертки* (действительно, плотность суммы двух независимых слагаемых есть свертка плотностей слагаемых).

- Упражнение: проверьте последнее утверждение прямым вычислением для случая $a_1 = a_2 = 0$

3.2.2 χ^2- распределение Пирсона

В силу той же ЦПТ многие формулы математической статистики хорошо приближаются гауссовым распределением. Рассмотрим суммы n квадратов независимых нормальных случайных величин $\mathcal{N}(0, 1)$, соответствующее распределение называется χ^2- распределением Пирсона с n степенями свободы. Вывод соответствующей формулы для χ^2-плотности имеет смысл рассмотреть (в следующем разделе) в более общем контексте, в примерах 3.2.5 имеется искомая формула.

3.2.3 Семейство гамма-распределений

Как известно из курса математического анализа гамма-функция от аргумента $t > 0$ определена следующей формулой

$$\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} du$$

в комплексном анализе рассказывается как она аналитически продолжается на другие комплексные значения аргумента. Известно, что гамма-функция интерполирует значения факториала, а именно при натуральном k:

$$\Gamma(k+1) = k!$$

Из определения гамма функции следует, что при любых $b > 0$,$r > 0$ интеграл функции положительного аргумента

$$f_{[b,r]}(x) = \frac{1}{\Gamma(r)} b^r x^{r-1} e^{-bx}$$

равен 1, а потому эта функция задает плотность распределения, называющуюся *гамма-плотностью с параметром r и масштабным параметром b.*

Упражнение по математическому анализу на вычисление несобственных интегралов показывает, что математическое ожидание для гамма-распределения с параметром r и масштабным параметром b — это r/b, а дисперсия, соответственно, r/b^2.

- Проверьте последнее утверждение прямым вычислением для случая $b = 3$ $r = 1$

Предложение 3.10 *Семейство гамма-плотностей замкнуто относительно операции свертки:*

$$f_{[b,r_1]} * f_{[b,r_2]} = f_{[b,r_1+r_2]}$$

Доказательство Действительно, вычисление свертки дает выражение

$$f_{[b,r_1]} * f_{[b,r_2]}(x) = \frac{b^{r_1+r_2}}{\Gamma(r_1)\Gamma(r_2)} e^{-x} \int_0^\infty (x-y)^{r_1-1} y^{r_2-1} dy$$

Подставляя $y = xt$ видим, что отличие свертки от $f_{[b,r_1+r_2]}$ заключено лишь в постоянном множителе. Однако свертка плотностей по смыслу также должна быть плотностью, поэтому этот множитель должен быть равен единице. ∎

3.2.4 Замечание

Интеграл, возникающий при вычислении свертки при $x = 1$ — это известная из математического анализа бета-функция $B(r_1, r_2)$. Возвращаясь к определению гамма-функции, видим из доказательства предложения, что для $u, v > 0$:

$$B(u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}$$

3.2.5 Примеры

- При $r = 1$ соответствующая гамма-плотность совпадает с плотностью показательного распределения. Объяснить.
- Рассмотрим распределение квадрата нормальной случайной величины $\mathcal{N}(0, 1)$. Проверьте (прямым вычислением), что плотность распределения квадрата нормальной случайной величины $\mathcal{N}(0, 1)$ при любом x совпадает с точностью до постоянного коэффициента с $f_{[\frac{1}{2}, \frac{1}{2}]}(x)$, а значит, этот коэффициент равен единице и мы вдобавок получаем, что $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. Воспользовавшись Предложением 3.10 мы сразу получаем, что сумма n квадратов независимых нормальных случайных величин распределена с плотностью $f_{[\frac{1}{2}, \frac{n}{2}]}(x)$ — это и есть формула плотности распределения χ^2 с n степенями свободы.
- Вывести формулу плотности распределения суммы n квадратов независимых гауссовых случайных величин с параметрами $(0, \sigma)$. Если независимых гауссовых слагаемых три, то ясна физическая интерпретация этой суммы как квадрата скорости трехмерной частицы.
- Используя формулу плотности для квадрата скорости трехмерной частицы, вывести формулу плотности распределения самой скорости — она имеет специальное название *плотности Максвелла* . Разумеется, эта задача сводится к связи плотностей для неотрицательной случайной величины и квадратного корня из нее. Ответ: $2x f_{[\frac{1}{2}, \frac{3}{2}]}(x^2)$.

3.2.6 Экспоненциальные семейства

Принято рассматривать достаточно общую конструкцию семейства (одномерных) распределений, плотности (или вероятности в дискретном случае семейства) $f_\xi(x)$ в котором могут быть представлены как произведение трех, зависящих от векторного параметра $\{\mathfrak{t}_1, \mathfrak{t}_2, \dots \mathfrak{t}_m\}$, сомножителей и достаточно хороших (фиксированных для данного семейства) функций R, T, S.

$$f_\xi(x, R, T, S, \vec{\mathfrak{t}}) = \exp\left[\sum_{i=1}^{m} \mathfrak{t}_i R_i(x) - T(\mathfrak{t}_1, \mathfrak{t}_2, \dots \mathfrak{t}_m) + S(x)\right]$$

Например, биномиальные распределения с параметрами n, p определены вероятностями вида

$$f(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x} = \exp\left\{\ln\left(\frac{p}{1-p}\right)x + n\ln(1-p) + \ln\binom{n}{x}\right\}$$

и если зафиксировать конкретное значение n, то получится экспоненциальное семейство по вещественному параметру $\ln\left(\frac{p}{1-p}\right)$.

- Прверьте, что рассмотренные выше семейства все экспоненциальные.

Свойство образовывать экспоненциальное семейство не связано с замкнутостью семейства по сложению, роль экспоненциальных семейств становится ясной в так называемом байесовском подходе в статистике, о нем пойдет речь в конце нашего вводного курса. При этом удобство рассмотрения именно экспоненциального семейства в свойстве факторизации плотности и в том, что носитель плотности у всех членов семейства одинаков.

3.2.7 Плотность отношения двух независимых случайных величин

Рассмотрим две независимые случайные величины α, β, причем $\beta > 0$. Плотность отношения $\theta = \alpha/\beta$ вычисляется обычным образом:

$$F_\theta(t) = P(\theta \leqslant t) = \int_0^\infty F_\alpha(ty) f_\beta(y) dy \tag{1}$$

$$f_\theta(t) = F'_\theta(t) = \int_0^\infty f_\alpha(ty) y f_\beta(y) dy \tag{2}$$

3.2.8 Смеси или рандомизации

На практике бывает так, что исследуемые данные поступают от нескольких источников, каждый из которых описан своей функцией распределения. Понятие смеси объясняет как именно представить случайную величину, лежащую в основе статистической модели (и тем самым, определяющей вероятностную меру на выборках).

Пусть задано семейство зависящих от параметра $y \in \mathbb{R}$ плотностей $f_\xi(x, y)$ и пусть на множестве параметров задана вероятностная мера, то есть y – значения некоторой случайной величины η. *Рандомизацией* посредством η или *смесью* называется плотность распределения вида

$$g(x) = \int_{-\infty}^{+\infty} f_\xi(x, y) f_\eta(y) dy$$

тем самым коэффициент $f_\eta(y)$ можно считать взвешивающим фактором в смеси разных плотностей.

- Объяснить, почему $g(x)$ — это плотность распределения.

В случае дискретного множества параметров соответствующая формула использует вероятности и знак суммы:

$$\tilde{g}(x) = \sum_y f_\xi(x, y) P(\eta = y)$$

Эта конструкция описывает указанную выше модель измерения с набором многих источников (проиндексированных параметром), когда в качестве окончательного измеренного значения в данном повторении опыта случайно выбирается один из источников. Такая модель характерна, например, когда в принципе ясно, что значение должно иметь гауссово распределение с нулевым математическим ожиданием, но значение σ случайно меняется в процессе эксперимента. Замечательно, что возникающая плотность смеси при этом может быть совсем не похожа на гауссов закон. Незнание того, как могут быть устроены смеси, при попытках объяснить удивительные формы экспериментальных гистограмм породило много бестолковых усложнений простых и ясных моделей (которые разумно предполагали справедливость ЦПТ для этих конкретных ситуаций). Мы рассмотрим несколько таких характерных форм, возникающих из смеси гауссовых семейств позже на упражнениях.

3.2.9 Распределения Фишера-Снедекора, Парето и Стьюдента

Для независимых нормально распределенных случайных величин $\xi_1, \xi_2, \ldots \xi_{m+n}$ положим $\zeta = \xi_1^2 + \xi_2^2 + \ldots + \xi_m^2$ и $\eta = \xi_{m+1}^2 + \xi_{m+2}^2 + \ldots + \xi_{m+n}^2$. Случайная величина вида $\frac{n\zeta}{m\eta}$ имеет плотность распределения, которая называется плотностью Фишера-Снедекора, то есть речь идет об отношении нормированных случайных величин с распределениями χ^2. Для этой же конструкции используют также термин *Z-распределение Фишера*, а также иногда термин *распределение Парето*[12], впрочем во всех этих случаях возможно появление перед отношением $\frac{\zeta}{\eta}$ зависящих от m, n дополнительных числовых коэффициентов. По-видимому, стоит запомнить лишь общую конструкцию с.в. Фишера-Снедекора, проверяя всякий раз какой именно коэффициент перед ней имеют в виду в конкретном выражении. Приведенные выше формулы позволяют найти явно аналитическое выражение плотности Фишера-Снедекора, действительно, в формулу для отношения 3.2.7 надо подставить плотности, равные соответственно $f_{[\frac{1}{2}, \frac{m}{2}]}(x)$ и $f_{[\frac{1}{2}, \frac{n}{2}]}(x)$, далее волшебной подстановкой переменных (изначально довольно сложное) выражение 3.2.7 упростится, позволив избавиться от интеграла в 3.2.7 и записать все в виде:

$$\frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)} \frac{x^{v-1}}{(1+x)^{u+v}}, \quad x > 0, \quad u = n/2, v = m/2$$

[12] В честь экономиста Парето, который интуитивно полагал, что аппроксимация при $x \to \infty$ плотности распределения дохода x должна бы иметь степенное убывание, а указанная конструкция с.в. обеспечивает это свойство.

В том случае, когда ζ – нормальная с.в. с плотностью $\mathcal{N}(0,1)$, а $\eta = \sqrt{\xi_{m+1}^2 + \xi_{m+2}^2 + \ldots + \xi_{m+n}^2}$ (то есть плотность η^2 совпадает с $f_{[\frac{1}{2},\frac{n}{2}]}$) распределение соответвующего отношения называется **распределением Стьюдента**[13]:

$$\frac{\Gamma((n+1)/2)}{\sqrt{\pi}\Gamma(n/2)}(1+x^2)^{-(n+1)/2}$$

- Выяснить, как обстоит дело с моментами распределения Стьюдента.

Детали как избавиться от знака интеграла в каждом из случаев см.в В.Феллер "Теория вероятностей", том 2, Глава II, §3, на практике окончательные аналитические выражения этих плотностей давно реализованы библиотечными функциями, поскольку запись в терминах гамма-функций, все равно без компьютера малополезна.

3.2.10 Бета-распределения

В связи с плотностью с.в. Фишера-Снедекора упомянем еще сходную формулу плотности $f_{\{B,u,v\}}(x)$ двупараметрического семейства Бета-распределений на $(0,1)$ с двумя произвольными параметрами $u > 0$, $v > 0$:

$$f_{\{B,u,v\}}(x) = \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}\frac{x^{v-1}}{(1-x)^{u-1}}, \quad x \in (0,1)$$

Название семейства происходит из нормализационного коэффициента, равного значению Бета-функции B. Формула плотности Снедекора есть частный случай преобразованной плотности

$$\frac{1}{(1+x)^2}f_{\{B,u,v\}}\left(\frac{1}{1+x}\right) = \frac{\Gamma(u+v)}{\Gamma(u)\Gamma(v)}\frac{x^{u-1}}{(1+x)^{u+v-1}}, \quad x \in (0,\infty)$$

которая, как можно показать, отвечает преобразованию $\xi \mapsto \frac{1}{\xi} - 1$ Бета-распределенной случайной величины ξ. Бета-распределения бывают нужны при байесовских оценках в статистике, пример которых будет далее.

3.2.11 Плотность Коши

Для двух независимых нормальных (то есть распределенных как $\mathcal{N}(0,1)$) случайных величин α, β отношение $\alpha/|\beta|$ распределено по закону Коши. Общее же определение плотности Коши (с масштабным параметром t) таково:

$$f_\xi(x) = \frac{1}{\pi}\frac{t}{t^2+x^2}$$

В контексте математической статистики для нас важно, что свертка плотности Коши с параметром u и плотности Коши с параметром v явно вычисляется несложным интегрированием и в ответе получается свертка плотности Коши с параметром $u+v$. Применительно к статистическим формулам это означает, что среднее значение $\bar{x}$ выборки значений $x_1, x_2, \ldots x_n$ распределенной по закону Коши случайной величины также распределено по закону Коши! В частности, множество стандартных формул статистики оказываются бессмысленными при их применении к данным, распределенным по закону Коши и аналогичным.

- Выяснить, как обстоит дело с моментами распределения Коши.
- Вычислением проверить, что плотность распределения суммы двух независимых случайных величин Коши с параметром 1 такая же как и плотность удвоенной случайной величины Коши с параметром 1.

Усилиями известного математика П.Леви описание подобных распределений систематизировано, что позже привело к созданию специализированных методов работы с такими данными (в первую очередь здесь надо упомянуть математика Б.Мандельброта)

[13] автор формулы английский статистик Госсет, под псевдонимом «студент» опубликовал ставшую очень популярной научную работу по оцениванию параметров гауссовых распределений

3.2.12 Двустороняя показательная плотность

Плотность распределения разности двух одинаково распределенных показательных случайных величин называется двусторонней показательной плотностью или **плотностью Лапласа**.

- Вывести прямым вычислением формулу плотности Лапласа и указать ее математическое ожидание и дисперсию. Ответ: плотность $\frac{b}{2}\exp -|b|x$, математическое ожидание 0, дисперсия $2b^{-2}$

По ряду причин (относящихся к свойствам распределений смесей) очень часто прямой предварительный (например, графический) анализ данных провоцирует гипотезу о распределении Лапласа вместо естественно ожидаемого гауссова распределения, что породило и порождает множество диссидентствующих публикаций в духе "а вот как оно на самом деле-то устроено в этой вашей реальности". Соответствующее теоретическое объяснение еще с 1970-х годов математикам известно, но в учебную литературу оно, увы, не попало.

3.2.13 Логистический закон

Мы включили в рассмотрение этот тип распределений как пример, к сожалению, слишком часто встречавшейся в практических применениях статистики ситуации. Речь идет о функции распределения вида

$$F_\xi(x) = \frac{1}{1+e^{-ax+b}}, \qquad a>0$$

Ссылки на логистический закон почему-то очень распространены, для внесения ясности приведем цитату из книги В.Феллер "Теория вероятностей", том 2, Глава II. *Существует невероятно большая литература, где делаются попытки доказать трансцендентный «закон логистического развития». Предполагалось возможным представить практически все процессы развития (измеренные в соответствующих единицах) функцией указанного типа с t, изображающим время. Весьма длинные таблицы с χ^2-критериями подтверждали это положение для человеческих популяций, бактериальных колоний, развития железных дорог и т. д. Было обнаружено, что как высота, так и вес растений и животных подчиняются логистическому закону, хотя из теоретических соображений ясно, что эти две величины не могут подчиняться одному и тому же распределению. Лабораторные эксперименты на бактериях показали, что даже систематические нарушения не могут привести к другим результатам. Теория популяций была основана на логистических экстраполяциях (даже если они оказывались очевидным образом ненадежными). Единственное затруднение «логистической» теории заключается в том, что не только логистическое распределение, но также нормальное, Коши да и другие распределения могут быть подогнаны под тот же самый материал с тем же или лучшим согласием. В этой конкуренции логистическое распределение не играет никакой выдающейся роли: самые противоречивые теоретические модели могут быть подтверждены на том же наблюдательном материале.*

Теории этого рода недолговечны, так как они не открывают новые пути, а новые подтверждения одних и тех же старых вещей очень скоро становятся надоедливыми. Однако наивное рассуждение само по себе не было заменено здравым смыслом, и поэтому может быть полезно иметь очевидное доказательство того, как могут вводить в заблуждение взятые сами по себе критерии согласия.

3.2.14 Дополнение

3.2.15 Многомерное гауссово распределение

Напомним конструкцию многомерного гауссова распределения и связи его с одномерными характеристиками. Одномерное гауссово распределение, определено своей функцией плотностей распределения, включающей два параметра a и $\sigma>0$, причем смысл этих параметров устанавливается прямым вычислением: a равно математическому ожиданию σ^2 – дисперсии. Многомерное гауссово распределение обобщает этот принцип: d- мерное распределение задается своей плотностью, зависящей от произвольных параметров $m_1, m_2, \dots m_d$ и симметрической, положительно определенной (обязательно вспомните, что это значит) матрицы $\mathbf{\Lambda}$ порядка d.

Например, в трехмерном $d=3$ случае:

$$\mathfrak{g}(x_1,x_2,x_3) = \sqrt{\frac{\det\mathbf{\Lambda}}{(2\pi)^3}}\exp\left[-\frac{1}{2}\sum_{i,j=1}^{3}\Lambda^{ij}(x_i-m_i)(x_j-m_j)\right], \tag{3}$$

Прямым вычислением устанавливается вероятностный смысл параметров: матрица $\mathbf{\Lambda}$ является обращением матрицы, составленной из всевозможных попарных ковариаций компонент вектора, а набор $m_1, m_2, \dots m_d$ образует вектор математических ожиданий компонент.

Если определить «$\mathbf{\Lambda}$-скалярное произведение векторов» $(\mathbf{x},\mathbf{y})_{\mathbf{\Lambda}} = (\mathbf{\Lambda x},\mathbf{y}) = \sum_{i,j=1}^{3}\Lambda^{ij}x_iy_j$ и соответствующую $\mathbf{\Lambda}$-«длину» $|\mathbf{x}|_{\mathbf{\Lambda}} = \sqrt{(\mathbf{x},\mathbf{x})_{\mathbf{\Lambda}}}$, то формула $d=3$ при станет напоминать формулу одномерного гауссовского распределения:

$$\mathfrak{g}(\mathbf{x}) = \sqrt{\frac{\det\mathbf{\Lambda}}{(2\pi)^3}}\exp\left[-\frac{1}{2}\Big(\mathbf{\Lambda}(\mathbf{x}-\mathbf{m}),\mathbf{x}-\mathbf{m}\Big)\right] = \sqrt{\frac{\det\mathbf{\Lambda}}{(2\pi)^3}}e^{-\frac{1}{2}|\mathbf{x}-\mathbf{m}|^2_{\mathbf{\Lambda}}} \tag{4}$$

Если матрица ковариаций диагональная, то и $\mathbf{\Lambda}$ диагональная, например, в указанном выше трехмерном случае на диагонали стоят три числа $\frac{1}{\sigma_1^2},\frac{1}{\sigma_3^2},\frac{1}{\sigma_3^2}$, от этого формула 3 дополнительно упрощается:

$$\mathfrak{g}(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}\sigma_1\sigma_2\sigma_3}e^{-\frac{1}{2}\left(\frac{x_1-m_1}{\sigma_1}\right)^2-\frac{1}{2}\left(\frac{x_2-m_2}{\sigma_2}\right)^2-\frac{1}{2}\left(\frac{x_3-m_3}{\sigma_3}\right)^2} = \prod_j\frac{1}{\sqrt{2\pi}\sigma_j}e^{-\frac{1}{2}\left(\frac{x_j-m_j}{\sigma_j}\right)^2}$$

Определение плотности многомерного гауссова вектора использет вид последнего выражения в формуле 4, а именно, многомерная плотность при $d=n$ полагается равной:

$$\sqrt{\frac{\det\mathbf{\Lambda}}{(2\pi)^n}}e^{-\frac{1}{2}|\mathbf{x}-\mathbf{m}|^2_{\mathbf{\Lambda}}}$$

3.2.16 Важная характеризация гауссова распределения

Приведем очень любопытное свойство гауссовых законов, во многих случаях влияющее на выбор гипотезы о распределении.

Пусть есть случайный вектор с компонентами ξ_1,ξ_2 и линейное преобразование с единичным детерминантом $\left\|\begin{matrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{matrix}\right\|$. Если ξ_1,ξ_2 независимые и гауссовы, то, преобразуя вектор-столбец этой матрицей, получится опять же вектор с гауссовыми компонентами ζ_1,ζ_2, причем существуют нетривиальные матрицы такие, что ζ_1,ζ_2 снова независимы. Оказывется, что *это свойство одномерного гауссова распределения уникально и не выполняется ни для какого иного распределения.* Ниже приведено формулировка лишь для распределений с непрерывными плотностями, более общее доказательство требует использования характеристических функций. Справедлива также более общая теорема Дармуа–Скитовича 1953г., в которой аналогичное свойство доказано и для любого конечного набора $(\xi_1,\xi_2,\dots\xi_n)$. Здесь мы приведем для простоты только версию утверждения с сильными условиями на законы распределения, менее ограничительные условия потребуют рассмотрений не дифференциальных соотношений, а разностных.

Теорема 3.11 *Пусть как и выше ξ_1,ξ_2 независимы и имеют гладкие плотности и ζ_1,ζ_2 независимы. Тогда либо все эти четыре случайные величины гауссовы либо преобразование тривиально: то есть с.в. ζ_i пропорциональны с.в.ξ_j, $i,j=1,2$*

Доказательство Для соответствующих плотностей имеем функциональное соотношение:

$$f_{\zeta_1}(a_{11}x+a_{12}y)f_{\zeta_1}(a_{21}x+a_{22}y) = f_{\xi_1}(x)f_{\xi_2}(y) \tag{5}$$

Лемма 3.12 утверждает, что такое соотношение *между гладкими функциями* влечет формулу $f_{\zeta_1}(x) = \pm\exp[P_1(x)]$, где $P_1(x)$ полином степени $\leqslant 2$. Отсюда и вытекает утверждение теоремы, так как гауссовы плотности характеризуются экспонентой квадратичной формы.

Лемма 3.12 *Пусть функции $f_{\zeta_1}(\cdot), f_{\zeta_2}(\cdot)$ дважды дифференциируемы и строго положительны и выполняется функциональное соотношение (5). Если элементы матрицы $\begin{pmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}$ все ненулевые и определитель отличен от нуля, то $f_{\zeta_1}(t) = \pm\exp[P_1(t)]$. Аналогичное выражение справедливо и для $f_{\zeta_2}(t)$.*

Доказательство Для $g_k = \ln f_{\zeta_k}$ и аргументов $z_k = a_{k1}x + a_{k2}y$ имеем $g_1(z_1)+g_2(z_2) = \ln f_{\xi_1}(x) + \ln f_{\xi_2}(y)$. Далее посредством дифференциирования по x и по y получаем соотношение $a_{11}a_{12}g_1''(z_1)+a_{21}a_{22}g_2''(z_2) = 0$. Варьируяя аргументы x,y так, чтобы z_2 оставалось постоянным, получаем $g_1''(z_1)\equiv\text{const}$. ∎

Предположение о линейной связи ζ_i и ξ_j существенно. Несложно показать, например, что для двух независимых показательно распределенных ξ_1,ξ_2 случайные величины $\frac{\xi_1}{\xi_1+\xi_2}$ и $\xi_1+\xi_2$ также независимы.

Заметим, что задолго до появления характеризационной теоремы Дармуа – Скитовича физик Максвелл, рассуждая о законе распределения молекул по скоростям (см. последний пример в разделе 3.2.5), исходил из следующего сходного по смыслу соображения: для случайного вектора, закон распределение которого зависит только от квадратичной формы $x^2+y^2+z^2$, предположение о взаимной независимости трех проекций скорости на ортогональные оси влечет, что распределения каждой из проекций гауссово.

3.3 Формулы статистики. Эмпирические соображения

Эмпирический аналог теоретического математического ожидания $E(\xi) = \int x dF_\xi$ получается с использованием эмпирической функции распределения $F_n^*(x)$ вместо $F_\xi(x)$, он также называется *выборочным средним* и очевидно вычисляется по формуле

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Теорема Хинчина показывает, что выборочное среднее (понимаемое как случайная величина) сходится по вероятности к математическому ожиданию. При этом, вообще говоря, важен и вопрос о том, как именно распределено выборочное среднее при данном распределении с.в. ξ и при конкретном n.

Аналогичным образом эмпирическую дисперсию можно определить формулой

$$S^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - (\bar{x})^2$$

вместо которой на практике употребляют несколько иную запись (чуть ниже объясним почему), а именно:

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \frac{n}{n-1}S^2$$

В целом этот подход показывает *одну из версий* происхождения формул статистики: для числового функционала T от с.в. ξ (например, математического ожидания $a = \int x dF_\xi(x)$, дисперсии $\sigma_\xi^2 = \int (x-a)^2 dF_\xi(x)$, медианы $m = F_\xi^{-1}(1/2)$) часто рассматривают *выборочную статистику* получающуюся из применения функционала к $F_n^*(x)$. В англоязычной литературе это называют plug-in estimator. Важно понимать, что выборочная оценка задает случайную величину, распределение которой вычисляется через распределение с.в. ξ.

Достаточно часто пытаются оценить значения теоретических параметров с помощью статистик. Вообще же статистика называется **состоятельной** если при $n \to \infty$ ее значения сходятся *по вероятности* к величине параметра теоретического распределения. В частности, оценка мат.ожидания состоятельна в силу теоремы Хинчина. Проверка состоятельности оценок дисперсии S^2 и s^2 уже несколько более громоздкая вещь, связанная с рутинным исследованием соответствующих формул, для экономии времени в состоятельность этих формул предлагается поверить (или прочесть в учебнике).

Альтернативное использование второй формулы для оценки дисперсии связано с другим важным понятием **несмещенности**, заключающемся в том, что математическое ожидание несмещенной оценки должно быть *при любом n равно* соответствующему параметру теоретического распределения.

- Покажите, что оценка s^2 несмещенная, а S^2 — смещенная. При этом разница их имеет порядок $1/n$ и потому сколь-либо важна лишь в случае небольших значений n.

Важно: откуда берутся формулы К сожалению, из-за того что изложение статистических методов принято вести перечислением примеров и формул, слишком часто встречается мнение, что указанные выше формулы $\bar{x}$, s^2 и S^2 универсальны, применяются везде и всегда, то есть имеют смысл для всех ситуаций.

На самом деле, тонкость заключена в том, какое именно предположение о выборке было дано изначально. В предыдущих разделах было пояснение о *гипотезах*, которые, в свою очередь подразделяются на *простые*, *сложные*, формулируются чаще не по-одиночке, а списком и т д. Иными словами речь идет о задании структуры того семейства $\mathfrak{F}$, в котором (мы предполагаем) лежит неизвестная нам функция распределения $F_\xi(x)$. Если такое семейство задано в виде множества зависящих от численного или векторного параметра w функций распределения $F_\xi(x, w)$, то такая задача, как уже говорилось, называется *задачей параметрической статистики* и тогда оценочные формулы для, например, моментов распределения могут быть *совсем не такими,* как приведенные выше среднее арифметическое $\bar{x}$ для математического ожидания или s^2 и S^2 для дисперсий. В методах параметрической статистики важное место занимает иной способ построения оценочных формул — Метод Максимального Правдоподобия, о нем рассказано будет далее. Сейчас надо отметить, что статистики $\bar{x}$, s^2 и S^2 получены из метода plug-in estimator, который по сути опирается лишь на теорему Гливенко (а в ней про $\mathfrak{F}$ не было сказано ничего конкретного). К этому обстоятельству следует отнестись внимательно, поскольку в справочниках и учебниках обычно это излагают крайне нечетко.

3.3.1 Для запоминания

Вот достаточно обширный список общеупотребительных формул для выборок размера n (большинство из них возникло из метода plug-in) :

1. выборочное среднее $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$ (эмпирический аналог математического ожидания),
2. выборочная дисперсия (эмпирический аналог дисперсии) $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2$, в случае смещенной оценки (см.выше) это
$$S^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - (\bar{x})^2$$
3. выборочный момент j-го порядка $M_j = \frac{1}{n}\sum_{k=1}^{n} x_k^j$.
4. выборочный центральный момент j-го порядка $\overset{0}{M}_j = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})^j$. Здесь, также как и в случае дисперсии для $j = 2$, различают смещенный $\overset{0}{M}_j$ и несмещенный $\overset{0}{m}_j$ моменты, связь между ними зависит от значения j, например,
$$\overset{0}{m}_3 = \frac{n^2}{(n-1)(n-2)}\overset{0}{M}_3$$
$$\overset{0}{m}_4 = \frac{n}{(n-1)(n-2)(n-3)}\left[(n^2-2n+3)\overset{0}{M}_4 + 3(2n-3)\overset{0}{M}_2^{\,2}\right]$$
$$\dots$$
5. среднеквадратичное (стандартное) отклонение s
6. коэффициент вариации $s/\bar{x}$
7. t-коэффициент Стьюдента с $n-1$ степенью свободы (применяется для гауссовых выборок с предположительно известным значением математическим ожиданием a)
$$t = \frac{|\bar{x} - a|}{s}\sqrt{n}$$
8. порядковые статистики, основанные на вариационном ряде $y_1 \leqslant y_2 \leqslant \dots y_n$ —- упорядочении исходной выборки $x_1, x_2, \dots x_n$
   (a) размах выборки $y_n - y_1$
   (b) выборочный q-квантиль для $0 < |q| < 1$ строится подходящей интерполяцией значений соседних членов вариационного ряда $y_{[qn]}$ и $y_{[qn]+1}$
   (c) выборочная медиана это $\frac{1}{2}$-квантиль

3.3.2 Упражнение

Важное, но не слишком простое упражнение (это так называемая Лемма Фишера):

- Пусть $\xi_1, \xi_2, \dots \xi_n$ – независимые одинаково распределенные по гауссовому закону $\mathcal{N}(0,1)$ случайные величины. Подставляя их в формулы выборочного среднего и выборочной дисперсии получим две случайные величины. Указать, зависимы ли они.

4 Структура математической статистики. Продолжение

После ознакомления на конкретных примерах как работают статистические методы и откуда берутся формулы попробуем разобраться в основной терминологии, которая используется при описании имеющихся методов. Не следует думать, что статистические методы организованы в регулярный список, из которого любой желающий быстро подберет оптимальный для своей задачи вариант.

Поскольку статистика используется как прикладная наука изменяются и задачи и методы, а главное — технологии применения этих методов. До середины XX века статистический справочник состоял из методов (как правило с примерами), которые вполне можно было реализовать вычислениями с использованием бумаги и карандаша. Эти методы сопровождались также таблицами значений тех функций распределения, которые не имели простой аналитической формулы. Обучение статистическим методам в значительной степени состояло из умения воспользоваться этими таблицами, причем точность таблиц была умеренной поскольку рассматриваемые выборки в приложениях были сравнительно невелики. Например, в конце каждого учебника или справочника имелись таблицы, связанные с функцией распределения стандартной нормальной гауссовой случайной величины η, однако затабулированы были необязательно значения

$$F_\eta(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt,$$

но чаще значения следующих интегралов (из которых сегодня лишь последний включен в мало-мальски продвинутые калькуляторы)

$$\frac{1}{\sqrt{2\pi}} \int\limits_{0}^{x} e^{-\frac{1}{2}t^2} dt, \quad \frac{1}{\sqrt{2\pi}} \int\limits_{-x}^{x} e^{-\frac{1}{2}t^2} dt, \quad \frac{2}{\sqrt{\pi}} \int\limits_{x}^{+\infty} e^{-t^2} dt, \quad \mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \int\limits_{0}^{x} e^{-t^2} dt$$

Впрочем, умение выразить значения одних интегралов через другие, осталось необходимым.

Широкое распространение компьютерной техники сместило деятельность в сторону больших объемов данных и сделало общеупотребительными совершенно новые методы, например бутстрэп (bootstrap) и «складной нож» (jackknife), а промежуточные вычисления давно реализованы в большинстве математических пакетов стандартными библиотечными функциями. Однако проблема систематизации методов никуда не делась. Попытка посильно систематизировать терминологию время от времени предпринималась разными авторами; ниже мы будем придерживаться варианта изложения из главы 6 книги Larry Wasserman "All of Statistics. A Concize Course in Statistical Inference", Springer, 2003.

4.1 Статистическая терминология. Продолжение

В целом принято различать методы трех видов: *точечные оценки, доверительные интервалы и проверка гипотез* применительно к статистическим задачам двух типов: параметрическим и непараметрическим.

Статистическая модель $\mathfrak{F}$ — это класс возможных распределений (в том числе распределений случайных векторов) к которому относится рассматриваемая выборка. **Параметрическая модель** задается указанием на конечный набор определяющих распределение параметров. Например, гауссова параметрическая модель – это двупараметрическая модель: все распределения $\mathcal{N}(a, \sigma)$ характеризуются парами a, σ. Вообще, параметрическая модель это $\mathfrak{F} = \left\{ f(x, \mathfrak{t}) \middle| \mathfrak{t} \in \Theta \right\}$, где Θ называется пространством параметров. Если класс возможных распределений не удается явно описать конечными средствами, то модель считается **непараметрической**.

Непараметрическая оценка функционала Задача оценки для момента распределения (например, для математического ожидания $E(\xi^k) = \int x^k dF_\xi$ состоит в поиске адекватной формулы, определенной на непараметрическом семействе $\mathfrak{F}$ таких распределений, где соответствующий интеграл сходится.

4.2 Оценивания: основные идеи

Задача оценивания может быть понята следующим образом: изучают измерениями реальное явление, предположительно результаты измерения $(x_1, \dots x_n)$ отвечают случайной величине из *параметризованного* (векторным или скалярным) параметром $\mathfrak{t}$ семейства $\mathfrak{F} = \left\{ f_\xi(x, \mathfrak{t}) \middle| \mathfrak{t} \in \Theta \right\}$. По данным схемы повторений ставится вопрос о значении конкретного параметра $\hat{\mathfrak{t}}$. Здесь речь не идет о проверке (простой) гипотезы о конкретном распределении, а о выборе $T(x_1, \dots x_n)$ значения $\mathfrak{t}$ для конкретного распределения $f_\xi(x, \mathfrak{t})$, разумеется, этот выбор может быть лишь приблизительным в силу конечности числа повторений n.

Естественные вопрос – «какое предсказание лучше» — упирается в необходимость договориться как вообще сравнивать разные оценки.

Общая точка зрения: поскольку для каждого метода (предсказывающей формулы) предсказание окажется случайной величиной, то методы надо сравнивать при их многократном применении и преде всего надо ввести

меру сравнения: функцию потерь $L(\mathfrak{t}, d) \geqslant 0$, $d = T(x_1, \dots x_n)$. В это выражение уже вошел неизвестный истинный параметр $\mathfrak{t}$ — конкретное значение функции потерь на конкретной выборке едва ли посчитаешь! В статистике принято характеризовать статистические правила средними результатами, достигаемыми при многократном применении, поэтому надо рассматривать усреднения *случайной величины* $L(\mathfrak{t}, d)$. Важный обстоятельство: усреднения в силу ЗБЧ должны быть близки к математическим ожиданиям $E\left(L(\mathfrak{t}, d)\right) = E\left(L(\mathfrak{t}, T(x_1, \dots x_n))\right)$, поскольку есть мера $f_\xi(x, \mathfrak{t})$ на n-выборках $(x_1, \dots x_n)$. **Эти математические ожидания, зависящие от выбранного метода, от выбранного способа сравнения L и от значения параметра $\mathfrak{t}$, называются *функциями риска*. Один метод лучше другого, если при всех $\mathfrak{t} \in \Theta$ у него риск меньше.** А наилучший метод тогда тот, у которого риск вообще самый меньший из возможных, да только обычно минимального элемента в возможных функциях риска не существует. Как же найти оптимально-лучший метод?

Есть специальный байесовский подход, состоящий в том, чтобы изначально считать Θ вероятностным пространством с равномерной мерой [14] и далее связывать выборки с изменениями этой меры, по аналогии с формулой Байеса апостериорной вероятности. В этих подходах рассматривают осреднения функции риска по $\mathfrak{t}$, а сам поиск ведется с целью найти экстремум этих осреднений. Эти соображения в нашем вводном курсе затрагиваться не будут, но могут встретиться в справочной литературе.

Другая возможность - поиск наилучших правил, но в более узком множестве методов, чаще всего для функции потерь вида $L(\mathfrak{t}, d) = |\mathfrak{t} - d|^2$ и в классе *несмещенных* оценок $\theta_n = T(x_1, \dots x_n)$, то есть таких, что $E(\theta_n)$ при любом n совпадает с истинным значением $\hat{\mathfrak{t}}$ параметра. На этом пути можно достигнуть успеха.

При этом надо помнить, что «оптимальное оценивание» является таковым лишь в строго определенном способе сравнения разных методов, никакого абсолютного подхода не просматривается.

4.3 Точечные оценки

Речь идет о поиске подходящей формулы в ситуациях: неизвестного выражения для функции распределения, неизвестного выражения регрессионной функции r итп. Такая оценка будет обозначаться, например, θ, обозначение как с.в. соответствует ее сущности: на выборках имеется мера, определенная независимыми повторениями рассматриваемой случайной величины ξ и формула (применимая к выборкам из генеральной совокупности) задает случайную величину.

Напомним, что оценка $\theta = \theta_n$ (индекс поставлен, чтобы подчеркнуть роль длин выборок в формуле) параметра t

состоятельная если при увеличении размера n выборки она *сходится по вероятности* к истинному значению $\hat{t}$ параметра (то есть к числу);

несмещенная если $E(\theta_n)$ при любом n совпадает с истинным значением $\hat{t}$ параметра.

4.3.1 Совсем простые упражнения на свойства оценок

- Возможны ли несмещенные, но несостоятельные оценки?
- Величину $(E\xi)^2$ оценивают посредством $(\bar{x})^2$ – состоятельная ли это оценка в классе равномерно распределенных на $[0, \mathfrak{t}]$ случайных величин.
- Величину $(E\xi)^2$ оценивают посредством $(\bar{x})^2$ – состоятельная ли это оценка в классе корней квадратных из равномерно распределенных на $[0, \mathfrak{t}]$ случайных величин.
- Величину $(E\xi)^2$ оценивают посредством $(\bar{x})^2$ – смещенная ли это оценка в классе корней квадратных из равномерно распределенных на $[0, \mathfrak{t}]$ случайных величин.
- Величину $\left(E\xi^2\right)$ оценивают посредством $(\bar{x})^2$ – смещенная ли это оценка в классе корней квадратных из равномерно распределенных на $[0, \mathfrak{t}]$ случайных величин.

4.3.2 Показатели оценок

Для смещенной оценки возникает показатель $\mathrm{bias}(\theta) = E(\theta) - \hat{t}$, а также показатель $\mathrm{se} = \sqrt{D(\theta)}$, называющийся **стандартной ошибкой**. Дисперсия в этой формуле отвечает использованию истинной меры на выборках (то есть той, которая соответствует неизвестному значению параметра $\hat{t}$). Прямое вычисление стандартной ошибки

[14] лично я понятия не имею, как это в принципе возможно в достаточно общем случае множества Θ!

как правило невозможно: для этого нужна в явном виде функция распределения для θ, которая связана с плотностью меры на выборках длины n

$$f(u_1, u_2, \dots u_n, \hat{t}) = \prod_{i=1}^{n} f_\xi(u_i, \hat{t})$$

здесь $f_\xi(x, \hat{t})$ истинная плотность ξ.

Поэтому рассматривают **оценочную стандартную ошибку** — корень из оценочной дисперсии $\widetilde{\text{se}}$.

Пример Для выборки распределения Бернулли с параметром t имеется состоятельная и несмещенная оценка параметра $\theta = \bar{x}$. Стандартная ошибка $\text{se} = \sqrt{t(1-t)/n}$ выражается через истинное значение параметра t, а оценочная стандартная ошибка будет соответственно $\widetilde{\text{se}} = \sqrt{\theta(1-\theta)/n} = \sqrt{\bar{x}(1-\bar{x})/n}$.

Наряду со стандартной ошибкой рассматривают **среднеквадратичную ошибку** $\text{mse} = E\left((\theta - t)^2\right)$; как обычно математическое ожидание требует знания распределения $F_\xi(x, \hat{t})$ или плотности $f_\xi(x, \hat{t})$.

Предложение 4.1 $\text{mse}_n = \text{bias}^2(\theta_n) + D(\theta_n)$

Доказательство Обозначим $\bar{\theta}_n = E(\theta_n)$, тогда

$$\begin{aligned} E\left((\theta_n - t)^2\right) &= E\left((\theta_n - \bar{\theta}_n + \bar{\theta}_n - t)^2\right) && (6)\\ &= E\left((\theta_n - \bar{\theta}_n)^2\right) + 2(\bar{\theta}_n - t)E\left(\theta_n - \bar{\theta}_n\right) + E\left((\bar{\theta}_n - t)^2\right) && (7)\\ &= E\left((\theta_n - \bar{\theta}_n)^2\right) + (\bar{\theta}_n - t)^2 && (8)\\ &= D(\theta_n) + \text{bias}^2(\theta_n) && (9) \end{aligned}$$

Определение Оценка θ_n параметра t называется **асимптотически нормальной**, если случайная величина

$$\frac{\theta_n - t}{\text{se}}$$

при $n \to \infty$ сходится по распределению к $\mathcal{N}(0, 1)$

Например, в силу ЦПТ оценка параметра p в схеме Бернулли асимптотически нормальна, то же самое верно для оценки математического ожидания случайной величины с конечными первым и вторым моментами.

Две оценки одного и того же параметра сравнивают *по эффективности*: из двух оценок более эффективна та, у которой дисперсия меньше.

4.3.3 Упражнения

- Дана выборка $x_1, x_2, \dots x_n$ равномерной на $[0, t]$ случайной величины, здесь t – неизвестный параметр. Рассматривается оценка $\hat{\theta}_n$ этого параметра вида $\max(x_1, x_2, \dots x_n)$, проверить ее состоятельность и несмещенность. Найти $\text{bias}(\hat{\theta}_n)$ и se.
- Дана выборка $x_1, x_2, \dots x_n$ пуассоновской случайной величины с неизвестным параметром t. Рассматривается оценка $\hat{\theta}_n$ этого параметра вида $\hat{\theta}_n = \bar{x}$, проверить ее состоятельность и несмещенность. Найти $\text{bias}(\hat{\theta}_n)$ и se. Является ли она асимптотически нормальной?

4.4 Наивные соображения: метод моментов

Этот метод строит систему уравнений на неизвестные параметры, коэффициенты системы вычисляются довольно просто. Для $x_1, x_2, \dots x_n$ распределения, зависящего от неизвестного параметра $\mathfrak{t} = (t_1, t_2, \dots t_k) \in \mathbb{R}^k$, $k \ll n$ рассмотрим набор plugin-оценок $\widehat{\mu_j} = \frac{1}{n}\sum_i^n x_i^j$ для моментов $\mathfrak{m}_j(\xi)$ и связанную с ним систему из k уравнений с k неизвестными:

$$\mathfrak{m}_j(\xi) = E_\mathfrak{t}(\xi^j) = \int x^j dF_\mathfrak{t}(x) = G_j(t_1, t_2, \dots t_k)$$

$$\begin{cases} G_j(t_1, t_2, \dots t_k) &= \widehat{\mu_1} \\ & \dots \\ G_j(t_1, t_2, \dots t_k) &= \widehat{\mu_k} \end{cases}$$

Определение Оценкой $\widehat{\mathfrak{t}_n}$ параметра $\mathfrak{t}$ по методу моментов называется решение $\widehat{t_1}, \widehat{t_2}, \dots \widehat{t_k}$ этой системы.

Явное указание на длину n выборки часто не пишут.

- Объясните, почему оценка параметра $\widehat{\mathfrak{t}_n}$ есть случайный вектор, хотя сам параметр — это обычный вектор из $\mathbb{R}^k$.

Резонная критика такого построения оценочных формул: система уравнений вполне может быть нелинейной, ее решения могут быть не единственны, какую тогда брать формулу в общем случае — совершенно непонятно! Зато в «хороших» случаях[15], когда функции $G_j(t_1, t_2, \dots t_k)$ достаточно просты, единственное решение этой системы может быть найдено несложным вычислением.

Пример. Схема Бернулли Выборка состоит из нулей и единиц, если параметр p этого распределения неизвестен, то достаточно единственного уравнения $\widehat{p} = \bar{x}$

Пример. Гауссово распределение Если параметры a, σ распределения неизвестны, то воспользовавшись формулой $D(\xi) = E(\xi^2) - (E(\xi))^2$ для дисперсии, получим систему двух уравнений:

$$\begin{cases} \widehat{a} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i \\ (\widehat{a})^2 + (\widehat{\sigma})^2 = \frac{1}{n}\sum\limits_{i=1}^{n} x_i^2 \end{cases}$$

Решения, впрочем, не выглядят чем-то неожиданным: $\widehat{a} = \bar{x}$, $(\widehat{\sigma})^2 = S^2$ (ранее приведенное обозначение для оценки дисперсии)

В случае «хороших» распределений оценка $\mathfrak{t}$ по методу моментов оказывается состоятельной $\widehat{\mathfrak{t}_n} \xrightarrow{P} \mathfrak{t}$ и, более того, *асимптотически нормальной* то есть имеет место сходимость по распределению разности $\mathfrak{t}_n - \mathfrak{t}$ к гауссову распределению.

4.4.1 Упражнения

Записать системы уравнений метода моментов и их решения (через значения выборки $x_1, x_2, \dots x_n$) для

- равномерного на $[0, b]$ распределения с неизвестным параметром b с использованием второго момента.
- показательного распределения с неизвестным параметром λ с использованием второго момента;
- гамма-распределения с неизвестными параметрами r, b;
- равномерного на $[a, b]$ распределения с неизвестными параметрами $a < b$.

4.4.2 Состоятельность метода моментов

Предложение 4.2 *Пусть есть соотношение, связывающая параметр и момент $H(t) = E_t(\xi^k)$, причем функция $H : \mathbb{R} \to \mathbb{R}$ непрерывна и имеет обратную . Тогда соответствующая оценка параметра состоятельна.*

Доказательство В силу ЗБЧ Хинчина $\frac{1}{n}\sum_1^n x_i^k \xrightarrow{P} E_t(\xi^k) = H(t)$. По свойству непрерывности

$$H^{-1}\left(\frac{1}{n}\sum_1^n x_i^k\right) \xrightarrow{P} H^{-1}\left(E_t(\xi^k)\right) = H^{-1} \circ H\,(t) = t$$

∎

Что же касается свойства несмещенности, то оно потребовало бы при всех допустимых значениях параметра выполнения равенства

$$E_t\left[H^{-1}\Big(\frac{1}{n}\sum_1^n \xi_i^k\Big)\right] = t = H^{-1} \circ H(t) = H^{-1}\left[E_t\Big(\frac{1}{n}\sum_1^n \xi_i^k\Big)\right] = H^{-1}\left[\frac{1}{n}\sum_1^n E_t(\xi^k)\right]$$

Для непостоянных независимых случайных величин ξ_i их значения на исходах разные и при выпуклых или вогнутых функциях H такое требование войдет в противоречие с неравенством Йенсена. Таким образом, на несмещенность оценок по методу моментов можно надеяться лишь в случае линейной функции H^{-1}.

[15] Обычно речь идет о распределениях из «короткого» списка, стандартно изучаемого в курсе Теория Вероятностей