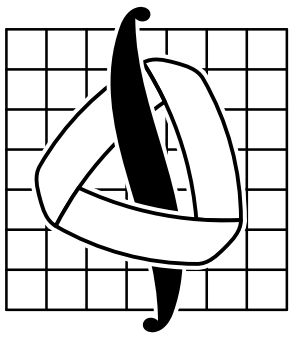
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА
Механико-математический факультет



Лекции по математической статистике

Лектор — Юрий Николаевич Тюрин

III курс, 5 семестр, поток математиков

Москва, 2004 г.

Предисловие

Данный документ представляет собой исправленную версию лекций по статистике, первоначально набранную автором курса. Огромная благодарность объявляется следующим людям: Евгению Гречникову, который исправил много ошибок и опечаток, а также провёл структурирование документа, а также Кириллу Никитину и Сергею Захарову.

В данной версии сделана еще одна серия исправлений, в основном типографского характера, а также устранены ошибки, привнесённые предыдущей редакцией.

Последнее обновление: 9 февраля 2006 г.

Оглавление

1. Введение

1.1. Статистическая модель

1.1.1. Простейшая модель: выборка

К общей (абстрактной) статистической модели мы придем, рассмотрев несколько примеров. Заодно укажем, каким образом статистический материал можно представить наглядно.

Пример 1.

Этот пример я заимствовал из старой книги А. Хальда, которую далее цитирую.

Хальд приводит результаты измерений 200 головок заклепок. Эти числа записаны в том порядке, в котором они поступали.

Первичное представление данных — таблица:

Исходные данные

Диаметры 200 головок заклепок, *мм*							
13,39	13,43	13,54	13,64	13,40	13,55	13,40	13,26
13,42	13,50	13,32	13,31	13,28	13,52	13,46	13,63
13,38	13,44	13,52	13,53	13,37	13,33	13,24	13,13
13,53	13,53	13,39	13,57	13,51	13,34	13,39	13,47
13,51	13,48	13,62	13,58	13,57	13,33	13,51	13,40
13,30	13,48	13,40	13,57	13,51	13,40	13,52	14,56
13,40	13,34	13,23	13,37	13,48	13,48	13,62	13,35
13,40	13,36	13,45	13,48	13,29	13,58	13,44	13,56
13,28	13,59	13,47	13,46	13,62	13,54	13,20	13,38
13,43	13,35	13,56	13,51	13,47	13,40	13,29	13,20
13,46	13,44	13,42	13,29	13,41	13,39	13,50	13,48
13,53	13,34	13,45	13,42	13,29	13,38	13,45	13,50
13,55	13,33	13,32	13,69	13,46	13,32	13,32	13,48
13,29	13,25	13,44	13,60	13,43	13,51	13,43	13,38
13,24	13,28	13,58	13,31	13,31	13,45	13,43	13,44
13,34	13,49	13,50	13,38	13,48	13,43	13,37	13,29
13,54	13,33	13,36	13,46	13,23	13,44	13,38	13,27
13,66	13,26	13,40	13,52	13,59	13,48	13,46	13,40
13,43	13,26	13,50	13,38	13,43	13,34	13,41	13,24
13,42	13,55	13,37	13,41	13,38	13,14	13,42	13,52
13,38	13,54	13,30	13,18	13,32	13,46	13,39	13,35
13,34	13,37	13,50	13,61	13,42	13,32	13,35	13,40
13,57	13,31	13,40	13,36	13,28	13,58	13,58	13,38
13,26	13,37	13,28	13,39	13,32	13,20	13,43	13,34
13,33	13,33	13,31	13,45	13,39	13,45	13,41	13,45

Таблица 1.

Для получения более ясного представления о данных, результаты измерений располагаются в соответствии с их величиной следующим образом: на бумагу, разграфленную в клетку (обычно — на миллиметровую бумагу), наносится горизонтальная прямая и на ней специальным образом размечается шкала. Результаты наблюдений отмечаются тогда точками над соответствующими числами.

Правильность заполнения точечной диаграммы может быть проверена посредством суммарного подсчета общего числа наблюдений, произведенного как по таблице исходных данных, так и по точечной диаграмме.

Для того, чтобы иметь возможность проследить появление возможных ошибок, перечисление и суммирование должно выполняться по группам, в каждой из которых содержится не более 100 наблюдений.

Контроль суммированием, разумеется, не является вполне надежным, так как при этом способе контроля противоположные по знаку ошибки могут скомпенсировать друг друга.

Результаты наблюдений могут перечисляться также при помощи *карт*, причем результат каждого из наблюдений наносится на карту, а карты сортируются по величине указанных в них результатов.

Рисунок 1. дает более ясное представление данных, чем первоначальный список из 200 чисел. Представление можно сделать еще более наглядным при помощи группировки наблюдений и построения т.н. *гистограмм*. Обратите внимание, как изменение интервалов группировки отражается на форме гистограмм. На последующих

рисунках можно видеть влияние изменения длины интервала группировки на внешний вид гистограммы. Частоты и другие величины, связанные с распределением, используются в качестве ординат после их деления на длину соответствующих интервалов группировки; поэтому единица ординаты обратно пропорциональна длине интервала группировки, а 1см^2 представляет на всех фигурах одно и то же число наблюдений.

Если длина интервала группировки мала, то влияние случайных колебаний начинает преобладать, так как каждый интервал содержит при этом лишь небольшое число наблюдений; если же длина интервала велика, то скрадываются основные характерные черты распределения.

Гистограммы и точечная диаграмма показывают, что данные из *таблицы 1* ведут себя как совокупность реализаций некой случайной величины.

В статистике совокупность независимых одинаково распределенных случайных величин часто называют одним словом: *выборка*. (Слово *выборка* имеет в статистике и другое, буквальное значение). Так что данные из *таблицы 1* похожи на выборку. Если разобраться в деле получше (например, с помощью выборочной функции распределения и нормальной вероятностной бумаги), то можно убедиться, что эти двести чисел можно считать выборкой из нормального распределения (выборкой из нормальной совокупности).

Итак, статистическая модель для 200 чисел этого примера — это выборка из нормального распределения. Параметры этого нормального распределения при этом не уточняются; они остаются неопределенными.

Можно указать и некоторые задачи, естественные для этого примера:

- оценить неизвестные параметры упомянутого нормального распределения;
- указать пределы, в которые укладывается предписанная доля изделий;
- проверить высказанное утверждение (предположение), что данная выборка извлечена из нормальной совокупности;
- и т.д.

1.1.2. Простая линейная регрессия

Рассмотрим данные из статьи Э. Хаббла (*E. Hubble*) 1929 года, где впервые была подтверждена мысль о расширении вселенной (о «разбегании галактик»).

Эти данные связывают расстояние от Земли до ближайших туманностей с лучевыми скоростями этих туманностей.

Рисунок 6. Данные из статьи E. Hubble 1929 года, связывающие удаления и лучевые скорости 24 туманностей.

Рисунок наводит нас на мысль (так же, как и Э. Хаббла семьдесят лет назад), что лучевые скорости «в целом» пропорциональны удалениям:

$$y_i = \theta x_i + \varepsilon_i, \qquad i = \overline{1, 24}.$$

Здесь:

- x_i — удаление i-й туманности;
- y_i — ее лучевая скорость;
- ε_i — отступление от линейной зависимости. Эти отступления, возможно, объясняются собственными движениями туманностей в пространстве, а также ошибками в измерении скоростей и удалений;
- коэффициент θ, определяющий скорость расширения пространства, сейчас называют *постоянной Хаббла*.

Величины удалений и скоростей для тех туманностей, которые отражены на *рисунке 6*, впоследствии были пересмотрены и уточнены, поэтому оценка для θ сильно изменилась по сравнению с той, которую нашел сам Хаббл.

Были также измерены удаления и скорости для многих других туманностей, находящихся на гораздо больших расстояниях от Земли, чем первые двадцать четыре, о которых написал Хаббл. Линейный характер зависимости, тем не менее, сохранился, был убедительно подтвержден и, в настоящее время, выражен из основных законов астрономии. Впрочем, численное значение θ все еще вопрос дискуссионный (и чрезвычайно важный для теорий возникновения и эволюции вселенной).

Основное статистическое предположение о $\varepsilon_1, \ldots, \varepsilon_n$: это реализация *независимых случайных величин*. Именно предположение о *случайности* отклонений от определенной закономерности позволяет называть обсуждаемую модель явления *статистической*. Дальнейшее накопление данных и/или более глубокий их анализ, а также опыт других задач обычно позволяет сказать больше о свойствах случайных ошибок $\varepsilon_1, \ldots, \varepsilon_n$:

(a) $\mathsf{E}\varepsilon_i = 0$, или $P(\varepsilon_i > 0) = P(\varepsilon_i < 0) = \frac{1}{2}$;

(Эти предположения выражают мысль об отсутствии в ошибках систематической составляющей.)

(b) Случайные величины $\varepsilon_1, \ldots, \varepsilon_n$ одинаково распределены;

(c) Случайные величины $\varepsilon_1, \ldots, \varepsilon_n$ распределены по (общему) нормальному закону;

(d) и т.д.

1.1.3. Общая (абстрактная) статистическая модель

Имеется наблюдение X. Это наш *статистический материал*. Все выводы мы будем делать, основываясь на наблюдении X. Его математическая природа несущественна: X может быть совокупностью чисел, вектором, матрицей, функцией времени (например, кривой, записанной самописцем) или пространства и.т.д.

Мы рассматриваем X как точку некоего множества $\mathcal{X}$, называемого *пространством наблюдений*, *выборочным пространством*, *генеральной совокупностью* и т.д.

Выборочное пространство $\mathcal{X}$ мы *примысливаем* к нашему реальному наблюдению X, собирая вместе все значения, которые, по нашему мнению, могли появиться вместо конкретного X.

Мы предполагаем, что данное значение X появилось как результат *случайного выбора* элемента из $\mathcal{X}$. Этот случайный выбор был произведен в соответствии с некоторым распределением вероятностей P на $\mathcal{X}$. Как правило, это конкретное распределение P нам не известно. Однако мы можем указать какие-то свойства, которыми P обладает. Иначе говоря, нам известно (мы можем указать) некоторое множество $\mathcal{P}$ вероятностных распределений на $\mathcal{X}$, которому принадлежит неизвестное истинное распределение P.

Задача статистики — выводы о P на основании X. Например, основываясь на X, вычислить приближенные значения функционалов от P или ответить, совместимы ли с наблюденным X предположения о тех или иных свойствах P.

Множество $\mathcal{P}$ в практических задачах часто оказывается параметризованным с помощью некоторого параметра θ, который меняется в заданной области Θ. Обычно Θ — интервал числовой прямой (когда θ — одномерный параметр) или область конечномерного пространства (когда θ — многомерный параметр).

В параметрическом случае:

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}.$$

В этом случае нас обычно интересует значение θ, отвечающее истинному распределению P_θ (истинное значение θ) либо значение тех или иных функций $\tau(\theta)$ при истинном θ. Основываясь на X, мы должны найти для них приближенные значения.

1.2. Теорема Гливенко

(Пример того, как по выборке устанавливаются свойства распределения вероятностей.)

Пусть $x_1, \ldots, x_n$ — независимые одинаково распределенные случайные величины. Их (общую) функцию распределения обозначим через $F(x)$:

$$F(x) = P(x_i \leqslant x)$$

Обозначим через $F_n(X)$ так называемую *эмпирическую* функцию распределения, которая строится по выборке. Для этого в каждую из точек $x_1, \ldots, x_n$ поместим вероятность, равную $\frac{1}{n}$. На числовой прямой возникнет новое распределение вероятностей. Его функцию распределения и обозначим через $F_n(X)$. $F_n(X)$ называют *функцией распределения выборки*. С помощью индикаторов событий $I(x_i \leqslant x)$ функцию $F_n(X)$ можно записать в виде:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leqslant x)$$

Замечание: Часто функцию распределения определяют чуть иначе, чем сказано выше, посредством строгих неравенств:

$$F(x) = P(x_i < x).$$

В этом случае аналогично изменяется и определение функции распределения выборки. Различие между этими двумя определениями несущественны:

- для непрерывных распределений они совпадают;
- для других различие состоит лишь в том, с какой стороны (слева или справа) функция распределения оказывается непрерывной.

Следующая ниже формулировка теоремы Гливенко не зависит от того, какой вариант определения мы принимаем.

Теорема 1.1 (Гливенко). *Последовательность случайных величин ($n = 1, 2, \ldots$)*

$$D_n = \sup_x |F_n(x) - F(x)|$$

сходится к нулю по вероятности при $n \to \infty$.

Другими словами: для любых $\varepsilon > 0$, $\delta > 0$ найдется номер $N = N(\varepsilon, \delta)$ такой, что для всех $n \geqslant N$

$$P\{\sup_x |F_n(x) - F(x)| < \varepsilon\} > 1 - \delta.$$

□ Предварительное замечание: для всякого x

$$F_n(x) \longrightarrow F(x), \qquad n \to \infty$$

Это всего лишь переформулировка теоремы Бернулли (о сходимости частоты события к его вероятности в последовательности независимых испытаний) для события $\{x_i \leqslant x\}$.

Сначала доказательство проведем для непрерывной функции $F(\cdot)$. С небольшими изменениями это окажется справедливым и для разрывных функций распределения, о чем будет сказано ниже.

1° Пусть R — натуральное число. Его выбор уточним позже. Разобьем отрезок $[0, 1]$ оси ординат на R равных частей. Одновременно, на R отрезков $\Delta_1, \ldots, \Delta_R$ будет разделена и ось абсцисс точками

$$-\infty = a_0 < a_1 < \ldots < a_R = +\infty,$$

где $\Delta_k = [a_{k-1}, a_k]$, $F(a_k) = k/R$, $k = 0, 1, \ldots, R$.

Пусть $\varepsilon' > 0$, выбор ε' уточним позже. Рассмотрим событие

$$\Omega_n = \left\{ \max_{1 \leqslant k \leqslant R-1} |F_n(a_k) - F(a_k)| < \varepsilon' \right\}$$

По теореме Бернулли существует $N = N(\varepsilon', \delta)$ такое, что для всех $n \geqslant N$

$$P(\Omega_n) > 1 - \delta.$$

(Другими словами: следствием сходимости в каждой точке является равномерная сходимость на каждом конечном множестве точек.)

2° Теперь покажем, что, если произошло событие Ω_n, то (при правильном выборе ε' и R)

$$\sup_{-\infty < x < \infty} |F_n(x) - F(x)| < \varepsilon.$$

Ясно, что

$$\sup_{-\infty < x < \infty} |F_n(x) - F(x)| = \max_{k=1,\ldots,R} \sup_{x \in \Delta_k} |F_n(x) - F(x)|$$

Поэтому достаточно показать, что если произошло событие Ω_n, то для каждого $k = \overline{1, R}$

$$\sup_{x \in \Delta_k} |F_n(x) - F(x)| < \varepsilon. \qquad (*)$$

Поскольку для любой функции $f(\cdot)$

$$\sup |f(x)| = \max \left[\sup f(x), \sup(-f(x)) \right],$$

для доказательства $(*)$ достаточно оценить сверху порознь

$$\sup_{x \in \Delta_k} [F_n(x) - F(x)] \text{ и } \sup_{x \in \Delta_k} [F(x) - F_n(x)]$$

Оценим только первое из двух выражений, поскольку вторая оценка получается аналогично.

3° В силу того, что функции распределения $F(\cdot)$ и $F_n(\cdot)$ монотонно неубывают, при $x \in \Delta_k = [a_{k-1}, a_k]$:

$$F_n(x) - F(x) \leqslant F_n(a_k) - F(a_{k-1}) = [F_n(a_k) - F(a_k)] + [F(a_k) - F(a_{k-1})] = [F_n(a_k) - F(a_k)] + \frac{1}{R}$$

Если произошло событие Ω_n, то цепочку можно продолжить и написать:

$$F_n(x) - F(x) \leqslant \varepsilon' + \frac{1}{R}$$

Причем это верно для каждого отрезка Δ_k.

Если R и ε' выбрать так, что $\frac{1}{R} + \varepsilon' < \varepsilon$, то получим, что (при $n \geqslant N(\varepsilon', \delta)$)

$$\Omega_n \subset \{\sup_x |F_n(x) - F(x)| < \varepsilon\}.$$

Для непрерывных $F(\cdot)$ доказательство окончено, поскольку $P(\Omega_n) > 1 - \delta$ для всех достаточно больших n.

Для функций с разрывами то же доказательство проходит с некоторыми изменениями.

4° Взамен последовательности $(a_0, a_1, \ldots, a_R)$ рассмотрим последовательность

$$-\infty = b_0 < b_1 < \ldots < b_K = +\infty$$

такую, что приращение $F(\cdot)$ на каждом интервале (b_{k-1}, b_k), $k = \overline{1, K}$, не превосходит $\varepsilon/2$:

$$|F(b_k - 0) - F(b_{k-1} + 0)| \leqslant \frac{\varepsilon}{2}.$$

(Пишем пределы слева и пределы справа вместо того, чтобы в одном случае написать значение функции в точке, с тем, чтобы выкладка годилась для обоих определений функции распределения: для $P(x_i \leqslant x)$ и для $P(x_i < x)$.)

Как можно построить такую последовательность, показано на рисунке.

В частности, в последовательность $(b_0, b_1, \ldots, b_K)$ войдут все точки скачков функции F, которые превосходят $\varepsilon/2$ (их конечное число).

5° Событие Ω_n, которое ранее было связано с последовательностью $a_0, a_1, \ldots, a_R$, теперь определим так:

$$\Omega_n = \left\{ \max_{1 \leqslant k \leqslant K-1} \left[|F_n(b_k + 0) - F(b_k + 0)|, |F_n(b_k - 0) - F(b_k - 0)| \right] < \frac{\varepsilon}{2} \right\}.$$

По теореме Бернулли (как и раньше), для достаточно больших n

$$P(\Omega_n) > 1 - \delta$$

С этим изменением доказательство проходит также, как и раньше. ■

Мы доказали, что F_n равномерно сходится к F по вероятности. Более сильная форма этой теоремы (которая и была доказана ее авторами: Гливенко — для непрерывного случая, Кантелли — для общего) утверждает сходимость с вероятностью 1.

Соотношение между этими двумя теоремами о сходимости F_n к F такое же, как между просто законом больших чисел и усиленным законом больших чисел. (Теорема Гливенко – Кантелли и есть закон больших чисел в функциональном пространстве).

Впрочем, для практики, имеющей дело с конечными выборками, сходимость с вероятностью 1 дает не больше, чем сходимость по вероятности:

- Если $\xi_n \longrightarrow \xi$ (почти наверно или по вероятности), то для данной нам выборки (для данного n) это означает лишь, что ξ_n приближенно равна ξ (если, к тому же, «n достаточно велико»).

Поэтому мы будем рассматривать только «слабые» предельные теоремы, утверждающие сходимость по вероятности, даже если известны их усиленные варианты.

2. Статистические оценки

2.1. Абстрактная статистическая модель, решающие правила

Имеется наблюдение X (так мы обозначаем имеющийся статистический материал. Его математическая природа не важна: это может быть набор чисел; числовая последовательность; запись, сделанная самописцем, и т.п.), К имеющемуся наблюдению X мы примысливаем множество $\mathcal{X}$, $X \in \mathcal{X}$, называемое *выборочным пространством*. Выборочное пространство — это совокупность таких исходов, которые могли бы появиться в нашем опыте вместо X. Мы предполагаем, что элемент X был выбран из множества $\mathcal{X}$ 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 $\mathcal{P}$ 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 (но не всех!) практически важных случаях множество $\mathcal{P}$ имеет естественную параметризацию, так что $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, где заданное параметрическое множество Θ принадлежит конечномерному (арифметическому) пространству.

Статистические задачи часто представляют в параметрической форме. В этом случае нас интересуют выводы о значении θ.

2.2. Постановка задачи

В этой главе мы будем обсуждать задачу оценивания параметра θ и/или функций от θ. «Оценить» здесь означает «указать приближенное значение, опираясь на наблюдение X». Надо найти правило $\delta(\cdot)$, по которому каждое возможное наблюдение $X \in \mathcal{X}$ пересчитывается в значение $\delta(X)$, которое далее выступает как приближенное значение неизвестного параметра θ : $\delta(X) \approx \theta$. [Либо как приближенное значение для $\tau(\theta)$, если нас интересует не сам параметр θ, а некоторая функция от него. В этом случае функция $\tau(\cdot)$ должна быть задана.] Задача статистики: выбрать правило $\delta(\cdot)$ так, чтобы оценить θ как можно лучше (точнее).

Можно предложить очень много способов, измеряющих близость $\delta(X)$ и θ. Общая точка зрения: есть функция потерь $L(\theta, d) \geqslant 0$, принимающая определенное числовое значение, когда в качестве оценки истинного θ выступает величина d. В случае наблюдения X и правила оценивания $\delta(\cdot)$ величина потерь составляет $L(\theta, \delta(X))$. Например, может быть

$$L(\theta, \delta(X)) = |\theta - \delta(X)| \text{ или } L(\theta, \delta(X)) = |\theta - \delta(X)|^2.$$

В каждом отдельном опыте величина потерь случайна. В статистике принято характеризовать статистические правила средними результатами, достигаемыми при многократном применении.

По закону больших чисел это:

$$\mathsf{E}_\theta L(\theta, \delta(X))$$

(Разъяснение обозначений: так как мы должны держать в уме все возможные значения параметра $\theta \in \Theta$, нам следует указывать, по какой именно мере мы производим усреднение, т. е. вычисляем математическое ожидание. Индекс θ около символа усреднения E или вероятности P явно указывает на это) Таким образом, точность (а, скорее, неточность) правила δ описывает теперь *функция риска*

$$R(\theta, \delta) := \mathsf{E}_\theta L(\theta, \delta)$$

Ясно, что правило $\delta_1(\cdot)$ лучше, чем правило $\delta_2(\cdot)$, если

$$R(\theta, \delta_1) \leqslant R(\theta, \delta_2) \qquad (*)$$

при всех $\theta \in \Theta$ (а для некоторых значений θ это соотношение есть строгое неравенство). Наилучшим следует назвать такое правило $\delta(\cdot)$, которое превосходит любое другое правило.

К сожалению, наилучшего в этом смысле правила обычно не существует, ибо здесь речь идет о сравнении функций. В множестве функций от θ вида $R(\theta, \delta)$ (где $\delta(\cdot)$ — функция от наблюдений) обычно нет минимального элемента. (Хотя бы потому, что правило $\delta(X) = \theta_0$, где θ_0 — фиксированное значение, нельзя улучшить в точке $\theta = \theta_0$. Хотя при других θ это правило никуда не годится.)

Для преодоления этого затруднения есть две главные возможности. Первая — это изучение *допустимых* правил.

Определение: Правило $\delta_1(\cdot)$ называют *допустимым*, если не существует $\delta_2(\cdot)$, для которого выполняется $(*)$.

Допустимые правила, по существу, совпадают с так называемыми *байесовскими правилами*.

Определение: *Байесовские правила* — это оптимальные правила в ситуации, когда неизвестный параметр θ получен путем случайного выбора.

В этом случае риск $R(\theta, \delta)$ естественно усреднить еще и по θ по той (вероятностной) мере, которая управляла выбором θ. Риск правила $\delta(\cdot)$ после этого превращается в число. Поэтому задача о минимуме имеет решение.

Взгляд на θ как на случайную величину называют *байесовским подходом* к статистике. Он имеет как горячих сторонников, так и противников. Мы не будем касаться его в этом курсе.

Другая возможность — продолжение поиска *оптимальных* (т.е. равномерно наилучших правил) $\delta(\cdot)$, но в более узком множестве возможностей. Сужение поля выбора достигаются путем наложения на оценку $\delta(\cdot)$ каких-либо дополнительных (и естественных) требований. Наиболее важные результаты получены для *несмещённых* правил.

Определение: Оценка $\delta(\cdot)$ параметра θ (либо функции $\tau(\theta)$) называется *несмещенной*, если $\mathsf{E}_\theta \delta(X) = \theta$ (либо $\mathsf{E}_\theta \delta(X) = \tau(\theta)$) для всех $\theta \in \Theta$.

Для важной с прикладной точки зрения линейной статистической модели удается найти наилучшие несмещённые оценки, если выбрать квадратичную функцию потерь $L(\theta, d) = |\theta - d|^2$ (или даже функцию потерь с матричными значениями $L(\theta, d) = (\theta - d)(\theta - d)^T$ — считая θ и d векторами-столбцами). В следующей главе линейная модель будет изучена нами подробно.

Из этого короткого рассказа видно, насколько неопределенным и зависящим от нашего произвола является путь к оптимальным статистическим решениям. На его выбор влияют не только логические соображения (они недостаточны), но и (в основном) конечный результат: удается ли получить в его конце явные и разумные статистические правила.

Для несмещенных оценок и квадратичной функции потерь функция риска оценки $\delta(\cdot)$ превращается в дисперсию (в векторном случае — в матрицу ковариаций): $R(\theta, \delta) = \mathsf{E}_\theta(\delta(X) - \theta)^2 = \mathsf{E}_\theta(\delta(X) - \mathsf{E}_\theta\delta(X))^2$.Задача теперь выглядит очень естественно: надо найти несмещенную оценку с наименьшей дисперсией. Однако подробнее этой задачей мы займемся несколько позже.

А сейчас приведем важные для теории неравенства, которые в так называемом «регулярном случае» ограничивают снизу дисперсию (для многомерного параметра — матрицу ковариаций) каждой оценки. Для несмещенных оценок именно дисперсия (матрица ковариаций) служит естественной мерой точности оценивания. Поэтому обсуждаемые неравенства показывают, что для точности оценивания есть граница снизу. Эта граница зависит от структуры статистической модели (и ее параметризации).

2.3. Неравенство Крамера – Рао для одномерного параметра

Это неравенство еще называется неравенством информации или неравенством Фреше.

Так называют неравенство для дисперсии статистических оценок одномерного параметра, которое можно вывести при многочисленных условиях гладкости, налагаемых на зависимость вероятностного распределения от меняющегося параметра. Такой тип зависимости от параметра, который ниже будет описан подробнее, часто называют *регулярным*. Впрочем, содержание этого термина будет меняться от задачи к задаче.

Пусть X — наблюдение (конечномерный вектор), распределение которого зависит от неизвестного параметра θ, причем $\theta \in \Theta \subset \mathbb{R}^1$, где θ — заданное открытое множество.

Отдельно будем рассматривать две возможности:

(a) X имеет плотность $p(x, \theta)$ (относительно меры Лебега)

(b) наблюдение X имеет дискретное распределение; в этом случае $p(x, \theta)$ означает вероятность события X; $p(x, \theta) > 0$ только для счетного множества значений x.

Выкладки в обоих случаях идут одинаково — с той разницей, что в случае (a) для математических ожиданий мы пишем интегралы, а случае (b) — суммы (ряды). Поэтому достаточно разобрать в подробностях какую-либо одну из этих двух возможностей, скажем, (a).

Пусть $T(X)$ — некоторая статистика, принимающая значения в $\mathbb{R}^1$, для которой существует математическое ожидание и дисперсия.

Пусть

$$\tau(\theta) := \mathsf{E}_\theta T(X)$$

Предположения о плотности $p(x, \theta)$

(взятые вместе они и составляют условия регулярности).

(a) Множество $A = \{x : p(x, \theta) > 0\}$ не зависит от θ (это наиболее важное условие).

(b) При всех $x \in A$, $\theta \in \Theta$ существует

$$\lambda(x,\theta) := \frac{\partial}{\partial\theta}\ln p(x,\theta)$$

(c) (Возможность дифференцирования под знаком интеграла)

$$\frac{\partial}{\partial\theta}\int\limits_A p(x,\theta)dx = \int\limits_A \frac{\partial}{\partial\theta}p(x,\theta)dx (= 0),$$

$$\frac{\partial}{\partial\theta}\int\limits_A T(x)p(x,\theta)dx = \int\limits_A T(x)\frac{\partial}{\partial\theta}p(x,\theta)dx (= \tau'(\theta)).$$

Введем важное понятие информации по Фишеру, точнее, количества информации о параметре θ, содержащейся в наблюдении X:

$$I(\theta) := \mathsf{E}_\theta\left[\frac{\partial}{\partial\theta}\ln p(X,\theta)\right]^2 = \mathsf{E}_\theta\lambda^2(X,\theta)$$

Последнее из условий регулярности:

(d)

$$0 < I(\theta) < \infty$$

Теорема 2.1 (неравенство Крамера – Рао). *В перечисленных условиях (a)-(d)*

$$\mathsf{D}_\theta T(X) \geqslant \frac{[\tau'(\theta)]^2}{I(\theta)} \qquad (*)$$

Для несмещенных оценок параметра θ, когда $\tau(\theta) = \theta$, из этого неравенства следует, что

$$\mathsf{D}_\theta T(X) \geqslant \frac{1}{I(\theta)}$$

□

1° Заметим, что $\mathsf{E}_\theta\lambda(X,\theta) = 0$. Действительно, из (c) мы заключаем, что:

$$0 = \int\limits_A \frac{\partial}{\partial\theta}p(x,\theta)dx = \int\limits_A \left[\frac{\partial}{\partial\theta}\ln p(x,\theta)\right]p(x,\theta)dx = \mathsf{E}_\theta\lambda(X,\theta).$$

2° Аналогично, из второго равенства (c) мы получаем, что

$$\tau'(\theta) = \int\limits_A T(x)\left[\frac{\partial}{\partial\theta}\ln p(x,\theta)\right]p(x,\theta)dx = \mathsf{E}_\theta T(X)\lambda(X,\theta) = \mathsf{E}_\theta[T(X)-\tau(\theta)]\lambda(X,\theta).$$

Последнее равенство — благодаря тому, что $\mathsf{E}_\theta\lambda(X,\theta) = 0$.

3° Неравенство Коши – Буняковского:

$$(\mathsf{E}\xi\eta)^2 \leqslant \mathsf{E}\xi^2\mathsf{E}\eta^2$$

применим к полученному в 2) равенству, полагая $\xi = T(X) - \tau(\theta)$, $\eta = \lambda(X,\theta)$. Получим, что:

$$[\tau'(\theta)]^2 \leqslant I(\theta)\mathsf{D}_\theta T(X).$$

Отсюда и следует указанное в теореме неравенство. ■

Замечание 1. Пусть $X = (X_1,\ldots,X_n)$ — выборка. Можно говорить о количестве информации, заключённой в выборке X — пусть это $I_X(\theta)$, и о количестве информации, содержащейся в отдельных наблюдениях (элементах выборки) пусть это $i(\theta)$.

В этих условиях

$$I_X(\theta) = ni(\theta)$$

Доказательство:

Совместная плотность $X = (X_1, \ldots, X_n)$ равна $\prod_{i=1}^{n} f(X_i, \theta)$, где через $f(\cdot, \theta)$ обозначена плотность вероятностей отдельных X_i.

Отсюда:
$$\lambda(X, \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(X_i, \theta);$$

$$\mathsf{D}_\theta \lambda(X, \theta) = \sum_{i=1}^{n} \mathsf{E} \left[\frac{\partial}{\partial \theta} \ln f(X_i, \theta) \right]^2 = ni(\theta). \qquad \square$$

Из сказанного можно вывести важное качественное следствие о возможной скорости уменьшения дисперсии несмещенной оценки при возрастании числа независимых наблюдений n:

$$\mathsf{D}_\theta T(X) \geqslant C/n, \qquad \text{где } C = [i(\theta)]^{-1}.$$

Замечание 2.

$$I(\theta) = -\mathsf{E}_\theta \frac{\partial^2}{\partial \theta^2} \ln p(X, \theta).$$

2.4. Экспоненциальные семейства

Случай, когда неравенство Крамера – Рао $(*)$ выполняется в виде равенства, заслуживает особого рассмотрения. При выводе $(*)$ мы применили неравенство Коши:

$$(\mathsf{E}\xi\eta)^2 \leqslant \mathsf{E}\xi^2 \mathsf{E}\eta^2 \tag{2.4.1}$$

в котором равенство достигается т. и т.т., когда между случайными величинами ξ и η существует линейная связь. Иначе говоря, когда существуют такие постоянные (такие числа) A, B, C, что с вероятностью 1 выполняется равенство

$$A\xi + B\eta + C = 0 \tag{2.4.2}$$

В нашем случае $\xi = \lambda(X, \theta)$, $\eta = T(X) - \tau(\theta)$. Для них приведенное выше равенство превращается в

$$T(X) = \tau(\theta) + a(\theta)\lambda(X, \theta) \tag{2.4.3}$$

где $a(\theta)$ — некоторая функция θ. Постоянная $C = 0$, т.к. здесь математические ожидания ξ и η равны нулю.

Оценка $T(X)$, для которой в $(*)$ (или, что эквивалентно, в (2.4.3)) имеет место равенство (при всех $\theta \in \Theta$), называется *эффективной*. Существуют эффективные оценки лишь для особых параметрических семейств распределений и лишь для некоторых функций τ.

Вид этих параметрических семейств мы сейчас установим. Исходим из равенства (2.4.3). Это равенство для плотности (вероятности) $p(x, \theta)$ дает уравнение

$$\frac{\partial}{\partial \theta} \ln p(x, \theta) = \frac{1}{a(\theta)} T(X) + \frac{\tau(\theta)}{a(\theta)}$$

для всех $x \in A$ (см. условия регулярности) и всех $\theta \in \Theta$. Интегрируя, для $p(x, \theta)$ получаем выражение:

$$p(x, \theta) = \exp\{(c(\theta)T(x) + d(\theta) + S(x))\} I_A(x) \tag{2.4.4}$$

Здесь $c(\theta), d(\theta), S(X)$ — некоторые функции, зависящие только от указанных аргументов, $I_A(x)$ — индикаторная функция множества A. (Заметим, что представление плотности в виде (2.4.4) не единственно).

Семейство распределений, плотности (вероятности) которого имеют вид (2.4.4), называют *экспоненциальным семейством*. Для экспоненциального семейства эффективная оценка существует для функции $\tau(\theta) = -\frac{d'(\theta)}{c'(\theta)}$.

Распределение выборки из экспоненциального семейства, т.е. распределение совокупности n независимых реализаций $(X_1, \ldots, X_n)$ случайной величины, принадлежащей экспоненциальному семейству (2.4.4), очевидным образом, в свою очередь, принадлежит экспоненциальному семейству с плотностью (вероятностью):

$$p(x_1, \ldots, x_n, \theta) = \exp\left[c(\theta) \sum_{i=1}^{n} T(x_i) + nd(\theta) + \sum_{i=1}^{n} S(x_i) \right] I_{A \times \ldots \times A}(x_1, \ldots, x_n).$$

Многие практически важные параметрические распределения входят в этот класс. Например:

- **Биномиальное распределение**, где:

$$p(x,\theta) = \mathbf{C}_n^x \theta^x (1-\theta)^{n-x} = \exp\left(x \ln \frac{\theta}{1-\theta} + n \ln(1-\theta) + \ln \mathbf{C}_n^x\right)$$

для $x = 0, 1, \ldots, n; 0 < \theta < 1$. Есть эффективная оценка X/n для параметра θ.

- **Показательное распределение** с параметром $\theta > 0$, где

$$p(x,\theta) = \begin{cases} \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) & \text{для } x \geqslant 0 \\ 0 & \text{для } x \leqslant 0 \end{cases}$$

Для выборки

$$p(x_1, \ldots, x_n, \theta) = \left(\frac{1}{\theta}\right)^n \exp\left(-\frac{1}{\theta}\sum_{i=1}^n x_i\right), x \geqslant 0, \theta > 0.$$

Для параметра θ есть эффективная оценка $\sum\limits_{i=1}^n X_i/n$.

В заключение отметим, что эффективная оценка может быть только одна (то есть только для одной функции $\tau(\theta)$ и ее линейных комбинаций). Чтобы доказать это, допустим противоположное: для некоторого параметрического семейства есть два равенства вида (2.4.3):

$$\begin{cases} T_1 = \tau_1(\theta) + a_1(\theta)\lambda(X,\theta); \\ T_2 = \tau_2(\theta) + a_2(\theta)\lambda(X,\theta). \end{cases} \tag{2.4.5}$$

Умножив второе равенство на $\frac{a_1(\theta)}{a_2(\theta)}$ и вычтя результат из первого, получим, что:

$$T_1(X) = \tau_1(\theta) - \frac{a_1(\theta)}{a_2(\theta)}\tau_2(\theta) + \frac{a_1(\theta)}{a_2(\theta)}T_2(X) \tag{2.4.6}$$

Равенство (2.4.6) возможно, только если:

$$\frac{a_1(\theta)}{a_2(\theta)} = \text{const}, \tau_1(\theta) - \frac{a_1(\theta)}{a_2(\theta)}\tau_2(\theta) = \text{const} \tag{2.4.7}$$

Действительно, $T_1(X) - \frac{a_1(\theta)}{a_2(\theta)}T_2(X)$ не должно изменяться, когда изменяется $X, X \in A$. Это возможно, только если $a_1(\theta)/a_2(\theta)$ не изменяется, когда изменяется $\theta \in \Theta$.

Из (3.6) следует, что все эффективные оценки линейно выражаются одна через другую (см. (3.5)), как и соответствующие функции $\tau(\theta)$.

2.5. Статистические оценки для многомерных параметров

2.5.1. Случайные векторы, их средние и дисперсии

Пусть X — случайный объект (случайная величина, случайный вектор и т.п.), распределение которого определено параметром θ.

Предположим, что θ — r-мерный параметр, который мы будем представлять в виде столбца: $\theta = (\theta_1, \ldots, \theta_r)^T$, $\theta \in \Theta \subset \mathbb{R}^r$, где Θ — заданное открытое множество. Рассмотрим задачу оценивания θ или функций от θ по наблюдению X. Ясно, что в качестве оценки θ или $\tau(\theta)$ должны выступать случайные векторы соответствующей размерности (функции от X).

Поэтому предварительно надо напомнить, что такое случайный вектор, случайная матрица, их математические ожидания и ковариации, вместе с некоторыми свойствами этих объектов. Случайный вектор при этом есть частный случай случайной матрицы.

Определение 1. Случайная матрица Z есть матрица, элементы z_{ij} которой суть случайные величины, заданные на общем пространстве элементарных исходов, т.е. имеющие совместное распределение вероятностей.

Определение 2. Математическое ожидание случайной матрицы $Z = \|z_{ij}\|$ есть

$$\mathsf{E}Z = \|\mathsf{E}z_{ij}\|$$

Утверждение 1. Пусть Z — случайная матрица, а постоянные матрицы A, B и C таковы, что матрица $AZB+C$ существует, то есть Размерности матриц A, B, Z и C согласованы. Тогда:

$$\mathsf{E}(AZB + C) = A(\mathsf{E}Z)B + C$$

В частности, если Y — случайный вектор, A — постоянная матрица и b — постоянный вектор, то

$$\mathsf{E}(AY + b) = A(\mathsf{E}Y) + b,$$

когда указанные операции (умножения и сложения) осуществимы.

Утверждение 2. Пусть Z_1 и Z_2 — две случайные матрицы, определенные на общем пространстве элементарных исходов. Пусть их размерности совпадают, так что матрица $Z_1 + Z_2$ существует. Тогда:

$$\mathsf{E}(Z_1 + Z_2) = \mathsf{E}Z_1 + \mathsf{E}Z_2.$$

Утверждения 1 и 2 вместе показывают, что операция взятия математического ожидания для случайных матриц обладает привычными для этой операции над случайной величиной линейными свойствами. Правда, с учетом того, что умножение матриц не коммутативно.

Пусть X и Y — два случайных вектора (произвольных размерностей, не обязательно одинаковых), имеющие совместное распределение. Векторы мы предпочтительно будем представлять в виде векторов-столбцов (одностолбцовых матриц).

Определение 3. Ковариационная матрица (она же — дисперсионная матрица) векторов X и Y есть

$$\operatorname{cov}(X, Y) = \mathsf{E}(X - \mathsf{E}X)(Y - \mathsf{E}Y)^T.$$

Если $X = (\xi_1, \xi_2, \ldots)^T$, $Y = (\eta_1, \eta_2, \ldots)^T$, то элемент (i, j) матрицы $\operatorname{cov}(X, Y)$ есть ковариация случайных величин ξ_i и η_j:

$$\mathsf{E}(\xi_i - \mathsf{E}\xi_i)(\eta_j - \mathsf{E}\eta_j).$$

Ясно, что:

$$\operatorname{cov}(X, Y) = \mathsf{E}XY^T - (\mathsf{E}X)(\mathsf{E}Y)^T.$$

Определение 4. Ковариационная матрица случайного вектора X определяется как:

$$\operatorname{cov}(X, X) = \mathsf{E}(X - \mathsf{E}X)(X - \mathsf{E}X)^T = \mathsf{E}XX^T - (\mathsf{E}X)(\mathsf{E}X)^T.$$

Диагональные элементы этой матрицы суть дисперсии случайных величин ξ_i. Обозначение $\operatorname{cov}(X, X)$ мы будем заменять коротким $\mathsf{D}X$.

Утверждение 3. Пусть X — случайная величина, A — неслучайная (постоянная) матрица, b — неслучайный (постоянный) вектор. Тогда:

$$\mathsf{D}(AX + b) = A(\mathsf{D}X)A^T,$$

если $AX + b$ существует (если указанные операции осуществимы, т.е. размерности A, X и b согласованы).

Частный случай: скалярное произведение. Пусть A — матрица, состоящая из одной строки. Рассмотрим A как результат транспонирования некоторого вектора a (вектора-столбца): $A = a^T$. При этом $AX = a^T X$ — есть скалярное произведение векторов a и X.

Утверждение 4.

$$\mathsf{D}(a^T X) = a^T(\mathsf{D}X)a$$

2.5.2. Квадратичный риск в многомерном случае

Вернемся к поставленной в начале этого параграфа задаче. Пусть $\varphi(\cdot)$ — некоторая вектор-функция, $\varphi(X)$ — оценка $\tau(\theta)$ (это векторы-столбцы), и пусть $\mathsf{E}_\theta\varphi(X) = \tau(\theta)$, где $\tau(\theta) = (\tau_1(\theta), \ldots, \tau_d(\theta))^T$, $\theta \in \Theta \subset \mathbb{R}^r$.

Как и в одномерном (однопараметрическом) случае мы готовимся указать границу снизу для квадратичного риска несмещенной оценки. Но прежде надо уточнить, что такое квадратичный риск в многомерном случае и как следует сравнивать квадратичные риски — например, двух разных оценок.

Пусть $\varphi(X)$, $\psi(X)$ — две несмещенные оценки $\tau(\theta)$. Какая из них лучше? Попробуем найти ответ, обратившись к уже изученному одномерному случаю. Выберем произвольный неслучайный вектор. Перейдем от $\varphi(X)$,

$\psi(X)$, $\tau(\theta)$ к линейным формам (скалярным произведениям) $\xi := z^T\varphi(X)$, $\eta := z^T\psi(X)$, $t(\theta) := z^T\tau(\theta)$. Ясно, что

$$\mathsf{E}_\theta\xi = \mathsf{E}_\theta\eta = t(\theta),$$

так что ξ и η суть несмещенные (одномерные) оценки $t(\theta)$. В одномерном случае (при квадратичной функции потерь) из двух несмещенных оценок лучше та, чья дисперсия меньше. В частности, ξ не хуже, чем η, если $\mathsf{D}\xi \leqslant \mathsf{D}\eta$ или:

$$z^T[\mathsf{D}_\theta\varphi(X)]z \leqslant z^T[\mathsf{D}_\theta\psi(X)]z \qquad (*)$$

Мы можем принять такое определение: $\varphi(X)$ лучше, чем $\psi(X)$, если $(*)$ выполняется для <u>любого</u> вектора $z \in \mathbb{R}^d$ (и для некоторых z это неравенство строгое).

По отношению к переменному $z \in \mathbb{R}^d$ выражения $z^T[\mathsf{D}_\theta\varphi(X)]z$ и $z^T[\mathsf{D}_\theta\psi(X)]z$ представляют собой квадратичные формы (неотрицательно определенные). Неравенство $(*)$, если оно выполняется для всех z, линейная алгебра истолковывает как соотношение между матрицами квадратичных форм. В данном случае, между матрицами ковариаций $\mathsf{D}_\theta\varphi(X)$ и $\mathsf{D}_\theta\psi(X)$: $\mathsf{D}_\theta\varphi(X) \leqslant \mathsf{D}_\theta\psi(X)$

Итак, мы пришли к заключению, что *квадратичным риском* статистики $\varphi(X)$, несмещённо оценивающей $\tau(\theta)$, можно назвать ее матрицу ковариаций:

$$\mathsf{D}_\theta\varphi = \mathsf{E}_\theta[\varphi(X) - \tau(\theta)][\varphi(X) - \tau(\theta)]^T.$$

Из двух несмещенных оценок лучше та, чья матрица ковариаций меньше (в указанном выше смысле). Заметим, что две оценки могут быть несравнимы.

Теперь понятно, что многомерное обобщение неравенства Крамера – Рао должно устанавливать границу снизу для матрицы ковариаций несмещенной оценки.

2.5.3. Многомерное неравенство Крамера – Рао

Переходим к выводу неравенства.

Введем оператор частного дифференцирования по θ, который (в виде исключения) запишем как строку:

$$\frac{\partial}{\partial\theta} = \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_r}\right)$$

Определим матрицу информации (обобщение количества информации $I(\theta)$):

$$I(\theta) = \mathsf{E}_\theta\left[\frac{\partial}{\partial\theta}\ln p(X,\theta)\right]^T\left[\frac{\partial}{\partial\theta}\ln p(X,\theta)\right]$$

Легко видеть, что $I(\theta)$ — неотрицательно определенная матрица, что мы будем записывать в виде $I(\theta) \geqslant 0$. Предположим, что $I(\theta)^{-1}$ существует для всех $\theta \in \Theta$.

Введем матрицу $\left(\frac{\partial\tau}{\partial\theta}\right)$ размера $d \times r$, положив:

$$\frac{\partial\tau}{\partial\theta} = \begin{pmatrix} \frac{\partial\tau_1}{\partial\theta_1} & \frac{\partial\tau_1}{\partial\theta_2} & \cdots & \frac{\partial\tau_1}{\partial\theta_r} \\ \frac{\partial\tau_2}{\partial\theta_1} & \frac{\partial\tau_2}{\partial\theta_2} & \cdots & \frac{\partial\tau_2}{\partial\theta_r} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\tau_d}{\partial\theta_1} & \frac{\partial\tau_d}{\partial\theta_2} & \cdots & \frac{\partial\tau_d}{\partial\theta_r} \end{pmatrix}$$

Покажем, что при принятых в пункте 1 «условиях регулярности»:

$$\mathsf{E}_\theta(\varphi(X) - \tau(\theta))(\varphi(X) - \tau(\theta))^T \geqslant \frac{\partial\tau}{\partial\theta}I^{-1}\left(\frac{\partial\tau}{\partial\theta}\right)^T. \qquad (1)$$

Доказательство:

Рассмотрим вектор-строку:

$$\lambda(X,\theta) = \frac{\partial}{\partial\theta}\ln p(X,\theta).$$

Так же, как и в одномерном случае, находим, что

$$\mathsf{E}_\theta\lambda(X,\theta) = 0. \qquad (2)$$

Дифференцируем по θ тождество

$$\int_A \varphi(x)p(x,\theta)dx = \tau(\theta);$$

получаем, что:

$$\int\limits_A \varphi(x)\frac{\partial}{\partial\theta}p(x,\theta)dx = \frac{\partial\tau}{\partial\theta},$$

или

$$\int\limits_A \varphi(x)\left[\frac{\partial}{\partial\theta}\ln p(x,\theta)\right]p(x,\theta)dx = \frac{\partial\tau}{\partial\theta}.$$

Последнее равенство означает, что:

$$\mathsf{E}_\theta\varphi(X)\lambda(X,\theta) = \frac{\partial\tau}{\partial\theta}. \tag{3}$$

Теперь рассмотрим (неотрицательно определенную) матрицу ковариаций вектора

$$\varphi(X) - \tau(\theta) - \frac{\partial\tau}{\partial\theta}I^{-1}(\theta)\lambda^T(X,\theta).$$

(Обратите внимание на то, что размерности перемножаемых матриц согласованы таким образом, что умножение возможно).

Рассмотрим очевидное неравенство:

$$\mathsf{E}_\theta\left[(\varphi-\tau) - \frac{\partial\tau}{\partial\theta}I^{-1}\lambda^T\right]\left[(\varphi-\tau) - \frac{\partial\tau}{\partial\theta}I^{-1}\lambda^T\right]^T \geqslant 0$$

Левую часть тождественно преобразуем:

$$\mathsf{E}_\theta(\varphi-\tau)(\varphi-\tau)^T - \mathsf{E}_\theta(\varphi-\tau)\left[\frac{\partial\tau}{\partial\theta}I^{-1}\lambda^T\right]^T - \mathsf{E}_\theta\left[\frac{\partial\tau}{\partial\theta}I^{-1}\lambda^T\right](\varphi-\tau)^T + \mathsf{E}_\theta\left[\frac{\partial\tau}{\partial\theta}I^{-1}\lambda^T\right]\left[\frac{\partial\tau}{\partial\theta}I^{-1}\lambda^T\right]^T \geqslant 0 \tag{4}$$

Второе слагаемое в (4):

$$\mathsf{E}_\theta(\varphi-\tau)\lambda I^{-1}\left(\frac{\partial\tau}{\partial\theta}\right)^T = \frac{\partial\tau}{\partial\theta}I^{-1}\left(\frac{\partial\tau}{\partial\theta}\right)^T, \tag{5}$$

ибо $\mathsf{E}_\theta\varphi\lambda = \frac{\partial\tau}{\partial\theta}$ (см. (3)), $\mathsf{E}_\theta\lambda = 0$ (см. (2)).

Третье слагаемое отличается от второго лишь транспонированием (третье слагаемое — это транспонированное второе). А так как (5) симметрично, то третье слагаемое тоже равно (5).

Наконец, четвертое слагаемое даст:

$$\frac{\partial\tau}{\partial\theta}I^{-1}\left[\mathsf{E}_\theta\lambda^T\lambda\right]I^{-1}\left(\frac{\partial\tau}{\partial\theta}\right)^T = \frac{\partial\tau}{\partial\theta}I^{-1}\left(\frac{\partial\tau}{\partial\theta}\right)^T$$

Приведя в (4) подобные члены, получим отсюда (1), что и требовалось. □

Заключим тему неравенств информации и эффективных оценок определением многопараметрических экспоненциальных семейств. Плотность (вероятность) для них имеет вид:

$$p(x,\theta) = \exp\left[\sum_{i=1}^r c_i(\theta)T_i(x) + d(\theta) + S(X)\right]I_A(x).$$

Наиболее важный пример — гауссовское распределение, где плотность зависит от двумерного параметра (a,σ^2):

$$p(x,a,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\}.$$

Вопрос: Для какой (двумерной) функции $\tau(\theta) = (\tau_1(a,\sigma^2),\tau_2(a,\sigma^2))^T$ существует эффективная оценка?

2.6. Достаточные статистики

Напомним, что мы рассматриваем следующую статистическую модель: наблюдение X получено случайным выбором из множества $\mathcal{X}$; случайный выбор управляется распределением вероятностей P_θ, где θ — некоторый (неизвестный) параметр, причем $\theta\in\Theta$; Θ — заданное множество возможных значений этого параметра.

2.6.1. Определение

Статистика $T = T(X)$ называется *достаточной* для параметра θ, $\theta\in\Theta$, если условное распределение X при данном значении $T(X)$ одно и то же для всех $\theta\in\Theta$. (Иначе говоря, если упомянутое условное распределение не меняется (не зависит от θ), когда θ пробегает множество Θ).

2.6.2. Дискретный случай

Когда распределение X дискретно, понятие условного распределения X вводится элементарно:

$$P_\theta(X = x|T(X) = t) = \frac{P_\theta(X = x, T(X) = t)}{P_\theta(T(X) = t)} = \begin{cases} \frac{P_\theta(X=x)}{P_\theta(T(X)=t)}, & \text{если } T(X) = t \\ 0, & \text{если } T(X) \neq t \end{cases}$$

Пример: испытания Бернулли.

Пусть $X = (X_1, \ldots, X_n)$ — результаты испытаний Бернулли, в которых вероятность успеха есть θ, $\theta \in (0, 1)$. В качестве статистики $T(X)$ возьмем $T = \sum\limits_{i=1}^{n} X_i$.

Здесь X_i принимает значения 0 или 1 (число успехов в испытании номер i), T — общее число успехов в n испытаниях.

Элементарная выкладка показывает, что в этом примере (где $x = (x_1, \ldots, x_n)$ — заданная последовательность нулей и единиц):

$$P_\theta(X = x|T(X) = t) = \begin{cases} \frac{1}{C_n^t}, & \text{если } \sum\limits_{i=1}^{n} x_i = t \\ 0, & \text{если } \sum\limits_{i=1}^{n} x_i \neq t \end{cases}$$

Как видно из формулы, $T = \sum\limits_{i=1}^{n} X_i$ есть достаточная статистика для θ, $\theta \in (0, 1)$.

2.6.3. Непрерывный случай

Так, для краткости, назовем статистическую модель, в которой распределение P_θ может быть задано с помощью плотности $p(x, \theta)$ относительно некоторой меры. Для простоты предположим, что X принимает значения в конечномерном пространстве и что $p(x, \theta)$ — плотность относительно лебеговской меры. В этом случае значения статистики T выделяют *множества уровня* $\{x : T(x) = t\}$.

Условное распределение X на множестве уровня $\{x : T(x) = t\}$ в этом случае можно задать с помощью плотности (относительно меры Лебега на множестве уровня). Эта условная плотность пропорциональна $p(x, \theta)$. Поскольку интеграл от плотности составляет 1, эта условная плотность X при данном $T(X) = t$, т.е. на множестве уровня $\{x : T(x) = t\}$, равна

$$\frac{p(x, \theta)}{\int\limits_{\{y:T(y)=t\}} p(y, \theta)dy}$$

(Выражение в знаменателе — это интеграл по поверхности уровня).

2.6.4. Достаточные разбиения

Из определения достаточной статистики следует, что, если случайная функция $S = S(T)$ находится во взаимно однозначном соответствии с достаточной статистикой $T = T(X)$, то S тоже является достаточной статистикой. Поэтому правильнее было бы говорить не о достаточных статистиках, а о производимых ими разбиениях выборочных пространств (разбиениях на множества уровня достаточных статистик). Условные распределения X на элементах этих разбиений одинаковы для всех распределений θ, когда $\theta \in \Theta$. Достаточная статистика $T = T(X)$ разбивает выборочное пространство $\mathcal{X}$ на множества уровня $\{x : T(x) = \text{const}\}$.

Пример Пусть $X = (X_1, \ldots, X_n)$ — выборка из показательного распределения, где плотность отдельного наблюдения X_i равна

$$f(u, \theta) = \begin{cases} \frac{1}{\theta} \exp\left(-\frac{u}{\theta}\right) & \text{для } u \geqslant 0 \\ 0 & \text{для } u < 0 \end{cases}$$

Параметр θ — неотрицательное число, т.е. $\theta \in (0, \infty)$. Покажем, что $T = \sum\limits_{i=1}^{n} X_i$ — достаточная статистика для θ в этой модели. Плотность X в точке $u = (u_1, \ldots, u_n)$ есть:

$$\prod_{i=1}^{n} f(u, \theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n \exp\left(-\frac{T}{\theta}\right) & \text{где } T = \sum\limits_{i=1}^{n} u_i, \text{ и } u_i \geqslant 0; \\ 0 & \text{в противном случае.} \end{cases}$$

В следующей формуле $S := \sum\limits_{i=1}^{n} y_i$. Условная плотность X при фиксированном T равна (в точке u такой, что $\sum\limits_{i=1}^{n} u_i = T$ и $u_1, \ldots, u_n \geqslant 0$):

$$\frac{\left(\frac{1}{\theta}\right)^n \exp\left(-\frac{T}{\theta}\right)}{\int\limits_{\{y\colon S=T, y\geqslant 0\}} \left(\frac{1}{\theta}\right)^n \exp\left(-\frac{S}{\theta}\right) dy} = \frac{\left(\frac{1}{\theta}\right)^n \exp\left(-\frac{T}{\theta}\right)}{\left(\frac{1}{\theta}\right)^n \exp\left(-\frac{T}{\theta}\right) \int\limits_{\{y\colon S=T, y\geqslant 0\}} dy} = \text{const}\,.$$

Здесь оказалось, что условная плотность (на множестве уровня) не только не зависит от θ, — что доказывает, что статистика T достаточна, но не зависит и от координаты y. Это означает, что указанное условное распределение X равномерно.

Выкладки, которые мы проделали в двух рассматриваемых примерах, по существу повторяются при доказательстве следующей теоремы:

2.6.5. Теорема факторизации

Теорема 2.2. *Статистика $T = T(X)$ достаточна для параметра θ, $\theta \in \Theta$, тогда и только тогда, когда существуют функции $g(t, \theta)$ и $h(x)$ такие, что*

$$p(x, \theta) = g\big(T(X), \theta\big) h(x) \qquad (*)$$

при всех $\theta \in \Theta$.

Замечание:

Величина $p(x, \theta)$ обозначает либо плотность наблюдения X в точке x, если модель непрерывна, либо вероятность точки x, если модель дискретна.

□ Доказательство проведем отдельно для дискретного случая; в непрерывном случае оно слабо отличается.

1° Если выполнено $(*)$, то $T = T(X)$ — достаточная статистика для θ. Надо показать, что $P(X|T(X))$ не зависит от $\theta \in \Theta$.

Сначала вычислим:

$$P_\theta(T = t) = \sum_{x:T(x)=t} p(x, \theta) = \sum_{x:T(x)=t} g(T(x), \theta) h(x) = g(t, \theta) \sum_{x:T(x)=t} h(x).$$

Теперь для x такого, что $T(x) = t$ получаем, что:

$$P_\theta(X = x | T(X) = t) = \frac{P_\theta(X=x, T(X)=t)}{P_\theta(T(X)=t)} = \frac{P_\theta(X=x)}{P_\theta(T(X)=t)} = \frac{g(T(x),\theta)h(x)}{g(t,\theta) \sum\limits_{y:T(y)=t} h(y)} = \frac{h(x)}{\sum\limits_{T(x)=t} h(x)} - \text{результат не зависит от } \theta \in \Theta.$$

Если же x таково, что $T(x) \neq t$, то обсуждаемая условная вероятность равна 0, вне зависимости от θ. Достаточность условия $(*)$ доказана.

2° Если T — достаточная статистика, то $(*)$ выполнено. Если T достаточна, то для таких x, что $T(x) = t$, и для всех $\theta \in \Theta$

$$P_\theta(X = x | T(X) = t) = h(x) \text{- результат не зависит от } \theta, \text{обозначим его через } h(x),$$

или

$$\frac{P_\theta(X = x, T(X) = t)}{P_\theta(T(X) = t)} = h(x).$$

Поскольку $T(x) = t$, то дробь в левой части есть:

$$\frac{P_\theta(X = x)}{P_\theta(T(X) = t)}.$$

Отсюда

$$P_\theta(X = x) = P_\theta(T(X) = t) h(x)$$

Обозначив $P_\theta(T(X) = t)$ через $g(t, \theta)$, получим то, что и требовалось доказать. ■

Заметим, что $h(x)$ — это условная вероятность X при данном T (в точке x), либо $h(x)$ пропорциональна этой условной вероятности. Аналогично $g(x, \theta)$ лишь постоянным множителем может отличаться от вероятности $P_\theta(T(X) = t)$.

2.6.6. Пример: линейная модель

(a) Линейная (гауссовская) модель — важный объект исследований и приложений. Сначала будет дана ее абстрактная формулировка, а затем одна из конкретных форм.

Наблюдаемый объект — вектор X. Сейчас мы считаем его n-мерным: $X = (X_1, \dots, X_n)^T$ — вектор-столбец. Его координаты считаем независимыми случайными величинами, распределенными по нормальному закону, причем $\mathsf{D}X_i = \sigma^2$, $i = \overline{1,n}$. Значение σ^2 неизвестно.

Относительно $\mathsf{E}X$ предположим, что $\mathsf{E}X$, будучи неизвестным, принадлежит заданному линейному подпространству L, $L \subset \mathbb{R}^n$.

Если обозначить $\mathsf{E}X = l$, $\mathsf{E}(X - \mathsf{E}X)(X - \mathsf{E}X)^T = \mathsf{D}_\theta X = \sigma^2 I$ (I — единичная матрица), то $X \sim N(l, \sigma^2 I)$, причем $l \in L$, L — задано.

(b) Покажем, что достаточной статистикой для (составного) параметра $\theta = (l, \sigma^2)$, причем $l \in L$, служит пара $(\mathrm{proj}_L X, |\,\mathrm{proj}_{L^\perp} X|^2)$. Здесь через proj_M обозначен оператор проектирования (в евклидовой метрике) на подпространство $M \subset \mathbb{R}^n$; $L^\perp$ обозначает ортогональное дополнение L до $\mathbb{R}^n$, т.е. $\mathbb{R}^n = L \oplus L^\perp$.

Для доказательства достаточно указать плотность X и затем ее преобразовать:

$$p(X,\theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - l_i)^2\right\} =$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}|X-l|^2\right\} =$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\big|(\mathrm{proj}_L X - l) + \mathrm{proj}_{L^\perp} X\big|^2\right\}$$

По теореме Пифагора:

$$|(\mathrm{proj}_L X - l) + \mathrm{proj}_{L^\perp} X|^2 = |\,\mathrm{proj}_L X - l|^2 + |\,\mathrm{proj}_{L^\perp} X|^2,$$

ибо $(\mathrm{proj}_L X - l) \perp \mathrm{proj}_{L^\perp} X$, т.к. $l \in L$.

Поэтому плотность X равна

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}|\,\mathrm{proj}_L X - l|^2\right\} \exp\left\{-\frac{1}{2\sigma^2}|\,\mathrm{proj}_{L^\perp} X|^2\right\}$$

Мы видим, что плотность зависит от статистик $\mathrm{proj}_L X$ и $|\,\mathrm{proj}_{L^\perp} X|^2$, но не от X непосредственно. Эта пара и составляет достаточную статистику. (Заметим, что функция $h(X)$ здесь равна 1, точнее — постоянна по отношению к X. Это означает, что условное распределение X при фиксированном значении достаточной статистики — равномерное.)

(c) Линейная регрессия. Задача *линейной регрессии* — одна из частных форм линейной модели. В простейшем случае это задача о подборе функции одного переменного — подборе по неточным наблюдениям (измерениям).

Предположим, что две переменные t и x связаны соотношением $x = f(t)$, где $f(\cdot)$ — некоторая функция. При некоторых значениях переменной t (называемой часто фактором) $t_1, \dots, t_n$ были произведены измерения переменной x (называемой откликом). Они дали значения $x_1, \dots, x_n$. При этом $x_i = f(t_i) + \varepsilon_i$, где $\varepsilon_1, \dots, \varepsilon_n$ — некоторые ошибки, сопровождающие измерения. Основное предположение состоит в том, что мы считаем упомянутые $\varepsilon_1, \dots, \varepsilon_n$ независимыми случайными величинами. Менее важные предположения: ε_i распределены одинаково и распределены по нормальному закону $N(0, \sigma^2)$. Предположение $\mathsf{E}\varepsilon_i = 0$ отражает представление о том, что систематических ошибок при измерении отклика в нашей схеме нет. Величина σ обычно считается неизвестной (необязательно). Она численно выражает неточность (изменчивость) измерений, т.е. масштаб случайных ошибок.

Последнее предположение, превращающее задачу регрессии в линейную: считаем, что $f(\cdot)$ можно (с достаточной аккуратностью) выразить в виде линейной комбинации заданного конечного набора функций (скажем $\varphi_1, \dots, \varphi_m$): существуют параметры $\theta_1, \dots, \theta_m$ такие, что

$$f(t) = \theta_1\varphi_1(t) + \dots + \theta_m\varphi_m(t).$$

В этом случае вектор $X = (x_1, \ldots, x_n)^T$ представляется в виде линейной комбинации векторов:

$$\Phi_j = \big(\varphi_j(t_1), \ldots, \varphi_j(t_n)\big)^T, j = \overline{1, m}$$

и вектора ε случайных ошибок: $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$:

$$X = \sum_{j=1}^{m} \theta_j \Phi_j + \varepsilon.$$

Линейное подпространство L, которому заведомо принадлежит вектор $\mathsf{E}X$, в данном случае порождено векторами $\Phi_1, \ldots, \Phi_m$.

(d) Нормальная выборка. Рассмотрим выборку $x_1, \ldots, x_n$ из нормальной совокупности $N(a, \sigma^2)$, где параметры $a \in \mathbb{R}$, $\sigma^2 \in (0, \infty)$ неизвестны. Теорема факторизации помогает найти достаточные статистики для (a, σ^2). Выпишем плотность этой модели (пользуясь независимостью гауссовских случайных величин $x_1, \ldots, x_n$) и преобразуем ее:

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - a)^2}{2\sigma^2}\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} x_i^2 - 2a\sum_{i=1}^{n} x_i + na^2\right]\right\}.$$

Поскольку плотность зависит от переменных $x_1, \ldots, x_n$ лишь посредством статистик $\sum\limits_{i=1}^{n} x_i$ и $\sum\limits_{i=1}^{n} x_i^2$, эта пара и является достаточной статистикой для (a, σ^2). Мы уже обращали внимание на то, что главным в определении достаточной статистики $T = T(X)$ является не ее конкретный вид, а то разбиение выборочного пространства на множества уровня вида $\{T(X) = \text{const}\}$, которое она производит. Любая другая статистика, если она порождает то же самое разбиение, тоже является достаточной. В частности, достаточной окажется любая статистика, находящаяся во взаимно однозначном соответствии с $T(X)$.

Для обсуждаемой нормальной выборки предпочитаемой достаточной статистикой служит:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Легко видеть, что $(\overline{x}, s^2)$ взаимно однозначно связана с $(\sum\limits_{i=1}^{n} x_i, \sum\limits_{i=1}^{n} x_i^2)$.

О преимуществах, которые дает статистика $(\overline{x}, s^2)$ перед другими статистиками для (a, σ^2), мы подробнее будем говорить позже. Сейчас же отметим лишь то, что $\overline{x}$ и s^2 несмещенно оценивают a и σ^2:

$$\mathsf{E}\overline{x} = a, \qquad \mathsf{E}s^2 = \sigma^2$$

Заметим, что эти соотношения справедливы для любой, не только гауссовской, выборки (если $\mathsf{D}x_i^2$ существуют).

Выборка из $N(a, \sigma^2)$ является частным случаем линейной модели. Рассмотрим вектор $X = (x_1, \ldots, x_n)^T$. Его математическое ожидание равно $(a, a, \ldots, a)^T$, и потому принадлежит линейному подпространству L, порожденному вектором $(1, \ldots, 1)^T$. Так как координаты вектора X независимы и одинаково распределены, то $\mathsf{D}X = \sigma^2 I$. Таким образом, предпосылки линейной модели соблюдены.

Достаточные статистики общей линейной модели в данном случае суть:

$$\operatorname{proj}_L X = \overline{x}(1, 1, \ldots, 1)^T,$$

$$|\operatorname{proj}_{L^\perp} X| = \sum_{i=1}^{n}(x_i - \overline{x})^2 = (n-1)s^2.$$

(e) При обсуждении гауссовской линейной модели мы отмечали, что условное распределение X при фиксированном значении достаточной статистики - равномерное. Из этого обстоятельства можно извлечь интересные следствия. В данном примере упомянутое условное распределение сосредоточено на $(n-2)$-мерной сфере:

$$\{y : y \in R^n, \sum_{i=1}^{n} y_i = \overline{x}, \sum_{i=1}^{n}(y_i - \overline{y})^2 = (n-1)s^2\}$$

Рассмотрим вектор

$$Y = \left(\frac{x_1 - \overline{x}}{s\sqrt{n-1}}, \frac{x_2 - \overline{x}}{s\sqrt{n-1}}, \ldots, \frac{x_n - \overline{x}}{s\sqrt{n-1}}, \right)^T$$

При фиксированном значении достаточной статистики $(\overline{x}, s^2)$ вектор Y является линейным (и взаимно-однозначным) преобразованием вектора X. Поэтому условное (при фиксированных $\overline{x}, s^2$) распределение Y тоже является равномерным. Это условное распределение сосредоточено на $(n-2)$-мерной единичной сфере

$$S_{n-2} = \left\{ y \colon y \in \mathbb{R}^n, \sum_{i=1}^n y_i = 0, \sum_{i=1}^n y_i^2 = 1 \right\}.$$

Теперь заметим, что сказанное условное распределение Y при данных $\overline{x}, s^2$ — одно и то же (а именно равномерное на S_{n-2}) при любых значениях $\overline{x}, s^2$. Значит:

1. вектор Y как случайный элемент не зависит от $\overline{x}, s^2$;
2. (безусловное) распределение Y совпадает с условным, т.е. является уже известным равномерным распределением на S_{n-2}.

Из сказанного следует, что для нормальной выборки такие (часто применяемые на практике) статистики, как выборочная асимметрия $\sum\limits_{i=1}^n \left(\frac{(x_i - \overline{x})}{s} \right)^3$ и выборочный эксцесс $\sum\limits_{i=1}^n \left(\frac{(x_i - \overline{x})}{s} \right)^4$ не зависят от $(\overline{x}, s^2)$. А их распределения не зависят от X и могут быть вычислены (табулированы).

Упомянутые статистики обычно в виде выборочного коэффициента асимметрии

$$\beta_1 = \frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^3 / s^3$$

и выборочного коэффициента эксцесса

$$\beta_2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{(x_i - \overline{x})}{s} \right)^4 - 3$$

могут служить для проверки нормальности имеющейся выборки, т. е. для проверки предположения о том, что данная выборка извлечена из некоторой нормальной совокупности. Нормальность выборки дает возможности для ее детального анализа (в дальнейшем будет видно, какие).

Для общей линейной гауссовской модели утверждение о равномерном распределении случайного вектора $(\mathrm{proj}_{L^\perp} X)/|\mathrm{proj}_{L^\perp} X|$ (на единичной сфере размерности $n - r$, где $r = \dim L$) и его статистической зависимости от пары $(\mathrm{proj}_L X, |\mathrm{proj}_{L^\perp} X|^2)$ доказывается аналогично.

Аналогичным порядком мы можем составить коэффициенты асимметрии и эксцесса, и тоже использовать их для проверки нормальности распределения X в линейной модели.

2.7. Наилучшие несмещенные оценки

2.7.1. Наилучшие несмещенные оценки

Так обычно называют несмещенные оценки с минимальным квадратичным риском.

Для скалярного параметра (и для скалярных функций от параметра) это несмещенные оценки с минимальной дисперсией; для векторного (конечномерного) параметра и функций от него — это несмещенные оценки с наименьшей матрицей ковариаций. В некоторых случаях указать наилучшую несмещенную оценку помогают неравенства Крамера – Рао: если оценка эффективная, то она и наилучшая в указанном выше смысле, так как имеет наименьшую возможную дисперсию.

Но даже для экспоненциальных семейств распределений, для которых только и существуют эффективные оценки, эффективно оценить можно лишь одну какую-то функцию от параметра. Скажем, для испытаний Бернулли, в которых параметром θ служит вероятность успеха, эффективная оценка есть только для θ (это частота успехов). Но каковы несмещенные оценки, например, для $\theta(1-\theta)$ или θ^2?

Вопрос тем более открыт для семейств распределений, не являющихся экспоненциальными.

Известные к настоящему времени обобщения неравенства Крамера – Рао расширяют наши возможности не слишком значительно.

Задачу о наилучших несмещённых оценках удается продвинуть (а часто — полностью решить), если для неизвестного параметра существует достаточная статистика. Несмещённое оценивание при достаточной статистике и будет нашей текущей темой. Для ее обсуждения нам понадобится понятие условного математического ожидания одной случайной величины при фиксированном значении другой. В полном объеме оно будет введено и изучено в следующей главе. А сейчас, чтобы завершить тему наилучшего несмещённого оценивания, мы ограничимся неформальным толкованием этого понятия. А также укажем некоторые его свойства, необходимые для упомянутой цели.

2.7.2. Условные математические ожидания: предварительные сведения

Пусть случайные величины X и Y заданы на одном вероятностном пространстве. (Содержательно это означает, что значения переменных X, Y получены в одном эксперименте). Понятие условного математического ожидания X при данном значении Y — далее $\mathsf{E}(X|Y)$ — можно ввести элементарными средствами, если при каждом (почти каждом) значении Y существует условное распределение X. Рассмотрим условное распределение X при данном Y. Усредним значения X (при данном Y) по этому условному распределению. Полученный результат (число, если X принимает числовые значения, вектор-столбец, если значения X суть векторы-столбцы и т.д.) зависит от фиксированного значения Y, т.е. является функцией Y. Его называют *условным математическим ожиданием* X при данном Y и обозначают как $\mathsf{E}(X|Y)$. Поскольку Y - случайная величина, $\mathsf{E}(X|Y)$ тоже является случайной величиной.

Если совместное распределение (X, Y) имеет плотность $p(x, y)$ (либо дискретно), то формулу для $\mathsf{E}(X|Y)$ можно получить явно. В этом случае условное распределение X при данном Y имеет плотность (в точке x), равную

$$\frac{p(x,Y)}{\int p(x,Y)\,dx}.$$

Отсюда

$$\mathsf{E}(X|Y) = \frac{\int xp(x,Y)\,dx}{\int p(x,Y)\,dx}.$$

Аналогичная формула (с заменой интегрирования суммированием) действует и в дискретном случае.

В общем случае соотношение между условным распределением и условным математическим ожиданием — обратное по отношению к описанному: $\mathsf{E}(X|Y)$ первично и вводится непосредственно, а понятие условного распределения X при данном Y может быть определено на его основе.

Укажем некоторые свойства условных математических ожиданий, которые нам сейчас понадобятся. Линейные свойства вполне ожидаемы и естественны:

1. $$\mathsf{E}(X_1 + X_2|Y) = \mathsf{E}(X_1|Y) + \mathsf{E}(X_2|Y)$$

   (Здесь случайные величины X_1 и X_2 должны быть заданы на том же пространстве элементарных исходов, что и Y).

2. $$\mathsf{E}(kX|Y) = k\mathsf{E}(X|Y),$$

   где k — постоянный (неслучайный) множитель.

3. $$\mathsf{E}[f(Y)X|Y] = f(Y)\mathsf{E}(X|Y),$$

   где $f(Y)$ — функция Y. Это свойство тоже естественно, ибо при фиксированном значении Y функция $f(Y)$ постоянна, а постоянный множитель можно выносить за знак математического ожидания.

   Надо оговорить, что перечисленные выше равенства выполняются с вероятностью 1, ибо они соединяют случайные величины. Нужно также, чтобы существовало $\mathsf{E}|X|$ (в первом пункте должны существовать $\mathsf{E}|X_1|$ и $\mathsf{E}|X_2|$).

   Наиболее важным является свойство

4. $$\mathsf{E}\{\mathsf{E}(X|Y)\} = \mathsf{E}X.$$

2.7.3. Улучшение несмещенных оценок

Вернемся к обсуждавшейся задаче о несмещенных оценках с минимальной дисперсией. В ее решении можно сделать шаг вперед, если в статистической модели есть достаточная статистика.

Пусть X — наблюдаемая случайная величина, распределенная по некоторому закону P_θ, где θ — неизвестный параметр, $\theta \in \Theta$, Θ — задано.

Пусть $d = d(X)$ — несмещенная оценка $\tau(\theta)$, где $\tau(\theta)$ — заданная функция, т.е.:

$$\mathsf{E}_\theta d(X) = \tau(\theta) \qquad \text{для всех } \theta \in \Theta,$$

причем $\mathsf{E}_\theta |d(X)|$ существует.

Пусть T — достаточная статистика для параметра θ.

Рассмотрим условное математическое ожидание $d(X)$ при данном T:

$$\varphi(T) = \mathsf{E}(d(X)|T)$$

Заметим, что $\mathsf{E}(d(X)|T)$ не зависит от θ, так как от θ не зависит условное распределение X при данном T — в силу определения достаточной статистики.

Теорема 2.3 (Blackwell – Rao, 1947-1949). *При указанных выше условиях*

$$\mathsf{E}_\theta \varphi(T) = \tau(\theta) \text{ и } \mathsf{D}_\theta \varphi(T) \leqslant \mathsf{D}_\theta d(X),$$

причем равенство достигается, если и только если $\varphi(T) = d(X)$ (с вероятностью 1, для каждого $\theta \in \Theta$).

□

1° Первое утверждение выполняется в силу свойства условных математических ожиданий $\mathsf{E}\mathsf{E}(X|Y) = \mathsf{E}X$:

$$\mathsf{E}_\theta \mathsf{E}[d(X)|T] = \mathsf{E}_\theta d(X) = \tau(\theta).$$

2° Доказательство второго свойства для одномерных φ, d и τ:

$$\mathsf{D}_\theta d(X) = \mathsf{E}_\theta [d(X) - \tau(\theta)]^2 = \mathsf{E}_\theta [(d(X) - \varphi(T)) + (\varphi(T) - \tau(\theta))]^2 =$$
$$= \mathsf{E}_\theta (d - \varphi)^2 + \mathsf{E}_\theta (\varphi - \tau)^2 + 2\mathsf{E}_\theta (d - \varphi)(\varphi - \tau) = \mathsf{E}_\theta (d - \varphi)^2 + \mathsf{D}_\theta \varphi,$$

поскольку

$$\mathsf{E}_\theta (d - \varphi)(\varphi - \tau) = \mathsf{E}_\theta \mathsf{E}[(d - \varphi)(\varphi - \tau)|T] = \mathsf{E}_\theta (\varphi - \tau)\mathsf{E}[(d - \varphi)|T] = 0,$$

ибо $\mathsf{E}_\theta [(d(T) - \varphi(X))|T] = \mathsf{E}(d|T) - \mathsf{E}(\varphi|T) = \varphi - \varphi = 0$. (Последнее равенство выполняется с вероятностью 1 для каждого распределения P_θ).

Равенство в (b) достигается, если и только если

$$\mathsf{E}_\theta [d(T) - \varphi(T)]^2 = 0 \quad \forall\, \theta.$$

Это возможно, если и только если

$$d(X) = \varphi(T) \qquad \text{с вероятностью 1}$$

для всех P_θ распределений.

3° Многомерный случай: пусть $d(X)$, $\tau(\theta)$ принимают значения в $\mathbb{R}^p$, записываем их в виде столбцов, $\mathsf{D}_\theta d < \infty$.

Пусть $z \in \mathbb{R}^p$, z — произвольный дисперсионный вектор. Рассмотрим скалярные величины:

$$\xi = \xi(X) := z^T d(X),$$
$$\eta = \eta(T) := \mathsf{E}[\xi(X)|T] = z^T \mathsf{E}[d(X)|T] = z^T \varphi(T),$$
$$t = t(\theta) := z^T \tau(\theta)$$

Ясно, что $\mathsf{E}_\theta \xi(X) = t(\theta) = \mathsf{E}_\theta \eta(T)$. По одномерной теореме Блекквелла – Рао

$$\mathsf{D}_\theta \eta(T) \leqslant \mathsf{D}_\theta \xi(X).$$

Отсюда

$$\mathsf{D}_\theta (z^T \varphi) \leqslant \mathsf{D}_\theta z^T d(X) \quad \Leftrightarrow \quad z^T (\mathsf{D}_\theta \varphi)\, z \leqslant z^T (\mathsf{D}_\theta d)\, z \quad \Leftrightarrow \quad \mathsf{D}_\theta \varphi \leqslant \mathsf{D}_\theta d,$$

что и требовалось доказать. Равенство будет, если

$$P_\theta [\eta(T) = \xi(X)] = 1$$

или

$$P_\theta \{ z^T \big(\varphi(T) - d(X)\big) = 0 \} \qquad \forall\, \theta \in \Theta \text{ и } \forall\, z \in \mathbb{R}^p.$$

■

2.7.4. Полные достаточные статистики

Из теоремы Блеквелла – Рао можно сделать, по меньшей мере, два вывода:

- Эта теорема дает способ улучшить несмещенную оценку, если мы такой оценкой уже располагаем;
- Она говорит, что при поиске наилучшей несмещенной оценки можно ограничить себя функциями от достаточной статистики. Если такая (зависящая от достаточной статистики) несмещенная оценка единственна, то она автоматически оказывается наилучшей.

Единственность зависящей от достаточной статистики несмещенной оценки обеспечивается так называемой полнотой достаточной статистики.

Определение Достаточная статистика $T = T(X)$ называется *полной*, если уравнение относительно функции f

$$\mathsf{E}_\theta f(T) = 0 \qquad \text{для всех } \theta \in \Theta$$

имеет только тривиальное $f \equiv 0$ решение.

Полнота очевидно является свойством семейства распределений статистики X. Поэтому часто говорят о полных семействах распределений (зависящих от θ, $\theta \in \Theta$).

Теорема(Леман, Шефаре, 1955)

Если $T = T(X)$ — полная достаточная статистика и $\varphi = \varphi(T)$ — несмещенная оценка θ, $\theta \in \Theta$, то $\varphi(T(X))$ — наилучшая несмещенная оценка $\tau(\theta)$.

Доказательство

Достаточно доказать единственность такой оценки φ.

Предположим, что существует другая (отличная от $\varphi(T)$) несмещенная оценка $\psi(T)$, так что

$$\mathsf{E}_\theta \psi(T) = \mathsf{E}_\theta \varphi(T) = \tau(\theta) \qquad \text{для всех } \theta \in \Theta$$

В этом случае

$$\mathsf{E}_\theta[\psi(T) - \varphi(T)] = 0 \qquad \text{для всех } \theta \in \Theta$$

Поскольку статистика T — полная, отсюда следует, что

$$\psi(T) - \varphi(T) = 0$$

почти наверное, для всех $\theta \in \Theta$.

Т.е. оценка φ единственна (с точностью до множества меры нуль), что и требовалось доказать. □

Пример 1. Испытания Бернулли.

Число успехов S_n (частота) в n испытаниях Бернулли является полной достаточной статистикой для вероятности успеха θ, когда эта вероятность θ рассматривается как неизвестный параметр, $\theta \in (0, 1)$.

Как известно, распределение S_n является биномиальным:

$$P_\theta(S_n = m) = C_n^m \theta^m (1-\theta)^{n-m} \qquad \text{для } m = \overline{0, n}.$$

Поэтому речь идет о полноте семейства биномиальных распределений, зависящих от параметра θ, $\theta \in (0, 1)$. Рассмотрим уравнение относительно $f(\cdot)$:

$$\forall \theta \in (0,1) \qquad \mathsf{E}_\theta f(S) = 0 \tag{$*$}$$

В данном случае функция $f(\cdot)$ должна быть определена на множестве $(0, 1, 2, \ldots, n)$, так что можно говорить о последовательности $f(0), f(1), \ldots, f(n)$.

Уравнение $(*)$ имеет вид:

$$\forall \theta \in (0,1) \qquad \sum_{m=0}^{n} C_n^m f(m) \theta^m (1-\theta)^{n-m} = 0. \tag{$**$}$$

Введем переменную $z = \frac{\theta}{1-\theta}$. Очевидно, что $z \in (0, \infty)$ и пробегает это множество, когда θ пробегает множество $(0, 1)$. Сократив $(**)$ на множитель $(1-\theta)^n$, получаем уравнение для последовательности $f(0), f(1), \ldots, f(n)$, т. е. для функции $f(\cdot)$:

$$\sum_{m=0}^{n} C_n^m f(m) z^m = 0, \qquad z \in (0, \infty).$$

Многочлен (от z) степени n может тождественно (на открытом множестве) обращаться в нуль, только если все его коэффициенты равны нулю.

Отсюда следует, что $f(0) = f(1) = \ldots = f(n) = 0$.

Таким образом, уравнение $(*)$ имеет лишь тривиальное решение, т.е. статистика S_n полная.

Получили, что частота $\frac{S_n}{n}$ является для испытаний Бернулли наилучшей несмещённой оценкой вероятности успеха.

Пример 2. Выборка из показательного распределения.

Пусть $x_1, \ldots, x_n$ — выборка из распределения с плотностью

$$p(x,\theta) = \begin{cases} \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right), & \text{для } x \geqslant 0 \\ 0, & \text{для } x < 0, \end{cases}$$

где $\theta \in (0,\infty)$ — неизвестный параметр. Нам уже известно, что $T = \sum\limits_{i=1}^{n} x_i$ является достаточной статистикой для θ. Покажем, что статистика T — полная.

Нетрудно показать, что T имеет плотность, задаваемую формулой:

$$q_n(x,\theta) = \left(\frac{1}{\theta}\right)^n \frac{x^{n-1}}{(n-1)!}\exp\left(-\frac{x}{\theta}\right) \qquad \text{для } x \geqslant 0.$$

Это распределение называют *гамма-распределением*, в котором θ служит масштабным параметром. (Случайная величина T по распределению совпадает со случайной величиной $\theta\gamma$, где случайная величина γ имеет так называемое «стандартное» гамма – распределение с плотностью

$$\frac{x^{n-1}}{(n-1)!}\exp(-x) \qquad \text{для } x \geqslant 0,$$

где n может принимать натуральные значения).

Полнота статистики T означает полноту относительно θ семейства гамма-распределений.

Рассмотрим уравнение

$$\mathsf{E}_\theta f(T) = 0 \qquad \text{для всех } \theta > 0$$

или

$$\int\limits_0^\infty f(x)\frac{x^{n-1}}{(n-1)!}\frac{1}{\theta^n}e^{-\frac{x}{\theta}}dx = 0 \qquad \text{для } \theta > 0$$

Введем новую переменную $t = \frac{1}{\theta}$; после сокращений получим уравнение

$$\int\limits_0^\infty \widetilde{f}(x)e^{-tx}dx = 0 \qquad \text{для всех } t > 0$$

$$(\text{где } \widetilde{f}(x) = x^{n-1}f(x)).$$

Левая часть этого уравнения — это преобразование Лапласа функции $\widetilde{f}(\cdot)$. Оно тождественно (относительно t) равно нулю только для $\widetilde{f}(\cdot) = 0$, или, что эквивалентно, $f(\cdot) = 0$. Отсюда следует, что статистика T — полная.

Пример 3.

Пусть $\{P_\theta, \theta \in \Theta\}$ — k-параметрическое экспоненциальное семейство распределений, где плотность

$$p(x,\theta) = \left\{\exp\left(\sum_{j=1}^{k} c_j(\theta)T_j(x) + d(\theta) + S(x)\right)\right\}I_A(x) \qquad (***)$$

По теореме факторизации $T(X) = (T_1(X), \ldots, T_k(X))$ есть достаточная статистика для θ, $\theta \in \Theta$.

Теорема

Если область значений векторной функции $(c_1(\theta), \ldots, c_k(\theta))$, которую она заполняет, когда θ пробегает параметрическое множество Θ, содержит какое-либо открытое множество, то статистика T — полная. (Семейство распределений с плотностями $(***)$ полное).

Доказательства этой теоремы мы не приводим. Оно может быть основано на свойствах преобразований Лапласа и Фурье (на обратимости этих преобразований), подобно Примеру 2.

Из этой теоремы можно извлечь много результатов, относящихся ко многим известным семействам распределений. В частности, утверждения Примеров 1 и 2. Еще одним следствием этой теоремы является полнота статистики $(\overline{x}, s^2)$, достаточной дл я параметров нормального распределения $N(a, \sigma^2)$ в случае выборки из этого распределения.

Пример 4. Линейная гауссовская модель.

Линейная гауссовская модель $X \sim N(l, \sigma^2 I), l \in L, L$ — задано. Следствием приведенной выше теоремы является утверждение о полноте достаточной статистики $(\mathrm{proj}_L X, |\mathrm{proj}_{L^\perp} X|^2)$ для $(l, \sigma^2 I)$.

3. Условное математическое ожидание

3.1. Сведения из других курсов

3.1.1. Вероятностное пространство и случайные величины

- *Вероятностной моделью*, или *вероятностным пространством* называют набор $(\Omega, \mathcal{A}, P)$, где Ω — это множество точек Ω; $\mathcal{A}$ — σ-алгебра подмножеств из Ω, а P — вероятностная мера на $\mathcal{A}$.
- Множество Ω называют *пространством элементарных исходов* (или элементарных событий).
- Множества из $\mathcal{A}$ называют *исходами* или *событиями*.
- Множество $A \subset \Omega$ называют $\mathcal{A}$-*измеримым*, если $A \in \mathcal{A}$.
- Для всякого A из $\mathcal{A}$ значение функции P на A, т.е. величину $P(A)$, называют *вероятностью события* A.

На числовой прямой выделяют σ-алгебру борелевских множеств $\mathcal{B}$. Это минимальная σ-алгебра подмножеств числовой прямой, которая содержит произвольные интервалы, полуинтервалы и отрезки числовой прямой.

- Действительная функция $\xi = \xi(\Omega)$, определенная на Ω, называется *случайной величиной*, если множества вида
$$\{\Omega : \xi(\Omega) \in B\} \tag{1}$$
являются событиями (т.е. принадлежат $\mathcal{A}$) для любых борелевских множеств B, $B \in \mathcal{B}$.

Каждая случайная величина ξ определяет в пространстве Ω некоторую совокупность подмножеств, образующих σ-алгебру, далее обозначаемую как A_ξ, состоящую из событий вида (1), когда B пробегает множество $\mathcal{B}$.

3.1.2. Производная Радона – Никодима

Пусть на некоторой σ-алгебре $\mathcal{F}$ подмножеств из Ω заданы меры μ и λ.

- Меру λ называют *абсолютно непрерывной* относительно меры μ, если из равенства $\mu(A) = 0$ следует, что и $\lambda(A) = 0$ (для множеств A из $\mathcal{F}$).
- Меру μ называют σ-*конечной*, если Ω можно представить в виде объединения счетной совокупности измеримых множеств, μ-меры которых конечны, т.е., если
$$\Omega = \bigcup_{i=1}^{\infty} A_i,$$
причем $\mu(A_i) < \infty, i = 1, 2, \ldots$

Теорема Радона-Никодима

Предположим, что на измеримом пространстве $(\Omega, \mathcal{F})$ задана σ-конечная мера μ и мера λ, абсолютно непрерывная относительно μ.

Тогда существует $\mathcal{F}$-измеримая функция $f(\Omega)$ такая, что для всякого $A \in \mathcal{F}$
$$\lambda(A) = \int_A f(\Omega)\mu(d\Omega).$$

С точностью до множества μ-меры нуль, функция $f(\Omega)$ единственная.

- Функцию $f(\Omega)$ называют *производной Радона – Никодима* меры λ по мере μ, или *плотностью* меры λ относительно меры μ:

$$f(\Omega) = \frac{d\lambda}{d\mu}(\Omega)$$

3.2. Определение условного математического ожидания

Пусть на вероятностном пространстве $(\Omega, \mathcal{A}, P)$ заданы две случайные величины $X = X(\Omega)$ и $Y = Y(\Omega)$.

Мы хотим определить математическое ожидание X при данном Y, в дальнейшем обозначаемое как $\mathsf{E}(X|Y)$.

Введем несколько более общее определение условного математического ожидания X относительно произвольной σ-подалгебры данной нам σ-алгебры $\mathcal{A}$. Это математическое ожидание мы затем свяжем с $\mathsf{E}(X|Y)$.

Пусть $\mathcal{G}$ — некоторая σ-подалгебра σ-алгебры $\mathcal{A}$. (Это означает, что, если множество A входит в $\mathcal{G}$, оно также входит и в $\mathcal{A}$). Определим условное математическое ожидание X относительно $\mathcal{G}$, в дальнейшем обозначаемое как $\mathsf{E}(X|\mathcal{G})$.

Представим X в виде

$$X = X^+ - X^-,$$

где $X^+ \geqslant 0$, $X^- \geqslant 0$.

Определим $\mathsf{E}(X^+|\mathcal{G})$ и $\mathsf{E}(X^-|\mathcal{G})$ и затем положим по определению:

$$\mathsf{E}(X|\mathcal{G}) = \mathsf{E}(X^+|\mathcal{G}) - \mathsf{E}(X^-|\mathcal{G}), \tag{1}$$

если хотя бы одно из этих условных математических ожиданий конечно.

Таким образом, $\mathsf{E}(X|\mathcal{G})$ может принимать значения $+\infty$ или $-\infty$. (Такую возможность имеет и $\mathsf{E}X$ при этом способе определения.) Впрочем, можно ограничиться случаем, когда $\mathsf{E}|X| < \infty$.

Итак, надо определить $\mathsf{E}(X|\mathcal{G})$ для $X \geqslant 0$.

На σ-алгебре $\mathcal{G}$ рассмотрим две меры: $P(\cdot)$ и $Q(\cdot)$, положив для произвольного $A \in \mathcal{G}$

$$Q(A) = \int\limits_A XP(d\Omega) \tag{2}$$

Ясно, что мера Q абсолютно непрерывна относительно меры P. Поэтому, по теореме Радона-Никодима, существует функция $f = f(\Omega)$, измеримая относительно $\mathcal{G}$ и такая, что

$$Q(A) = \int\limits_A f(\Omega)P(d\Omega) \tag{3}$$

Функцию $f(\Omega)$ из (3) назовем *условным математическим ожиданием* X (здесь $X \geqslant 0$) *относительно σ-алгебры* $\mathcal{G}$, т.е.:

$$\mathsf{E}(X|\mathcal{G})(\Omega) = f(\Omega).$$

Определив $\mathsf{E}(X^+|\mathcal{G})$ и $\mathsf{E}(X^-|\mathcal{G})$, по формуле (1) определим $\mathsf{E}(X|\mathcal{G})$ для произвольной случайной величины X.

Таким образом, $\mathsf{E}(X|\mathcal{G})$ — это случайная величина, измеримая относительно σ-алгебры $\mathcal{G}$. Она определена единственным образом, с точностью до множеств нулевой вероятности.

Пусть сейчас $\mathcal{G} = \mathcal{A}_Y$. Так как $\mathsf{E}(X|\mathcal{A}_Y)$ измерима относительно $\mathcal{A}_Y$, как функция от Ω, эта случайная величина с вероятностью 1 постоянна на множествах вида $\{\Omega : Y(\Omega) = \text{const}\}$. Поэтому $\mathsf{E}(X|\mathcal{A}_Y)$ можно рассматривать как функцию от $Y = Y(\Omega)$, и, по определению, можно положить

$$\mathsf{E}(X|Y) = \mathsf{E}(X|\mathcal{A}_Y).$$

3.3. Некоторые свойства условного математического ожидания

1. $$\int\limits_A \mathsf{E}(X|\mathcal{G})P(d\Omega) = \int\limits_A XP(d\Omega)$$

   для всякого $A \in \mathcal{G}$.

   Это свойство — всего лишь другая запись определения (3.2.3).

   Заметим различие между X и $\mathsf{E}(X|\mathcal{G})$: случайная величина X, вообще говоря, не измерима относительно $\mathcal{G}$ (она измерима относительно более «богатой» σ-алгебры $\mathcal{A}$, $\mathcal{G} \subset \mathcal{A}$).

2.
$$\mathsf{E}\mathsf{E}(X|\mathcal{G}) = \mathsf{E}X.$$

Для доказательства надо положить $A = \Omega$ в свойстве 1.

Тогда:
$$\mathsf{E}[\mathsf{E}(X|\mathcal{G})] = \int\limits_{\Omega} \mathsf{E}(X, \mathcal{G})\, dP = \int\limits_{\Omega} X\, dP = \mathsf{E}X,$$

что и требовалось.

3. Линейное свойство:
$$\mathsf{E}(aX + bY|\mathcal{G}) = a\mathsf{E}(X|\mathcal{G}) + b\mathsf{E}(Y|\mathcal{G})$$

для произвольных случайных величин X, Y и постоянных a, b. При этом левая часть существует, если существует правая часть.

Для доказательства достаточно показать, что для любого $A \in \mathcal{G}$
$$\int\limits_{A} \mathsf{E}(aX + bY|\mathcal{G})\, dP = \int\limits_{A} [a\mathsf{E}(X|\mathcal{G}) + b\mathsf{E}(Y|\mathcal{G})]\, dP \tag{1}$$

и что $a\mathsf{E}(X|\mathcal{G}) + b\mathsf{E}(Y|\mathcal{G})$ измеримо относительно $\mathcal{G}$. Последнее, впрочем, очевидно.

Преобразуем левую часть (1):
$$\int\limits_{A} \mathsf{E}(aX + bY|\mathcal{G})\, dP = \int\limits_{A} [aX + bY]\, dP =$$
$$a\int\limits_{A} [\mathsf{E}(X|\mathcal{G})]\, dP + b\int\limits_{A} [\mathsf{E}(Y|\mathcal{G})]\, dP = \int\limits_{A} [a\mathsf{E}(X|\mathcal{G}) + b\mathsf{E}(Y|\mathcal{G})]\, dP,$$

что и требовалось.

4. Если X измерима относительно $\mathcal{G}$, то $\mathsf{E}(X|\mathcal{G}) = X$.

В частности,
$$\mathsf{E}(X|\mathcal{A}_X) = X.$$

5. Если X и Y независимы, то
$$\mathsf{E}(X|Y) = \mathsf{E}X.$$

Для доказательства достаточно проверить, что для любого $A \in \mathcal{A}_Y$:
$$\int\limits_{A} \mathsf{E}(X|Y)\, dP = \int\limits_{A} (\mathsf{E}X)\, dP \tag{2}$$

Обозначим через $I_A = I_A(\Omega)$ индикаторную функцию множества A.

Как случайная величина, I_A измерима относительно $\mathcal{A}_Y$. При этом случайные величины X и I_A независимы, ибо независимы две σ-алгебры $\mathcal{A}_X$ и $\mathcal{A}_Y$.

Преобразуем левую часть (2), заметив предварительно, что правая часть (2) равна $(\mathsf{E}X)P(A)$.

Имеем:
$$\int\limits_{A} \mathsf{E}(X|Y)\, dP = \int\limits_{A} X\, dP = \mathsf{E}[XI_A] = (\mathsf{E}X)\mathsf{E}I_A = (\mathsf{E}X)P(A)$$

(в силу независимости X и I_A), что и требовалось.

6. Условные вероятности.

Как мы только что вспомнили, $P(A) = \mathsf{E}I_A$. По аналогии с этим равенством, условную вероятность события A относительно σ-алгебры $\mathcal{G}$ определим как $P(A|\mathcal{G}) = \mathsf{E}(I_A|\mathcal{G})$. Соответственно, условная вероятность события A относительно случайной величины Y (при данном Y) есть $P(A|Y) := P(A|\mathcal{A}_Y)$.

7. Условные распределения.

Напомним, что *распределением* случайной величины X мы называем совокупность вероятностей вида

$$P_X(B) := P(X \in B), B \in \mathcal{B},$$

когда B пробегает σ-алгебру борелевских множеств числовой прямой. При этом $P_X(B)$ как функция $B \in \mathcal{B}$ образует на $\mathcal{B}$ вероятностную меру.

По аналогии с этим условным распределением случайной величины X относительно σ-алгебры $\mathcal{G}$ естественно называть совокупность условных вероятностей

$$P_X(B|\mathcal{G}) := P(X \in B \mid \mathcal{G})(\Omega), B \in \mathcal{B}. \tag{3}$$

Из дальнейших свойств условного математического ожидания будет следовать, что с вероятностью 1 эти условные распределения вероятностей σ-аддитивны.

Не следует забывать, что (3) — это случайная величина, определенная с точностью до множества меры нуль. Можно показать, что существует такой вариант ее определения, что (3) как функция B, $B \in \mathcal{B}$, с вероятностью 1 образует на $\mathcal{B}$ (случайную) вероятностную меру. В этом случае

$$\mathsf{E}(X|\mathcal{G})(\Omega) = \int X(\Omega')P_X(d\Omega'|\mathcal{G})(\Omega)$$

почти наверное.

Впрочем, в простой ситуации, которую мы рассмотрим в следующем параграфе, мы определим условное математическое ожидание, отправляясь от условного распределения. Подобно тому, как математическое ожидание случайной величины мы обычно вводим, отправляясь от распределения.

3.4. Случай простых случайных величин

В этом параграфе мы рассмотрим $\mathsf{E}(X|Y)$ для простых случайных величин X и Y. В этом случае условное математическое ожидание можно ввести элементарными средствами.

Случайная величина Y называется *простой*, если Y можно представить в виде

$$Y = \sum y_j I(D_j), \tag{1}$$

где $I(D) = I_D(\Omega)$ — индикаторная функция множества D. (Для удобства обозначения $I(D)$ предпочтительнее здесь, чем $I_D = I_D(\Omega)$.) Можно считать, что числа $y_1, y_2, \ldots$ различны и что совокупность множеств D_j, $j = 1, 2, \ldots$ в (1) образует разбиение пространства Ω: $D_j \bigcap D_i = \varnothing$, если $j \neq i$; $\bigcup\limits_j D_j = \Omega$. Когда случайная величина Y простая, то порожденная ею σ-алгебра $\mathcal{A}_Y$ порождается разбиением $D_1, D_2, \ldots$. (Здесь D_j, $j = 1, 2, \ldots$ — это множества уровня функции $Y = Y(\Omega)$, $D_j = \{\Omega\colon Y(\Omega) = y_j\}$.)

Далее мы будем рассматривать σ-алгебры, порожденные конечными (или счетными) разбиениями.

Пусть $\mathcal{G}$ — такая σ-алгебра. Порождающее ее разбиение обозначим, как и выше, через $D_1, D_2, \ldots$

Пусть X — простая случайная величина. Тогда для $\mathsf{E}(X|\mathcal{G})$ можно дать элементарное определение.

Начнем с определения условной вероятности.

Положим по определению для всякого $A \in \mathcal{A}$

$$P(A|\mathcal{G}) = P(A|\mathcal{G})(\Omega) = \sum_j P(A|D_j)I(D_j). \tag{2}$$

Ясно, что $P(A|\mathcal{G})$ есть измеримая относительно $\mathcal{G}$ случайная величина.

Главное свойство условной вероятности (2):

$$\mathsf{E}P(A|\mathcal{G}) = P(A). \tag{3}$$

Доказательство очевидно :

$$\mathsf{E}P(A|\mathcal{G}) = \sum_j P(A|D_j)\mathsf{E}I(D_j) = \sum_j P(A|D_j)P(D_j) = P(A).$$

Пусть

$$X = \sum_i x_i I(A_i). \tag{4}$$

По аналогии с $\mathsf{E}X = \sum\limits_i x_i P(A_i)$, определим $\mathsf{E}(X|\mathcal{G})$ формулой:

$$\mathsf{E}(X|\mathcal{G}) = \sum_i x_i P(A_i|\mathcal{G}). \tag{5}$$

Отметим, что так определенное $\mathsf{E}(X|\mathcal{G})$ — измеримая относительно $\mathcal{G}$ случайная величина и что

$$\mathsf{E}\mathsf{E}(X|\mathcal{G}) = \mathsf{E}X. \tag{6}$$

Доказательство (6) очевидно:

$$\mathsf{E}\mathsf{E}(X|\mathcal{G}) = \sum_i x_i \mathsf{E}P(A_i|\mathcal{G}) = \sum_i x_i P(A_i).$$

Покажем, что определение (5) совпадает с общим определением математического ожидания из пункта 3.2. Для этого достаточно проверить, что для любого $B \in \mathcal{G}$:

$$\int\limits_B \mathsf{E}(X|\mathcal{G})\, dP = \int\limits_B X\, dP. \tag{7}$$

Так как $B \in \mathcal{G}$, то B можно представить в виде объединения некоторой совокупности множеств D_j:

$$B = \sum_{j\in K} D_j,$$

где K — некоторое множество индексов.

Далее заметим, что

$$\int\limits_B \mathsf{E}(X|\mathcal{G})\, dP = \sum_{j\in K} \int\limits_{D_j} \mathsf{E}(X|\mathcal{G})\, dP,$$

$$\int\limits_B X\, dP = \sum_{j\in K} \int\limits_{D_j} X\, dP.$$

Поэтому (7) достаточно доказать для множеств D_j, $j = 1, 2, \ldots$

Итак, положив $B = D_k$, преобразуем левую часть (7), используя (5) и (2):

$$\int\limits_{D_k} \mathsf{E}(X|\mathcal{G})\, dP = \sum_i x_i \int\limits_{D_k} P(A_i|\mathcal{G})\, dP = \sum_i x_i \sum_j P(A_i|D_j)\mathsf{E}I(D_k)I(D_j)$$

$$= \sum_i x_i P(A_i|D_k)P(D_k) = \sum_i x_i P(A_i D_k).$$

Преобразование правой части (7) дает тот же результат:

$$\int\limits_{D_k} X\, dP = \mathsf{E}I(D_k)[\sum_i x_i I(A_i)] = \sum_i x_i \mathsf{E}I(A_i D_k) = \sum_i x_i P(A_i D_k),$$

что и требовалось.

Усреднением набора чисел $x_1, \ldots, x_n$ с неотрицательными весами $p_1, \ldots, p_n$, $\sum\limits_{i=1}^n p_i = 1$ называют число $\sum\limits_{i=1}^n x_i p_i$. (С вероятностной точки зрения, усреднение — это математическое ожидание случайной величины, принимающей значения $x_1, \ldots, x_n$ с вероятностями $p_1, \ldots, p_n$.)

Покажем, что значения, которые принимает случайная величина $\mathsf{E}(X|\mathcal{G})$, суть усреднения значений X.

Действительно,

$$\mathsf{E}(X|\mathcal{G}) = \sum_i x_i P(A_i|\mathcal{G}) = \sum_i x_i \sum_j P(A_i|D_j) I(D_j) = \sum_j \left[\sum_i x_i P(A_i|D_j)\right] I(D_j).$$

На множестве D_j случайная величина $\mathsf{E}(X|\mathcal{G})$ принимает значение

$$y_j = \sum_i x_i P(A_i|\mathsf{D}_j).$$

Отметим, что $P(A_i|D_j) \geqslant 0$ и что $\sum\limits_i P(A_i|D_j) = 1$, ибо $A_1, A_2, \ldots$ — это разбиение всего пространства. Таким образом, y_j — это усреднение набора $x_1, \ldots, x_n$ значений, принимаемых X, с весами $p_i = P(A_i|D_j)$.

3.5. Вынесение множителя, постоянного при данном условии

Следующее свойство условных математических ожиданий — возможность вынести за знак математического ожидания случайный множитель, постоянный при данном условии:

$$\mathsf{E}[\varphi(Y)X|Y] \overset{\text{п.н.}}{=} \varphi(Y)\mathsf{E}(X|Y) \tag{1}$$

Предпочтительнее сформулировать это свойство в более общем виде:
Если Y измерима относительно $\mathcal{G}$, то

$$\mathsf{E}(XY|\mathcal{G}) \overset{\text{п.н.}}{=} Y\mathsf{E}(X|\mathcal{G}) \tag{2}$$

при условии, что эти математические ожидания существуют.

Доказательство этого равенства начнем с простых случайных величин.

3.5.1. Доказательство для случая простых случайных величин

Пусть Y — простая случайная величина, измеримая относительно σ-алгебры $\mathcal{G}$. Тогда верно (2).

Доказательство.
По предположению $Y = \sum\limits_i y_i I(B_i)$, причем $B_i \in \mathcal{B}$, $i = 1, 2, \ldots$ Теперь

$$\mathsf{E}(XY|\mathcal{G}) = \sum_i y_i \mathsf{E}(I(B_i)X|\mathcal{G}).$$

Чтобы получить (2), достаточно показать, что

$$\mathsf{E}(I(B)X|\mathcal{G}) = I(B)\mathsf{E}(X|\mathcal{G}),$$

если $B \in \mathcal{B}$.

Поскольку $I(B)\mathsf{E}(X|\mathcal{G})$ измерима относительно $\mathcal{G}$, для этого достаточно показать, что для любого $A \in \mathcal{G}$

$$\int\limits_A I(B)\mathsf{E}(X|\mathcal{G})\,dP = \int\limits_A I(B)X\,dP \tag{3}$$

Преобразуя левую часть, докажем, тем самым, (3):

$$\int\limits_A I(B)\mathsf{E}(X|\mathcal{G})\,dP = \int\limits_{A\cap B} \mathsf{E}(X|\mathcal{G})\,dP = \int\limits_{A\cap B} X\,dP = \int\limits_A I(B)X\,dP,$$

что и требовалось. □

3.5.2. Общий случай

Пусть Y измерима относительно $\mathcal{G}$, $\mathsf{E}|X| < \infty$, $\mathsf{E}|Y| < \infty$, $\mathsf{E}|XY| < \infty$. Тогда верно (2).

Доказательство. Основывается на пункте 3.5.1 и обобщенной теореме Лебега о мажорированной сходимости, которая будет дана позже.

Выбираем последовательность простых случайных величин Y_n так, чтобы $Y_n \uparrow Y$ п.н. при $n \to \infty$. В таком случае

$$\mathsf{E}(XY_n|\mathcal{G}) \overset{\text{п.н.}}{=} Y_n\mathsf{E}(X|\mathcal{G})$$

в силу 3.5.1.

По упомянутой теореме

$$\mathsf{E}(XY_n|\mathcal{G}) \xrightarrow{\text{п.н.}} \mathsf{E}(XY|\mathcal{G})$$

Кроме того,

$$Y_n\mathsf{E}(X|\mathcal{G}) \xrightarrow{\text{п.н.}} Y\mathsf{E}(X|\mathcal{G})$$

Это и доказывает (2). □

Следствие:

$$\mathsf{E}[\varphi(Y)X|Y] \overset{\text{п.н.}}{=} \varphi(Y)\mathsf{E}(X|Y).$$

Лемма (обобщенная теорема Лебега о мажорируемой сходимости).

Пусть $|\alpha_n| \leqslant \eta$, $\mathsf{E}\eta < \infty$ и $\alpha_n \overset{\text{п.н}}{\to} \alpha$ при $n \to \infty$.

Тогда

(a)
$$\mathsf{E}(\alpha_n|\mathcal{G}) \overset{\text{п.н.}}{\longrightarrow} \mathsf{E}(\alpha|\mathcal{G})$$

(b)
$$\mathsf{E}(|\alpha_n - \alpha| \,\Big|\, \mathcal{G}) \overset{\text{п.н.}}{\longrightarrow} 0$$

Сравним с (обычной) теоремой Лебега (о мажорированной сходимости):

Пусть $|\xi_n| \leqslant \eta$, $\mathsf{E}\eta < \infty$ и $\xi_n \overset{\text{п.н}}{\to} \xi$ при $n \to \infty$.

Тогда

(a)
$$\mathsf{E}\xi_n \overset{\text{п.н.}}{\longrightarrow} \mathsf{E}\xi \text{ } (\mathsf{E}\xi \text{ существует})$$

(b)
$$\mathsf{E}(|\xi_n - \xi|) \overset{\text{п.н.}}{\longrightarrow} 0$$

Доказательство.
Положим
$$\xi_n := \sup_{m:m \geqslant n} |\alpha_m - \alpha|.$$
Ясно, что $\xi_n \geqslant |\alpha_n - \alpha|$.
Так как $\alpha_n \overset{\text{п.н}}{\to} \alpha$, то $\xi_n \downarrow 0$ п.н.
Теперь
$$|\mathsf{E}(\alpha_n|\mathcal{G}) - \mathsf{E}(\alpha|\mathcal{G})| = |\mathsf{E}\big[(\alpha_n - \alpha) \,\Big|\, \mathcal{G}\big]| \leqslant \mathsf{E}(|\alpha_n - \alpha| \,\Big|\, \mathcal{G}) \leqslant \mathsf{E}(\xi_n|\mathcal{G}). \qquad (*)$$
Докажем, что
$$\mathsf{E}(\xi_n|\mathcal{G}) \overset{\text{п.н.}}{\longrightarrow} 0.$$
Из этого вытекает утверждение леммы.
Заметим, что
$$0 \leqslant \mathsf{E}(\xi_{n+1}|\mathcal{G}) \leqslant \mathsf{E}(\xi_n|\mathcal{G}) \text{ п.н.}$$
Поэтому существует предел (почти наверное):
$$h := \lim_{n\to\infty} \mathsf{E}(\xi_n|\mathcal{G}) \geqslant 0$$
Далее,
$$0 \leqslant \int\limits_\Omega h \; dP \leqslant \int\limits_\Omega \mathsf{E}(\xi_n|\mathcal{G}) \; dP = \int\limits_\Omega \xi_n \; dP = \mathsf{E}\xi_n \longrightarrow 0.$$
Последнее заключение есть следствие цитированной теоремы Лебега, ибо
$$0 \leqslant \xi_n \leqslant 2\beta, \mathsf{E}\beta < \infty, \xi_n \overset{\text{п.н.}}{\longrightarrow} 0.$$
Получили, что $\int\limits_\Omega h \, dP = 0$.
Т.к. $h \geqslant 0$, то $h = 0$ п.н. Следовательно:
$$\mathsf{E}(\xi_n|\mathcal{G}) \overset{\text{п.н.}}{\longrightarrow} 0$$
Это и доказывает лемму. □

3.6. σ-аддитивность условной вероятности

Пусть $A = \sum\limits_i A_i$, причем $A_i \bigcap A_j = \emptyset$, если $i \neq j$.

Тогда
$$P(A|\mathcal{G}) \overset{\text{п.н.}}{=} \sum_i P(A_i|\mathcal{G}).$$

Для доказательства достаточно положить в предыдущей лемме $\alpha_n = \sum\limits_{i=1}^{n} I(A_i)$, $\alpha = I(A)$ и заметить, что $\alpha_n \uparrow \alpha$ при $n \to \infty$. Прочие условия леммы тоже соблюдены.

3.7. Условная дисперсия

По аналогии с определением дисперсии $\mathsf{D}X = \mathsf{E}(X - \mathsf{E}X)^2$, введем условную дисперсию X относительно $\mathcal{G}$, положив, по определению,

$$\mathsf{D}(X|\mathcal{G}) = \mathsf{E}\{[X - \mathsf{E}(X|\mathcal{G})]^2|\mathcal{G}\}.$$

Покажите, что

$$\mathsf{D}X = \mathsf{E}\mathsf{D}(X|\mathcal{G}) + \mathsf{D}\mathsf{E}(X|\mathcal{G}).$$

(при условии, что $\mathsf{D}X$ существует).

3.8. Наилучший квадратичный прогноз

(Формулируется в виде задачи.)

Пусть случайные величины ξ и η заданы на одном вероятностном пространстве. Надо найти для η наилучший прогноз по наблюдаемой случайной величине ξ. Иначе говоря, надо найти такую функцию $f(\xi)$, что для любой функции $g(\xi)$:

$$\mathsf{E}(\eta - f(\xi))^2 \leqslant \mathsf{E}(\eta - g(\xi))^2.$$

Ответ:

$$f(\xi) = \mathsf{E}(\eta|\xi).$$

3.9. Пример вычисления условного математического ожидания

Рассмотрим пример одновременно типичный и вычислительно несложный. Пусть вероятностная тройка $(\Omega, \mathcal{A}, P)$ такова:

- $\Omega = \{\Omega : \Omega = (x, y), 0 \leqslant x \leqslant 1, 0 \leqslant y \leqslant 1\}$;
- $\mathcal{A}$ — σ-алгебра борелевских множеств Ω;
- P — мера Лебега на Ω.

Рассмотрим две случайные величины $\xi = \xi(\Omega)$ и $\eta = \eta(\Omega)$:

$$\xi = \xi(x, y) = x, \eta = \eta(x, y) = x + y.$$

Вычислим $\mathsf{E}(\xi|\eta)$.

Отметим, что A_ξ (σ-алгебра подмножеств Ω, порожденная случайной величиной ξ) - это совокупность цилиндрических множеств из Ω вида $B \times [0, 1]$, где B - произвольное борелевское множество из [0,1].

Сигма-алгебра A_η устроена схожим образом. Ее составляют (пересеченные с Ω) прямые произведения борелевских множеств, лежащих на прямой $x = y$, и прямой $\{(x, y) : x + y = 0\}$. (См. чертеж)

Хорошо видно, что ξ не измерима относительно A_η, и обратно.

По определению, $\mathsf{E}(\xi|\eta)$ - такая измеримая относительно A_η функция $f(x, y)$ от $\Omega = (x, y)$, для которой

$$\iint\limits_{(x,y)\in A} f(x, y)\, dP = \iint\limits_{(x,y)\in A} x\, dP \tag{1}$$

для любого $A \in A_\eta$.

Так как $f(x, y)$ измерима относительно A_η, она должна зависеть от (x, y) через посредство $\eta = x + y$. Это означает, что в качестве $f(x, y)$ здесь следует взять, пока произвольную, функцию $g(x + y)$, где $g(\cdot)$ - измеримая функция одного переменного.

В (1) достаточно рассматривать только множества A вида

$$A = \{(x, y) : x + y \leqslant z, (x, y) \in \Omega\},$$

где z - произвольно.

При таком выборе $f(x, y)$ и A условие (1) примет вид:

$$\iint\limits_{\substack{x+y\leqslant z\\(x,y)\in\Omega}} g(x + y)\, dP = \iint\limits_{\substack{x+y\leqslant z\\(x,y)\in\Omega}} x dx dy \tag{2}$$

В интегралах (2) следует сделать замену переменных $(x, y) \to (u, v)$, положив $u = x + y$.

Выбор второй переменной не очень важен, положим, например, $v = x - y$.

После этой замены двойные интегралы в (2) представим в виде повторных.

Для простоты возьмем $z \in [0, 1]$. (Случай $z \in [1, 2]$ легко сводится к рассматриваемому.) Получим уравнение для $g(\cdot)$

$$\frac{1}{2}\int_0^z \left(g(u) \int_{-u}^{u} dv \right) du = \frac{1}{2}\int_0^z \left(\int_{-u}^{u} \frac{u+v}{2} dv \right) du.$$

Отсюда

$$\int_0^z ug(u)\, du = \int_0^z \frac{1}{2}u^2 du,$$

или

$$g(z) = \frac{z}{2}.$$

Таким образом, здесь

$$\mathsf{E}(\xi|\eta) = \eta/2,$$

или

$$\mathsf{E}(x|x+y) = \frac{x+y}{2}.$$

Заметим, что при вычислении $\mathsf{E}(X|X+Y)$, если X и Y независимы и одинаково распределены (как в рассмотренном выше примере), можно обойтись практически без вычислений, если вспомнить некоторые из перечисленных выше свойств условных математических ожиданий.

Во-первых, в силу симметрии,

$$\mathsf{E}(X|X+Y) = \mathsf{E}(Y|X+Y).$$

Затем

$$X + Y = \mathsf{E}(X+Y|X+Y) = \mathsf{E}(X|X+Y) + \mathsf{E}(Y|X+Y).$$

Отсюда

$$\mathsf{E}(X|X+Y) = \frac{1}{2}(X+Y).$$

4. Линейная гауссовская модель

В абстрактной форме - это статистическая модель о (векторном) наблюдении X, $X \in \mathbb{R}^n$, X - вектор-столбец, $X = (X_1, \ldots, X_n)^T$.

Предположим, что X - случайный вектор, распределенный по нормальному закону, причем математическое ожидание X, т.е. вектор $\mathsf{E}X$, принадлежит заданному линейному подпространству L, $L \subset \mathbb{R}^n$, а матрица ковариаций вектора X равна $\sigma^2 I$ (скалярная матрица).

Вектор $l := \mathsf{E}X$ и скаляр σ^2, $\sigma^2 > 0$ неизвестны. Короткая запись: наблюдаемый вектор X случаен и $X \sim N(l, \sigma^2 I)$, причем $l \in L$, где L - заданное линейное подпространство.

Статистические задачи в этой модели - выводы о неизвестных параметрах l и σ^2.

4.1. Несмещенное оценивание параметров

В лекциях о достаточных статистиках было сказано, что для параметра $\theta := (l; \sigma^2)$ в этой модели есть достаточная статистика. Это пара $T = (\mathrm{proj}_L X; |\mathrm{proj}_{L^\perp} X|^2)$.

Согласно примеру 2.7.4.4 эта статистика T полна.

Поэтому наилучшая (имеющая наименьшую матрицу ковариаций) несмещенная оценка параметра θ должна быть функцией достаточной статистики (такая оценка единственна).

Заметим, что $\mathsf{E}\,\mathrm{proj}_L X = \mathrm{proj}_L \mathsf{E}X = \mathrm{proj}_L l = l$, ибо:

- операцию усреднения (вычисления математического ожидания) и проектирования X можно поменять местами (проектирование X на подпространство - линейная операция, а усреднение обладает свойствами линейности);
- так как $l \in L$, то $\mathrm{proj}_L l = l$.

Следовательно, наилучшая несмещенная оценка l уже найдена - это $\mathrm{proj}_L X$.

Чтобы найти наилучшую несмещенную оценку σ^2, надо подробнее изучить статистические свойства $\mathrm{proj}_L X$ и $\mathrm{proj}_{L^\perp} X$.

4.2. χ^2-распределение

Определение 1 (χ^2-распределение). Пусть $\eta_1, \ldots, \eta_r$ суть независимые случайные величины, распределенные каждая по стандартному нормальному закону $N(0,1)$. Случайной величиной *хи-квадрат с r степенями свободы* называют

$$\chi^2(r) := \eta_1^2 + \ldots + \eta_r^2.$$

Распределение случайной величины $\chi^2(r)$ для любого натурального r может быть вычислено во всех подробностях (плотность, функция распределения, квантили и т.д.). Явный его вид нам не понадобится. Достаточно сказать, что таблицы распределений и квантилей есть в сборниках статистических таблиц.

Случайную величину $\chi^2(r)$ можно толковать как квадрат длины случайного вектора $\vec{\eta} = (\eta_1, \ldots, \eta_r) \sim N_r(0, I)$, составленного из независимых одномерных стандартных гауссовских величин $\eta_i \sim N(0,1), i = \overline{1,r}$.

Распределение $N_r(0,I)$ часто называют стандартным r-мерным гауссовским распределением, а вектор $\vec{\eta}$ - r-мерным стандартным гауссовским вектором.

Определение 2 (нецентральное χ^2-распределение). Пусть $\vec{a} = (a_1, \ldots, a_r)$ - заданный вектор. Рассмотрим случайную величину

$$\chi^2(r, \Delta) := (\eta_1 + a_1)^2 + \ldots + (\eta_r + a_r)^2$$

Здесь $\Delta = a_1^2 + \ldots + a_r^2$. Из леммы 4.3.1 (которую мы докажем в следующем разделе) следует, что распределение случайной величины $\sum\limits_{i=1}^{r} (\eta_i + a_i)^2$ зависит от $\Delta := |\vec{a}|^2$, но не от $\vec{a}$. Это обстоятельство отражено в обозначении $\chi^2(r, \Delta)$. Величина $\Delta = \sum\limits_{i=1}^{r} a_i^2$ называется *параметром нецентральности* распределения хи-квадрат. Если $\Delta = 0$, распределение хи-квадрат называют *центральным*.

Нецентрально распределенную случайную величину $\chi^2(r, \Delta)$ можно толковать как квадрат длины r-мерного гауссовского вектора $\vec{\eta} + \vec{a}$, причем $\Delta = |a|^2$.

Нетрудно показать, что семейство случайных величин $\chi^2(r, \Delta)$ стохастически упорядочено по параметру $\Delta, \Delta > 0$, если r - фиксировано.

Иными словами, если $0 \leqslant \Delta_1 \leqslant \Delta_2$, то для любого $z > 0$

$$P(\chi^2(r, \Delta_1) \geqslant z) \leqslant P(\chi^2(r, \Delta_2) \geqslant z)$$

(о доказательстве скажем позже.)

Графики функций распределения $y = P(\chi^2(r, \Delta_1) \geqslant z)$ при разных $\Delta > 0$ выглядят примерно так:

4.3. Две леммы о круговых нормальных распределениях

Лемма 4.3.1.

Пусть $X \sim N(l, \sigma^2 I)$, C - ортогональная матрица. Тогда

$$Y := CX \sim N(Cl, \sigma^2 I)$$

(Словесная форма: при ортогональных преобразованиях круговое нормальное распределение остается круговым.)

Доказательство.

Для доказательства достаточно вычислить матрицу ковариации вектора $Y = CX$. Поскольку для любой матрицы A матрица ковариаций вектора AX есть

$$\mathsf{D}(AX) = A(\mathsf{D}X)A^T$$

(где $\mathsf{D}\xi$ обозначает матрицу ковариаций вектора ξ), то

$$\mathsf{D}Y = C(\mathsf{D}X)C^T = C(\sigma^2 I)C^T = \sigma^2 I,$$

что и требовалось. □

Следствие. Пусть $\eta_1, \ldots, \eta_r$ суть независимые $N(0,1)$. Тогда:

$$(\eta_1 + a_1)^2 + \ldots + (\eta_r + a_r)^2 \stackrel{d}{=} (\eta_1 + \sqrt{\Delta})^2 + \ldots + \eta_r^2,$$

где $\Delta = a_1^2 + \ldots + a_r^2$.

Это утверждение доказывает правильность употребления выражения $\chi^2(r, \Delta)$ для распределения квадрата длины вектора $\vec{\eta} + \vec{a}$. Здесь $\vec{\eta} = (\eta_1, \ldots, \eta_r)$.

Доказательство.

Доказательство основывается на том, что вектор $\vec{\eta} + \vec{a}$ можно ортогональным преобразованием (скажем, C) перевести в вектор с координатами $(\tilde{\eta}_1 + \sqrt{\Delta}, \tilde{\eta}_2, \ldots, \tilde{\eta}_r)^T$. При ортогональных преобразованиях длина вектора не меняется; распределение $C\eta$, так же как и распределение η, есть $N(0, I)$. □

Лемма 4.3.2.

Пусть $L_1, L_2, \ldots$ - попарно ортогональные подпространства, прямая сумма которых составляет все пространство $\mathbb{R}^n$:

$$L_1 \oplus L_2 \oplus \ldots = \mathbb{R}^n.$$

Пусть $\mathrm{proj}_L X$ обозначает проекцию вектора X на подпространство L (в евклидовой метрике). Пусть, скажем, $X \sim N(l, \sigma^2 I)$. Тогда:

(a) Случайные векторы $\mathrm{proj}_{L_1} X, \mathrm{proj}_{L_2} X, \ldots$ независимы (в совокупности) и распределены нормально, причем $\mathsf{E}\,\mathrm{proj}_{L_i} X = \mathrm{proj}_{L_i} l, i = 1, 2, \ldots$;

(b)
$$|\,\mathrm{proj}_{L_i} X|^2 = \sigma^2 \chi^2(r_i, \Delta_i)$$
где $r_i = \dim L_i, \Delta_i = |\frac{1}{\sigma}\,\mathrm{proj}_{L_i} l|^2$.

Доказательство.

Рассмотрим в $\mathbb{R}^n$ новый ортонормированный базис, который строим, объединяя ортонормированные базисы подпространств $L_1, L_2, \ldots$

Ради определенности введем соответствующие обозначения:

$$f_1, f_2, \ldots, f_{r_1} - \text{базис } L_1;$$

$$f_{r_1+1}, f_{r_1+2}, \ldots, f_{r_1+r_2} - \text{базис } L_2;$$

$$\ldots \qquad \text{и т.д.}$$

Рассмотрим координаты вектора $X = (X_1, \ldots, X_n)$ в базисе $\{f\}$. Обозначим их через $Y_1, \ldots, Y_n$.

Как известно векторы-столбцы $Y = (Y_1, \ldots, Y_n)$ связаны с помощью матрицы C перехода соотношением $Y = CX$. Заметим, что C — ортогональная матрица, и поэтому $Y \sim N(Cl, \sigma^2 I)$.

Следовательно, случайные величины $Y_1, \ldots, Y_n$ независимы, распределены нормально и имеют одну и ту же дисперсию σ^2.

Заметим, что

$$\mathrm{proj}_{L_1} X = Y_1 f_1 + \ldots + Y_{r_1} f_{r_1},$$
$$\mathrm{proj}_{L_2} X = Y_{r_1+1} f_{r_1+1} + \ldots + Y_{r_1+r_2} f_{r_1+r_2},$$

и т.д.

Из этих представлений для $\mathrm{proj}_{L_i} X, i = 1, 2, \ldots, n$ и отмеченных свойств случайных величин $Y_1, Y_2, \ldots$ следует утверждение (a).

Для доказательства (b) заметим, что

$$|\,\mathrm{proj}_{L_1} X|^2 = Y_1^2 + \ldots + Y_{r_1}^2 =$$
$$\sigma^2 \left(\frac{Y_1^2}{\sigma^2} + \ldots + \frac{Y_{r_1}^2}{\sigma^2} \right) = \sigma^2 \chi^2(r_1, \Delta_1),$$

ибо $\frac{Y_1}{\sigma}, \ldots, \frac{Y_{r_1}}{\sigma}$ суть независимые случайные величины, распределенные по нормальному закону.

Параметр нецентральности - это квадрат длины математического ожидания вектора

$$\frac{1}{\sigma} (Y_1, Y_2, \ldots, Y_{r_1})^T.$$

По сказанному выше,

$$\mathsf{E} \left[\frac{1}{\sigma} (Y_1, Y_2, \ldots, Y_{r_1})^T \right] = \frac{1}{\sigma}\,\mathrm{proj}_{L_1} \mathsf{E}X.$$

Лемма 4.3.2 доказана. □

4.4. Линейная модель

Вернемся к линейной модели $X \sim N(l, \sigma^2 I)$, причем $l \in L$, где L задано. Для оценивания σ^2 рассмотрим вторую составляющую достаточной статистики: случайную величину $|\operatorname{proj}_{L^\perp} X|^2$.

Согласно лемме 4.3.2,

$$|\operatorname{proj}_{L^\perp} X|^2 = \sigma^2 \chi^2(n-r, \Delta),$$

где $n - r = \dim L^\perp = n - \dim L$. Параметр нецентральности Δ здесь равен

$$\Delta = \frac{1}{\sigma^2} |\operatorname{proj}_{L^\perp} \mathsf{E}X|^2 = 0,$$

ибо $\mathsf{E}X \in L$ по условиям модели, так что $\operatorname{proj}_{L^\perp} \mathsf{E}X = 0$.

Поскольку $\mathsf{E}\chi^2(m) = m$, наилучшей несмещенной оценкой параметра σ^2 служит

$$\frac{1}{n-r} |\operatorname{proj}_{L^\perp} X|^2 = \frac{1}{n-r} |X - \operatorname{proj}_L X|^2.$$

Последнее выражение для $\operatorname{proj}_{L^\perp} X$ зачастую бывает удобнее (особенно когда подпространство L задано своим базисом).

Отметим также, что в силу леммы 4.3.2, статистически $\operatorname{proj}_L X$ и $\operatorname{proj}_{L^\perp} X$ независимы (как случайные векторы).

Замечание о вычислении $\operatorname{proj}_L X$ и $\operatorname{proj}_{L^\perp} X$.

По определению, проекцией точки (вектора) X на множество L называют такую точку множества L, на которой достигается минимум расстояния:

$$\operatorname{proj}_L X := arg \min_{Z \in L} |X - Z| = arg \min_{Z \in L} |X - Z|^2;$$

$$\operatorname{proj}_L X = arg \min_{Z \in L} \sum_{i=1}^{n} (X_i - Z_i)^2.$$

Последнее равенство объясняет название оценок в этой задаче: оценки наименьших квадратов (как и всего метода: метод наименьших квадратов).

Отметим также, что

$$\min_{Z \in L} \sum_{i=1}^{n} (X_i - Z_i)^2 = |\operatorname{proj}_{L^\perp} X|^2 = |X - \operatorname{proj}_{L^\perp} X|^2.$$

4.5. Выборка из нормального закона

Простой пример гауссовской модели - выборка из нормального закона $N(a, \sigma^2)$:

$$X = (X_1, \ldots, X_n)^T, \text{ где } X_i \sim N(a, \sigma^2 I).$$

При этом $X \sim N(l, \sigma^2 I)$, где $l = a(1, 1, \ldots, 1)^T$.

Таким образом, подпространство L здесь одномерное; оценивая l, мы, тем самым, оцениваем a.

Наилучшие несмещенные оценки a и σ^2 суть $\overline{X} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$ и $s^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - \overline{X})^2$. Эта же пара $(\overline{X}, s^2)$ и служит достаточной статистикой для (a, σ^2).

Статистики $\overline{X}$ и s^2 независимы, $\overline{X} \sim N(a, \frac{1}{n}\sigma^2)$;

$$(n-1)s^2 = \sigma^2 \chi^2(n-1).$$

4.6. Факторные модели (факторные эксперименты)

В этих экспериментах *отклик* (регистрируемый результат опыта), точнее - его неслучайная, закономерная часть, - есть результат действия одного или нескольких известных *факторов*.

Регистрируемый результат опыта может отличаться от ожидаемого благодаря присутствию случайной ошибки.

4.6.1. Однофакторная гауссовская модель

Некий фактор может принимать несколько различных значений, называемых уровнями: $A_1, \ldots, A_r$. При каждом значении $A_i, i = \overline{1,r}$ производится несколько (скажем n_i) независимых опытов. Их результаты обозначим через $X_{ij}, i = \overline{1,n_j}$ - это номер опыта в серии j, $j = \overline{1,r}$ - серия j соответствует уровню A_j.

Статистическая модель:

$$x_{ij} = a_j + \varepsilon_{ij}, \qquad j = \overline{1,r},$$

где $a_1, \ldots, a_r$ — некие числа (обычно неизвестные экспериментатору), ε_{ij} - независимые случайные величины («ошибки»).

В гауссовской модели дополнительно предполагается, что $\varepsilon_{ij} \sim N(0, \sigma^2)$; параметр σ (масштаб случайных отклонений) обычно неизвестен.

Представление однофакторной модели в каноническом виде $X \sim N(l, \sigma^2 I)$ очевидно: в качестве X можно взять столбец (размерности $n_1 + \ldots + n_r$), в котором последовательно записаны элементы всех r выборок:

$$X = (x_{11}, x_{21}, \ldots, x_{n_1 1}, x_{12}, x_{22}, \ldots, x_{n_2 2}, \ldots)^T.$$

Линейное подпространство L (которому принадлежит $\mathsf{E}X$), порождено r векторами вида:

$$(\underbrace{1, \ldots, 1}_{n_1}, 0, \ldots, 0)^T,$$

$$(\underbrace{0, \ldots, 0}_{n_1}, \underbrace{1, \ldots, 1}_{n_2}, 0, \ldots, 0)^T \text{ и т.д.}$$

Оценки параметров $a_1, \ldots, a_r$ и σ^2 мы получим в этой модели, применяя общие результаты. Здесь: $a_j^* = \frac{1}{n_j} \sum\limits_{i=1}^{n_j} x_{ij}$ для $j = \overline{1,r}$;

$$s^2 = \frac{1}{\sum\limits_{j=1}^{r} (n_j - 1)} \sum_{j=1}^{r} \sum_{i=1}^{n_j} (x_{ij} - a_j^*)^2$$

Статистики $a_1^*, \ldots, a_r^*, s^2$ независимы.

4.6.2. Аддитивная двухфакторная модель

К двух- (и более) факторной модели приходится прибегать, когда кроме главного фактора A приходится учитывать действие еще одного (или нескольких) факторов. Пусть, как выше, $A_1, \ldots, A_r$ — суть уровни фактора A, а фактор B принимает уровни $B_1, \ldots, B_s$.

Планы эксперимента в этой схеме могут быть более разнообразны, чем в факторной модели. В данном случае, план опыта указывает, какое количество независимых повторений n_{ij} надо произвести для комбинации A_i и B_j $i = \overline{1,r}, j = \overline{1,s}$ уровней факторов A и B.

Наиболее простой и популярный план: $n_{ij} = 1$. (Специальное выражение: «одно наблюдение в клетке»).

Результаты опыта можно записать таблицей

$A \backslash B$	B_1		B_j		B_s
A_1	x_{11}		x_{1j}		x_{1s}
A_i	x_{i1}		x_{ij}		x_{is}
A_r	x_{r1}		x_{rj}		x_{rs}

Статистическая модель (аддитивная):

$$x_{ij} = a_i + b_j + \varepsilon_{ij}, i = \overline{1,r}, j = \overline{1,s}.$$

Здесь a_i, b_j истолковываются как результаты действия факторов A и B, находящихся на уровнях A_i и B_j. Модель отражает представление о том, что факторы действуют на отклик, не взаимодействуя друг с другом, и что их воздействия суммируются. Величины ε_{ij} истолковываются как независимые случайные ошибки.

Если мы предполагаем, что $\varepsilon_{ij} \sim N(0, \sigma^2)$, модель называют *гауссовской* (хотя автор этого статистического направления отнюдь не К. Ф. Гаусс, а Р. Фишер).

В приведенном выше представлении аддитивной двухфакторной модели параметры (a_i, b_j) не идентифицируемы: даже если ошибки отсутствуют ($\varepsilon_{ij} \equiv 0$), по результатам опыта (в данном случае по суммам $a_i + b_j$) нельзя однозначно восстановить величины a_i, b_j.

Есть две возможности преодолеть это затруднение:

- Ставить вопросы и делать выводы только о таких функциях параметров, которые определяются однозначно.

  К таким относятся, например, попарные результаты $a_i - a_{i'}$, $b_j - b_{j'}$ и их комбинации.

- Но, по моему мнению, предпочтительней второй путь: иная параметризация модели. Представим ожидаемое значение отклика (ранее это было $a_i + b_j$) в виде:

$$\mathsf{E}x_{ij} = \mu + \alpha_i + \beta_j, i = \overline{1,r}, j = \overline{1,s},$$

  дополнительно наложив на параметры (α_i, β_j) связи:

$$\sum_{i=1}^{r} \alpha_i = 0, \quad \sum_{j=1}^{s} \beta_j = 0.$$

  С учетом связей параметры $\mu, \alpha_1, \ldots, \alpha_r, \beta_1, \ldots, \beta_s$ однозначно восстанавливаются по матрице $\|\mu+\alpha_i+\beta_j\|$.

В двухфакторной аддитивной модели (как и в однофакторной) результаты наблюдений можно представить в виде вектора-столбца.

Удобнее, впрочем, сохранить для (x_{ij}) естественную структуру матрицы (прямоугольной, размера $r \times s$).

Итак, пусть теперь:

$$X = \|x_{ij}, i = \overline{1,r}, j = \overline{1,s}\|.$$

Матрицы фиксированного размера образуют линейное подпространство. Подпространство L, которому принадлежит $\mathsf{E}X$, имеет размерность $r + s - 1$. Оно порождено $r + s$ матрицами (размера $r \times s$). Каждая из таких матриц имеет либо строку, либо столбец из единиц; прочие их параметры равны нулю. Симметрии ради (не изменяя L) к перечисленным матрицам можно присоединить матрицу, сплошь состоящую из единиц.

Оценки параметров μ, α, β получают, проецируя случайный вектор X на подпространство L, т.е. по методу наименьших квадратов. Иначе говоря, решая экстремальную задачу:

$$\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij} - \mu - \alpha_i - \beta_j)^2 \longrightarrow \min(\mu, \vec{\alpha}, \vec{\beta}),$$

$$\vec{\alpha} : \sum_{i=1}^{r} \alpha_i = 0, \quad \vec{\beta} : \sum_{j=1}^{s} \beta_j = 0.$$

Ответ можно записать в компактной форме, если употребить (широко принятую) символику:

$$x_{\cdot j} = \frac{1}{r}\sum_{i=1}^{r} x_{ij}, \qquad x_{i\cdot} = \frac{1}{s}\sum_{j=1}^{s} x_{ij}, \qquad x_{\cdot\cdot} = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s} x_{ij}.$$

(Точка замещает индекс, по которому произведено усреднение отклика).

В этих обозначениях наилучшие несмещенные оценки параметров суть:

$$\mu^* = x_{\cdot\cdot},$$

$$\alpha_i^* = x_{i\cdot} - x_{\cdot\cdot},$$

$$\beta_j^* = x_{\cdot j} - x_{\cdot\cdot},$$

$$s^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij} - x_{i\cdot} - x_{j\cdot} + x_{\cdot\cdot})^2/(r-1)(s-1).$$

При этом

$$(r-1)(s-1)s^2 = \sigma^2\chi^2(r-1)(s-1).$$

Указанные выше оценки можно получить как прямым решением приведенной ранее экстремальной задачи, так и на основе тождества

$$\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij} - \mu - \alpha_i - \beta_j)^2 =$$

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\left[(x_{ij} - x_{i\cdot} - x_{j\cdot} + x_{\cdot\cdot})^2 + (x_{i\cdot} - x_{\cdot\cdot} - \alpha_i)^2 + (x_{\cdot j} - x_{\cdot\cdot} - \beta_j)^2 + (x_{\cdot\cdot} - \mu)^2\right],$$

$$\text{если} \qquad \sum_{i=1}^{r} \alpha_i = 0, \sum_{j=1}^{s} \beta_j = 0.$$

4.7. Линейная регрессия

В линейной модели вычисление наилучших несмещенных оценок сводится к вычислению проекции вектора X на заданное линейное подпространство L. Ход вычислений зависит от того, каким образом задано (описано) подпространство L. Сейчас мы рассмотрим частый на практике случай, когда L порождено заданным набором векторов. Ради определенности, будем говорить о линейной модели в ее канонической форме, когда вектор наблюдений X и его ожидаемое значение $l = \mathsf{E}X$ - это n-мерные векторы-столбцы.

Пусть векторы (столбцы) $F_1, \ldots, F_r$ порождают подпространство L. Эта совокупность векторов может быть как линейно независимой (базис L), так и нет.

Так как $l \in L$, то

$$l = \theta_1 F_1 + \ldots + \theta_r F_r$$

при некоторых коэффициентах $\theta_1, \ldots, \theta_r \in \mathbb{R}$. Это представление l можно записать в матричной форме. Для этого введем матрицу F (размера $n \times r$), столбцами которой служат векторы $F_1, \ldots, F_r$:

$$F := \begin{pmatrix} \vec{F}_1 & \vec{F}_2 & \cdots & \vec{F}_r \end{pmatrix}.$$

Определим r-мерный вектор-столбец θ, положив $\theta := (\theta_1, \ldots, \theta_r)^T$. Тогда $l = F\theta$, а исходная линейная модель представима в виде

$$X = F\theta + \varepsilon,$$

где $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, \sigma^2)$, $\theta \in \mathbb{R}^r$, матрица F задана. Линейную модель в такой форме часто называют регрессионной моделью (задачей линейной регрессии).

В регрессионной модели достаточно оценить вектор параметров θ. Проекцию X на подпространство L теперь можно найти, решив экстремальную задачу

$$|X - F\theta|^2 \longrightarrow \min_{\theta \in \mathbb{R}^r}.$$

Для этого достаточно сначала найти градиент функции

$$Q(\theta) := |X - F\theta|^2 = (X - F\theta)^T (X - F\theta),$$

а затем, приравняв его к нулю, найти точку минимума функции $Q(\theta)$. Условимся считать оператор частного дифференцирования $\frac{\partial}{\partial\theta}$ строкой:

$$\frac{\partial}{\partial\theta} = \left(\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_r} \right).$$

При таком соглашении

$$Q(\theta + d\theta) = Q(\theta) + \frac{\partial Q}{\partial\theta} d\theta + o(d\theta).$$

Далее,

$$Q(\theta + d\theta) = [X - F(\theta + d\theta)]^T [X - F(\theta + d\theta)] = Q(\theta) - (X - F\theta)^T F d\theta - (Fd\theta)^T (X - F\theta) + o(d\theta).$$

Отсюда следует, что

$$\frac{\partial Q}{\partial\theta} = -2(X - F\theta)^T F.$$

По отношению к неизвестному вектору θ это дает уравнение

$$F^T X = (F^T F)\theta.$$

Это уравнение всегда имеет решение (по смыслу исходной задачи). Это решение единственно тогда и только тогда, когда система $F_1, \ldots, F_r$ линейно независима. В этом и только в этом случае матрица $F^T F$ невырождена и

$$\hat{\theta} = (F^T F)^{-1} F^T X;$$

при этом

$$\mathrm{proj}_L X = F\hat{\theta} = F(F^T F)^{-1} F^T X.$$

Можно указать и свойства $\hat{\theta}$ как оценки θ:

$$\hat{\theta} \sim N(\theta, \sigma^2 (F^T F)^{-1}).$$

Оценкой (несмещенной, наилучшей) σ^2 служит

$$s^2 = \frac{1}{n - r} |X - F\hat{\theta}|^2.$$

Статистики $\hat{\theta}$ и s^2 независимы.

Отметим, что вычисление $\hat{\theta}$ значительно упрощается, если базис подпространства L выбран ортогональным: в этом случае матрица $F^T F$ - диагональная. Важным достоинством ортогонального базиса служит также статистическая независимость оценок $\hat{\theta}_1, \ldots, \hat{\theta}_r$. Это облегчает интерпретацию результатов.

5. Доверительное (интервальное) оценивание

5.1. Введение

Знакомство с оцениванием завершим рассказом о доверительных границах, доверительных интервалах и доверительных областях для оцениваемых параметров. С прикладной точки зрения, статистическая оценка - это статистическое приближение к неизвестному параметру или его функции, это его приближенное значение, полученное из опыта. До сих пор мы стремились к тому, чтобы путем статистической обработки получить как можно более точное приближение. Однако способа измерить точность приближения у нас не было.

Между тем, точность приближения - это общенаучное понятие, так же, как и способ ее количественного выражения. Всякий раз, когда точное значение какой-либо величины мы замещаем приближенным значением, нам следует сопровождать такую замену также и сообщением о точности этого приближения.

К примеру, 288 приблизительно равно 300; но также 288 приблизительно равно 290. Однако точность этих приближений различна. Так, в первом случае, точность приближения не ниже 15, а во втором - меньше 5: $|288-300|<15$ и $|288-290|<5$.

В задачах статистического оценивания мы получаем аналогичное приближенное равенство $\hat\theta(X)\approx\theta$. (Либо $\hat\theta(X)\approx\varphi(\theta)$, если мы оцениваем функцию от параметра). Здесь θ - неизвестное истинное значение параметра, $\hat\theta(X)$ - его оценка по наблюдению X. Для статистического приближения, как правило, не существует гарантированной точности: нет такого $\varepsilon>0$, для которого бы достоверно выполнялось соотношение $|\hat\theta(X)-\theta|<\varepsilon$. Мы можем говорить лишь о вероятности, с которой выполняется это неравенство. Если эта вероятность близка к 1, можно говорить, что статистическая погрешность в определении θ не превосходит ε.

В этих примерах для неизвестной величины a мы указываем ее приближенное значение x, причем $|x-a|<\varepsilon$ для некоторого определенного $\varepsilon>0$. Здесь ε - гарантированная точность приближения $x\approx a$.

Рассмотрим на примере нормальной выборки, как реализуются эти соображения.

5.2. Нормальная выборка с известной дисперсией

Пусть $x_1,\dots,x_n$ суть независимые измерения некоторой величины a, причем $x_i\sim N(a,\sigma^2)$ для $i=\overline{1,n}$. Оценкой для a может служить $\overline{x}=\frac{1}{n}\sum\limits_{i=1}^{n}x_i$, так что $\overline{x}\approx a$. Как можно судить о точности этого приближения, то есть о $|\overline{x}-a|$? С какой вероятностью для данного $\varepsilon>0$ выполняется неравенство $|\overline{x}-a|<\varepsilon$? Каким надо взять ε, чтобы вероятность этого неравенства была бы 0.95? Или 0.99? И т.д.

Пусть, для начала, σ^2 известно. Рассмотрим случайную величину $\sqrt{n}\,\frac{\overline{x}-a}{\sigma}\sim N(0,1)$. Зададимся какой-либо (обычно близкой к 1) вероятностью, для удобства обозначив ее через $1-2\alpha$. Здесь α задано, $0<\alpha<\frac12$. Пусть $z_{1-\alpha}$ обозначает $(1-\alpha)$-квантиль стандартного нормального распределения. Иными словами, $z_{1-\alpha}$ удовлетворяет уравнению $\Phi(z_{1-\alpha})=1-\alpha$, где $\Phi(\cdot)$ - функция стандартного нормального распределения, или функция Лапласа.

Ввиду симметрии (относительно нуля) $z_{1-\alpha}=-z_\alpha$.

Поэтому для $\sqrt{n}\,\frac{\overline{x}-a}{\sigma}$ справедливо утверждение

$$P\{|\sqrt{n}\,\frac{\overline{x}-a}{\sigma}|<z_{1-\alpha}\}=1-2\alpha \qquad (*)$$

или

$$P\{|\overline{x}-a|<\frac{\sigma}{\sqrt{n}}z_{1-\alpha}\}=1-2\alpha.$$

Итак, с вероятностью $1-2\alpha$ точность приближения $\overline{x}\approx a$ не хуже, чем $\frac{\sigma}{\sqrt{n}}z_{1-\alpha}$.

Соотношение $(*)$ можно преобразовать далее и написать, что

$$P\{\overline{x}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha}<a<\overline{x}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha}\}=1-2\alpha.$$

Интервал (случайный)

$$(\overline{x}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha},\overline{x}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha}) \qquad (1)$$

содержит неизвестное a (часто говорят — «накрывает» неизвестное a) с вероятностью $1-2\alpha$.

Эту вероятность $1-2\alpha$ называют *доверительной вероятностью* (иногда - коэффициентом доверия), а упомянутый случайный интервал - *доверительным интервалом*.

На практике не следует ограничиваться одной какой-либо доверительной вероятностью и одним доверительным интервалом. Чтобы лучше передать, как связаны $\overline{x}$ и a, следует вычислить доверительные интервалы для нескольких доверительных вероятностей, скажем, для 0.50, 0.90, 0.95 и 0.99. Чертеж, на котором выделены эти доверительные интервалы, дает нам наглядное представление о точности статистического приближения $\overline{x}\approx a$.

Отметим некоторые очевидные, но важные свойства полученных доверительных интервалов.

- Эти интервалы тем лучше, чем меньше σ.

  В нашем примере σ^2 - дисперсия ошибки при измерении a. Ясно, что чем больше эта дисперсия, тем ниже и точность статистического вывода.

- Интервалы тем шире, чем больше $z_{1-\alpha}$, которая, в свою очередь, возрастает при приближении $1-\alpha$ к 1. (И скорость роста тем выше, чем ближе α к нулю).

  Это свойство тоже легко объяснимо: чем выше требования к достоверности суждения, тем менее содержательно и информативно самое это суждение.

- Наконец, на точность приближения $\overline{x} \approx a$ влияет число наблюдений n: чем больше n, тем у́же доверительный интервал, т.е. тем выше точность.

  Заметим, однако, что длина доверительного интервала пропорциональна $1/\sqrt{n}$. Так что если мы хотим повысить статистическую точность вдвое, нам придется увеличить количество независимых измерений вчетверо. (А если в 10 раз, то - в 100). Притом, все эти измерения надо проводить в неизменных условиях. На практике большие выборки встречаются не часто.

5.3. Нормальная выборка с неизвестной дисперсией. Распределение Стьюдента

Предыдущие результаты верны, но бесполезны, когда σ неизвестно, что чаще всего на практике и бывает.

Естественная мысль: заместить неизвестное σ его оценкой s, где $s^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$, и рассмотреть случайную величину

$$t = \sqrt{n}\,\frac{\overline{x} - a}{s} \tag{1}$$

Ее называют *отношением Стьюдента* (*Student's ratio* - стьюдентова дробь, стьюдентово отношение). Легко видеть, что распределение (1) не зависит от неизвестных параметров нормальной выборки (a, σ^2) и совпадает с распределением отношения стандартной нормальной величины $N(0,1)$ к случайной величине $\sqrt{\frac{1}{n-1}\,\chi^2(n-1)}$, причем эти случайные величины независимы (см. пункт 4.5).

Распределение случайной величины (1) называют *распределением Стьюдента с* $(n-1)$ *степенями свободы*. Приведем общее определение.

Определение: Пусть $\xi_0, \xi_1, \ldots, \xi_m$ (m - натуральное) суть независимые стандартные гауссовские случайные величины (т.е. $\xi_0, \xi_1, \ldots, \xi_m \sim N(0,1)$). *Стьюдентовским отношением* (*стьюдентовской дробью*) называют случайную величину

$$t = t(m, \mu) = \frac{\xi_0 + \mu}{\sqrt{\frac{1}{m}(\xi_1^2 + \cdots + \xi_m^2)}}, \tag{2}$$

где $\mu \in \mathbb{R}$ - произвольное число.

Распределение случайной величины $t(m, \mu)$ называют *распределением Стьюдента*; число m называют *числом степеней свободы*, а число μ - *параметром нецентральности* распределения Стьюдента.

Если $\mu = 0$, то распределение случайной величины $t(m) = t(m, 0)$ называют *центральным распределением Стьюдента*. Обычно этот эпитет опускают и распределение $t(m)$ называют просто распределением Стьюдента (с m степенями свободы).

Распределение Стьюдента (центральное) снабжено разнообразными и подробными таблицами. Есть, в частности, таблицы квантилей. Пакеты статистических программ содержат команды, позволяющие получить всю необходимую информацию о распределении $t(m)$.

Функции плотности вероятности для $t(m)$ и $t(m, \mu)$ известны (их можно найти в справочниках). Их аналитическими выражениями мы пользоваться не будем. Для информации, приведу формулу плотности для $t(m)$:

$$\frac{1}{\sqrt{m}B(\frac{1}{2}, \frac{m}{2})}(1 + \frac{x^2}{m})^{-\frac{m+1}{2}}$$

(Из этой формулы и (2) при $m = 1$ следует, что распределение Стьюдента с одной степенью свободы совпадает с распределением Коши).

Отметим важное для дальнейшего свойство распределений Стьюдента:

при каждом m семейство $t(m, \mu)$ стохастически упорядочено (стохастически монотонно возрастает) относительно μ. Это означает, что для любого $x \in \mathbb{R}$

$$P(t(m, \mu_1) > x) < P(t(m, \mu_2) > x),$$

если $\mu_1 < \mu_2$.

Доказательство.

Доказательство почти очевидно:

Из (2) следует, что

$$P(t(m,\mu_1) > x) = P(\xi_0 + \mu_1 > x\sqrt{\frac{1}{m}\chi^2(m)}) =$$

$$= \mathsf{E}(\xi_0 + \mu_1 > z | z = x\sqrt{\frac{1}{m}\chi^2(m)}).$$

Для завершения доказательства остается заметить, что для любого $z \in \mathbb{R}$

$$P(\xi_0 + \mu_1 > z) < P(\xi_0 + \mu_2 > z),$$

если $\mu_1 < \mu_2$, $\xi_0 \sim N(0,1)$. $\square$

Вернемся к поставленной задаче: построению доверительных интервалов для a (для среднего) по нормальной выборке (по выборке из $N(a,\sigma^2)$). Ее решение теперь почти не отличается от рассмотренного в первом пункте. Единственное, что надо изменить: вместо нормальных квантилей ввести квантили распределения Стьюдента.

Все же, повторим необходимые шаги.

Выбираем доверительную вероятность $1-2\alpha$. По таблицам находим $(1-\alpha)$-квантиль распределения Стьюдента, с $(n-1)$ степенями свободы, которую обозначим через $t_{1-\alpha}(n-1)$, т.е. решение уравнения

$$P(t(n-1) < t_{1-\alpha}) = 1-\alpha.$$

Ввиду симметрии распределения Стьюдента, можно утверждать, что

$$P\left(|\sqrt{n}\frac{\overline{x}-a}{s}| < t_{1-\alpha}\right) = 1-2\alpha.$$

Преобразуя, получаем оценку точности для приближения $\overline{x} \approx a$

$$P\left(|\overline{x}-a| < \frac{s}{\sqrt{n}}t_{1-\alpha}\right) = 1-2\alpha$$

и доверительный интервал (с доверительной вероятностью $1-2\alpha$) для a

$$P\left\{\overline{x} - \frac{s}{\sqrt{n}}t_{1-\alpha} < a < \overline{x} + \frac{s}{\sqrt{n}}t_{1-\alpha}\right\} = 1-2\alpha. \tag{3}$$

Все сделанные в предыдущем пункте замечания о свойствах доверительного интервала (5.2.1), остаются верными и для (3). Равно как и рекомендации не ограничиваться каким-либо одним доверительным интервалом (и какой-либо одной доверительной вероятностью), но вычислять их несколько - для нескольких коэффициентов доверия.

Тем же приемом можно выводить для a и другие доверительные утверждения.

Пример: *Доверительные пределы* (границы сверху или снизу).

Выбираем доверительную вероятность $1-\alpha$. Если мы хотим получить для a границу сверху, берем α-квантиль $t_\alpha = t_\alpha(n-1)$; для границы сверху берем $(1-\alpha)$-квантиль $t_{1-\alpha}(n-1)$. (Заметим, что из-за симметрии $t_\alpha = -t_{1-\alpha}$).

Далее, заметим, что для (1) выполняется соотношение

$$P\left(t_\alpha < \sqrt{n}\,\frac{\overline{x}-a}{s}\right) = 1-\alpha.$$

Отсюда, поскольку $t_\alpha = -t_{1-\alpha}$, следует, что

$$P\left\{a < \overline{x} + \frac{s}{\sqrt{n}}t_{1-\alpha}\right\} = 1-\alpha, \tag{2.9}$$

так что $\overline{x} + \frac{s}{\sqrt{n}}t_{1-\alpha}$ - это верхняя доверительная граница для a с коэффициентом доверия $1-\alpha$.

Нижняя $(1-\alpha)$-доверительная граница для a, равная $\overline{x} - \frac{s}{\sqrt{n}}t_{1-\alpha}$, получается аналогично.

Пересечение двух полученных доверительных областей дает для a уже известный доверительный интервал (3) с доверительной вероятностью $1-2\alpha$.

5.4. Центральные величины

Обсудим в общем виде тот прием, который мы применяли в пунктах 5.2 и 5.3.

Пусть распределение наблюдения X определяется неизвестным параметром $\theta, \theta \in \Theta$. Предположим, что существует случайная переменная $G(X,\theta)$, $G(\cdot,\cdot)$ - известная функция от X и θ, распределение которой нам известно и не зависит от θ, когда $\theta \in \Theta$. (В предыдущем примере это было $\sqrt{n}\,\frac{\overline{x}-a}{s}$). $G(X,\theta)$ называют *центральной случайной величиной*, а чаще (хоть и не совсем правильно) - *центральной статистикой*.

Предположим для простоты, что распределение $G(X,\theta)$ непрерывно и пусть g_α, $\alpha \in (0,1)$, обозначает α-квантиль $G(X,\theta)$. Теперь для всякого $\theta \in \Theta$ и $\alpha \in (0,1)$ справедливо соотношение

$$P(g_\alpha < G(X,\theta)) = 1-\alpha. \tag{1}$$

(Точнее было бы в этом равенстве употребить символ P_θ для распределения вероятностей, зависящих от $\theta, \theta \in \Theta$. Но, поскольку (1) выполняется для всех таких θ, индекс θ, которым мы обычно сопровождаем символ вероятности P, здесь и далее можно опустить, не опасаясь недоразумений.)

Решаем неравенство $g_\alpha < G(X,\theta)$ относительно θ. Получим зависящее от X множество

$$S_{1-\alpha}(X) := \{\theta \in \Theta : g_\alpha < G(X,\theta)\}. \tag{2}$$

Ясно, что для всякого $\theta \in \Theta$

$$P(\theta \in S_{1-\alpha}(X)) = 1-\alpha,$$

так что $S_{1-\alpha}(X)$ - это доверительная область для θ с доверительной вероятностью $1-\alpha$.

Если мы не собираемся ограничивать себя какой-либо одной доверительной областью (2), но использовать все семейство $S_{1-\alpha}(\cdot)$, $\alpha \in (0,1)$, тогда разумно потребовать от центральной величины $G(X,\theta)$, чтобы семейство $\{S_{1-\alpha}(X), \alpha \in (0,1)\}$ было бы монотонным по вложению:

$$\text{если } 0 < \alpha_1 < \alpha_2 < 1, \text{ то}$$

$$S_{1-\alpha_1}(X) \supset S_{1-\alpha_2}(X) \tag{3}$$

Когда θ - одномерный параметр, достаточным условием для (3) служит монотонность $G(X,\theta)$ по переменной θ (при каждом фиксированном X). Точнее: для (3) нужно, чтобы $G(X,\theta)$ монотонно убывала по θ. В этом случае $S_{1-\alpha}(X)$ - полуинтервал (точнее, это пересечение Θ с полуинтервалом); его правый конец - это верхняя $(1-\alpha)$-доверительная граница для $\theta, \theta \in \Theta$.

Другая система доверительных областей возникает из аналогичного (1) соотношения

$$P(G(X,\theta) < g_{1-\alpha}) = 1-\alpha. \tag{4}$$

Действуя как выше, т.е. решая неравенство относительно θ, получим для θ доверительную область

$$T_{1-\alpha}(X) = \{\theta \in \Theta : G(X,\theta) < g_{1-\alpha}\}.$$

В оговоренном выше одномерном монотонном случае множество $T_{1-\alpha}(X)$ - это полупрямая (пересеченная с Θ). Ее левый конец для θ дает $(1-\alpha)$-доверительную границу сверху. Пересечение областей $S_{1-\alpha}(X) \bigcap T_{1-\alpha}(X)$ дает для θ доверительную область (как правило, ограниченную) с доверительной вероятностью $1-2\alpha$.

5.5. Испытания Бернулли

В этом случае нет точной центральной величины, но есть случайная величина, распределенная асимптотически свободно (имеется в виду, что распределение не зависит от неизвестных параметров, свободно от их влияния).

Пусть θ - неизвестная вероятность успеха, $\theta \in (0,1)$; S_n - число успехов, случившееся в n проведенных испытаниях Бернулли.

По теореме Муавра - Лапласа, случайная величина

$$\frac{S_n - \theta n}{\sqrt{n\theta(1-\theta)}} \xrightarrow{d} N(0,1) \qquad \text{при } n \to \infty.$$

Как обычно, мы заключаем из этой теоремы, что для достаточно больших n и $z \in \mathbb{R}$

$$P\left(\left|\frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}}\right| < z\right) \approx \Phi(z) - \Phi(-z).$$

Пусть $1-2\alpha$ - выбранная нами доверительная вероятность, $z_{1-\alpha}$ означает $(1-\alpha)$-квантиль стандартного нормального распределения, так что $\Phi(z_{1-\alpha}) = 1-\alpha$.

Тогда

$$P\left(\left|\frac{S_n - n\theta}{\sqrt{n\theta(1-\theta)}}\right| < z_{1-\alpha}\right) \approx 1 - 2\alpha.$$

Это неравенство надо разрешить относительно θ, $\theta \in (0,1)$. После тождественных преобразований получим для этого неравенства эквивалентную форму

$$(S_n - n\theta)^2 - n\theta(1-\theta)z_{1-\alpha}^2 < 0. \tag{1}$$

Левая часть (1) - квадратный трёхчлен относительно θ, причем коэффициент при θ^2 положителен.

Поэтому решение имеет вид

$$\underline{\theta}(S_n) < \theta < \overline{\theta}(S_n),$$

где $\underline{\theta}(S_n), \overline{\theta}(S_n)$ - суть корни квадратного трехчлена в (1). Здесь

$$\underline{\theta}(S_n), \overline{\theta}(S_n) = \frac{S_n + \frac{z_{1-\alpha}^2}{2} \mp z_{1-\alpha}\sqrt{\frac{S_n(n-S_n)}{n} + \frac{z_{1-\alpha}^2}{4}}}{n + z_{1-\alpha}^2}. \tag{2}$$

Выражение (2) дает для θ доверительный интервал, доверительная вероятность которого приближенно равна $1-2\alpha$.

5.6. Регрессионная модель

Метод центральных величин пригоден для того, чтобы строить доверительные области для параметров гауссовских линейных моделей. Рассмотрим регрессионную модель

$$X = F\theta + \varepsilon, \tag{1}$$

где X - наблюдаемый n-мерный вектор (столбец); $\theta = (\theta_1, \ldots, \theta_m)^T$ - неизвестный параметр, $\theta \in \mathbb{R}^m$; F - заданная $n \times m$ матрица, $F = \|F_1, \ldots, F_m\|$; все ее столбцы $F_1, \ldots, F_m$ будем предполагать линейно независимыми; $\varepsilon \sim N(0, \sigma^2 I)$ - вектор случайных ошибок.

Как нам уже известно, в этой модели наилучшая несмещенная оценка $\hat{\theta}$ получается по методу наименьших квадратов и равна

$$\hat{\theta} = (F^T F)^{-1} F^T X. \tag{2}$$

Из теории гауссовских линейных моделей (точнее, из леммы 4.3.2 об ортогональных разложениях) вытекает, что $|X - F\hat{\theta}|^2$ и $\hat{\theta}$ статистически независимы, причем

$$|X - F\hat{\theta}|^2 = \sigma^2 \chi^2(n-m), \tag{3}$$

$$\hat{\theta} \sim N(\theta, \sigma^2 (F^T F)^{-1}).$$

Для построения центральной величины нам понадобится изложенная ниже лемма, а также еще одно семейство распределений.

Лемма.

Пусть $\xi \sim N_p(a, A)$, причем A^{-1} существует. Тогда

$$\xi^T A^{-1} \xi = \chi^2(p, \Delta),$$

где параметр нецентральности $\Delta = a^T A^{-1} a$.

Доказательство.

Из линейной алгебры известно, что квадратичную форму с матрицей A линейным невырожденным преобразованием можно привести к каноническому виду. В данном случае, преобразованная матрица квадратичной формы - единичная (ибо $A \geqslant 0$ и A невырожденная).

Иначе говоря, существует невырожденная квадратная матрица, скажем, B, такая, что

$$BAB^T = I.$$

Заметим, что

$$A^{-1} = B^T B.$$

Рассмотрим случайный вектор $\eta = B\xi$. Ясно, что

$$\eta \sim N_p(Ba, BAB^T) = N_p(Ba, I).$$

Поэтому

$$|\eta|^2 = \chi^2(p, \Delta), \text{ где } \Delta = |Ba|^2 = (Ba)^T Ba = a^T A^{-1} a.$$

С другой стороны:

$$|\eta|^2 = \eta^T \eta = (B\xi)^T B\xi = \xi^T A^{-1} \xi.$$

Лемма доказана. □

Применим эту лемму к гауссовскому вектору (2). Получим, что

$$(\hat{\theta} - \theta)^T (F^T F)(\hat{\theta} - \theta) = \sigma^2 \chi^2(m). \tag{4}$$

Эф - распределение.

Называемое также распределением Снедекора, распределением Фишера, распределением дисперсионного отношения Фишера, и т.д.

Определение Пусть случайные величины X_1 и X_2 независимы и распределены по закону хи-квадрат:

$$X_1 = \chi^2(m_1, \Delta), X_2 = \chi^2(m_2, 0).$$

Случайная величина

$$F = F(m_1, m_2, \Delta) = \frac{\frac{1}{m_1} X_1}{\frac{1}{m_2} X_2} \tag{5}$$

называется F - *отношением* (эф-отношением, дисперсионным отношением Фишера). Распределение (5) называют *нецентральным эф-распределением* с m_1 и m_2 степенями свободы и параметром нецентральности Δ. Если $\Delta = 0$, распределение называют *центральным*. Слово «центральное» часто опускают и говорят просто о эф-распределении с m_1 и m_2 степенями свободы и о случайной величине $F(m_1, m_2)$.

Плотность эф-распределения можем вывести из определения (5) и вида плотности хи-квадрат. Мы не будем к ней обращаться, полагаясь на то, что необходимые сведения об эф-распределении (например, квантили) можно найти в таблицах.

Все же приведем плотность $F(m_1, m_2, 0)$:

$$\frac{\left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}}}{B(\frac{m_1}{2}, \frac{m_2}{2})} \frac{x^{\frac{m_1}{2}-1}}{(1 + \frac{m_1}{m_2} x)^{\frac{m_1+m_2}{2}}} \text{ для } x \geqslant 0$$

Легко видеть, что семейство распределений $F(m_1, m_2, \Delta)$ стохастически упорядочено по Δ при любых m_1 и m_2. Доказывают этот факт тем же способом, то и упорядоченности семейства $\chi^2(m, \Delta)$ по Δ.

Вернемся к доверительному оцениванию θ в модели (1).

Из двух независимых случайных величин (3) и (4) составим эф-отношение

$$F(m, n-m) = \frac{\frac{1}{m}(\hat{\theta} - \theta)^T (F^T F)(\hat{\theta} - \theta)}{\frac{1}{n-m}|X - F\hat{\theta}|^2}. \tag{6}$$

Выбрав доверительную вероятность $1 - \alpha$, с помощью таблицы квантилей для $F(m, n-m)$ найдем $(1-\alpha)$-квантиль, которую обозначим как $F_{1-\alpha}(m, n-m)$.

Теперь

$$P\left\{ \frac{\frac{1}{m}(\hat{\theta} - \theta)^T (F^T F)(\hat{\theta} - \theta)}{\frac{1}{n-m}|X - F\hat{\theta}|^2} < F_{1-\alpha}(m, n-m) \right\} = 1 - \alpha.$$

Заметим, что

$$s^2 := \frac{1}{n-m}|X - F\hat{\theta}|^2 \tag{7}$$

- это несмещенная оценка для σ^2.

Теперь видно, что $(1-\alpha)$-доверительное множество для θ, заданное неравенством

$$\{\theta : (\hat{\theta} - \theta)^T (F^T F)(\hat{\theta} - \theta) < m s^2 F_{1-\alpha}(m, n-m)\}, \tag{8}$$

представляет собой внутреннюю часть (случайного) эллипсоида с центром в точке $\hat{\theta}$. Эта область (внутренность эллипсоида) накрывает неизвестное θ (точку $\theta \in \mathbb{R}^m$) с вероятностью $1-\alpha$.

Можно указать доверительные интервалы и для отдельных параметров θ_i, $\theta = (\theta_1, \ldots, \theta_m)^T$. Из (2) следует, что каждая координата $\hat{\theta}_i$ вектора $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)^T$ распределена по нормальному закону $N(\theta_i, \sigma^2 a_{ii})$, если положить $(F^T F)^{-1} = \|a_{ij}\|$.

С учетом (7) (и независимости $\hat{\theta}_i$ и s^2) можно утверждать, что случайная величина

$$t := \frac{\hat{\theta}_i - \theta}{s\sqrt{a_{ii}}} \tag{5.9}$$

распределена по Стьюденту, с $(n-m)$ степенями свободы. Исходя из этого, можно строить для θ_i доверительные интервалы так же, как мы делали это в пункте 5.3.

Отметим, что если матрица $F^T F$ не ортогональна, то координаты вектора оценок $\hat{\theta}$ не независимы. Поэтому не являются независимыми и доверительные утверждения для отдельных $\theta_1, \ldots, \theta_m$, когда эти утверждения основываются на центральных величинах (9). В этом случае вероятность того, что несколько доверительных утверждений выполняются одновременно, нельзя получить, перемножая их индивидуальные доверительные вероятности. Одновременные доверительные выводы о $\theta_1, \ldots, \theta_m$ надо получать иначе. Например, по методу Шеффа (или Тьюки).

6. Проверка статистических гипотез

6.1. Постановка задачи, основные понятия

Наблюдение X получено случайным выбором из генеральной совокупности $\mathcal{X}$ по некоторому вероятностному закону P, который нам не известен. Относительно распределения P известно лишь, что оно является элементом некоторого заданного множества $\mathcal{P}$ вероятностных распределений на измеримом пространстве $\mathcal{X}$. Относительно истинного распределения P высказано предположение, которое мы хотим проверить, опираясь на наблюдение X: P обладает некоторыми определенными свойствами. Эти свойства выделяют в множестве $\mathcal{P}$ некоторое подмножество $\mathcal{P}_0$. Поэтому упомянутое подлежащее проверке предположение H_0 (в дальнейшем - гипотеза H_0) звучит так: $P \in \mathcal{P}_0$, где $\mathcal{P}_0 \subset \mathcal{P}$.

Когда множество распределений $\mathcal{P}$ параметризовано с помощью какого-либо параметра θ, причем $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, тогда гипотеза H_0 тоже приобретает параметрическую форму

$$H_0 : \theta \in \Theta_0,$$

где $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\}$, Θ_0 задано и $\Theta_0 \subset \Theta$.

Гипотеза H_0 либо верна, либо нет. В последнем случае выполнено альтернативное предположение о распределении (альтернатива): $P \in \mathcal{P}_1$.

При этом $\mathcal{P}_0 \bigcap \mathcal{P}_1 = \emptyset$, $\mathcal{P}_1 \bigcup \mathcal{P}_0 = \mathcal{P}$. (Последнее, впрочем, не обязательно: гипотетическое и альтернативное множество распределений не всегда в своем объединении составляют все возможные вероятностные распределения).

В параметрической форме альтернатива H_1 имеет вид

$$H_1 : \theta \in \Theta_1,$$

где $\mathcal{P}_1 = \{P_\theta : \theta \in \Theta_1\}$, Θ_1 задано, $\theta_1 \in \Theta$ и $\Theta_0 \bigcap \Theta_1 = \varnothing$.

По наблюдению X мы должны либо принять H_0, либо H_0 отвергнуть (иногда в этом случае говорят: *принять* H_1). Мы расширяем эту задачу так: на множестве $\mathcal{X}$ мы должны определить функцию от x, $x \in \mathcal{X}$, значениями которой могут быть «отвергнуть H_0» или «не отвергать H_0». Затем мы применим эту функцию к наблюдаемому значению X и в результате получим конкретное решение.

Пусть

$$S = \{x : x \in \mathcal{X}, \text{ по наблюдаемому } x \text{ отвергаем } H_0\}.$$

Множество S, $S \subset \mathcal{X}$, называют *критическим множеством* для гипотезы H_0, или *критерием*.

Поскольку гипотезы, о которых мы говорили, касаются распределения вероятностей, такие гипотезы называются *статистическими*, а критерии для их проверки - *статистическими критериями*.

С любыми статистическими критериями неразрывно связаны возможные ошибки:

- ошибка рода I: отвергаем H_0, когда H_0 верна;
- ошибка рода II: не отвергаем H_0, когда H_0 неверна.

По своим последствиям эти ошибки обычно не равнозначны: 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 H_0 в результате дальнейших ее проверок. Поэтому при проверке статистических гипотез возможность ошибки первого рода стараются уменьшить. Желательно, впрочем, иметь такие статистические критерии, для которых малы (близки к 0) вероятности обеих ошибок. Но поскольку это обычно невозможно, к выбору критерия S выдвигают такие требования:

- Вероятность ошибки первого рода не должна превосходить выбранной (малой) величины, называемой уровнем значимости критерия S.
- При этом условии вероятность ошибки II рода надо сделать как можно меньше.

С большей определенностью говорить о свойствах статистического критерия помогает его функция мощности. Ее аргументом служит распределение вероятностей Q на $\mathcal{X}$, $Q \in \mathcal{P}$.

Определение.

Мощностью $\beta(Q)$ критерия S называют

$$\beta(Q, S) = \beta(Q) = Q\{X \in S\},$$

т.е. вероятность события $\{X \in S\}$, когда случайный выбор $X, X \in \mathcal{X}$, происходит согласно распределению вероятностей Q. (Напомним, что гипотезу H_0 мы отвергаем с помощью критерия S, если происходит событие $X \in S$). Функцию $\beta(\cdot)$, заданную на множестве распределений $\mathcal{P}$, называют *функцией мощности* (критерия S).

Согласно сказанному ранее, статистический критерий имеет уровень значимости α, если $\beta(Q) \leqslant \alpha$ для всех $Q \in \mathcal{P}_0$. Поскольку каждый критерий уровня α есть одновременно и критерий уровня α', если $\alpha < \alpha'$, то полезно определить для критерия его минимальный уровень значимости

$$\sup_{Q \in \mathcal{P}_0} \beta(Q).$$

Эту величину называют *размером* критерия.

Когда множество $\mathcal{P}$ параметризовано, т.е. когда $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, мощность можно считать функцией параметра θ:

$$\beta(\theta, S) = \beta(\theta) = P_\theta(X \in S).$$

В этом случае размер критерия S есть

$$\sup_{\theta \in \theta_0} P_\theta(X \in S).$$

Желательные свойства любого статистического критерия, предназначенного для проверки статистической гипотезы $P \in \mathcal{P}_0$:

(a) малый размер,

(b) быстрое возрастание его функции мощности (при удалении распределения Q от гипотетического множества распределений $\mathcal{P}_0$).

6.2. Пример реальной проверки статистической гипотезы

Математическая (статистическая) модель закона Менделя проста. Гибриды первого поколения имеют генотип Aa (и фенотип A). Они производят гаметы (зародышевые клетки) $\underline{A}$ и $\underline{a}$ в равных количествах. При слиянии гамет возникают соматические клетки четырех генотипов: $\underline{AA}, \underline{Aa}, \underline{aA}$ и $\underline{aa}$ (здесь первым указан генотип материнской клетки, вторым - отцовской, для определенности). Если в оплодотворении нет селективности, если жизнеспособность гамет одинакова, если жизнеспособность потомства (например, всхожесть семян) одинакова и т.д., то наудачу взятое растение второго поколения имеет один из трех генотипов $\underline{AA}, \underline{Aa}, \underline{aa}$ с вероятностями $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ соответственно. Отсюда следует, что вероятности фенотипов A и a суть $\frac{3}{4}$ и $\frac{1}{4}$. Поэтому в опыте частоты должны относиться (приблизительно) как 3:1.

Школа Т. Д. Лысенко в СССР в тридцатые годы пыталась бороться с менделевскими законами наследственности научными методами. Дальнейший рассказ — об одном из эпизодов этой борьбы — представляет собой извлечение из статьи А.Н. Колмогорова (1940) «Об одном новом подтверждении законов Менделя», ДАН СССР, том 27, стр.38 — 42.

См. также:

А. Н. Колмогоров. Теория вероятностей и математическая статистика. — М.: Наука, 1986.

В. Н. Тутубалин. Теория вероятностей и случайных процессов. — М.: изд-во МГУ, 1992 (ч. 2, гл. 3, §1).

Работа Колмогорова основывается на экспериментальных данных Н. И. Ермолаевой: «Еще раз о гороховых законах», Яровизация (1939), N 2 (23). Н. И. Ермолаева экспериментировала с томатами. В ее опытах результаты разделялись по семействам.

Например, семейство составляли все растения, выросшие в одном ящике. Семейства мы занумеруем индексом i, $i = 1, \ldots, N$; N - их общее число. Чистые линии, которые подвергались скрещиванию (гибридизации), отличались внешне: одни имели гладкие, а другие - морщинистые листья.

Пусть μ_i, $i = 1, \ldots, N$, обозначают частоты фенотипа $\underline{a}$ в каждой из N серий, а n_i обозначает число растений в серии.

Если численности n_i не слишком малы (порядка нескольких десятков), то по теореме Муавра-Лапласа и при справедливости законов Менделя нормированные частоты (где $p = \frac{1}{4}$)

$$\xi_i = \frac{\mu_i - n_i p}{\sqrt{n_i p(1-p)}},$$

или

$$\xi_i = \frac{\mu_i - \frac{n_i}{4}}{\sqrt{n_i \frac{1}{4}\frac{3}{4}}}$$

имеют (приближенно) распределение $N(0,1)$. Поэтому на совокупность $\xi_1, \xi_2, \ldots, \xi_n$ можно смотреть как на выборку (объема N) из $N(0,1)$. Все это - если верен закон Менделя.

Возникает естественная мысль сравнить выборочную функцию $F_N(x)$, построенную по этой выборке, и функцию стандартного нормального распределения (функцию Лапласа)

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du.$$

Согласно известной нам теореме Гливенко, случайная величина

$$\mathcal{D}_N = \sup_x |F_N(x) - \Phi(x)| \qquad (*)$$

при больших N должна быть малой, если верны законы Менделя, ибо в этом случае $\mathcal{D}_N \xrightarrow{P} 0$ при $N \to \infty$.

Если же закон Менделя в обсуждаемых опытах не действует, то вероятность появления фенотипа a отличается от $\frac{1}{4}$. В этом случае выборочная функция $F_N(\cdot)$ сходится не к $\Phi(\cdot)$, а к другому пределу.

В результате

$$\mathcal{D}_N \xrightarrow{P} c > 0,$$

если закон Менделя неверен.

Этих соображений, однако, недостаточно для точных статистических выводов. Надо привлечь следующую теорему.

Теорема Колмогорова (1933).

$$\mathcal{D}_N = \sup_{x \in R} |F_N(x) - \Phi(x)|$$

и для любого $z > 0$

$$P(\sqrt{N}\mathcal{D}_N < z) \longrightarrow K(z),$$

где

$$K(z) = 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2 z^2},$$

или $K(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}$ для $z > 0$. (Функцию $K(\cdot)$ называют *функцией Колмогорова*).

В случае же его нарушения

$$\sqrt{N}\mathcal{D}_N \longrightarrow \infty \text{ при } N \to \infty.$$

Это значит, что для конечных значений N статистика $\sqrt{N}\mathcal{D}_N$ должна принимать *большие значения*, если гипотеза неверна.

Таким образом, статистика $\sqrt{N}\mathcal{D}_N$ различно ведет себя при гипотезе и при ее нарушении (при альтернативе). Именно это позволяет по величине $\sqrt{N}\mathcal{D}_N$ сделать вывод о том, что же действует на самом деле: гипотеза или альтернатива.

В данном случае естественно следующее решающее правило: отвергать гипотезу о том, что выборка извлечена из распределения с функцией $F(\cdot)$, если статистика $\sqrt{N}\mathcal{D}_N$ приняла (в опыте) слишком большое значение. Т.е. столь большое значение, которое маловероятно, если гипотеза верна.

Дать точный смысл этому предложению можно так.

- Выбираем уровень значимости ε, $\varepsilon > 0$ - это вероятность отвергнуть гипотезу, когда она верна.
- По этому значению ε вычисляем критическое значение, скажем $\mathcal{C}_\varepsilon$,такое, что

$$K(\mathcal{C}_\varepsilon) = 1 - \varepsilon.$$

- Если наблюдаемое значение $\sqrt{N}\mathcal{D}_N$ превосходит $\mathcal{C}_\varepsilon$, мы проверяемую гипотезу отвергаем (как говорят - на уровне ε). В данном случае - это гипотеза (закон) Менделя.

Судить о том, совместимо ли наблюденное в опыте значение статистики $\sqrt{N}\mathcal{D}_N$ с проверяемой гипотезой, можно и иначе. Как было сказано, против гипотезы (закона Менделя) говорят большие значения $\sqrt{N}\mathcal{D}_N$, и тем сильнее, чем наблюденное значение выше.

Рассмотрим вероятность того, что в независимом повторении проведенного опыта мы получим такое же или даже большее значение статистики $\sqrt{N}\mathcal{D}_N$, чем наблюденное. (Вероятность эту вычисляем в предположении, что гипотеза верна). Наблюденное значение надо признать большим, если его трудно превзойти за счет случайности. То есть, если упомянутая вероятность - малая. И обратно: если эта вероятность не мала, то и наблюденное значение считать большим не следует; оно совместимо с проверяемой гипотезой.

Обсуждаемую вероятность называют p-значением (по-английски - *p-value*). Применять p-значения для проверки гипотез предложил Фишер (*R. Fisher*).

В данной задаче p-значение равно $1 - K(\sqrt{N}\mathcal{D}_N)$.

Вернемся к опытам Ермолаевой. Всего было две выборки: $N = 98$ и $N = 123$. В обеих выборках наблюденные значения $\mathcal{D}_N$ были далеки от критических: их p-значения были равны 0.51 и 0.63 соответственно. Таким образом, научная атака Т.Д. Лысенко на законы Менделя не удалась.

6.3. Оптимальный критерий Неймана-Пирсона

(*J. Neyman, S. Pearson, 1933*)

Статистический критерий S для проверки гипотезы $H_0\colon P \in \mathcal{P}_0$ против альтернативы $H_1\colon P \in \mathcal{P}_1$ естественно называть оптимальным, если среди всех критериев заданного уровня значимости критерий S имеет наибольшую мощность.

Чуть подробнее. Из двух критериев R и S данного уровня значимости критерий S называют более мощным, если

$$\beta(Q, S) \geqslant \beta(Q, R) \qquad \text{для всех } Q \in \mathcal{P}_1. \tag{1}$$

Критерий S называют *оптимальным критерием* уровня α, если для любого другого критерия R уровня α выполняется соотношение (1). Критерий S в этом случае называют также *равномерно наиболее мощным критерием* уровня α.

Оптимальный выбор критерия для проверки гипотезы $H_0 : P \in \mathcal{P}_0$ против альтернативы $H_1\colon P \in \mathcal{P}_1$ 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 к простейшей ситуации: и гипотеза H_0, и альтернатива H_1 — простые, то есть оба множества $\mathcal{P}_0$ и $\mathcal{P}_1$ — одноточечные, каждое из них состоит из одного распределения вероятностей P_0 и P_1 соответственно. (Если множества $\mathcal{P}_0$ и $\mathcal{P}_1$ состоят каждое более чем из одного распределения, гипотезу $H_0 : P \in \mathcal{P}_0$ и альтернативу $H_1 : P \in \mathcal{P}_1$ называют *сложными*).

Оптимальный критерий для проверки простой гипотезы против простой альтернативы мы построим в элементарной ситуации, когда распределения P_0 и P_1 либо оба дискретны, либо оба имеют плотности (относительно некоторой меры на $\mathcal{X}$).

Пусть $f_0(x)$ и $f_1(x)$, $x \in \mathcal{X}$, суть две плотности распределений на $\mathcal{X}$ (или два дискретных распределения на $\mathcal{X}$). Пусть наблюдение X получено выбором элемента из $\mathcal{X}$ согласно f_0 либо f_1. Рассмотрим гипотезу $H_0 : X$ имеет плотность (распределение) f_0 и альтернативу $H_1 : X$ имеет плотность (распределение) f_1.

Рассмотрим множество вида

$$S_\lambda = \{x : f_1(x) - \lambda f_0(x) \geqslant 0\}, \quad \lambda > 0 \tag{2}$$

как критерий для H_0 против H_1. [Точнее, мы рассмотрим всё семейство множеств указанного вида, параметризованное переменной $\lambda > 0$, как семейство критических множеств. Эти критические множества различаются уровнями значимости.]

Пусть R — какой-либо статистический критерий для проверки H_0 против H_1 по наблюдению X, $R \subset \mathcal{X}$.

Предположим, что

$$P_0(X \in R) \leqslant P_0(X \in S_\lambda). \tag{3}$$

То есть вероятность ошибки I рода для R не выше чем для S_λ. [В типичном случае для данного R можно подобрать критерий S_λ вида (b) с тем же уровнем значимости. Тогда в (3) стоит равенство.]

Тогда

(a)

$$P_1(X \in R) \leqslant P_1(X \in S_\lambda),$$

(b)

$$P_0(X \in S_\lambda) \leqslant P_1(X \in S_\lambda).$$

Пункт (a) означает, что критерий S_λ имеет наибольшую мощность среди всех критериев, уровень значимости которых не превосходит уровня значимости S_λ.

Пункт (b) касается свойств самого критерия S_λ и утверждает, что функция мощности критерия S_λ возрастает при переходе от гипотетического распределения P_0 к альтернативному P_1. [Такое свойство критерия называют *несмещенностью*. Оно означает, что более вероятно (с помощью этого критерия) отвергнуть проверяемую гипотезу, когда она неверна, чем когда она верна - весьма естественное качество для критерия.]

Критерии вида (2) называют *критериями Неймана – Пирсона*, а сформулированное выше утверждение об оптимальности критериев (2) — *леммой (теоремой) Неймана – Пирсона*.

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.

Записи будут компактными, если вместе с критериями R и S_λ рассмотреть их индикаторные функции $I_R(x)$ и $I_S(x)$:

$$I_R(x) = \begin{cases} 1, & \text{для } x \in R, \\ 0, & \text{для } x \notin R; \end{cases}, \qquad I_S(x) = \begin{cases} 1, & \text{для } x \in S_\lambda, \\ 0, & \text{для } x \notin S_\lambda. \end{cases}$$

С помощью I_R, I_S вероятности событий $(X \in R)$, $(X \in S_\lambda)$ можно записать в виде математических ожиданий. Усреднение (математическое ожидание) по P_0 обозначим через E_0, усреднение по P_1 - через E_1. Например, $P_0(X \in R) = \mathsf{E}_0 I_R(X)$, а предложение (3) имеет вид

$$\mathsf{E}_0 I_R(X) \leqslant \mathsf{E}_0 I_S(X). \tag{4}$$

Доказательство утверждения (a).

Легко проверить, что справедливо неравенство

$$I_R(x)[f_1(x) - \lambda f_0(x)] \leqslant I_S(x)[f_1(x) - \lambda f_0(x)]. \tag{5}$$

Действительно, если $f_1(x) - \lambda f_0(x) > 0$, то $I_S(x) = 1$ и (5) превращается в очевидное утверждение $I_R(x) \leqslant 1$. Если же $f_1(x) - \lambda f_0(x) < 0$, то $I_S(x) = 0$, и потому правая часть (5) обращается в нуль, а левая часть (5) при этом неположительна, так что (5) верно и в этом случае.

Интегрируем (5) по всему пространству. Результат запишем в виде математических ожиданий.

$$\mathsf{E}_1 I_R(X) - \lambda \mathsf{E}_0 I_R(X) \leqslant \mathsf{E}_1 I_S(X) - \lambda \mathsf{E}_0 I_S(X)$$

или

$$\mathsf{E}_1 I_S(X) - \mathsf{E}_1 I_R(X) \geqslant \lambda[\mathsf{E}_0 I_S(X) - \mathsf{E}_0 I_R(X)]. \tag{6}$$

В силу (4) и $\lambda > 0$ правая часть (6) неотрицательна, что и доказывает (a). □

Доказательство утверждения (b).

Для доказательства утверждения (b) надо порознь рассмотреть для λ, определяющего S_λ в (2), две возможности: $\lambda \geqslant 1$ и $\lambda < 1$.

- Допустим, что $\lambda \geqslant 1$. Тогда из (2) следует, что $f_1(x) \geqslant f_0(x)$ для $x \in S_\lambda$. Поэтому

$$P_0(X \in S_\lambda) = \int I_S(x) f_0(x) dx \leqslant \int I_S(x) f_1(x) dx = P_1(X \in S_\lambda),$$

что и требуется.

- Допустим, что $\lambda < 1$. Рассмотрим множество

$$\overline{S}_\lambda = \{x : f_1(x) \leqslant \lambda f_0(x)\}$$

Его индикатор есть $1 - I_S(x)$. При $\lambda < 1$ получаем, что $f_1(x) \leqslant f_0(x)$ для $x \in \overline{S}_\lambda$. Поэтому

$$P_1(X \in \overline{S}_\lambda) = \int [1 - I_S(x)] f_1(x) dx \leqslant \int [1 - I_S(x)] f_0(x) dx = P_0(X \in \overline{S}_\lambda).$$

Отсюда следует, что при $\lambda < 1$

$$1 - P_1(X \in S_\lambda) \leqslant 1 - P_0(X \in S_\lambda).$$

Это доказывает (b) и в этом случае. □

Доказанная теорема определяет вид наилучшего критерия. Если мы хотим остановиться на оптимальном критерии уровня ε, где ε задано, мы должны подобрать $\lambda > 0$ так, чтобы

$$P_0(X \in S_\lambda) = \varepsilon. \tag{7}$$

В случае плотности это означает, что мы должны решить относительно λ уравнение

$$\int\limits_{\{x : f_1(x) \geqslant \lambda f_0(x)\}} f_0(x) dx = \varepsilon.$$

В типичном случае решение существует (и единственно).

Для дискретно распределенных наблюдений X уравнение (7) разрешимо не для всех $\varepsilon > 0$. В таком случае в поисках оптимального критерия уровня ε либо останавливаются на критерии вида (2) с меньшим, чем назначенный ε, с вероятностью ошибки I рода (увеличивая тем самым вероятность ошибки II рода), либо изменяют выбор уровня значимости так, чтобы (7) стало разрешимо. Последнее правильнее, ибо назначение уровня значимости — решение в немалой степени произвольное.

6.4. Равномерно наиболее мощные критерии

Определение равномерно наиболее мощных критериев дано в начале пункта 3. Как правило, для сложных гипотез и/или сложных альтернатив равномерно наиболее мощных критериев не существует. Типично такое положение, когда для каждой пары распределений $P_0 \in \mathcal{P}_0$, $P_1 \in \mathcal{P}_1$ есть «свой» (определяемый леммой Неймана – Пирсона) оптимальный критерий, но нет единого оптимального критерия. Но есть важные (для практики) исключения из сказанного, когда равномерно наиболее мощные критерии существуют.

Пример:

проверка односторонних гипотез против односторонних альтернатив в схеме Бернулли.

Пусть проведено n, n задано, испытаний Бернулли. Пусть $\theta \in (0, 1)$ - неизвестная вероятность успеха. Обозначим результат испытаний через $X = (X_1, \ldots, X_n)$, где $X_i = 1$, если в i-ом испытании был успех, и $X_i = 0$ в противном случае.

По наблюденному X надо проверить гипотезу

$$H_0 : \theta \leqslant \theta^0$$

против альтернативы

$$H_1 : \theta > \theta^0,$$

где $\theta^0 \in (0, 1)$ задано.

Далее мы найдем р. н. м. критерий для проверки H_0 против H_1. Этот критерий будет найден с помощью правила Неймана – Пирсона.

Произвольно выберем два значения a и b параметра θ: a из гипотетического множества $(0, \theta^0]$, b из альтернативного множества $(\theta^0, 1)$:

$$0 < a \leqslant \theta^0 < b < 1 \tag{1}$$

Для проверки простой гипотезы $\theta = a$ против простой альтернативы $\theta = b$ применим правило Неймана – Пирсона. Здесь:

$$f_1(x) = b^{T_n(x)}(1 - b)^{n - T_n(x)},$$
$$f_0(x) = a^{T_n(x)}(1 - a)^{n - T_n(x)},$$

где $x = (x_1, \ldots, x_n)$ - точка выборочного пространства $\mathcal{X}$, x — произвольная последовательность из нулей и единиц, $T_n(x) = \sum\limits_{i=1}^{n} x_i$. (Заметим, что $T_n(X)$ - знакомая нам достаточная статистика, общее число успехов.)

Критические множества Неймана – Пирсона для пары a, b суть

$$S_\lambda = \{x : \frac{f_1(x)}{f_0(x)} \geqslant \lambda\}, \quad \lambda > 0$$

или

$$S_\lambda = \left\{ x : \left(\frac{b}{a}\frac{1-a}{1-b}\right)^{T_n(x)} \left(\frac{1-b}{1-a}\right)^{n} \geqslant \lambda \right\}, \quad \lambda > 0. \tag{2}$$

Мы уже отмечали, что критерии Неймана – Пирсона образуют семейство оптимальных критериев. Из этого семейства потом выбирают критерий заданного уровня значимости. Сейчас семейство (2) параметризовано параметром λ, $\lambda > 0$. Любая другая параметризация этого семейства ничуть не хуже.

В частности, семейство (2) можно записать в виде

$$\left\{ x : \left(\frac{b}{1-b}\frac{1-a}{a}\right)^{T_n(x)} \geqslant \lambda' \right\}, \quad \lambda' > 0,$$

где $\lambda' = \lambda \left(\frac{1-a}{1-b}\right)^n$. Впрочем, связь между новым параметром λ' и старым параметром λ не важна. При дальнейших изменениях параметризации мы такие связи отмечать не будем. В силу (1)

$$\frac{b}{1-b}\frac{1-a}{a} > 1.$$

Поэтому (2) можно еще упростить:

$$\{x : T_n(x) \geqslant t\}, \ t > 0. \tag{3}$$

Отметим главную особенность (3) как статистического критерия: его вид не зависит от конкретных $a \leqslant \theta^0$, $b > \theta^0$. Этот критерий - общий для всех $a \in (0, \theta^0]$, $b \in (\theta^0, 1)$. Это означает, что критерий (3) в рассматриваемой задаче является равномерно наиболее мощным.

Статистическое правило теперь таково:

Отвергать гипотезу $H_0 : \theta \leqslant \theta^0$ против альтернативы $H_1 : \theta > \theta^0$, если произошло событие

$$T_n(X) \geqslant t, \tag{4}$$

где t - некоторое критическое значение. (Это значение t еще предстоит уточнить). Заметим, что решение основывается на достаточной статистике $T_n(X) = \sum\limits_{i=1}^{n} X_i$ (суммарном числе успехов), а не на самом наблюдении X. Эта черта характерна для всякого критерия в тех статистических моделях X, где существуют достаточные статистики.

Остается определить критическое значение t в (4).

Для этого зададимся некоторым уровнем значимости ε. Для t должно выполняться условие

$$P_\theta(T_n \geqslant t) \leqslant \varepsilon \qquad \text{для всех } \ \theta \leqslant \theta^0.$$

Из утверждения (b) леммы Неймана – Пирсона следует, что

$$\sup_{\theta : \theta \leqslant \theta^0} P_\theta(T_n \geqslant t) = P_{\theta^0}(T_n \geqslant t).$$

Поэтому условие для выбора t упрощается:

$$P_{\theta^0}(T_n \geqslant t) \leqslant \varepsilon. \tag{5}$$

Ради достижения наибольшей мощности против альтернативы $\theta > \theta^0$ в качестве критического значения следует взять наименьшее t, удовлетворяющее (5). Выбор t при заданных θ и n помогают осуществить таблицы для вероятности

$$P_\theta(T_n \geqslant t) = \sum_{k \geqslant t} C_n^k \theta^k (1-\theta)^{n-k}$$

как функции от θ и t; $\theta \in (0,1), t = \overline{1,n}$.

Можно не связывать себя заранее выбранным уровнем значимости и принимать решения на основе p-значения (p-*value*) критической статистики. В нашем случае против проверяемой гипотезы говорят большие значения критической статистики T_n. p-значение определяется как вероятность получить (при независимом повторении опыта) не меньшее, чем получено, значение критической статистики (не менее сильное, чем получено, свидетельство против проверяемой гипотезы).

Если наблюденное значение значение статистики T_n обозначить как T_n(набл.), сохранив за T_n смысл случайной переменной, то p-значением T_n(набл.) служит

$$P_{\theta^0}(T_n \geqslant T_n(\text{набл.})). \tag{6}$$

Сопоставляя это выражение с (5), видим, что p-значение - это наименьший уровень значимости, на котором еще можно опровергнуть гипотезу H_0 по правилу (5).

Испытания Бернулли служат статистической моделью для многих реальных процессов. В частности, при (массовом) производстве изделие может оказаться негодным (брак). Если предположить, что появление брака - дело случая, что бракованными различные изделия могут оказаться независимо друг от друга и что, наконец, вероятность появления бракованного изделия постоянна, для описания процесса мы можем применить схему Бернулли. Присутствие среди изделий некоторой доли θ бракованных неизбежно для любого производства.

Величина θ^0 может служить границей для все еще допустимой доли брака; если эта доля выше, в производство требуется вмешательство (наладка станков, например).

Для контроля за долей текущего брака нужно производить регулярные проверки: нужно проверять гипотезу $H_0 : \theta \leqslant \theta^0$ против $H_1 : \theta > \theta^0$. Выше мы установили, как это следует делать наилучшим образом при простейшем плане эксперимента - выборке.

Как объем выборки n, так и частота описанных проверок в нашей постановке не определяются. Их устанавливают, исходя из расходов на организацию и проведение контроля, от потерь от увеличения доли брака, скорости изменения θ в течение работы и т.д.

Планы выборочного контроля, реально применяемые на массовых производствах, могут быть значительно сложнее, чем изученная нами простая выборка и контроль по качественному признаку (когда изделие либо годно, либо нет). Научная и техническая литература, посвященная контролю качества продукции, необъятна.

Равномерно наиболее мощные критерии (для проверки односторонних гипотез против односторонних альтернатив) типичны для однопараметрических экспоненциальных семейств распределений. Об этих семействах мы упоминали в связи с неравенством Крамера – Рао и эффективными оценками. Плотность (вероятность) наблюдения X при этом равна

$$p(x, \theta) = \exp\{c(\theta)T(x) + S(x) + d(\theta)\}I_A(x).$$

Биномиальные распределения, которые мы исследовали выше, принадлежат этому классу. Если функция $c(\theta)$ монотонно зависит от θ, все проведенные выше выкладки повторяются практически без изменений и приводят к решающим правилам вида

$$T_n \geqslant t \qquad \text{либо} \qquad T_n \leqslant t.$$

6.5. Проверка линейных гипотез

6.5.1. Выбор степени многочлена

В задачах регрессии y по x функциональный вид зависимости ожидаемого значения отклика $\mathsf{E}(y|x)$, как функции x, бывает известен далеко не всегда. В таких случаях аппроксимирующее выражение для $\mathsf{E}(y|x)$ подбирают эмпирически. Часто для приближения выражения $\mathsf{E}(y|x)$ используют многочлены от x.

Пусть, для простоты, x - скалярная переменная. Предположим, что:

$$\mathsf{E}(y|x) = a_0 + a_1 x + \cdots + a_p x^p \tag{1}$$

для некоторой степени $p \geqslant 0$ и некоторых коэффициентов $a_0, a_1, \ldots, a_p$.

Далее предположим, что при некоторых заданных значениях $x_1, \ldots, x_n$ фактора x проведены независимые измерения $y_1, \ldots, y_n$ отклика y, так что

$$y_i = a_0 + a_1 x_i + \cdots + a_p x_i^p + \varepsilon_i, \quad i = \overline{1, n}, \tag{2}$$

где $\varepsilon_1, \ldots, \varepsilon_n$ суть независимые случайные величины (ошибки). Мы предположим, что $\varepsilon_i \sim N(0, \sigma^2), i = \overline{1, n}$, причем дисперсия ошибки σ^2 неизвестна.

Выбор степени p аппроксимирующего многочлена в формуле (1) всегда представляет определенную проблему. Эту степень надо выбрать так, чтобы погрешность в (1) (она же - систематическая ошибка в (2)) не влияла на статистические выводы о $\mathsf{E}(y|x)$, которые мы сумеем сделать по наблюдениям (x_i, y_i), $i = \overline{1, n}$. Чем ниже эта степень, тем легче интерпретировать результаты опытов. На практике эта степень редко превышает 3.

Особенно часто приходится отвечать на вопрос: можно ли для аппроксимации $\mathsf{E}(y|x)$ обойтись многочленом первой степени, т.е. простой линейной регрессией, или же надо обратиться к параболической регрессии, т.е. к многочлену второй степени?

Статистически проблема выглядит так.

Предположим, что наблюдения y удовлетворяют статистической модели

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \varepsilon_i, \quad i = \overline{1,n}, \tag{3}$$

где $\varepsilon_1, \ldots, \varepsilon_n$ — суть неизвестные $N(0, \sigma^2)$, a_0, a_1, a_2 - неизвестные коэффициенты. По наблюдениям (3) надо проверить гипотезу

$$H_0 : a_2 = 0 \tag{4}$$

против альтернативы

$$H_1 : a_2 \neq 0$$

Гипотеза H_0 (4) состоит в том, что зависимость отклика от фактора можно передать моделью

$$y_i = a_0 + a_1 x_i + \varepsilon_i, i = \overline{1,n}$$

при тех же, что и выше, предположениях об ошибках $\varepsilon_1, \ldots, \varepsilon_n$.

6.5.2. Однофакторный дисперсионный анализ

Пусть наблюдаются $k \geqslant 2$ независимых выборок, объемы которых обозначим через $n_1, \ldots, n_k$. Элементы выборки номер j, $j = \overline{1,k}$, обозначим через x_{ij}, i меняется от 1 до n_j. Предположим, что

$$x_{ij} = a_j + \varepsilon_{ij}, \tag{5}$$

где ε_{ij} $(j = \overline{1,k}, i = \overline{1,n_j})$ суть независимые одинаково распределенные случайные величины. Всюду в дальнейшем $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Такая модель возникает, например, при сравнении нескольких способов обработки, нескольких условий хранения, нескольких мест размещения и т.д. Модель (5) возникает также при любой классификации объектов по одному признаку (однофакторная классификация).

При сравнении способов обработки часто бывает нужно выделить лучший (или группу лучших, или группу наихудших и т.п.) способов обработки. Целесообразно, однако, прежде задаться вопросом: дают ли наши данные основания для такого выбора? По-видимому, нет, если с наблюдениями (5) совместима гипотеза

$$H_0 : a_1 = a_2 = \cdots = a_k. \tag{6}$$

Легко видеть, что гипотеза (4) в модели (3) и гипотеза (6) в модели (5) являются частными формами общей линейной гипотезы в линейной модели, как она формулируется в следующем разделе.

6.5.3. Общая линейная гипотеза

Мы говорим, что в отношении наблюдения X (X - элемент линейного пространства, в наших рассмотрениях $X \in \mathbb{R}^n$) действует линейная модель, если наблюдение X имеет структуру $X = l + \xi$, где

- l - неслучайный неизвестный вектор, который заведомо принадлежит некоторому заданному линейному подпространству L;
- ξ - случайный вектор (вектор ошибок).

Модель называют *гауссовской*, если ξ имеет гауссовское распределение. В большинстве приложений $\mathsf{E}\xi = 0$, $\mathsf{D}\xi = \sigma^2 I$, причем σ^2 неизвестно. (Такая форма матрицы ковариаций ξ означает, что компоненты вектора X независимы и имеют одинаковые дисперсии.)

Линейная гипотеза: гипотеза $H_0 : l \in L_0$, где L_0 - заданное линейное подпространство, причем $L_0 \subset L$. Альтернативой к H_0 выступает отрицание H_0 в рамках линейной модели: $H_1 : l \notin L_0$, но при этом $l \in L$.

Линейную гипотезу можно рассматривать как частный случай общей параметрической гипотезы о распределении наблюдения X.

6.5.4. Критерий отношения правдоподобий

Предположим, что случайная величина X имеет плотность $f(x,\theta)$, где $\theta \in \Theta$ — неизвестный параметр. Плотность берётся относительно некоторой меры, в нашем случае — относительно меры Лебега в $\mathbb{R}^n$.

Гипотеза H_0 состоит в том, что параметр θ принадлежит заданному множеству Θ_0, более узкому, чем Θ: $\Theta_0 \subset \Theta$. Критерий, предлагаемый для проверки $H_0 : \theta \in \Theta_0$ против $H_1 : \theta \in \Theta \setminus \Theta_0$, строится по образцу критерия Неймана – Пирсона.

- Пусть $\hat\theta$ обозначает оценку параметра θ, вычисленную по наблюдению X в предположении, что $\theta \in \Theta$.
- Пусть $\hat{\hat\theta}$ обозначает аналогичную оценку, но вычисленную в предположении, что $\theta \in \Theta_0$.
- Критические события теперь имеют вид

$$S_\lambda = \left\{ X : \frac{f(X,\hat\theta)}{f(X,\hat{\hat\theta})} \geqslant \lambda \right\}. \tag{1}$$

Параметр λ, как обычно, выбирают по заданному уровню значимости ε из условия

$$P(S_\lambda | H_0) \leqslant \varepsilon.$$

Критерий (1) называют *критерием отношения правдоподобий.*

В рассматриваемой нами линейной модели оценки $\hat\theta$, $\hat{\hat\theta}$ (для пары (l,σ^2)) нам известны, и вскоре мы к ним обратимся. В общей задаче в качестве $f(x,\hat\theta)$ и $f(x,\hat{\hat\theta})$ обычно берут

$$f(X,\hat\theta) = \max_{\theta \in\ \theta} f(X,\theta),$$

$$f(X,\hat{\hat\theta}) = \max_{\theta\in\Theta_0} f(X,\theta).$$

Получаемые по такому правилу оценки θ

$$\hat\theta = arg\max_{\theta\in\Theta} f(X,\theta) \text{ и } \hat{\hat\theta} = arg\max_{\theta\in\Theta_0} f(X,\theta)$$

называют *оценками наибольшего правдоподобия* (при условиях $\theta \in \Theta$ и $\theta \in \Theta_0$).

Критерий отношения правдоподобий теперь имеет такие критические события:

$$S_\lambda = \left\{ X : \frac{\max\limits_{\theta \in\ \theta} f(X,\theta)}{\max\limits_{\theta\in\Theta_0} f(X,\theta)} > \lambda \right\}.$$

Само выражение $f(X,\theta)$, рассматриваемое как функция θ, называют *правдоподобием* θ.

Отсюда и названия: оценки наибольшего правдоподобия и критерий отношения правдоподобий. Свойства оценок наибольшего правдоподобия мы еще будем изучать, но позже.

6.5.5. Применение критерия отношения правдоподобий к проверке линейных гипотез

Применим критерий отношения правдоподобий к проверке линейных гипотез. В рассматриваемой гауссовской модели правдоподобие есть

$$f(X,\theta) = \left(\frac{1}{\sqrt{2\pi}}\right)^n (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2}|X-l|^2 \right\}. \tag{1}$$

При условии, что $l \in L$, оценки $\hat l, \hat\sigma^2$ суть

$$\hat l = \operatorname{proj}_L X, \tag{2}$$

$$\hat\sigma^2 = \frac{1}{n-m}|\operatorname{proj}_{L^\perp} X|^2 = \frac{1}{n-m}|X - \operatorname{proj}_L X|^2,$$

где $m = \dim L$.

При условии, что $l \in L_0$, оценки $\hat{\hat l}, \hat{\hat\sigma}^2$ суть

$$\hat{\hat l} = \operatorname{proj}_{L_0} X, \tag{3}$$

$$\hat{\hat{\sigma}}^2 = \frac{1}{n-m_0}|\operatorname{proj}_{L_0^\perp} X|^2 = \frac{1}{n-m_0}|X - \operatorname{proj}_{L_0} X|^2,$$

где $m_0 = \dim L_0$.

В обоих случаях показатель экспоненты $-\frac{|X-l|^2}{2\sigma^2}$ при подстановке вместо l, σ^2 их оценок превращается в постоянную, не зависящую от X величину: в первом случае это - $-(n-m)/2$, во втором - $-(n-m_0)/2$.

Поэтому семейство критических событий (6.5.4.1) для проверки гипотезы H_0 имеет вид

$$\{X : \frac{|X - \operatorname{proj}_{L_0} X|^2}{|X - \operatorname{proj}_L X|^2} \geqslant \lambda\}. \tag{4}$$

(Параметр λ в (4) не тождественен параметру λ в (6.5.4.1); несмотря на это мы употребили для них один и тот же символ. Как уже отмечалось в 6.4, для нас важна параметризация семейства критических событий, но не связь между различными возможными параметризациями. Поэтому соотношение между параметрами в (6.5.4.1) и (4) мы можем оставить без внимания.)

Ради дальнейшего упрощения (4) введем в рассмотрение еще одно линейное подпространство: ортогональное дополнение L_0 до L. Обозначим его через L_1. Итак, $L_1 \perp L_0$, $L_0 \oplus L_1 = L$. Теперь $\mathbb{R}^n$ представимо в виде суммы трех попарно ортогональных подпространств L_0, L_1 и $L^\perp$. (Как обычно, $L^\perp$ обозначает ортогональное дополнение L до всего пространства $\mathbb{R}^n$):

$$\mathbb{R}^n = L_0 + L_1 + L^\perp.$$

В связи с этим для X действует разложение

$$X = \operatorname{proj}_{L_0} X + \operatorname{proj}_{L_1} X + \operatorname{proj}_{L^\perp} X,$$

причем

$$|X - \operatorname{proj}_{L_0} X|^2 = |\operatorname{proj}_{L_1} X|^2 + |\operatorname{proj}_{L^\perp} X|^2. \tag{5}$$

В силу (5) критерий отношения правдоподобий (4) можно преобразовать:

$$\frac{\frac{1}{m_1}|\operatorname{proj}_{L_1} X|^2}{\frac{1}{n-m}|\operatorname{proj}_{L^\perp} X|^2} \geqslant \lambda \tag{6}$$

с учетом замечаний к (4).

Вспомним, что оценкой для σ^2 при условии, что $l \in L$, служит

$$\frac{1}{n-m}|\operatorname{proj}_{L^\perp} X|^2. \tag{7}$$

Это несмещенная оценка для σ^2, вне зависимости от того, верна или нет гипотеза $H_0 : l \in L_0$. Если же H_0 верна, то для σ^2 можно предложить еще одну несмещенную оценку, притом статистически независимую от первой: это

$$\frac{1}{m_1}|\operatorname{proj}_{L_1} X|^2. \tag{8}$$

Если гипотеза H_0 неверна, оценка (8) преображает смещение - тем больше, чем больше $|\operatorname{proj}_{L_1} l|^2$. (Но о смещении - чуть позже, когда будем говорить о распределениях (7) и (8)). Поэтому критериальная статистика в (6) - это отношение двух независимых оценок дисперсии. Если гипотеза H_0 верна, это отношение отличается от 1 только за счет случайных колебаний. Представление об их размере дает распределение статистики (6) при гипотезе.

Обсудим распределение статистики из (6) при гипотезе и при альтернативе. Лемма об ортогональном разложении 4.3.2 говорит, что

$$|\operatorname{proj}_{L^\perp} X|^2 \stackrel{d}{=} \sigma^2\chi^2(n-m),$$

$$|\operatorname{proj}_{L_1} X|^2 \stackrel{d}{=} \sigma^2\chi^2(m_1, \Delta),$$

где параметр нецентральности $\Delta = \frac{1}{\sigma^2}|\operatorname{proj}_{L_1} l|^2$. Если верна гипотеза H_0, то $\Delta = 0$.

Критериальная статистика из (6) распределена как $F(m_1, n-m, \Delta)$:

$$\frac{\frac{1}{m_1}|\operatorname{proj}_{L_1} X|^2}{\frac{1}{n-m}|\operatorname{proj}_{L^\perp} X|^2} \stackrel{d}{=} F(m_1, n-m, \Delta). \tag{9}$$

(Соотношение (9) объясняет, между прочим, и принятое для эф-отношения название дисперсионного отношения Фишера.)

Примечательно, что при гипотезе H_0 статистика (9) распределена свободно (от влияния неизвестных параметров $l, l \in L_0$, и σ^2). (Это свойство получено нами сверх ожиданий. Ничто в наших выкладках того не обещало.) Поэтому выбор критического значения λ в (6) очень упрощается: для этого надо (с помощью таблиц распределения, например) решить уравнение

$$P\{F(m_1, n-m, \Delta) \geqslant \lambda\} = \varepsilon.$$

В качестве критического значения (для проверки H_0 на уровне ε) в (6) надо взять $(1-\varepsilon)$-квантиль эф-распределения с m_1, $n-m$ степенями свободы. Которую мы уже когда-то обозначили $F_{1-\varepsilon}(m_1, n-m)$.

С вычислительной точки зрения более удобной формой для статистики (9) может быть выражение

$$\frac{\frac{1}{m-m_0}|\operatorname{proj}_L X - \operatorname{proj}_{L_0} X|^2}{\frac{1}{n-m}|X - \operatorname{proj}_L X|^2}. \tag{10}$$

Итак, получили статистическое правило:

- Отвергаем гипотезу H_0 на уровне ε, если статистика (9) или (10) превосходит $F_{1-\varepsilon}(m_1, n-m)$.

Из свойств эф-отношения следует, что мощность этого критерия монотонно возрастает вместе с ростом параметра нецентральности $\Delta = \frac{1}{\sigma^2}|\operatorname{proj}_{L_1} X|^2$.

6.5.6. Пример: две нормальные выборки.

Рассмотрим две независимые нормальные выборки

- $x_1, \dots, x_m$, где $x_i \sim N(a, \sigma^2)$
  и
- $y_1, \dots, y_n$, где $y_i \sim N(b, \sigma^2)$,
  параметры a, b и σ^2 неизвестны.

Подлежащая проверке гипотеза

$$H_0 : a = b. \tag{1}$$

Альтернатива к ней

$$H_1 : a \neq b.$$

В $(n+m)$-мерном пространстве рассмотрим векторы

$$Z = (x_1, \dots, x_m, y_1, , \dots, y_n)^T,$$

$$e_1 = (\underbrace{1, \dots, 1}_{m}, \underbrace{0, \dots, 0}_{n})^T,$$

$$e_2 = (\underbrace{0, \dots, 0}_{m}, \underbrace{1, \dots, 1}_{n})^T,$$

$$\varepsilon = (\xi_1, \dots, \xi_m, \xi_{m+1}, \dots, \xi_{m+n})^T,$$

где $\xi_1, \xi_2, \dots$ суть независимые $N(0, \sigma^2)$.

Вектор Z можно представить в виде

$$Z = ae_1 + be_2 + \varepsilon.$$

Ясно, что Z следует линейной гауссовской модели, причем $\mathsf{E}Z \in L(e_1, e_2)$, где $L(e_1, e_2)$ обозначает (двумерное) линейное подпространство с базисом e_1, e_2. При гипотезе H_0 вектор $\mathsf{E}Z$ лежит в одномерном линейном подпространстве L_0, порожденном единственным вектором $e_1 + e_2$.

Для проверки H_0 против H_1 с помощью статистики (6.5.4.10) надо вычислить

$$|\operatorname{proj}_L Z - \operatorname{proj}_{L_0} Z|^2 \text{ и } |Z - \operatorname{proj}_L Z|^2.$$

Будем использовать обозначения

$$\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \qquad s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2,$$

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j, \qquad s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2.$$

Легко видеть, что

$$\mathrm{proj}_L Z = \overline{x}e_1 + \overline{y}e_2,$$

$$\mathrm{proj}_{L_0} Z = \left(\frac{m}{m+n}\overline{x} + \frac{n}{m+n}\overline{y}\right)(e_1 + e_2).$$

Отсюда

$$|Z - \mathrm{proj}_L Z|^2 = (m-1)s_x^2 + (n-1)s_y^2,$$

$$|\,\mathrm{proj}_L Z - \mathrm{proj}_{L_0} Z|^2 = \frac{mn}{m+n}(\overline{x} - \overline{y})^2.$$

В этих обозначениях статистика (6.5.5.10) и последующее статистическое правило таковы:

- Отвергать $H_0 : a = b$ на уровне ε, если

$$\frac{mn(m+n-2)}{m+n}\,\frac{(\overline{x} - \overline{y})^2}{\sum\limits_{i=1}^{m}(x_i - \overline{x})^2 + \sum\limits_{j=1}^{n}(y_j - \overline{y})^2} > F_{1-\varepsilon}(1, m+n-2). \tag{2}$$

Обычно вместо эф-статистики (2) рассматривают статистику Стьюдента t, причем $t^2 = F$:

$$t = \frac{\sqrt{\frac{mn}{m+n}}(\overline{x} - \overline{y})}{\sqrt{\frac{1}{m+n-2}[(m-1)s_x^2 + (n-1)s_y^2]}}. \tag{3}$$

При гипотезе H_0 статистика (3) распределена по Стьюденту с $m+n-2$ степенями свободы.
С помощью (3) можно отдельно проверять H_0 против односторонних альтернатив: против правосторонней

$$H^+ : a > b$$

или левосторонней

$$H^- : a < b.$$

6.5.7. Заключение

Теория гауссовских линейных моделей составляет классическую главу математической статистики, ее большое достижение и достояние. Вместе с тем, с прикладной точки зрения, гауссовские методы не свободны от недостатков и ограничений.

Эти методы не следует применять, если распределение наблюдений (или ошибок) определенно не гауссовское. В статистических задачах за пределами геодезии, астрономии и т.п. негауссовские ошибки - это скорее правило, чем исключение.

Гауссовские методы (к которым я здесь отношу и метод наименьших квадратов) применять опасно, если распределения близки к гауссовским, но не исключают появления далеко отстоящих от центра наблюдений. (Их называют грубыми ошибками или *выбросами*.) Статистические оценки (и другие правила), оптимальные для гауссовских распределений, оказываются чувствительными к выбросам. Даже небольшая доля таких «засоряющих» значений в общем массиве данных может радикально изменить результаты статистического анализа.

Поэтому для приложений нужны и другие статистические методы. Об одном из них, не опирающемся на какую-либо параметрическую форму распределений (и поэтому называемом параметрическим), простом математически и достаточно универсальном, будем рассказывать далее.

7. Ранговые методы

7.1. Общее определение рангов

От любой числовой последовательности (в которой нет повторяющихся чисел) можно перейти к последовательности их номеров, если указан принцип их линейного упорядочения (нумерации). Обычно числовые совокупности упорядочивают от меньшего к большему, т.е. в возрастающем порядке. (Но бывает и по-другому.)

Номера, которые получили элементы числовой последовательности при упорядочении, называют их *рангами*.

(Понятно требование, чтобы в совокупности не было одинаковых чисел: неясно, как упорядочить одинаковые числа. Им надо бы дать одинаковые номера). Как бы ни проводилось упорядочение числовой совокупности, совокупность их рангов - это одна из перестановок натуральных чисел $1, 2, \ldots, n$, где n - размер исходной совокупности.

Пусть теперь исходная совокупность $X = (x_1, \ldots, x_n)$ - выборка из некоторого непрерывного распределения. С вероятностью 1 эта выборка не имеет одинаковых элементов.

Рассмотрим ранги величин $x_1, \ldots, x_n$. Для определенности, при упорядочении в порядке возрастания. Обозначим из через $R(x_1), \ldots, R(x_n)$.

Основное свойство случайных рангов:

$$P(\vec{R}(X) = \vec{r}) := P(R(x_1) = r_1, R(x_2) = r_2, \ldots) = \frac{1}{n!},$$

где $(r_1, \ldots, r_n)$ - произвольная перестановка чисел $(1, 2, \ldots, n)$.

Заметим, что распределение рангов - равномерное и не зависит от того, каким было исходное распределение случайных величин $(x_1, \ldots, x_n)$, т.е. выборки X. (Если исходное распределение - непрерывное).

7.2. Сравнение двух выборок, могущих отличаться сдвигом: постановка задачи

Пусть:

- $X = (x_1, \ldots, x_n)$ - выборка, функция распределения $P(x_i \leqslant u) = F(u)$;
- $Y = (y_1, \ldots, y_n)$ - выборка из $F(u - \theta)$, независимая от X;
- $\theta \in \mathbb{R}$ - параметр сдвига, $F(\cdot)$ - непрерывная функция, в остальном - неизвестная.

В этой постановке надо

(a) Проверить гипотезу $H : \theta = 0$ против лево- и правосторонних альтернатив $H^- : \theta < 0$, $H^+ : \theta > 0$;

(b) Построить доверительные интервалы для θ;

(c) Указать точечную оценку θ.

Все это возможно с помощью ранговых средств.

7.3. Критерий ранговых сумм (Wilcoxon)

Ранговый метод (проверки гипотезы H)

Рассмотрим объединенную совокупность (X, Y):

$$x_1, \ldots, x_m, y_1, \ldots, y_n.$$

От чисел $\{x\}$, $\{y\}$ перейдем к их рангам в объединенной совокупности (X, Y). Обозначим ранги игреков через $\vec{S} : R(y_j) = S_j$.

Ясно, что при гипотезе H в качестве $(S_1, \ldots, S_n)$ с одинаковыми вероятностями может появиться любая совокупность n чисел, взятых из отрезка натуральной последовательности $1, 2, \ldots, N$, где $N = m + n$.

Эта вероятность равна $1/[N(N-1)\ldots(N-n+1)]$.

В частности, $P(R(y_j) = S) = \frac{1}{N}$ для любого $S = 1, 2, \ldots, N$.

Чтобы понять, каково распределение рангов игреков $(S_1, \ldots, S_n)$ при альтернативах H^- или H^+, представим выборку из Y как продолжение выборки из X, но «со сдвигом»:

$$y_1 = \theta + x_{m+1}, \ldots, y_n = \theta + x_{m+n}.$$

Здесь $x_{m+1}, x_{m+2}, \ldots, x_{m+n}$ - независимые (в совокупности) и не зависящие от $x_1, \ldots, x_n$ случайные величины, имеющие ту же, что и $x_1, \ldots, x_n$, функцию распределения $F(\cdot)$.

Теперь ясно, что:

- При альтернативе $H^+(\theta > 0)$: $P(y_j > x_i) > \frac{1}{2}$.
- При альтернативе $H^-(\theta < 0)$ верно противоположное неравенство $P(x_i > y_j) > \frac{1}{2}$.

Поэтому при H^+ для игреков, т.е. для случайных величин $(S_1, \ldots, S_n)$, более вероятны значения из правой части ряда $1, 2, \ldots, N$, чем из левой.

При H^- - наоборот, для рангов $(S_1, \ldots, S_n)$ более вероятны малые числа из $1, 2, \ldots, N$.

Выявленное различие в распределениях $\vec{S}$ при гипотезе и при альтернативах можно усилить, если в качестве критериальной статистики взять их сумму. Это - так называемая *статистика Уилкоксона*, или, чуть пространнее, *статистика ранговых сумм Уилкоксона* (Wilcoxon):

$$W_{m,n} := \sum_{j=1}^{n} S_j.$$

Как следует из сказанного ранее, при гипотезе H (т.е. в случае однородности выборок X и Y) статистика $W_{m,n}$ распределена свободно: ее распределение не зависит от того, какова (непрерывная) функция F; распределение $W_{m,n}$ одинаково для всех них. Поэтому распределение $W_{m,n}$ при гипотезе H можно вычислить для любой пары натуральных чисел m и n. Эти распределения табулированы.

При альтернативе H^+ для $W_{m,n}$ становятся более вероятными большие значения: для $z > 0$

$$P(W_{m,n} \geqslant z \mid H^+) > P(W_{m,n} \geqslant z \mid H).$$

При H^- справедливо противоположное неравенство:

$$P(W_{m,n} \leqslant z \mid H^-) > P(W_{m,n} \leqslant z \mid H).$$

Приняв во внимание эти различия в статистическом поведении $W_{m,n}$ при гипотезе и альтернативах, можно предложить правило проверки H против H^- либо H^+.

Правило проверки H против H^+

1. Выбираем уровень значимости $\varepsilon > 0$.
2. По заданному $\varepsilon > 0$ (с помощью таблицы распределения $W_{m,n}$ при гипотезе) находим $(1-\varepsilon)$-квантиль $W_{m,n}$ - т.е. такое число $w(\varepsilon, m, n)$, что

   $$P(W_{m,n} \geqslant w(\varepsilon, m, n) \mid H) = \varepsilon.$$

   (Лучше выбрать ε так, чтобы это уравнение имело решение — из-за дискретности распределения $W_{m,n}$ это возможно только для некоторых значений ε).
3. Опровергаем гипотезу H в пользу H^+ на уровне ε, если наблюденное значение $W_{m,n}$ равно или превосходит $w(\varepsilon, m, n)$, т.е. если

   $$\text{набл.}W_{m,n} \geqslant w(\varepsilon, m, n).$$

Правило проверки H против H^- выглядит аналогично, с естественными изменениями.

Если же с гипотезой H конкурирует двусторонняя альтернатива $\overline{H} : \theta \neq 0$, то правило выглядит так:

- опровергать H в пользу $\overline{H}$, если наблюденное значение $W_{m,n}$ далеко (легко уточнить, что это значит) отклоняется от центра распределения $W_{m,n}$ при H.

Так как это распределение симметричное (проверьте!), то упомянутый центр равен $\mathsf{E}_0 W_{m,n}$. (Индексом ноль отмечаем распределения, соответствующие $\theta = 0$). Проверьте, что

$$\mathsf{E}_0 W_{m,n} = \frac{n(m+n+1)}{2}.$$

Можно показать, что функции мощности этих критериев возрастают по мере удаления значения θ от 0.

7.4. Связь доверительного оценивания и проверки гипотез

Пусть X - наблюдение, P_θ - распределение X, θ - неизвестный параметр.

Предположим, что для проверки гипотезы $H_t : \theta = t$ мы располагаем статистическим критерием, уровень которого $\leqslant \varepsilon$. Пусть $\delta(X, t)$ - индикаторная функция критерия. (Отвергаем $H_t : \theta = t$, если $\delta(X, t) = 1$.)

Доверительное множество для параметра θ с доверительной вероятностью $\geqslant 1 - \varepsilon$

$$C(X) = \{t : \delta(X, t) = 0\}.$$

Т.е. доверительное множество образуют те значения параметра, которые совместимы с наблюдением X (точнее, с X совместимы распределения вероятностей).

Легко видеть, что

$$P_\theta(\theta \in C(X)) \geqslant 1 - \varepsilon.$$

Ибо событие $\theta \in C(X)$ означает, что $\delta(X, t) = 0$, т.е. гипотеза, что истинное значение параметра есть θ, не отвергнута - а при параметре θ эта вероятность $\geqslant 1 - \varepsilon$.

Пример: (доверительная) оценка сдвига одной параметрической выборки относительно другой.

Пусть

- $X = (x_1, \ldots, x_n)$- выборка из $N(a, \sigma^2)$,
- $Y = (y_1, \ldots, y_n)$- выборка из $N(b, \sigma^2)$.

Здесь $\theta = (b - a)$ - сдвиг выборки Y относительно X.

Для проверки гипотезы $H_0 : a = b$, т.е. $H_0 : \theta = 0$ мы располагаем статистикой

$$F = \frac{mn}{m+n} \frac{(\overline{x} - \overline{y})^2}{s^2}.$$

Рассмотрим гипотезу $\theta = t$, t - задано. Можно свести задачу к предыдущей, если выборку Y преобразовать в $Z = (z_1, \ldots, z_n)$, где $z_j = y_j - t$.

Критериальная статистика для проверки $H_t : \theta = t$ теперь равна

$$\frac{mn}{m+n} \frac{\big(\overline{x} - (\overline{y} - t)\big)^2}{s^2}.$$

(Заметим, что при таком преобразовании Y в Z оценка дисперсии s^2 не изменяется).

Решающее правило для проверки $H_t : \theta = t$ на уровне значимости ε: не опровергать H_t, если

$$\sqrt{\frac{mn}{m+n}} \frac{\big|\overline{x} - (\overline{y} - t)\big|}{s} < t_{1-\varepsilon/2}.$$

Решая это неравенство относительно t, получим для θ доверительный интервал

$$\left\{ \overline{y} - \overline{x} - \sqrt{\frac{m+n}{mn}} s t_{1-\varepsilon/2} < \theta < \overline{y} - \overline{x} + \sqrt{\frac{m+n}{mn}} s t_{1-\varepsilon/2} \right\}.$$

Критическое значение $t_{1-\,\varepsilon/2}$ находим с помощью таблиц распределения Стьюдента с $m + n + 2$ степенями свободы.

7.5. Доверительная оценка параметра сдвига одной выборки относительно другой

Доверительную оценку параметра сдвига одной выборки относительно другой можно получить и для выборок, распределенных не по нормальному, но по произвольному закону (лишь бы непрерывному). Для этого надо воспользоваться статистическим критерием, действенным в этих условиях. Скажем, критерием Уилкоксона. Критерий Уилкоксона надо применять для проверки гипотезы однородности выборок

$$x_1, \ldots, x_m, \quad y_1 - t, \ldots, y_n - t \quad \forall\, t \in \mathbb{R}. \tag{1}$$

Обозначим статистику Уилкоксона для (1) через $W_{m,n}(t)$:

$$W_{m,n}(t) = \sum_{j=1}^{n} R(y_j - t).$$

Теперь доверительное множество для неизвестного истинного значения параметра сдвига θ (доверительная вероятность которого равна $1 - 2\alpha$) есть

$$\{t : nN - w(\alpha, m, n) < W_{m,n}(t) < w(\alpha, m, n)\}. \tag{**}$$

Остается дать явный вид этому доверительному множеству.

Рассмотрим статистику $W_{m,n}(t)$ как функцию переменного $t \in \mathbb{R}$. При $t \to -\infty$ (т.е.для значений t, больших по модулю и отрицательных) каждое значение $y_j - t$, $j = \overline{1,n}$, превосходит любое значение x_i, $i = \overline{1,m}$. Поэтому здесь

$$W_{m,n}(t) = N + (N-1) + \ldots + (N-(n-1)),$$

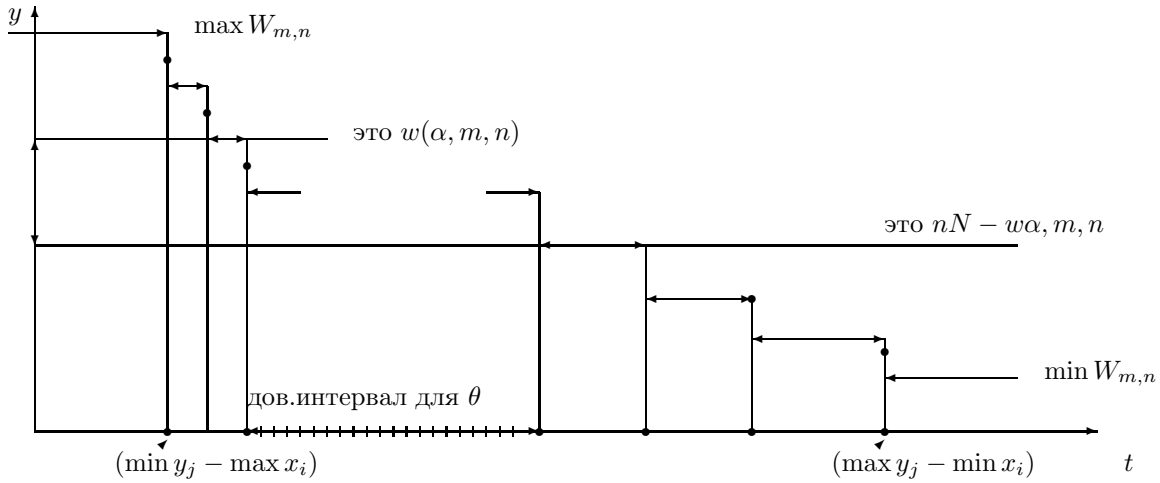
т.е. равно $\max W_{m,n} = nN - \frac{n(n-1)}{2} = \frac{1}{2}n(n+2m+1)$. При $t \to +\infty$ по противоположным соотношениям между $y_j - t$ и x_i находим, что здесь

$$W_{m,n}(t) = 1 + 2 + \ldots + n = \frac{n(n+1)}{2} = \min W_{m,n}.$$

Далее отметим, что $W_{m,n}(t)$ монотонно не возрастает (убывает), когда t растет, и что каждое уменьшение величины $W_{m,n}$ происходит скачком на единицу, когда t переходит через одно из mn чисел $x_i - y_j$ ($i = \overline{1,m}, j = \overline{1,n}$).

(Для контроля: $\max W_{m,n} + \min W_{m,n} = 2\mathsf{E}_0 W_{m,n}$, $\max W_{m,n} - \min W_{m,n} = mn$, т.е. равен количеству единичных скачков).

График функции $y = W_{m,n}(t)$, $t \in \mathbb{R}$:



Ради некоторых дальнейших удобств при $t = y_j - x_i$ положим $W_{m,n}(t)$ равным полусумме пределов справа и слева. Это равносильно соглашению, что при ранжировании совпадающих значений мы приписываем всем им одинаковые (средние) ранги.

Из свойств функции $W_{m,n}(t)$ и ее графика следует, что доверительное множество (**) есть интервал; его концами служат некоторые элементы из множества $\{x_i - y_j, i = \overline{1,m}, j = \overline{1,n}\}$, которые нетрудно указать точно. Для этого сказанное множество нужно упорядочить, а затем выбрать порядковые статистики с нужными номерами. (Из рисунка видно, какие это номера).

7.6. Точечная оценка сдвига (величины θ)

Статистика $W_{m,n}(t)$ количественно выражает степень согласия (однородности) двух выборок: $x_1, \ldots, x_m$ и $y_1 - t, \ldots, y_n - t$. Чем более отклоняется $W_{m,n}(t)$ от $\mathsf{E}_0 W_{m,n}$ (от ожидаемого значения $W_{m,n}$ при полной однородности), тем больше (сильнее) различаются выборки. Эти две выборки тем ближе к однородным (если мерить с помощью статистики Уилкоксона), чем ближе $W_{m,n}(t)$ к $\mathsf{E}_0 W_{m,n}$.

Отсюда вытекает предложение: выбрать в качестве точечной оценки неизвестного сдвига θ величину $\hat{\theta}$ такую, что

$$W_{m,n}(\hat{\theta}) = \mathsf{E}_0 W_{m,n}(\text{ т.е. } = \frac{n(n+m+1)}{2}).$$

Из графика видно, что

$$\hat{\theta} = med(\{x_i - y_j, i = \overline{1,m}, j = \overline{1,n}\}).$$

($\hat{\theta}$ - так называемая медиана Ходжеса-Лемана).

7.7. Асимптотическая нормальность статистики ранговых сумм Уилкоксона

7.7.1. Формулировка теорем

Теорема 1.

Пусть $(x_1, \ldots, x_m)$ и $(y_1, \ldots, y_n)$ суть независимые выборки из непрерывных распределений, статистика $W_{m,n}$ ранговых сумм Уилкоксона вычислена по этим выборкам. Тогда при $m, n \to \infty$

$$\frac{W_{m,n} - \mathsf{E}W_{m,n}}{\sqrt{\mathsf{D}W_{m,n}}} \xrightarrow{d} N(0,1).$$

Введем статистику Манна – Уитни (*Mann – Whitney*):

$$H_{m,n} := \sum_{i=1}^{m} \sum_{j=1}^{n} I(x_i < y_j).$$

С вероятностью 1

$$W_{m,n} = H_{m,n} + \frac{n(n+1)}{2}.$$

Поэтому для доказательства теоремы 1 достаточно доказать

Теорема 2.

В условиях теоремы 1

$$\frac{H_{m,n} - \mathsf{E}H_{m,n}}{\sqrt{\mathsf{D}H_{m,n}}} \xrightarrow{d} N(0,1).$$

Теорема 2 — это частный случай теоремы 3 об асимптотическом поведении так называемой *U-статистики* (*U*-statistics). (В данном случае, $H_{m,n}$ - это двувыборочная U-статистика.)

$$U_{m,n} := \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i, y_j).$$

Теорема 3.

Пусть $(x_1, \ldots, x_m)$ и $(y_1, \ldots, y_n)$ — две независимые выборки, функция $f(x, y)$ такова, что

$$\mathsf{E}f^2(x_1, y_1) < \infty, \mathsf{E}(\mathsf{E}[f(x_1, y_1)|x_1])^2 > 0, \mathsf{E}[\mathsf{E}(f(x_1, y_1)|y_1)]^2 > 0.$$

Тогда при $m, n \to \infty$

$$\frac{U_{m,n} - \mathsf{E}U_{m,n}}{\sqrt{\mathsf{D}U_{m,n}}} \xrightarrow{d} N(0,1).$$

Мы докажем теорему 3, ограничиваясь случаем $\mathsf{E}f = 0$ (что соответствует однородной выборке в теореме 2), поскольку этот случай для нас более важен и поскольку в этом случае легко вычислить $\mathsf{E}_0 U_{m,n}$ и $\mathsf{D}_0 U_{m,n}$.
По ходу доказательства нам будет необходима так называемая

Теорема Слуцкого.

Пусть:

- случайная последовательность $\{\xi_n\}$ сходится по распределению к случайной величине ξ;
- случайная последовательность $\{\eta_n\}$ сходится по вероятности к постоянной величине C.

Тогда при $n \to \infty$

(a)

$$\xi_n + \eta_n \xrightarrow{d} \xi + C.$$

(b)

$$\xi_n \eta_n \xrightarrow{d} C\xi.$$

7.7.2. Доказательство теоремы 3: начало

Вместо x_i, y_j будем писать X, Y. Мы предполагаем, что $\mathsf{E}f(X,Y)=0$ и, следовательно, $\mathsf{E}U_{m,n}=0$. Введем случайные величины $\alpha(X)$ и $\beta(Y)$:

$$\alpha(X)=\mathsf{E}[f(X,Y)|X],\qquad \beta(Y)=\mathsf{E}[f(X,Y)|Y].$$

Представим $U_{m,n}$ в виде

$$U_{m,n}=\sum_{i=1}^{m}\sum_{j=1}^{n}[f(x_i,y_j)-\alpha(x_i)-\beta(y_j)]+\sum_{i=1}^{m}\sum_{j=1}^{n}[\alpha(x_i)+\beta(y_j)]=$$

$$=n\sum_{i=1}^{m}\alpha(x_i)+m\sum_{j=1}^{n}\beta(y_j)+\Delta_{m,n},$$

где

$$\Delta_{m,n}=\sum_{i=1}^{m}\sum_{j=1}^{n}[f(x_i,y_j)-\alpha(x_i)-\beta(y_j)].$$

Далее дробь $U_{m,n}/\sqrt{\mathsf{D}U_{m,n}}$, предельное поведение которой и есть предмет теоремы 3 ($\mathsf{E}U_{m,n}=0$), представляем в виде:

$$\frac{U_{m,n}}{\sqrt{\mathsf{D}U_{m,n}}}=\underbrace{\frac{n\sum\limits_{i=1}^{m}\alpha(x_i)+m\sum\limits_{j=1}^{n}\beta(y_j)}{\sqrt{\mathsf{D}[n\sum\limits_{i=1}^{m}\alpha(x_i)+m\sum\limits_{j=1}^{n}\beta(y_j)]}}}_{\xi_{m,n}}\underbrace{\sqrt{\frac{\mathsf{D}[n\sum\limits_{i=1}^{m}\alpha(x_i)+m\sum\limits_{j=1}^{n}\beta(y_j)]}{\mathsf{D}U_{m,n}}}}_{C_{m,n}}+\underbrace{\frac{\Delta_{m,n}}{\sqrt{\mathsf{D}U_{m,n}}}}_{\eta_{m,n}}$$

или, коротко:

$$\frac{U_{m,n}}{\sqrt{\mathsf{D}U_{m,n}}}=\xi_{m,n}C_{m,n}+\eta_{m,n}.$$

Для доказательства теоремы 3 достаточно показать, что

(a) $C_{m,n}\longrightarrow 1$,

(b) $\xi_{m,n}\xrightarrow{d}N(0,1)$,

(c) $\eta_{m,n}\xrightarrow{p}0$.

Затем применить теорему Слуцкого.

7.7.3. Вычисление дисперсии U-статистик.

Ключевую роль играет вычисление дисперсии U-статистик. Поэтому мы выделяем это в отдельный пункт. Так как $\mathsf{E}f=0$, то

$$\mathsf{D}U_{m,n}=\mathsf{E}U_{m,n}^2=\sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n}\sum_{j'=1}^{n}\mathsf{E}f(x_i,y_j)f(x_{i'},y_{j'}).$$

Стоящую в правой части сумму представим в виде четырех слагаемых, каждое из которых есть сумма, где индексы удовлетворяют условиям:

$$\sum\nolimits_1=\sum\dots\sum(i\neq i',j\neq j'),$$
$$\sum\nolimits_2=\sum\dots\sum(i=i',j\neq j'),$$
$$\sum\nolimits_3=\sum\dots\sum(i\neq i',j=j'),$$
$$\sum\nolimits_4=\sum\dots\sum(i=i',j=j').$$

1. $\sum_1=0$, т.к. $\mathsf{E}f=0$.

2. $\sum_2 = mn(n-1)\mathsf{E}f(x_1,y_1)f(x_1,y_2) = mn(n-1)\mathsf{D}\alpha$, так как

$$\mathsf{E}f(x_1,y_1)f(x_1,y_2) = \mathsf{EE}[f(x_1,y_1)f(x_1,y_2)|x_1] = \mathsf{E}\{\mathsf{E}[f(x_1,y_1)|x_1]\mathsf{E}[f(x_1,y_2)|x_1]\}$$

$$= \mathsf{E}\alpha(x_1)\alpha(x_1) = \mathsf{D}\alpha,$$

ибо $\mathsf{E}\alpha(x_1) = 0$.

3. $\sum_3 = mn(m-1)\mathsf{D}\beta$ - аналогично.

4. $\sum_4 = mn\mathsf{E}f^2 = mn\mathsf{D}f(x,y)$.

Поэтому

$$\mathsf{D}U_{m,n} = mn(n-1)\mathsf{D}\alpha + nm(m-1)\mathsf{D}\beta + mn\mathsf{D}f.$$

7.7.4. Доказательство теоремы 3: окончание

- Утверждение (c) есть следствие неравенства Чебышева для $\eta_{m,n}$, ибо

$$\mathsf{D}\eta_{m,n} = \frac{\mathsf{D}\Delta_{m,n}}{mn[n\mathsf{D}\alpha + m\mathsf{D}\beta + \text{const}]} \longrightarrow 0,$$

так как (в силу 7.7.3):

$$\mathsf{D}\Delta_{m,n} = mn\mathsf{D}[\tilde{f}(x_1,y_1) - \tilde{\alpha}(x_1) - \tilde{\beta}(y_1)]$$

ибо для функции $\tilde{f} = f(x_1,y_1) - \alpha(x_1) + \beta(y_1)$ слагаемые $\sum_3 = 0$ и $\sum_2 = 0$.

- Утверждение (a) очевидно, ибо

$$\mathsf{D}[n\sum_{i=1}^{m}\alpha(x_i) + m\sum_{j=1}^{n}\beta(y_j)] = n^2m\mathsf{D}\alpha + m^2n\mathsf{D}\beta.$$

- Утверждение (b) есть одна из форм центральной предельной теоремы. Ее легко доказать методом характеристических функций, по аналогии с доказательством центральной предельной теоремы для суммы независимых одинаково распределенных случайных величин.

7.7.5. Доказательство теоремы Слуцкого.

Ограничимся доказательством утверждения (a).
Надо показать, что для любой непрерывной ограниченной функции $f(\cdot)$ справедливо утверждение

$$\mathsf{E}f(\xi_n + \eta_n) \to \mathsf{E}f(\xi + C).$$

Докажем это утверждение. Точнее, мы покажем, что при $n \to \infty$

$$\mathsf{E}[f(\xi_n + \eta_n) - f(\xi + C)] \to 0.$$

Заметим, что для любого $\varepsilon > 0$ существует $A > 0$, такое, что $P(|\xi| > A) < \varepsilon$.
Поскольку $\xi_n \xrightarrow{d} \xi$, то для достаточно больших n

$$P(|\xi_n| > A) < 2\varepsilon.$$

Далее: для любого $\delta > 0$ для достаточно больших n

$$P(|\eta_n - C| > \delta) < \varepsilon,$$

т.к. $\eta_n \xrightarrow{p} C$.
Поскольку

$$\mathsf{E}[f(\xi_n + \eta_n) - f(\xi + C)] =$$

$$= \mathsf{E}[f(\xi_n + \eta_n) - f(\xi_n + C)] + \mathsf{E}[f(\xi_n + C) - f(\xi + C)], \qquad (*)$$

достаточно показать, что каждое из двух слагаемых в правой части $(*)$ для достаточно больших n становится меньше любого наперед заданного числа.

Начнем с первого из них.

Рассмотрим

$$\mathsf{E}[f(\xi_n+\eta_n)-f(\xi_n+C)]=$$

$$=\mathsf{E}[f(\xi_n+\eta_n)-f(\xi_n+C)][I(|\xi|\leqslant A)+I(|\xi|>A)][I(|\eta_n-C|\leqslant\delta)+I(|\eta_n-C|>\delta)]$$

$$=\mathsf{E}[f(\xi_n+\eta_n)-f(\xi_n+C)]I(|\xi|\leqslant A)I(|\eta_n-C|\leqslant\delta)+R_n$$

Через R_n обозначена сумма, составленная из прочих слагаемых, которые получаются, когда мы раскроем скобки. В каждом из этих слагаемых есть либо $I(|\xi|>A)$, либо $I(|\eta_n-C|>\delta)$, либо оба.

Каждое из упомянутых слагаемых можно оценить по модулю сверху как $2M\varepsilon$, где $M=\max\limits_u f(u)$.

Обратимся к главному слагаемому и заметим, что ξ_n и η_n в нем ограничены. Поэтому значения $\xi_n+\eta_n$ принадлежат компакту.

Так как функция $f(\cdot)$ непрерывна, на этом компакте она равномерно непрерывна. Это означает, что $\forall\varepsilon>0\quad\exists\delta>0$: если $|u-v|<\delta$, то $|f(u)-f(v)|<\varepsilon$.

Здесь

$$|(\xi_n+\eta_n)-(\xi_n-C)|=|\eta_n-C|\leqslant\delta,$$

так что

$$|f(\xi_n+\eta_n)-f(\xi_n-C)|<\varepsilon.$$

В итоге получаем, что для произвольного ε и достаточно больших n

$$|\mathsf{E}[f(\xi_n+\eta_n)-f(\xi_n+C)]|<K\varepsilon,$$

где K - некоторая постоянная.

Обратимся ко второму слагаемому в $(*)$.

Из сходимости $\xi_n\xrightarrow{d}\xi$ следует, что

$$|\mathsf{E}f(\xi_n+C)-\mathsf{E}f(\xi+C)|<\varepsilon$$

для $\varepsilon>0$ и достаточно больших n.

Возвращаясь к $(*)$, заключаем, что для достаточно больших n верно

$$|\mathsf{E}f(\xi_n+C)-\mathsf{E}f(\xi+C)|<\tilde{K}\varepsilon$$

где $\tilde{K}$ - некоторая постоянная. Поскольку ε может быть выбрано произвольно малым, утверждение теоремы доказано.

7.7.6. Применение теоремы 1 для вычисления статистики Уилкоксона

Теорема 1 бывает полезна для вычисления критических значений статистики $W_{m,n}$ при больших m,n.

Чтобы воспользоваться теоремой 1, надо вычислить $\mathsf{D}_0W_{m,n}$ (дисперсию при гипотезе, т.е. для однородных выборок $(x_1,\ldots,x_m)$ и $(y_1,\ldots,y_n)$).

Мы вычислим $\mathsf{D}_0H_{m,n}=\mathsf{D}_0W_{m,n}$.

Воспользуемся результатом пункта 7.7.3, положив

$$f(x_1,y_1)=I(x_1<y_1)-\mathsf{E}I(x_1<y_1)=I(x_1<y_1)-P(x_1<y_1).$$

Как сказано, ограничимся однородными выборками. Тогда

$$P(x_1<y_1)=P(x_1>y_1)=1/2.$$

Общую функцию распределения (непрерывную!) обозначим через

$$F(u)=P(x_i<u)=P(y_j<u).$$

Вычисляем

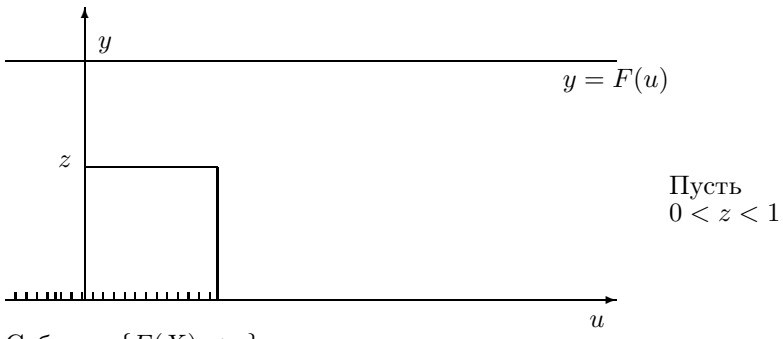
$$\alpha(x_i)=$$

$$\mathsf{E}\{[I(x_i<y_j)-1/2]|x_i\}=P(x_i<y_j|x_i)-1/2=1-P(y_j<x_i|x_i)-1/2=1/2-F(x_i).$$

Аналогично: $\beta(y_j)=F(y_j)-1/2$.

Заметим, что для случайной величины X, имеющей непрерывную функцию распределения $F(u)=P(X<u)$, «новая» случайная величина $\xi=F(X)$ распределена равномерно на [0, 1].

Доказательство следует из чертежа:



Пусть
$0 < z < 1$

Событие $\{F(X) < z\}$

Ясно, что $P(F(X) < z) = z$.

Получили, что при гипотезе (однородности)

$$\alpha(x_i) = 1/2 - U_i, \qquad \beta(y_j) = V_j - 1/2,$$

где $U_1, \ldots, U_m$, $V_1, \ldots, V_n$ суть независимые случайные величины, равномерно распределенные на [0, 1].

Очевидно, что

$$\mathsf{E}_0\alpha^2 = \mathsf{D}U_i = \frac{1}{12} = \mathsf{E}_0\beta_j^2 = \mathsf{D}V_j = \frac{1}{12}.$$

Поэтому

$$\mathsf{D}_0 H_{m,n} = \mathsf{D}_0 W_{m,n} = \frac{mn(n-1)}{12} + \frac{m(m-1)n}{12} + \frac{mn}{4} = \frac{mn(m+n+1)}{12}.$$

ибо $\mathsf{D}_0 f$ здесь равно

$$\mathsf{D}_0 f = \mathsf{D}_0 I(x_1 < y_1) = P(x_1 < y_i)[1 - P(x_1 < y_1)] = 1/4.$$

Итак, для непрерывных однородных выборок теорема 1 дает:

$$W^*_{m,n} = \frac{W_{m,n} - n\frac{n+m+1}{2}}{\sqrt{mn\frac{m+n+1}{12}}} \xrightarrow{d} N(0,1) \text{ при } m, n \to \infty.$$

Пользоваться этим нормальным приближением (для не слишком малых или больших вероятностей) можно при $m, n \geqslant 10$. Центральная предельная теорема не дает нам оценок для скорости сходимости. Сказанное правило подтверждается сравнением точного распределения $W_{m,n}$ и его нормальной аппроксимации.

8. Метод наибольшего правдоподобия

8.1. Определения

Пусть X - наблюденная случайная величина, распределение которой принадлежит параметрическому семейству P_θ, $\theta \in \Theta$; пусть θ^0 обозначает истинное значение параметра.

Предположим, что распределения P_θ имеют плотность (обозначаемую $p(x, \theta)$) относительно какой-либо меры. (Если эта мера считающая, то $p(X, \theta)$ - это вероятность события $X = x$).

Правдоподобием значения параметра θ называют (случайную величину) $p(X, \theta)$.

То значение параметра θ, для которого правдоподобие принимает наибольшее значение, называют *оценкой наибольшего правдоподобия* θ:

$$\hat\theta = \arg\max_{\theta \in\, \theta} p(X, \theta). \tag{1}$$

Асимптотические свойства оценок наибольшего правдоподобия мы изучим для выборки, объем которой неограниченно возрастает.

Итак, пусть $X = (x_1, \ldots, x_n)$ - выборка из распределения, обладающего плотностью $f(x, \theta)$, где $\theta \in \Theta$ — неизвестный параметр; его истинное значение (при котором получена выборка X) есть $\theta^0 \in \Theta$.

Относительно оценки (1) мы докажем - при определенных условиях на $f(\cdot, \theta)$, что

a) $\hat\theta_n$ - состоятельная оценка для θ^0,

b) $\hat\theta_n$ распределена асимптотически нормально.

Дадим определения.
Определение 1.

Оценка $t = t(X)$ параметра θ называется *состоятельной*, если $t(X) \xrightarrow{p} \theta^0$ при $n \to \infty$.

Определение 2. (упрощенное)

Мы говорим, что статистика $\hat{\theta}_n$ распределена *асимптотически нормально*, когда

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, \sigma^2)$$

для некоторых θ^0 и σ^2.

При этом θ^0 называют *асимптотическим средним*, а σ^2/n - *асимптотической дисперсией* $\hat{\theta}_n$.

8.2. Состоятельность оценок наибольшего правдоподобия

Начнем с леммы (варианта т.н. неравенства информации).
Лемма.

Пусть $f(\cdot)$, $g(\cdot)$ - две плотности вероятности.

Тогда

$$\int f(x) \ln f(x) dx \geqslant \int f(x) \ln g(x) dx, \tag{1}$$

причем равенство возможно, только если $f = g$ почти всюду.

Соглашения:

1. Для интегралов допускается значение $-\infty$.
2. Будем считать, что
$$\int\limits_A f(x) \ln g(x) dx = 0, \qquad \text{если } f(x) = 0 \text{ для } x \in A$$
вне зависимости от значений $g(\cdot)$.

Доказательство.
Достаточно показать, что

$$\int f(x) \ln \frac{g(x)}{f(x)} dx \leqslant 0.$$

Заметим, что $\ln(1 + x) \leqslant x$ для $x \geqslant -1$. (См. на рисунке 1 графики функций $y = x$ и $y = \ln(1 + x)$)
Рассмотрим множество $A = \{x : f(x) > 0\}$.
Для $x \in A$:

$$\ln \frac{g(x)}{f(x)} \equiv \ln \left[1 + \left(\frac{g(x)}{f(x)} - 1\right)\right] \leqslant \frac{g(x)}{f(x)} - 1.$$

Умножив обе части неравенства на $f(\cdot)$, интегрируем:

$$\int f(x) \ln \frac{g(x)}{f(x)} dx \leqslant \int [g(x) - f(x)] dx = 0,$$

ч.т.д. $\square$.

8.3. Почему оценка наибольшего правдоподобия состоятельна - правдоподобное рассуждение.

Если $X = (x_1, \ldots, x_n)$ - выборка из распределения с плотностью $f(x, \theta)$, то правдоподобие X имеет вид $\prod_{i=1}^n f(x_i, \theta)$, а оценка наибольшего правдоподобия (8.1.1) есть

$$\arg\max_{\theta \in \Theta} \prod_{i=1}^n f(x_i, \theta),$$

или

$$\arg\max_{\theta\in\Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta)\right].$$

(Точка экстремума не изменяется при переходе от функции к ее логарифму и при умножении на положительное число.)

В силу закона больших чисел при $n\to\infty$

$$\frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta)\xrightarrow{p}\mathsf{E}_0\log f(x_i,\theta), \tag{1}$$

где E_0 означает усреднение по плотности $f(x,\theta^0)$, где θ^0 - истинное значение θ.

Поэтому естественно ожидать, что

$$\arg\max_{\theta\in\Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta)\right]\xrightarrow{p}\arg\max_{\theta\in\Theta}\mathsf{E}_0\log f(x_i,\theta).$$

Согласно лемме из 8.2 справедливо (8.2.1); это неравенство для $g(x)=f(x,\theta)$, $f(x)=f(x,\theta^0)$ дает:

$$\mathsf{E}_0\log f(x_i,\theta)\equiv\int[\log f(x,\theta)]f(x,\theta^0)dx\leqslant\int[\log f(x,\theta^0)]f(x,\theta^0)dx.$$

Следовательно,

$$\arg\max_{\theta\in\Theta}\mathsf{E}_0 f(x_i,\theta)=\theta^0.$$

Доказательство сходимости $\hat\theta_n\xrightarrow{P}\theta^0$ надо проводить, учитывая свойства $\mathsf{E}_0\log f(x_1,\theta)$, как функции θ, $\theta\in\Theta$. Если эта функция непрерывна по θ, обычно удается такой план:

- Показать, что сходимость в (1) равномерна по θ на компакте, содержащем θ^0.
- В этом случае можно утверждать, что существует последовательность локальных экстремумов функции $\hat\theta_n$, по вероятности сходящаяся к θ^0:

$$\hat\theta_n\xrightarrow{P}\theta^0,\quad n\to\infty. \tag{2}$$

8.4. Доказательство сходимости $\hat\theta_n\xrightarrow{P}\theta^0$ для одномерного случая

В одномерном случае доказательство очевидно.

Чтобы доказать (8.3.2), мы покажем, что (локальный) экстремум функции

$$\frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta)$$

при достаточно больших n со сколь угодно близкой к 1 вероятностью лежит внутри интервала (θ^0 - h, θ^0+h), где h - произвольное число.

Рисунок 2.

Так как

$$\mathsf{E}_0\log f(x_1,\theta^0)>\mathsf{E}_0\log f(x_1,\theta^0\pm h),$$

то можно подобрать такое $\varepsilon>0$, что

$$\mathsf{E}_0\log f(x_1,\theta^0)-\varepsilon>\mathsf{E}_0\log f(x_1,\theta^0\pm h)+\varepsilon.$$

Для произвольного, но фиксированного $\delta>0$, в силу упомянутого в 8.3 закона больших чисел (8.3.1) для достаточно больших n выполняется неравенство:

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta^0)-\mathsf{E}_0\log f(x_1,\theta^0)\right|<\varepsilon\right\}>1-\delta,$$

$$P\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta^0\pm h)-\mathsf{E}_0\log f(x_1,\theta^0\pm h)\right|<\varepsilon\right\}>1-\delta.$$

Поэтому

$$P\left\{\frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta^0) > \frac{1}{n}\sum_{i=1}^{n}\log f(x_i,\theta^0\pm h)\right\} > 1-2\delta.$$

Поэтому (при достаточно больших n) экстремум (локальный) функции правдоподобия из (8.1.1) лежит в сколь угодно узкой окрестности точки θ^0. Поэтому последовательность этих локальных экстремумов сходится (по вероятности) к θ^0, что и требовалось доказать. □

8.5. Асимптотическая нормальность оценок наибольшего правдоподобия (по выборке из регулярного семейства)

(См. Ивченко, Медведев, §2.4.)

Пусть $X=(x_1,\ldots,x_n)$ - выборка из распределения с плотностью (вероятностью) $p(x,\theta)$, $\theta\in\Theta\subset\mathbb{R}$. (После того, как мы закончим исследование одномерного параметра θ, мы обсудим, какие изменения надо сделать, когда $\theta\in\Theta\subset\mathbb{R}^r$.) Множество Θ будем считать открытым.

В рассматриваемом случае оценка наибольшего правдоподобия есть решение уравнения правдоподобия

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log p(x_i,\theta^0)=0 \tag{1}$$

Считая, что $p(x,\theta)$ трижды дифференцируема по θ, предположим, что существует функция $M(x)$ такая, что

1.
$$\left|\frac{\partial^3}{\partial\theta^3}\log p(x,\theta)\right| < M(x),$$

2.
$$\mathsf{E}_\theta M(x)<\infty$$

для всех $\theta\in\Theta$.

В дальнейшем ради краткости будем писать

$$l(x,\theta)=\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log p(x,\theta).$$

Введем новую переменную τ, положив

$$\theta=\theta^0+\frac{\tau}{\sqrt{n}}.$$

Теперь уравнение правдоподобия (1) имеет вид

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}l(x_i,\theta^0+\frac{\tau}{\sqrt{n}})=0. \tag{2}$$

Разлагаем левую часть (2) по формуле Тейлора в точке 0. Получим:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}l(x_i,\theta^0)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}l'_\theta(x_i,\theta^0)\frac{\tau}{\sqrt{n}}+\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}l''_{\theta\theta}(x_i,\tilde\theta_n)\left(\frac{\tau}{\sqrt{n}}\right)^2=0, \tag{3}$$

где $\tilde\theta_n$ - некая промежуточная точка между θ^0 и θ.

Заметим, что если ограничить область изменения переменной τ произвольным компактом, т.е. предположить, что $|\tau|<C$ для некоторого C, то третье слагаемое окажется (при $n\to\infty$) бесконечно малым.

Действительно:

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}l''_{\theta\theta}(x_i,\tilde\theta_n)\left(\frac{\tau}{\sqrt{n}}\right)^2\right| < \frac{C^2}{\sqrt{n}}\left|\frac{1}{n}\sum_{i=1}^{n}M(x_i)\right| \xrightarrow{P} 0,$$

т.к. по закону больших чисел

$$\frac{1}{n}\sum_{i=1}^{n}M(x_i)\xrightarrow{P}\mathsf{E}_\theta M(x_1).$$

Сопоставим решение уравнения (2), левая часть которого представлена в форме (3), и решение уравнения

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l(x_i,\theta^0)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l'_\theta(x_i,\theta^0)\frac{\tau}{\sqrt{n}}=0. \tag{4}$$

(Левая часть в (3), но без третьего слагаемого).

Решение (4) очевидно:

$$\tau_n^*=\frac{-\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n} l(x_i,\theta^0)}{\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n} l'_\theta(x_i,\theta^0)\frac{1}{\sqrt{n}}}. \tag{5}$$

При этом легко увидеть, что при $n\to\infty$

$$\tau_n^* \xrightarrow{d} N(0,(i(\theta^0))^{-1}).$$

Здесь $i(\theta^0)$ - количество информации (по Фишеру) о θ, содержащейся в одном наблюдении x_1.

Действительно, числитель (5) есть сумма независимых случайных величин $\frac{\partial}{\partial\theta}\sum\limits_{i=1}^{n}\log p(x_i,\theta^0)$, $\qquad i=\overline{1,n}$.

При обсуждении неравенств Крамера – Рао мы отметили, что

$$\mathsf{E}_\theta\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log p(x_i,\theta)=0$$

для $\theta\in\Theta$, и что

$$\mathsf{E}_\theta\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log p(x_i,\theta)\right]^2=i(\theta).$$

По центральной предельной теореме числитель (5) по распределению сходится к $N(0,i(\theta^0))$, когда $n\to\infty$.

Знаменатель (5) по закону больших чисел сходится (по вероятности) к $\mathsf{E}_\theta l'_\theta(x_i,\theta^0)$, где $\theta{=}\theta^0$.

Мы (при упомянутых выше обсуждениях) отмечали, что

$$\mathsf{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log p(x_i,\theta)\right]=-i(\theta).$$

Поэтому (по теореме Слуцкого) при $n\to\infty$

$$\tau_n^* \xrightarrow{d} N(0,(i(\theta^0))^{-1}). \tag{6}$$

Разумеется, надо проверить отдельно (дополнительно), что

$$0<i(\theta)<\infty.$$

Остается убедиться, что решение уравнения (2) асимптотически эквивалентно решению уравнения (4) - эквивалентно в том смысле, что при $n\to\infty$ разность между ними стремится к нулю (по вероятности).

Мы уже отмечали, что левые части (2) и (4) отличаются бесконечно мало (и притом равномерно по τ), когда $|\tau|<C$, C - произвольная постоянная.

Рассмотрим левую часть (4) как функцию от τ: $y=\psi_n(\tau)$.

Для достаточно больших n график левой части (2), скажем $y=\varphi_n(\tau)$, будет - при $\tau<C$ - проходить в ε-окрестности графика $y=\psi_n(\tau)$.

Поскольку $\varepsilon>0$ может быть выбрано сколь угодно малым, у уравнения правдоподобия (2) найдется решение $\hat\tau_n$, такое, что $\hat\tau_n-\tau_n^* \xrightarrow{P} 0$ - при том дополнительном условии, что уравнение (4) имеет решение, принадлежащее компакту $\{\tau:|\tau|<C\}$.

Остается сделать последнее замечание, чтобы завершить исследование (2). Так как τ_n^* (решение (4)) асимптотически нормально, можно выбрать упомянутый выше компакт $\{\tau:|\tau|<C\}$ так, чтобы для произвольно выбранного $\delta>0$

$$P(|\tau_n^*|<C)>1-\delta$$

для достаточно больших n.

Таким образом, со сколь угодно близкой к 1 вероятностью, уравнение (4) имеет корень на выбранном компакте (для достаточно больших n), притом этот корень $\hat{\tau}_n$ распределен также, как и τ_n^*, т.е. при $n \to \infty$

$$\hat{\tau}_n \xrightarrow{d} N(0, (i(\theta^0))^{-1}). \tag{7}$$

Вернемся к переменной $\theta = \theta^0 + \frac{\tau}{\sqrt{n}}$, $\hat{\theta}_n = \theta^0 + \frac{\hat{\tau}_n}{\sqrt{n}}$, $\hat{\tau}_n = \sqrt{n}(\hat{\theta}_n - \theta^0)$.
Утверждение (7) означает, что

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, (i(\theta^0))^{-1}).$$

Следовательно, мы доказали, что (при наложенных на $p(x, \theta)$ выше условиях) существует решение (точнее, последовательность решений) уравнения правдоподобия (1) $\hat{\theta}_n$, сходящееся к θ^0 и распределенное асимптотически нормально $N(\theta^0, \frac{1}{ni(\theta^0)})$.

8.6. Многомерный случай

Для многомерного параметра θ все исследование проходит так же, как и для одномерного, с очевидными изменениями.

- Так, уравнение правдоподобия (1) теперь - векторное (т.е. (1) представляет собой систему уравнений).
- Условие о третьих производных формулируется так:

$$\left| \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_k} \log p(x, \theta) \right| < M(x),$$

  и т.д.
- Место количества информации $i(\theta)$ займет теперь матрица информации $\mathcal{J}(\theta)$, а окончательный результат примет вид

$$\hat{\theta}_n \xrightarrow{d} N(\theta^0, n^{-1}\mathcal{J}^{-1}(\theta^0)).$$

9. Асимптотическая нормальность оценок (статистических функций фон Мизеса)

Пусть $x_1, \ldots, x_n$ - выборка из распределения (для простоты - одномерного), неизвестную функцию распределения которого обозначим через $F(\cdot)$. По этой выборке мы хотим оценить функционал $T(F)$, $T(\cdot)$ задан (на множестве или подмножестве функций распределения). Воспользуемся для этого близостью к $F(\cdot)$ выборочной (эмпирической) функции распределения $F_n(\cdot)$:

- Если $T(\cdot)$ - непрерывный функционал, то $T(F_n) \approx T(F)$, т.е.

$$T(F_n) \xrightarrow{P} T(F)$$

  при $n \to \infty$.

Примеры функционалов $T(F)$:

1. Моменты и квантили функции распределения $F(\cdot)$.
2. Второй центральный момент случайной величины X, функция распределения которой есть $F(\cdot)$:

$$\mathsf{E}(X - \mathsf{E}X)^2 = \frac{1}{2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - y)^2 dF(x) dF(y).$$

3. Оценка наибольшего правдоподобия для семейства распределений $F(x, \theta)$, $\theta \in \Theta$, соответствует функционалу

$$T(F) = \arg\max_{\theta \in \Theta} \int \log f(x, \theta) dF(x).$$

4. Для оценивания неизвестного параметра θ по выборке, распределение которой принадлежит семейству $F(x,\theta)$, $\theta \in \Theta$, можно применять (использовать) различные функционалы $T(F)$.

   Пусть, например, по выборке из $N(a,\sigma^2)$ мы хотим оценить неизвестное a. В качестве $T(F)$ мы можем взять, например, функционалы

$$T(F) = \int x dF(x)$$

   и

$$T(F) = med\ F.$$

Если функционал $T(F)$ в каком-либо смысле дифференцируем, то

$$T(G) = T(F + (G - F)) = T(F) + \int a(x) d(G - F) + R, \tag{1}$$

где $a(\cdot)$ - некоторая функция, R - остаточный член, который убывает быстрее, чем $G - F$. Эта формула - аналог формулы Тейлора для функции нескольких переменных.

Пусть, действительно, $f(X) = f(x_1, \ldots, x_p)$, $dX = (dx_1, \ldots, dx_p)$. Тогда

$$f(X + dX) = f(X) + \sum_{i=1}^{p} \frac{\partial f}{\partial x_i} dx_i + o(dX)$$

$$= f(X) + [\nabla f(X)][dX] + o(dX).$$

В бесконечномерном варианте суммирование заменяется интегрированием.

Заметим, что линейный функционал $\int a d(G - F)$ задан на линейном пространстве разностей функций распределения. Мы можем распространить его определение и на множество функций распределения. Для этого достаточно указать значение $\int a(X) dF(X)$. Фактически, речь идет о выборе функции $a(X)$: возьмем $a(\cdot)$ таким, чтобы

$$\int a(x) dF(x) = 0. \tag{2}$$

Тогда (1) превращается в

$$T(G) = T(F) + \int a(x) dG(x) + R. \tag{3}$$

(Впрочем, можно было бы начать сразу с (3). Тогда (2) стало бы следствием (3).)

Функционал $\int a(X) dG(X)$ - это производная $T(F)$ в направлении G по Gâteaux. По определению, производная по Gâteaux равна

$$\lim_{\lambda \to +0} \lambda^{-1} [T((1 - \lambda)F + \lambda G) - T(F)], \tag{4}$$

если этот предел существует.

Возьмем в (3) в качестве G функцию распределения

$$F + \lambda(G - F) = (1 - \lambda)F + \lambda G, \qquad \text{где } \lambda \in [0, 1].$$

Подставив это выражение в (3) и перейдя к пределу по $\lambda \to +0$, получим сказанное.

Для функции $a(\cdot)$ можно получить явное (более явное) выражение, если в (4) в качестве $G(\cdot)$ взять функцию распределения вероятностей, сосредоточенную в одной точке.

Пусть $\Delta_u(x)$ - функция распределения, сосредоточенная в точке u:

$$\Delta_u(x) = \begin{cases} 0, & \textit{если } x < u, \\ 1, & \textit{если } x > u. \end{cases}$$

Тогда

$$\lim_{\lambda \to +0} \lambda^{-1} [T((1 - \lambda)F + \lambda \Delta_u) - T(F)] = \int a(x) d\Delta_u(x) = a(u). \tag{5}$$

9.1. Функция влияния

Формулу (5) можно трактовать и так:

$$T((1-\lambda)F + \lambda\Delta_u) \approx T(F) + \lambda a(u). \tag{6}$$

Функция $(1-\lambda)F+\lambda\Delta_u$ соответствует смеси двух распределений: $F(\cdot)$, взятого с весом $(1-\lambda)$, и Δ_u, взятого с весом λ.

Такое распределение возникает, когда к (основному) распределению $F(\cdot)$ «примешана» некоторая доля λ наблюдений, происходящих не из $F(\cdot)$, а просто равных числу u. Соответственно в выборке, извлеченной из такого распределения, помимо наблюдений, подчиняющихся распределению $F(\cdot)$, есть доля λ «посторонних» величин u. В статистике такие посторонние значения называют *выбросами*, *грубыми ошибками*, *засорениями* и т.д.

Формулы (5) и (6) показывают, какое влияние на $T(F)$ оказывают эти засоряющие значения. Поэтому другое название для $a(x)$ - *функция влияния*. За ней закрепилось довольно неуклюжее обозначение:

$$a(x) = IF(x, T, F)$$

(IF - первые буквы *«influence function»*).

Функция влияния показывает, сколь значительно может измениться $T(F)$ из-за появления «засоряющих» элементов в выборке. Таким образом, она служит одной из количественных характеристик устойчивости $T(F)$ к засорению.

9.2. Асимптотическая нормальность $T(F_n)$

Теперь применим (3), чтобы получить асимптотическое распределение оценки $T(F)$ по выборке. В силу (3):

$$T(F_n) = T(F) + \int a(x) dF_n(x) + R_n. \tag{7}$$

Ясно, что

$$\int a(x)\, dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} a(x_i),$$

причем $\mathsf{E}a(x_i) = 0$. (Ибо $\mathsf{E}a(x_i) = \int a(x)\, dF(x) = 0$.)

Заметим, что случайные величины $a(x_i)$ независимы и одинаково распределены. Поэтому

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} a(x_i) \xrightarrow{d} N(0, \mathsf{E}a^2(x_1)),$$

если

$$\mathsf{E}a^2(x_1) = \int a^2(x)\, dF(x) < \infty.$$

Таким образом,

$$\int a(x) dF_n(x) = O(\frac{1}{\sqrt{n}}).$$

Если остаточный член R_n в (7) стремится к нулю быстрее, чем $\frac{1}{\sqrt{n}}$, то

$$\sqrt{n}[T(F_n) - T(F)] \xrightarrow{d} N(0, Var\ IF(x_1, T, F)). \tag{8}$$

Это и есть утверждение об асимптотической нормальности случайной величины (статистики) $T(F_n)$ при $n \to \infty$.

Приведенное рассуждение не является доказательством. Однако оно приводит к правильным результатам (во всех известных случаях). Оно позволяет предугадать ответ (который затем можно доказывать отдельно).

Задачи.

- Вычислите функции влияния для двух функционалов: $T(F) = \int x dF$ и $T(F) = med\, F$.
- Убедитесь, что первая из них неограниченно возрастает по u, так что $\overline{x} = \sum\limits_{i=1}^{n} x_i$ (как статистическая оценка) неустойчива (not robust).
- С помощью функции влияния найдите асимптотическое распределение $med(x_1, \ldots, x_n)$ - медианы выборки $x_1, \ldots, x_n$ из распределения F, считая, что $F'(\mu) > 0$, где μ - медиана F, т.е. $F(\mu) = 1/2$.

9.3. Асимптотическое неравенство Крамера – Рао

Пусть распределение $F(x,\theta)$ имеет плотность $f(x,\theta)$, $\theta \in \Theta$, $\Theta \subset \mathbb{R}$ - открытое множество.

Пусть

$$T(F(\cdot,\theta)) = \theta$$

для $\theta \in \Theta$.

(Такие функционалы называют *состоятельными по Фишеру.*)

Пусть θ^0 - некоторое фиксированное значение параметра θ, $\theta^0 \in \Theta$. Положим (для краткости) $F^0(x) = F(x,\theta^0)$.

Применим (3), полагая $G(x) = F(x,\theta)$, $F(x) = F^0(x) = F(x,\theta^0)$.

Получаем

$$\theta \equiv T(F(\cdot,\theta)) = T(F^0) + \int IF(y,T,F^0)dF(y,\theta) + R. \tag{9}$$

В данном случае остаточный член R есть функция от θ, θ^0. Предположим, что

$$R = (\theta - \theta^0)g(\theta,\theta^0).$$

Примем все предположения о семействе $F(\cdot,\theta)$, которые мы сделали ранее, при доказательстве неравенства Крамера – Рао. Да и действовать будем по той же схеме.

Дифференцируя (9) по θ в точке $\theta = \theta^0$, найдем:

$$1 = \int IF(y,T,F^0)\frac{\partial}{\partial\theta}f(y,\theta)\Big|_{\theta=\theta^0}dy,$$

или

$$1 = \int IF(y,T,F^0)\left[\frac{\partial}{\partial\theta}\ln f(y,\theta^0)\right]f(y,\theta^0)dy.$$

Правую часть можно записать как математическое ожидание, если ввести случайную величину X, которая имеет плотность $f(y,\theta^0)$:

$$1 = \mathsf{E}_0[IF(X,T,F^0)][\frac{\partial}{\partial\theta}\ln f(X,\theta^0)].$$

Согласно неравенству Коши-Римана-. . . ,

$$(\mathsf{E}\xi\eta)^2 \leqslant (\mathsf{E}\xi^2)(\mathsf{E}\eta^2).$$

Отсюда

$$1 \leqslant \underbrace{\mathsf{E}_0[IF(X,T,F^0)]^2}_{\text{Это } Var\ IF(X,T,F^0)} \cdot \underbrace{\mathsf{E}_0[\frac{\partial}{\partial\theta}\ln f(X,\theta^0)]^2}_{\text{Это } I(\theta^0)}.$$

Следовательно, для любого $\theta^0 \in \Theta$

$$Var\ IF(X,T,F^0) \geqslant \frac{1}{I(\theta^0)}. \tag{10}$$

Левая часть (10) - это асимптотическая дисперсия $T(F_n)$, согласно (8). Следовательно, состоятельная по Фишеру оценка параметра θ не может иметь асимптотическую дисперсию (будучи асимптотически нормальной), меньшую, чем $I^{-1}(\theta)$.

10. Критерии согласия типа Пирсона-Фишера

Они относятся к независимым испытаниям с несколькими исходами и гипотезам об их вероятностях.

Рассмотрим независимые испытания с m ($m \geqslant 2$) исходами. Обозначим их исходы через $A_1,\dots,A_m$.

Вероятности этих исходов неизменны во всех испытаниях. Обозначим эти вероятности через $p_1,\dots,p_m$, причем $\sum\limits_{i=1}^{m} p_i = 1$. Описанные испытания будем называть*испытаниями Бернулли* (даже в случае $m > 2$).

Предположим, что в n испытаниях Бернулли были зарегистрированы частоты (количества осуществлений) $\mu_1,\dots,\mu_m$ исходов $p_1,\dots,p_m$; при этом $\sum\limits_{i=1}^{m}\mu_i = n$. Теоремы, обсуждаемые в этой главе, касаются проверок гипотез о $\vec{p} = (p_1,\dots,p_m)^T$ по частотам $\vec{\mu} = (\mu_1,\dots,\mu_m)^T$.

Начнем с первого критерия такого рода, установленного К. Пирсоном (Karl Pearson) к 1900 году. (Теорему Пирсона, которая будет сформулирована чуть позже, можно считать первой значительной теоремой математической статистики). Критерий Пирсона относится к проверке простой гипотезы о вероятностях:

$$H_0 : \vec{p} = \vec{p}^{\,0}$$

или, подробнее,

$$H_0 : p_1 = p_1^0, \ldots, p_m = p_m^0,$$

где $p_1^0, \ldots, p_m^0$ - заданные положительные вероятности, $\sum\limits_{i=1}^{m} p_i^0 = 1$. Правило Пирсона имеет асимптотический характер и может корректно применяться лишь при достаточно большом количестве испытаний n (что это означает — обсудим позже).

10.1. Правило К. Пирсона

Отвергнуть $H_0 : \vec{p} = \vec{p}^{\,0}$ на (приближенном) уровне ε, $\varepsilon > 0$, если

$$\sum_{i=1}^{m} \frac{(\mu_i - np_i^0)^2}{np_i^0} > \chi^2_{1-\varepsilon}(m-1).$$

Здесь $\chi^2_{1-\varepsilon}(m-1)$ обозначает $(1-\varepsilon)$-квантиль распределения хи-квадрат с $(m-1)$ степенью свободы.

Вопрос о том, какие численности n достаточно велики для того, чтобы, при необходимости, можно было обращаться к этому правилу, довольно темен, несмотря на долгую его историю.

Осторожная (консервативная) рекомендация: должны выполняться соотношения $np_i^0 \geqslant 5$ для всех $i = \overline{1,m}$.

Сказанное правило основано на асимптотических свойствах статистики Пирсона

$$X_n^2 := \sum_{i=1}^{m} \frac{(\mu_i - np_i^0)^2}{np_i^0}$$

при гипотезе (когда истинные вероятности $\vec{p} = \vec{p}^{\,0}$) и альтернативе (когда $\vec{p} \neq \vec{p}^{\,0}$).

- Начнем со случая $\vec{p} \neq \vec{p}^{\,0}$. Перепишем X_n^2 в виде

$$X_n^2 = n \sum_{i=1}^{m} \Big(\frac{\mu_i}{n} - p_i^0 \Big)^2 / p_i^0.$$

По закону больших чисел (в данном случае - это теорема Бернулли)

$$\frac{1}{n}\vec{\mu} \longrightarrow \vec{p}.$$

Поэтому

$$\sum_{i=1}^{m} \Big(\frac{\mu_i}{n} - p_i^0 \Big)^2 / p_i^0 \xrightarrow{P} \sum_{i=1}^{m} \frac{(p_i - p_i^0)^2}{p_i^0}.$$

Этот предел положителен, если и только если $\vec{p} \neq \vec{p}^{\,0}$.

Отсюда следует, что при альтернативе статистика X_n^2 неограниченно возрастает:

$$X_n^2 \xrightarrow{P} \infty$$

при $n \to \infty$.

- Асимптотическое поведение X_n^2 при гипотезе $\vec{p} = \vec{p}^{\,0}$:

Теорема (*Karl Pearson*, 1900 г.).
При $n \to \infty$

$$\sum_{i=1}^{m} \frac{(\mu_i - np_i^0)^2}{np_i^0} \xrightarrow{d} \chi_0^2(m-1).$$

(Случайная величина χ_n^2 при $n \to \infty$ сходится по распределению к хи-квадрат с $(m-1)$ степенями свободы).

Таким образом, большие значения X_n^2, маловероятные при гипотезе H_0, оказываются в области больших вероятностей при альтернативе $\overline{H}_0$.

На этом свойстве X_n^2 и основано приведенное выше правило проверки гипотезы $H_0 : \vec{p} = \vec{p}^{\,0}$.

Мы докажем эту теорему чуть позже.

10.1.1. Многомерная теорема Муавра-Лапласа

В описанной выше схеме испытаний Бернулли с m исходами при $n \to \infty$:

$$\sqrt{n}(\frac{1}{n}\vec{\mu} - \vec{p}) \xrightarrow{d} N(0, \mathcal{P} - \vec{p}\vec{p}^{\,T}),$$

где $\mathcal{P} = diag(p_1, \ldots, p_m)$ - диагональная матрица.

Доказательство.

Доказательство этой теоремы можно провести методом характеристических функций практически так же, как и доказательство классической теоремы Муавра-Лапласа, когда $m = 2$.

В этом последнем случае обычно рассматривают не весь вектор частот (двумерный), но лишь одну его координату, ибо вторая при этом полностью определяется первой (их сумма равна n).

Представляем вектор $\vec{\mu} = (\mu_1, \ldots, \mu_m)^T$ в виде суммы n независимых и одинаково распределенных случайных векторов $\vec{x}_j$, $j = \overline{1,n}$, j - номер испытания.

Все координаты m-мерного вектора $\vec{x}_j$ равны 0, за исключением одной, которая равна 1. Единица стоит на том месте, номер которого соответствует осуществившемуся в j-ом испытании исходу из ряда $A_1, \ldots, A_m$.

Ясно, что

$$\vec{\mu} = \sum_{j=1}^{n} \vec{x}_j$$

и что случайные векторы $\vec{x}_1, \ldots, \vec{x}_j, \ldots$ независимы и одинаково распределены.

Согласно центральной предельной теореме для независимых и одинаково распределенных случайных слагаемых при $n \to \infty$

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\vec{x}_j - \mathsf{E}\vec{x}_j) \xrightarrow{d} N(0, \Sigma),$$

где

$$\Sigma = \mathsf{E}\vec{x}_j\vec{x}_j^{\,T} - (\mathsf{E}\vec{x}_j)(\mathsf{E}\vec{x}_j)^T.$$

Очевидный подсчет дает

$$\mathsf{E}\vec{x}_j = \vec{p}, \qquad \mathsf{D}\vec{x}_j = \mathcal{P} - \vec{p}\vec{p}^{\,T}.$$

Заметим, что матрица $\mathcal{P} - \vec{p}\vec{p}^{\,T}$ вырождена. Ее ранг равен $(m-1)$.

Если бы не это обстоятельство, предельное распределение хи-квадрат для нормы вектора

$$\xi_n \xrightarrow{d} N(0, \Sigma)$$

мы могли бы получить немедленно. Ибо очевидно, что

$$\xi_n^T \Sigma^{-1} \xi_n \xrightarrow{d} \chi^2(m).$$

□

10.1.2. Доказательство теоремы Карла Пирсона

Введем в рассмотрение вектор

$$\xi_n := \sqrt{n}\mathcal{P}^{-1/2}(\frac{1}{n}\vec{\mu} - \vec{p}).$$

Легко видеть, что при $n \to \infty$

$$\xi_n \xrightarrow{d} N(0, I - zz^T),$$

где I - единичная матрица, $z = (\sqrt{p_1}, \ldots, \sqrt{p_m})^T$.

Ведем ортогональную матрицу V, первая строка которой есть $(\sqrt{p_1}, \ldots, \sqrt{p_m})^T$, а прочие строки произвольны. Заметим, что при $n \to \infty$

$$V\xi_n \xrightarrow{d} N(0, I_1),$$

где I_1 - матрица $(m \times m)$, которая получена из единичной заменой левой верхней единицы нулем:

$$I_1 = \begin{pmatrix} 0 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

Это доказывает простая выкладка:

$$\mathsf{D}(V\xi_n) = V(\mathsf{D}\xi_n)V^T = V(1 - zz^T)V^T =$$

$$VV^T - (Vz)(Vz)^T = I - \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix},$$

ибо $Vz = (1, 0, \dots, 0)^T$.

Теперь

$$|\xi_n|^2 = \sum_{i=1}^{m} \Big(\frac{1}{\sqrt{p_i}} \sqrt{n} (\frac{1}{n}\mu_i - p_i) \Big)^2 = \sum_{i=1}^{m} \frac{(\mu_i - np_i)^2}{np_i},$$

а также

$$|\xi_n|^2 = |V\xi_n|^2 \xrightarrow{d} |N(0, I_1)|^2 = \chi^2(m-1).$$

Здесь через $|N(0, I_1)|^2$ мы обозначили квадрат длины, т.е. сумму квадратов координат гауссовского вектора

$$(0, \eta_2, \dots, \eta_m)^T,$$

где $\eta_2, \dots, \eta_m$ суть независимые стандартные гауссовские случайные величины $N(0, 1)$.

По определению,

$$\eta_2^2 + \ldots + \eta_m^2 = \chi^2(m-1).$$

□

10.2. Сложные гипотезы

Здесь мы рассмотрим гипотезы о $\vec{p}$ вида

$$H : \vec{p} \in Q,$$

где Q - некоторое заданное гладкое многообразие, принадлежащее симплексу $\{\vec{p} : \sum_{i=1}^{m} p_i = 1, p_1 \geqslant 0, \dots, p_m \geqslant 0\}$. «Гладкое» здесь означает, что в каждой точке $\vec{p} \in Q$ существует касательное линейное многообразие. Размерность $\vec{p}$ обозначим через r.

Справедлива

Теорема 1 (*J. Neyman, E. Pearson*, 1928):

При $n \to \infty$

$$\min_{\vec{p} \in Q} \sum_{i=1}^{m} \frac{(\mu_i - np_i)^2}{np_i} \xrightarrow{d} \chi^2(m - r - 1). \qquad (*)$$

Заметим, что при вычислении статистики из $(*)$ обычно находят и то значение $\vec{p} \in Q$, при котором достигается минимум в $(*)$. Это минимизирующее значение часто называют *оценкой* $\vec{p} \in Q$, полученной по *«методу минимума хи-квадрат»*.

Другая формулировка той же теоремы возникает, когда многообразие Q задано параметрически, т.е., когда гипотеза $\vec{p} \in Q$ представима в виде

$$\vec{p} = \vec{p}(\theta),$$

где θ - r-мерный параметр.

Пусть $\hat{\theta}_n$ - оценка наибольшего правдоподобия для неизвестного θ, основанная на частотах $\mu_1, \dots, \mu_m$. (Либо иная оценка, но с теми же асимптотическими свойствами, что и $\hat{\theta}_n$). Тогда справедлива

Теорема 2.

При $n \to \infty$

$$\sum_{i=1}^{m} \frac{(\mu_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \xrightarrow{d} \chi^2(m - r - 1). \qquad (**)$$

Эти теоремы и другие, подобные, часто связывают с именем Р. Фишера (R. A. Fisher).

Фишер был первым, кто заметил уменьшение числа степеней свободы предельного распределения хи-квадрат, когда параметры оцениваются по выборке, и ровно настолько, сколько независимых параметров пришлось оценить. Он обнаружил это при проверке гипотезы о независимости признаков в таблицах сопряженности. Мы будем говорить об этом в 10.5.

А сейчас, чтобы окончить, сформулируем правило проверки $H : \vec{p} \in Q$, основанное на приведенных выше теоремах. А также на том факте, что статистики $(*)$ и $(**)$ неограниченно возрастают при $n \to \infty$, если истинное значение $\vec{p} \notin Q$.

Правило проверки $H : \vec{p} \in Q$ против $\overline{H} : \vec{p} \notin Q$.

> Отвергаем H на (приближенном) уровне $\varepsilon > 0$, если статистика $(*)$ или $(**)$ превосходит $(1-\varepsilon)$-квантиль $\chi^2(m-r-1)$.

Это правило применимо «для достаточно больших n». Осторожная (консервативная) практическая рекомендация: $\mu_i \geqslant 5$. (Впрочем, разные авторы говорят несколько различное на эту тему.)

10.3. Таблицы сопряженности.

Предположим, что каждый объект некоторой (бесконечной) совокупности может быть классифицирован по двум признакам A и B. Признак A при этом принимает r значений, признак B - s значений, соответственно $A_1, \ldots, A_r$ и $B_1, \ldots, B_s$. Каждый объект обладает некоторой комбинацией A_iB_j, $i = \overline{1,r}$, $j = \overline{1,s}$, значений признаков A и B.

Пусть p_{ij} обладает комбинацией признаков A_iB_j.

Пусть μ_{ij} - число комбинаций A_iB_j, зарегистрированное при случайном выборе n объектов из генеральной совокупности (μ_{ij} - выборочные частоты). Таблицу частот $\|\mu_{ij}, i = \overline{1,r}, j = \overline{1,s}\|$ называют *таблицей сопряженности* признаков A и B.

Важная статистическая гипотеза - гипотеза о независимости признаков A и B.

В этом случае

$$p_{ij} \equiv P(A_iB_j) = P(A_i)P(B_j).$$

Вероятность появления A_i и вероятность появления B_j обозначим через $p_{i\cdot}$ и $p_{\cdot j}$ соответственно. При этом

$$p_{i\cdot} = \sum_{j=1}^{s} p_{ij}, p_{\cdot j} = \sum_{i=1}^{r} p_{ij}.$$

Гипотеза независимости признаков теперь может быть выражена так:

$$H : p_{ij} = p_{i\cdot}p_{\cdot j} \qquad \text{для всех } i = \overline{1,r}, j = \overline{1,s}.$$

Каждое извлечение объекта из генеральной совокупности - это испытание Бернулли, которое оканчивается одним из $m = rs$ исходов A_iB_j.

При гипотезе $H : p_{ij} = p_{i\cdot}p_{\cdot j}$, вероятности этих исходов выражаются через параметры $p_{i\cdot}$, $p_{\cdot j}$. Поэтому вектор вероятностей (в данном случае - матрица размера $(r \times s)$) $\vec{p} = |p_{ij}|$ принадлежит $(r+s-2)$-мерному многообразию. (Размерности именно $r+s-2$, так как параметры подчиняются связям $\sum\limits_{j=1}^{r} p_{i\cdot} = 1$, $\sum\limits_{i=1}^{s} p_{\cdot j} = 1$.)

Поскольку мы имеем дело с испытаниями Бернулли и гипотезой о вероятностях в этих испытаниях, мы можем воспользоваться результатами пункта 10.4.

Для этого найдем оценки наибольшего правдоподобия для $p_{i\cdot}$ и $p_{\cdot j}$ и затем применим теорему 2.

Правдоподобие $\|p_{ij}\|$, основанное на таблице $|\mu_{ij}|$, равно

$$n! \prod_{i=1}^{r} \prod_{j=1}^{s} \frac{1}{(\mu_{ij})!} (p_{ij})^{\mu_{ij}}.$$

При гипотезе независимости правдоподобие упрощается: правдоподобие $|p_{i\cdot}, p_{\cdot j}, i = \overline{1,r}, j = \overline{1,s}|$ равно

$$Const \prod_{i=1}^{r} (p_{i\cdot})^{\mu_{i\cdot}} \prod_{j=1}^{s} (p_{\cdot j})^{\mu_{\cdot j}}.$$

где $\mu_{i\cdot} = \sum\limits_{j=1}^{s} \mu_{ij}, \mu_{\cdot j} = \sum\limits_{i=1}^{r} \mu_{ij}$, *Const* означает множитель, не содержащий параметров $p_{i\cdot}$, $p_{\cdot j}$ (и поэтому не влияющий на оценки наибольшего правдоподобия).

Далее легко находим оценки наибольшего правдоподобия:

$$\hat{p}_{i\cdot} = \frac{\mu_{i\cdot}}{n}, \hat{p}_{\cdot j} = \frac{\mu_{\cdot j}}{n} \qquad \text{для } i = \overline{1,r}, j = \overline{1,s}.$$

Статистика X_n^2 из теоремы 2 здесь

$$X_n^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(\mu_{ij} - n\frac{\mu_{i\cdot}}{n}\frac{\mu_{\cdot j}}{n})^2}{n\frac{\mu_{i\cdot}}{n}\frac{\mu_{\cdot j}}{n}}.$$

При гипотезе независимости признаков

$$X_n^2 \xrightarrow{d} \chi^2((r-1)(s-1)),$$

ибо $rs - (r+s-2) - 1 = (r-1)(s-1)$.

Гипотезу независимости признаков следует отвергать, если наблюденное (вычисленное) значение статистики X_n^2 слишком велико (по сравнению с квантилями распределения хи-квадрат с указанным числом степеней свободы).