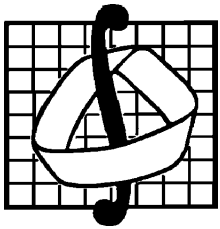
МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ
имени М.В.ЛОМОНОСОВА



Механико-математический факультет
Кафедра теории вероятностей

Математическая статистика. Записки лекций.

Ю.Н. Тюрин

Под редакцией Г.И. Симоновой

Москва 2003 год

Ю.Н. Тюрин
Под редакцией Г.И. Симоновой

Математическая статистика.
Записки лекций.

Курс математической статистики является обязательным для студентов отделения математики механико-математического факультета МГУ. По действующему учебному плану курс читается в пятом семестре. Настоящее издание представляет собой несколько расширенные записки лекций этого курса.

Издание может быть полезным для студентов математических факультетов и других ВУЗов.

Рецензент — д.ф.-м.н., проф. В. Н. Тутубалин

Содержание

Предисловие

Курс представляет собой введение в математическую статистику в её традиционном понимании, когда отклонения от закономерностей (ошибки) толкуют как случайные величины, в духе частотной теории вероятностей.

Мы стремились изложить в курсе основные идеи (оценивание, проверка гипотез, робастность, статистическая оптимальность и т. д.) и основные методы (наименьшие квадраты, максимальное правдоподобие, непараметрический подход и т. д.) математической статистики. Мы знакомим с ними наших слушателей на примере линейной статистической модели как в гауссовской, так и в непараметрической постановках. Первая часть курса посвящена оптимальным статистическим выводам для конечного числа наблюдений. Несколько особняком в ней стоят две лекции, где рассказано об условных математических ожиданиях (по Колмогорову). Эта тема относится, скорее, к теории вероятностей, но по особенностям учебного плана перенесена в курс математической статистики. Условные математические ожидания необходимы для изложения несмещенного оценивания с минимальной дисперсией — теоретического основания для метода наименьших квадратов, и для асимптотической теории U-статистик.

Во второй части на примере выборки изложен асимптотический подход к анализу данных, когда объем выборки неограниченно возрастает.

Большим недостатком курса, который его автор полностью сознает, является как отсутствие задач с числовыми (реальными) данными, так и отсутствие примеров, где бы такие данные обрабатывались. Но опыт преподавания убедил автора в том, что среди студентов факультета теоретической математики (каков мех-мат) подобные задачи не находят много приверженцев.

Автор и редактор благодарят Е. Сафонову за помощь в подготовке электронного варианта рукописи.

Программа курса лекций

I. Начала теории оценивания.

1. Понятие статистической модели. Примеры: выборка, линейная гауссовская модель.
2. Теорема Гливенко.
3. Некоторые понятия теории оценивания: функции ущерба и риска, допустимые и байесовские оценки. Несмещенное оценивание, квадратичный риск.
4. Неравенство информации (неравенство Крамера-Рао) для регулярных однопараметрических семейств: непрерывные и дискретные распределения.
5. Эффективные оценки, экспоненциальные семейства распределений.
6. Многомерное неравенство Крамера-Рао.
7. Достаточные статистики: определение, примеры.
8. Достаточные статистики в случае нормальной выборки. Выборочные коэффициенты асимметрии и эксцесса.
9. Теорема факторизации (доказательства для элементарных случаев).
10. Улучшение несмещенных оценок путем их усреднения по достаточным статистикам: одномерная теорема Блеквелла-Рао.
11. Многомерная теорема Блеквелла-Рао.
12. Полные достаточные статистики: определение, примеры, единственность наилучшей несмещенной оценки.

II. Условные математические ожидания и условные вероятности.

13. Напоминания: вероятностные пространства, случайные величины, абсолютная непрерывность мер, производная Радона-Никодима.
14. Условное математическое ожидание случайной величины относительно сигма-алгебры: определение.
15. Некоторые простейшие свойства условного математического ожидания.
16. Условные вероятности, условные распределения.
17. Специальный случай: $E(X|Y)$ для простых случайных величин X и Y.
18. Некоторые дальнейшие свойства условных математических ожиданий. В частности: $E(\varphi(Y)X|Y) = \varphi(Y)E(X|Y)$; условная

дисперсия, наилучший квадратичный прогноз.
19. Примеры вычисления $E(X|Y)$.

Ш. Оценивание в линейной модели.
20. Регрессионные и факторные (с одним и двумя факторами) линейные модели.
21. Достаточные статистики в линейной гауссовской модели.
22. Лемма об ортогональных разложениях случайного гауссовского вектора. Распределения хи-квадрат (центральные и нецентральные).
23. Наилучшие несмещенные оценки параметров в линейной гауссовской модели, их распределения.
24. Вычисление оценок наименьших квадратов в модели линейной регрессии.
25. Интервальные оценки для параметров нормальной выборки. Распределения Стьюдента (центральное и нецентральное).
26. Доверительные границы для вероятности успеха в испытаниях Бернулли.
27. Доверительные эллипсоиды для параметров линейной гауссовской модели. Эф-отношения и эф-распределения.

IV. Проверка статистических гипотез.
28. Проверка статистических гипотез: общие принципы и основные понятия (критическое множество, уровень значимости, альтернативы, ошибки первого и второго родов, функция мощности).
29. Лемма Неймана-Пирсона.
30. Понятие о равномерно наиболее мощных критериях и пример: проверка гипотезы $\theta \leq \theta_0$ против альтернативы $\theta > \theta_0$ (по результатам испытаний Бернулли, здесь θ — вероятность успеха).
31. Связь между проверкой гипотез и доверительным оцениванием.
32. Проверка линейных гипотез в гауссовских линейных моделях с помощью критерия отношения правдоподобий.
33. Однофакторный дисперсионный анализ: проверка нулевой гипотезы (о равенстве эффектов обработки).
34. Двухфакторный дисперсионный анализ (аддитивная модель, одно наблюдение в клетке): проверка нулевой гипотезы (о равенстве эффектов обработки).
35. Две гауссовские выборки, могущие отличаться сдвигом: про-

верка гипотезы о их однородности. Доверительные интервалы для параметра сдвига.
36. Ранги наблюдений. Статистика ранговых сумм $W_{m,n}$ (Уилкоксон) для проверки гипотезы об однородности двух выборок. Вычисление $EW_{m,n}$ и $DW_{m,n}$ при гипотезе.
37. Точечные и интервальные оценки для сдвига одной выборки относительно другой с помощью статистики ранговых сумм Уилкоксона.
38. Статистика Манна-Уитни $U_{m,n}$ и ее связь с $W_{m,n}$.
39. Теорема Слуцкого.
40. Теорема об асимптотической нормальности двухвыборочных U- статистик.
41. Асимптотические распределения статистик Манна-Уитни $U_{m,n}$ и Уилкоксона $W_{m,n}$ при m, $n \to \infty$.

V. Асимптотические методы оценивания и проверка гипотез.
42. Метод наибольшего правдоподобия. Неравенства теории информации.
43. Состоятельность оценок наибольшего правдоподобия (выборка, одномерный параметр).
44. Асимптотическая нормальность оценок наибольшего правдоподобия (выборки из регулярных семейств распределений).
45. Устойчивость оценок, функции влияния, примеры.
46. М-оценки и их функции влияния.
47. Асимптотические свойства М-оценок — эвристический вывод с помощью функций влияния.
48. Критерий согласия К.Пирсона для простой гипотезы; теорема К.Пирсона.
49. Критерии типа Пирсона-Фишера для сложных гипотез (без доказательства). Проверка гипотезы о независимости признаков по таблице сопряженности.

Лекция 1. Вступление: чем и как занимается математическая статистика

Математическую статистику можно описать как науку о том, как делать выводы из наблюдений (измерений, сообщений и т.п.), которые подвержены действию случая, а также и о том, как собирать такие сведения (т. е. наблюдать, измерять и т. д.). Математическая статистика, будучи наукой математической, ориентирована на приложения. Из приложений приходят задачи, которыми занимается математическая статистика и для которых она создает свои теории. Статистические методы нужны всюду, где есть измерения (в широком смысле) и где есть необходимость действий с ними (их обработка). В первую очередь это естественные науки и техника, но также медицина, психология, науки о сельском хозяйстве, экономика и т.д. Эконометрия, будучи частью экономической науки, по сути представляет собой специализированный раздел математической статистики.

§ 1. Статистические модели и задачи

В этом разделе мы рассмотрим несколько простых примеров. На этих примерах мы познакомимся с тем, каким может быть статистический материал (упомянутые выше наблюдения), как ставятся задачи математической статистики и какими они бывают. Так мы придем к общей (абстрактной) статистической модели. Попутно мы увидим, каким образом статистический материал можно представить наглядно.

1.1. Простейшая модель: выборка

Этот пример я заимствую из старой книги А. Хальда [Хальд А. Математическая статистика с техническими приложениями. — М.: Изд-во Иностранной литературы, 1956. — 664 с.]. Хальд приводит результаты измерений 200 заклепок, случайно отобранных для этого в процессе производства.

Совокупность объектов, выбранных из общей совокупности для последующего изучения, называют *выборкой* (*sample*). (У этого слова несколько значений. Другие нам встретятся позже). Совокупность, из которой извлечена выборка, называют *генеральной*

совокупностью, *популяцией* (*population*) и т. д. Статистика употребляет эти термины как для конечных, так и для бесконечных совокупностей, как реально существующих, так и примысливаемых к выборкам. Хальд приводит данные в том порядке, как они поступали. Первичное представление данных — таблица.

Таблица 1.1.1.

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

Табличное представление данных не дает о них быстрого и ясного понятия. Для этого нужны иные формы — более наглядные и обобщенные. Приемы такого рода образуют так называемую описательную статистику. Кое-какие из них мы сейчас используем. *Точечная диаграмма* (*scatter diagram*, рис. 1.1.1) в нашем случае строится так. На оси абсцисс выбираем такой масштаб, чтобы на

листе бумаги (на экране) можно было удобно разместить все (или почти все — см. ниже) наблюдения. Каждое наблюденное значение далее представляем точкой плоскости, которая располагается над осью абсцисс и абсцисса которой равна этому значению. Рис. 1.1.1, полученный таким способом, дает о данных таблицы гораздо более ясное представление. При построении мы обнаруживаем, что одно из чисел выборки — именно, 14.56 — далеко отступает от основной массы наблюдений. Это число — явный "выброс". На него следует обратить внимание. Некоторое количество выбросов присутствует в выборках довольно часто. Порой они столь велики, что при разумно выбранном масштабе им не остается места на графике (как это случилось и у нас). О том, как относиться к выбросам, мы будем говорить в своем месте. В данном случае выброс — явная опечатка. Ее надо либо исправить (на 13.56, по-видимому), либо это число просто удалить.

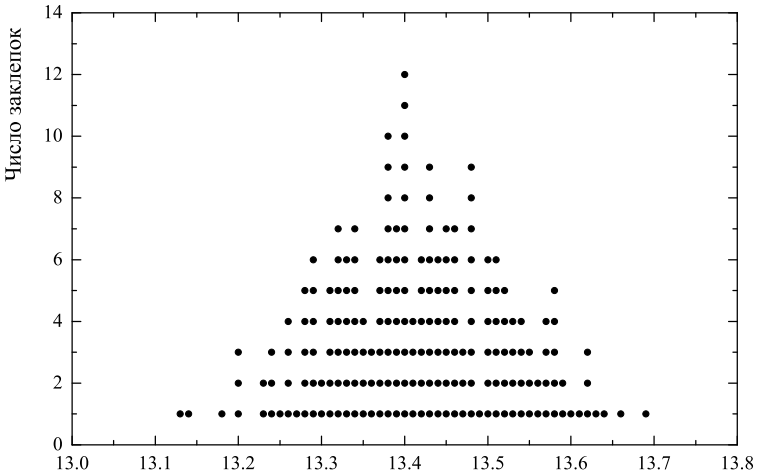


Рис. 1.1.1. Точечная диаграмма диаметров головок заклепок (по данным табл. 1.1.1)

Представление числового массива можно сделать еще более наглядным при помощи его группировки и построения *гистограмм* (*histogram*, рис. 1.1.2). Ось абсцисс в диапазоне наблюденных значений разбивают на некоторое число непересекающихся интерва-

лов (обычно равных). Над каждым интервалом "надстраивают" прямоугольник, площадь которого пропорциональна числу попавших в его основание наблюдений. В случае равных интервалов высоты прямоугольников пропорциональны этим численностям (частотам интервалов). Рис. 1.1.2 показывает гистограммы с разными длинами интервалов.

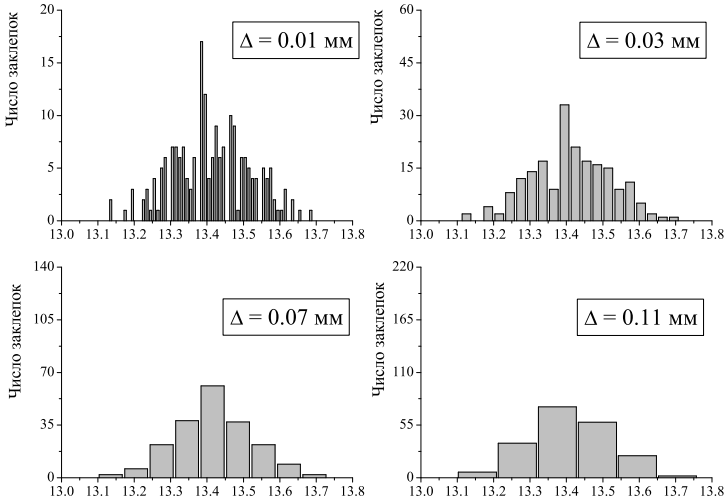


Рис. 1.1.2. Гистограммы диаметров головок заклепок при разных длинах интервалов группировки Δ

По этим рисункам можно понять, как длина интервала группировки влияет на форму гистограммы. Если длина интервала группировки мала, то случайные колебания оказывают сильное влияние на форму гистограммы (зубчатую), так как при этом каждый интервал содержит лишь небольшое число измерений. Если же длина интервала велика, то скрадываются характерные черты распределения. Гистограммы и точечная диаграмма наводят нас на *статистическую модель*: данные в табл. 1.1.1 — это независимые реализации какой-то случайной величины. Более подробный анализ (например, с помощью выборочной функции распределения и нормальной вероятностной бумаги — см. § 3) показывает, что это гауссовская случайная величина (упомянутое число 14.56 в рамках гауссовской модели появиться не может и, действитель-

но, должно рассматриваться как грубая ошибка наблюдения, как выброс).

В статистике совокупность независимых реализаций случайной величины также называют *выборкой* — выборкой из некоторого распределения вероятностей. Название выборки часто переносят и на совокупность независимых случайных величин.

Итак, статистическая модель для этого примера: 200 чисел табл. 1.1.1 — это выборка из нормального распределения (содержащая один выброс, если не исправлять упомянутую опечатку). Как еще говорят, выборка из нормальной (генеральной) совокупности $N(a, \sigma^2)$. Параметры a и σ^2 этого нормального распределения в модели не уточняются, они остаются неопределенными.

Статистические задачи для модели выборки:

- оценить параметры распределения, в данном случае гауссовского, т.е. найти для a и σ^2 приближенные значения;
- проверить предположение, что данная выборка извлечена из нормальной совокупности (проверить согласие гауссовского распределения с распределением выборки).

1.2. Простая линейная регрессия

Рассмотрим данные из статьи Э. Хаббла [Hubble E. A relation between distance and radial velocity among extra-galactic nebulae //Astronomy. -1929.- v.15. - p. 168-173], 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 приводит нас к мысли (так же, как и Э. Хаббла более семидесяти лет назад), что лучевые скорости "в целом" пропорциональны удалениям, и к следующей модели:

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

Здесь:
x_i — расстояние до i-ой туманности (удаление),
y_i — её лучевая скорость,
ε_i — отступление от линейной зависимости. Их называют ошибками. Эти отступления, возможно, объясняются собственными движениями туманностей в пространстве, а также ошибками в измерении скоростей и удалений.

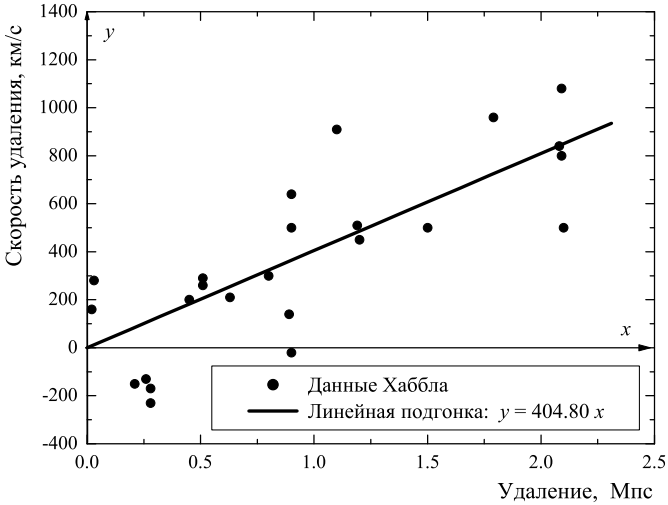


Рис. 1.1.3. Данные из статьи E. Hubble 1929 года, связывающие удаления и лучевые скорости 24 туманностей.

Коэффициент θ, определяющий скорость расширения Вселенной, сейчас называют *постоянной Хаббла*.

Величины удалений и скоростей для тех туманностей, которые отражены на рис. 1.1.3, впоследствии были пересмотрены и уточнены, поэтому численное значение θ сильно изменилось по сравнению с тем, которое нашел сам Хаббл (см. рис. 1.1.4).

Были также измерены удаления и скорости для многих других туманностей, находящихся на гораздо бóльших 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, составляет один из основных законов астрономии. Впрочем, численное значение θ — вопрос все еще дискуссионный (и чрезвычайно важный для теорий возникновения и эволюции Вселенной).

Каждое из последующих предположений или их сочетания позволяют делать более содержательные и определенные выводы о свойствах модели, её параметрах, её адекватности, о предсказаниях на её основе и т. д. Одновременно принятие более содержательных предположений о модели делает эти выводы более подвержен-

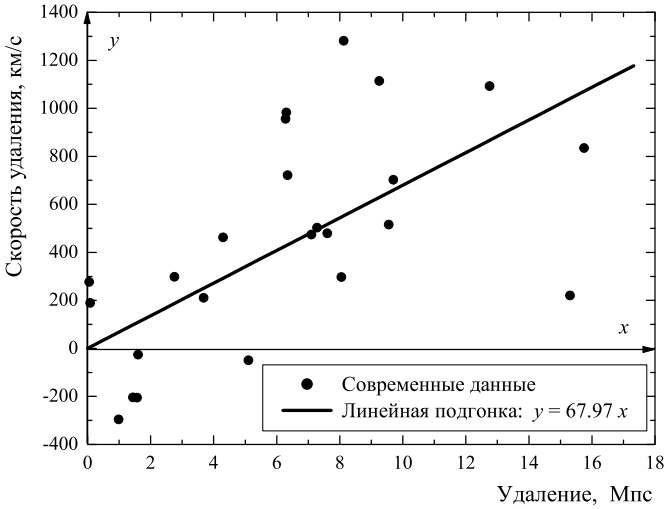


Рис. 1.1.4. Современные данные, связывающие удаления и лучевые скорости туманностей.

ными риску ошибки, если сделанные предположения неверны.

Приведем — в порядке убывания важности — предположения об ошибках, которые делает математическая статистика.

(1) $\varepsilon_1, \ldots, \varepsilon_n$ — случайные величины (в том смысле, который дает этому понятию теория вероятностей). Предположение о случайности ошибок — главное для математической статистики.

(2) $\varepsilon_1, \ldots, \varepsilon_n$ — независимые случайные величины. Это очень важное предположение. Содержательно оно означает, что ошибки, сделанные в одном измерении, не влияют на ошибки других измерений, т.е. каждое измерение делается наново. На практике встречаются задачи, где предположение о независимости ошибок неприемлемо. Но тогда надо указать, какова структура их зависимости (т.е. уточнить статистическую модель).

(3) В измерениях нет систематических ошибок. Математически эту мысль выражают разными способами. Наиболее употре-

бительно предположение

(a) $$E\varepsilon_1 = E\varepsilon_2 = \ldots = E\varepsilon_n = 0.$$

Другая возможная форма: для каждого i

(b) $$P\{\varepsilon_i > 0\} = P\{\varepsilon_i < 0\} = 0.5.$$

Статистические методы позволяют уменьшить влияние случайных ошибок (при увеличении числа наблюдений) лишь тогда, когда систематические ошибки отсутствуют. Реально, в приложениях, предположение об отсутствии систематических ошибок означает, что такие ошибки пренебрежимо малы по сравнению со случайными.

(4) $\varepsilon_1, \ldots, \varepsilon_n$ — одинаково распределенные случайные величины. Если это предположение неприемлемо, желательно указать, как точность измерения в опыте i связана с другими переменными модели (с x_i, например). Т.е. уточнить модель.

(5) $\varepsilon_1, \ldots, \varepsilon_n$ суть независимые случайные величины, распределенные по нормальному закону $N(0, \sigma^2)$. Дисперсию σ^2 обычно считают неизвестной. Это предположение об ошибках долгое время для статистики было основным. Его можно назвать классическим. Предположение о нормальном распределении ошибок хорошо описывает характер таких ошибок в геодезии, астрономии и других науках с высокой культурой измерения.

Все вместе — формула $y_i = \theta x_i + \varepsilon_i$ и предположения об ошибках — образуют статистическую модель для наблюдений (x_i, y_i), $i = 1, \ldots, n$. Если, например, мы принимаем предположение (5), мы получаем т.н. гауссовскую модель. А, скажем, предположения (1) и (3b) дают одну из т.н. непараметрических моделей.

На этом примере можно перечислить и основные задачи, которые стоят перед математической статистикой.

• О ц е н и в а н и е. Надо предложить такой метод обработки наблюдений (x_i, y_i), $i = 1, \ldots, n$, т.е. такую функцию $f(.)$, чтобы $f(x_1, y_1, x_2, y_2, \ldots)$ приближенно равнялось неизвестному θ. Величину $\hat{\theta} = f(x_1, y_1, \ldots)$ называют *оценкой* параметра θ. Оценкой называют и самою функцию $f(.)$. Функцию $f(.)$ надо выбирать так, чтобы приближенное равенство $\hat{\theta} \approx \theta$ было как можно более

точным. Как измерить близость $\hat{\theta}$ к θ и как выбрать функцию $f(.)$ — предметы дальнейшего изучения.

• Точность оценивания. Надо дать количественную характеристику для точности приближенного равенства $\hat{\theta} \approx \theta$. Эта задача решается в рамках т.н. *доверительного* (*интервального*) оценивания.

• Проверка статистических гипотез. Возможно, что о величине неизвестного θ существуют какие-то предположения (не основанные на наблюдениях $(x_1, y_1), \ldots, (x_n, y_n)$). Например, предположение, что $\theta = \theta_0$ или $\theta \geq \theta_0$, где θ_0 — заданное значение. В таком случае естественно поставить вопрос о проверке этого предположения по имеющимся наблюдениям.

Замечание о терминологии. При формировании статистической модели обычно приходится делать те или иные предположения о свойствах наблюдений, связи переменных и параметров, свойствах случайных ошибок и т.д. Чтобы выделить те предположения, которые мы собираемся проверить по наблюдениям, эти предположения мы будем называть гипотезами, точнее — статистическими гипотезами.

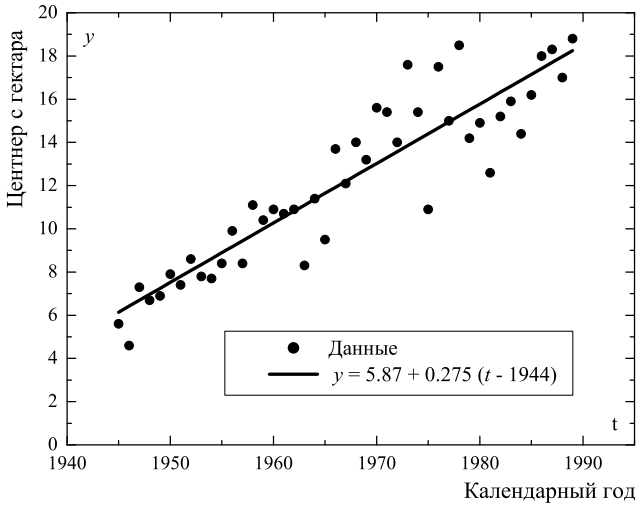


Рис. 1.1.5. Динамический ряд урожайности зерновых культур в СССР за 1945–1989 гг.

Второй пример линейной регрессии: урожайность зерновых в СССР, по данным А. Манелли. Модель: $y_t = a + bt + \varepsilon_t$, t — календарный год.

1.3. Финансовые данные [1]

1.3.1. Что такое финансовые данные? Вероятностно-статистические методы, по своему историческому происхождению — это определенная группа методов математической физики. Впрочем, и при самом их возникновении совершались попытки применения в областях гораздо более широких, чем фундаментальная или прикладная физика. Страхование, демографию и даже вероятности судебных приговоров можно найти в старинных трактатах по теории вероятностей. Но в последние несколько десятилетий бóльший научный интерес и лучшую зарплату обещают приложения к экономике, чем к классическим областям физики или техники. Поэтому многие специалисты, учившиеся и работавшие как математики или физики, переключаются на работу в области экономики, в частности, на финансы. Интересно представить себе реальные возможности применения мышления в стиле математической физики в этих вещах.

По-русски, "финансы" и "деньги" — почти синонимы. Небольшое размышление показывает, что мы на самом деле плохо знаем, что такое деньги. Еще сто лет назад в ходу были золотые монеты (вроде бы как основа денежной системы), но они дополнялись ассигнациями. Специалисты по денежному обращению заметили, что государство, в сущности, не имеет способа ограничить массу денег, обращающихся в стране, потому что при ограничении, скажем, количества монет и ассигнаций в ход идут различные векселя, расписки или билеты, которые обращаются примерно на тех же правах, что и такие деньги, которые монопольно выпускаются государством. До известных пор, конечно, пока не возникает кризис доверия к суррогатам и весь народ начинает жаждать почему-то именно "настоящих" денег. Тогда (как это было в середине девятнадцатого века) Английский банк хоть и не перестает совсем выдавать деньги по вкладам, но делает это ... трехпенсовыми монетами, пока жаждущим не надоест стоять в очереди и кризис не успокоится.

[1] Этот раздел написан В.Н. Тутубалиным

Ценные бумаги, такие как акции или государственные обязательства, надо, очевидно, считать разновидностями денег. Они могут существовать и в безбумажном виде, т. е. в виде компьютерных кодов. При сколько-нибудь нормальных условиях ценные бумаги ликвидны, т. е. могут быть быстро проданы на рынке (т. е. превращены в ту или иную валюту), а следовательно (через валюту) и друг в друга. Вот и возникает всемирный финансовый рынок, который занимается в огромных масштабах тем, что превращает одни виды денег в другие. Он живет бурной жизнью благодаря спекулянтам, заветной мечтой которых является выгадать что-нибудь на колебаниях курсов, по которым одни деньги превращаются в другие. Для этого хотелось бы находить в общем хаосе какие-то закономерности, в том числе путем применения вероятностно-статистических методов.

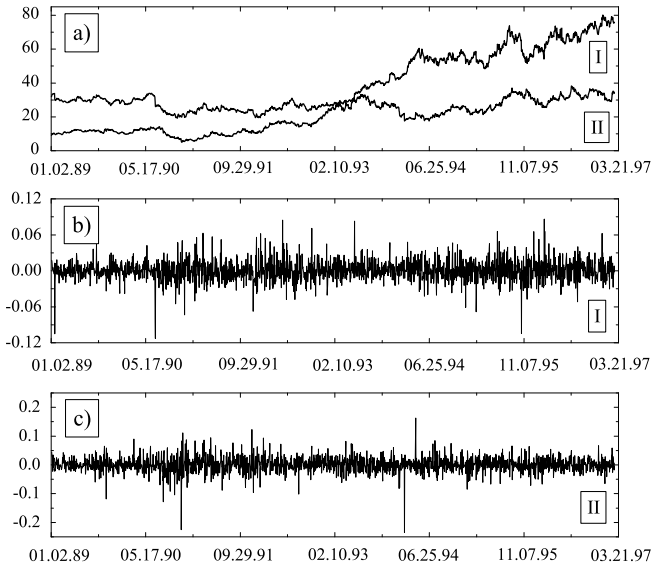


Рис. 1.1.6. a) Динамика цен акций двух американских компаний I и II; b) приращения логарифмов цен акций I-ой компании; c) приращения логарифмов цен акций II-ой компании

Рассмотрим рис. 1.1.6a), на котором представлен курс акций двух наудачу выбранных американских компаний. Для начала

следует аккуратно сказать, какие именно данные представлены на этом рисунке. Это так называемые цены закрытия торгов. Каждый день (исключая выходные и праздники) на биржах Соединенных Штатов совершаются сделки с акциями многих различных компаний (кто-то продает акции, а кто-то другой их одновременно покупает). Данные о цене, по которой совершилась сделка, попадают в биржевую информационную систему, а затем делаются доступными всем желающим. Особенное значение придается цене последней сделки с акциями данной компании в данный торговый день: это как бы итог, на котором остановился рынок в данный день в своем процессе оценки акций этой компании (на момент закрытия торгов). Кроме того, могут быть установлены те или иные правила игры, в которых участвует именно цена последней сделки. Например, с учетом этой цены происходит исполнение опционов. Но как быть, если в какой-нибудь день сделок с акциями данной компании вообще не производилось? Тогда нужно брать цену закрытия предыдущего дня и т.д.

Таким образом, следует помнить о том, что при имитации тех или иных стратегий биржевой игры на основе прошлых данных о ценах закрытия мы не в состоянии точно воссоздать, какими были бы настоящие цены сделок при применении этой стратегии в реальных условиях. По порядку величины разница может составлять единицы процентов от цены акций, что в одних условиях может быть существенно, а в других — нет. Но в дальнейшем при описании имитаций действия стратегий мы не будем упоминать об этой неизбежной условности.

Есть еще одно обстоятельство, о котором следует упомянуть. Время от времени та или иная компания производит со своими акциями так называемый "сплит" (split), при котором старые акции обмениваются на новые в определенном отношении, например 3:2, т.е. две старых акции обмениваются на три новых. Автоматически цена новой акции равна 2/3 цены старой. Понятно, что при рассмотрении динамики цен за какой-то длительный период нужно привести все цены к каким-то определенным акциям, обычно тем, которые существуют на конец периода. На американском рынке укрепилась немного странная традиция, когда целые доллары в значении цены считаются, естественно, по десятичной системе, но дроби после запятой — по двоичной, например, 5/32 доллара (дальше 1/32, кажется, не идут). В файлах же данных, конечно, всё пишется по десятичной системе. Казалось бы, могут

возникать лишь такие десятичные дроби, которые соответствуют степеням двойки, но из-за сплита это не всегда так.

Итак, по оси ординат на рис. 1.1.6а) отложены цены последних сделок каждого дня с акциями данной компании, приведенные к тем акциям, которые были на конец рассматриваемого календарного периода. Что же отложено по оси абсцисс? Дело в том, что в файлах данных стоят календарные даты, но те (выходные или праздничные) дни, в которые не было торгов, пропущены. Таким образом, по оси абсцисс фактически отложен номер дня торгов, считая за нулевой день 2 января 1989 года, но (по понятным соображениям) оцифровка дана в календарных датах. Иными словами, в дни, когда нет торгов, считается, что время как бы не идет, и по-видимому, это достаточно правильно для финансовых данных. В году примерно 250 торговых дней, и в связи с этим возникает вопрос, как разумнее пересчитывать годовые проценты в дневные: конкретно 5% годовых — это 0.05/365 или 0.05/250 в день? Видимо, более правилен второй способ (если, конечно, речь идет о каких-то расчетах на финансовом рынке).

Теперь можно сказать кратко: на рис. 1.1.6а) представлены данные о ценах акций двух американских компаний примерно за 8 календарных лет: с начала января 1989 г. по конец января 1997 г. В каждом файле данных наблюдений несколько более, чем $8*250 = 2000$. Календарные даты обозначены по американской системе: 05.17.91 означает 17 мая 1991 года. Данные представлены с некоторым огрублением в соответствии с возможностями компьютерной графики.

Первое наблюдение, которое можно сделать, глядя на рис. 1.1.6а), состоит в том, что курсы акций чрезвычайно динамичны. Пусть целью спекулянта является приобретение возможно большего количества долларов (это не обязательно так: доллары лишь один из видов денег и можно было бы стремиться приобрести, наоборот, побольше акций или чего-нибудь еще). Тогда важно оценить, за сколько времени можно (в принципе) получить тот или иной процент прибыли на вложенный капитал. Оцифровка оси абсцисс на рис. 1.1.6а) произведена с интервалом примерно в 15 месяцев. Мы видим, что в ряде случаев курс акций за половину этого срока меняется в 1.5–2 раза. Итак, если удачно (дешево) купить и тоже удачно (дорого) продать, то можно менее чем за год заработать 50-100% прибыли. Вот и возникает племя биржевых игроков, которые рассчитывают на свои способности удачно выбирать мо-

менты покупки и продажи. Посмотрим, какое отношение к этому могут иметь вероятностные методы.

1.3.2. Вероятностные модели динамики курсов акций. При одном взгляде на рис. 1.1.6а) становится ясным, что в вероятностном смысле речь идет о нестационарных случайных процессах. Но в понятии нестационарного случайного процесса пользы мало: нестационарность означает, что распределения вероятностей, соответствующие процессу, как-то меняются со временем. Чтобы эти распределения каким-то образом узнать путем обработки фактических данных, следовало бы создать ансамбль идентичных в вероятностном смысле реализаций данного случайного процесса. Но как создать ансамбль коммерческих компаний, идентичных данной? В этом и разница между миром экономики и миром физики, в котором ансамбль идентичных процессов, вообще говоря, возможен.

Следующим ходом мысли является разложение наблюдаемого процесса на сумму детерминированной и случайной составляющих. Детерминированная составляющая в экономике называется трендом, и вопрос состоит в том, чтобы так определить и вычесть тренд, чтобы оставшаяся случайная составляющая оказалась, по меньшей мере, стационарным случайным процессом (стационарность делает возможным определение вероятностных характеристик по единственной реализации процесса — за счет усреднения каких-то статистик по времени). Рассматривая более или менее произвольный чертеж, такое разделение можно с той или иной степенью надежности проделать, но мы должны насторожиться при мысли о том, что оно далеко не однозначно (в случае курсов акций).

Например, рассматривая акции первой компании, можно сказать, что до конца 1993 года тренд был близок к нулю, а потом вдруг стал очень даже положительным, так что акции в конце концов выросли в 2.5 раза. Но можно сказать и так, что до конца 1991 года тренд был отрицательным (акции упали в 1.5 раза), а потом стал положительным, да так, что акции выросли в 4 раза. Для акций второй компании можно выделить один общий тренд за все 8 лет или два разных тренда (один с 1989 по конец 1991 года, другой, более крутой, с начала 1992 года до начала 1997 года), а при желании — три тренда или более. Если же позво-

лить выделять тренды не только в виде монотонно растущих или убывающих функций, но также в виде тех или иных периодичностей (гармоник), то имя всем этим трендам будет — легион. При любом выделении тренда можно добиться, чтобы модель неплохо описывала имеющиеся данные наблюдений, но будет совершенно неизвестно, продолжится ли такой тренд хоть какое-нибудь время в будущем.

Наконец, существует еще третий способ выделения статистически стационарных явлений в нестационарных случайных процессах. Этот способ использован в применении к финансовым данным Башелье на рубеже 19-го и 20-го веков, а в применении к теории турбулентности Колмогоровым и его учениками в середине 20-го века. Он состоит в рассмотрении приращений процесса за не слишком большое время (отсюда возникло понятие случайного процесса со стационарными приращениями).

Пусть S_t — курс акций в момент t (в случае, когда речь идет о ценах закрытия, t дискретно и обозначает номер дня торгов). Если следовать подходу Башелье, то нужно рассмотреть разности (приращения)

$$\Delta_h S_t = S_{t+h} - S_t. \tag{1.1.1}$$

Если законы распределения таких разностей оказываются не зависящими от t, то процесс S_t называется процессом со стационарными приращениями. Впрочем, для финансовых данных считается полезным прибегать к логарифмированию

$$x_t = \ln S_t, \tag{1.1.2}$$

так что рассматриваются приращения логарифма

$$\Delta_h x_t = \ln S_{t+h} - \ln S_t = \ln(S_{t+h}/S_t). \tag{1.1.3}$$

З а м е ч а н и е. Если $S_t = 100$ долларов, то что такое $\ln S_t$? Иначе говоря, в какую степень нужно возвести число e, чтобы возникли 100 долларов? Ответ на этот бессмысленный вопрос состоит в том, что любые финансовые данные — это данные об отношении, в котором обмениваются друг на друга единицы двух различных "денег", например акции и доллары. Иначе говоря, данные о курсах безразмерны (потому на рис. 1.1.6а) шкала ординат оцифрована в безразмерных единицах, а не в долларах).

Переход к приращениям в случае теории турбулентности приводит к некоторым достаточно интересным и даже удивительным закономерностям в смысле корреляционных и спектральных свойств этих приращений (речь идет, например, о приращениях проекции скорости турбулентного потока на какую-то из осей координат). Но для финансовых данных — как в случае Башелье (1.1.1), так и в случае более современного подхода (1.1.2), (1.1.3) — чрезвычайно трудно пойти дальше некоторой тривиальности, суть которой станет ясной из рассмотрения рисунков 1.1.6b) и 1.1.6c).

На этих рисунках представлены при $h = 1$ день приращения логарифмов цен акций тех же двух компаний, что и на рис. 1.1.6a). (Обратите внимание на то, что чертежи выполнены в разном масштабе по оси ординат). Мы видим типичную картину белого шума, т.е. последовательности независимых случайных величин. Получается, что разности

$$\delta_t = \Delta_1 \ln S_t = \ln(S_{t+1}/S_t) \tag{1.1.4}$$

похожи (при различных t) на независимые случайные величины.

В каком смысле трудно пойти дальше этой довольно тривиальной модели? Теоретически предложить какие-либо модели с зависимыми величинами δ_t, разумеется, вполне возможно. Но нужно на конкретном материале доказать пользу этих теоретически мыслимых моделей для лучшего понимания фактических данных (а желательно также — и для каких-то практических целей). И вот это оказывается очень трудным, как мы частично увидим ниже.

Допустив, что $\delta_t = \ln(S_{t+1}/S_t)$ — независимые случайные величины, мы получим, что предположение стационарности по t сведется просто к тому, что распределение δ_t при всех t одинаково. В таком случае

$$E\delta_t = a,\ D\delta_t = E(\delta_t - a)^2 = \sigma^2$$

не зависят от t. Для приращения

$$\Delta_h x_t = \ln S_{t+h} - \ln S_t = \delta_t + \cdots + \delta_{t+h-1}$$

получаем

$$E\Delta_h x_t = ah,\ D\Delta_h x_t = h\sigma^2. \tag{1.1.5}$$

Пока что единицей времени у нас являлся один (торговый) день, t и h принимали целые значения. Но можно время выражать в других единицах, например в годах, и тогда t и h будут

меняться на решетке с шагом $1/250$. Психологически естественно перейти в таком случае к модели с непрерывным временем: считать, что S_t и $x_t = \ln S_t$ определены для непрерывного времени, причем x_t является процессом с независимыми приращениями, а равенства (1.1.5) сохраняются. Простейшим из таких процессов является процесс броуновского движения с коэффициентом сноса a и коэффициентом диффузии σ^2. Вот мы и пришли к знаменитой модели геометрического (или экономического) броуновского движения, которую можно записать в следующем виде:

$$\ln S_t - \ln S_0 = at + \sigma w(t),$$

где $w(t)$ — винеровский процесс.

З а м е ч а н и е. Формулы (1.1.5) считаются справедливыми для не слишком больших значений h. В какой именно области значений h справедлив линейный рост дисперсии приращения, следует выяснять по фактическим данным.

1.4. Общая (абстрактная) статистическая модель

Ради единообразия будем далее говорить, что имеющийся в нашем распоряжении статистический материал образует одно наблюдение X. В примере из раздела 1.1 X — это двести чисел из табл. 1.1.1, что можно записать как $X = (x_1, \ldots, x_n)$, где $n = 200$. В первом примере из раздела 1.2 в качестве единого наблюдения выступают $n = 24$ пары (x_i, y_i) — удаления и лучевые скорости. Во втором примере из раздела 1.2 — это данные об урожайности — X есть последовательность чисел x_t, где t (календарная дата) изменяется от $T_1 = 1945$ до $T_2 = 1989$, пробегая 45 лет. В примере из раздела 1.3 наблюдение X — это последовательность цен акций данной компании за период наблюдения. На рис. 1.1.6а) изображены две такие последовательности (две траектории), т.е. имеются два наблюдения.

Формирование общей статистической модели начнем с того, что скажем, что мы располагаем наблюдением X. Это наш *статистический материал*. Все выводы мы будем делать, основываясь на наблюдении X. Его математическая природа не существенна: 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.

Во всех рассмотренных примерах в качестве выборочного пространства $\mathcal{X}$ можно взять n-мерное арифметическое пространство R^n (с разными n), хотя в некоторых случаях можно было бы ограничиться определенными подмножествами R^n.

Ключевое предположение состоит в том, что данное значение X появилось как результат некоего *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.

В наших примерах P — это распределение вероятностей на R^n. Для модели выборки, когда наблюденные значения рассматриваются как реализации независимых одинаково распределенных случайных величин, вероятностная мера P — это произведение n одномерных одинаковых распределений. Множество $\mathcal{P}$ — это совокупность таких n-мерных распределений. За счет дальнейших предположений об одномерных распределениях эта совокупность может быть сделана более узкой. Если, например, мы предположим, что упомянутое одномерное распределение — гауссовское (с неопределенными параметрами), то в качестве $\mathcal{P}$ мы получим двухпараметрическое семейство n-мерных гауссовских распределений.

В регрессионных моделях наблюденные отклики $y_1, \ldots, y_n$ тоже рассматривают как реализации независимых случайных переменных, так что P — снова произведение мер. Но здесь одномерные распределения не одинаковы: каждое из них зависит от соответствующего значения фактора x (и от распределения случайных ошибок, если последние не предполагаются одинаково распределенными).

Наконец, в примере из раздела 1.3 цены акций в последовательные моменты времени рассматриваются как случайные величины, не являющиеся независимыми. Поэтому распределение P в этом случае устроено более сложно, чем произведение мер. Исследованием таких моделей занимается статистика временных рядов. Мы

же в курсе будем заниматься лишь *статистикой независимых наблюдений*.

Задача статистики — выводы о P или свойствах P на основании X. Например, основываясь на X, надо 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 (если θ — одномерный параметр) или область конечномерного пространства (когда θ — многомерный параметр). В параметрическом случае:

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

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

§ 2. Теорема Гливенко

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

Пусть $x_1, x_2, \ldots, x_n$ — независимые, одинаково распределенные случайные величины (выборка). Их (общую) функцию распределения обозначим через $F(x)$:

$$F(x) = P\{x_i \le x\}.$$

Обозначим через $F_n(x)$ так называемую *эмпирическую* функцию распределения, которая строится по выборке. Для этого в каждую из точек $x_1, x_2, \ldots, x_n$ поместим вероятность, равную $\frac{1}{n}$. Так на числовой прямой возникает новое распределение вероятностей, случайное. Его функцию распределения и обозначим через $F_n(x)$. Поэтому $F_n(x)$ называют еще и *функцией распределения выборки*. С помощью индикаторов событий $I(x_i \le x)$, функцию $F_n(x)$ можно записать в виде:

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

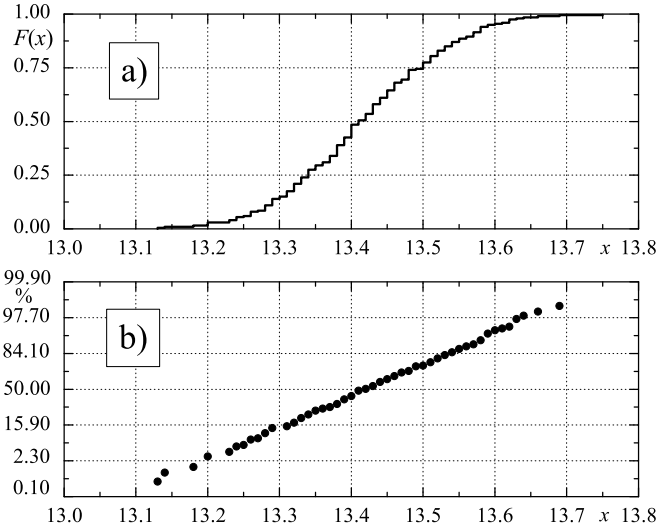


Рис. 1.2.1. a) Эмпирическая функция распределения для данных из табл. 1.1.1; b) Её изображение на нормальной бумаге

З а м е ч а н и е. 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.

Т е о р е м а Г л и в е н к о. *Последовательность случайных величин* $D_n = \sup\limits_x |F_n(x) - F(x)|$ $(n = 1, 2, \ldots)$ *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*

$$P\{\sup_x |F_n(x) - F(x)| < \varepsilon\} > 1 - \delta \quad \textit{для всех} \quad n \geq N.$$

Д о к а з а т е л ь с т в о. Предварительные замечания: для всякого x

$$F_n(x) \xrightarrow{P} F(x).$$

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

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

Пусть $\varepsilon > 0$ и $\delta > 0$ заданы. Выберем натуральное число R так, чтобы $1/R < \varepsilon/2$. 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 $-\infty = a_0 < a_1 < a_2 < \ldots < a_R = \infty$, где $\Delta_k = [a_{k-1}, a_k]$, $F(a_k) = k/R$, $k = 1, \ldots, R$.

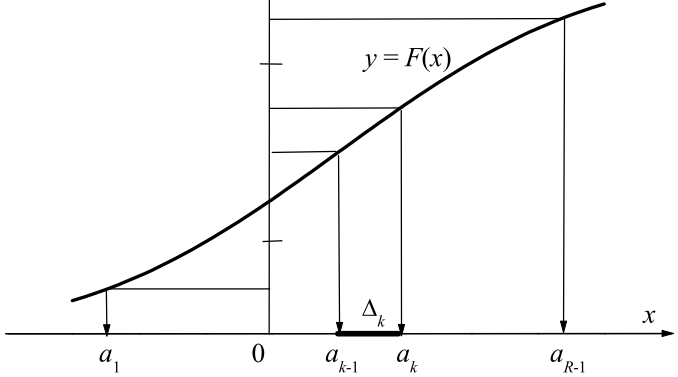


Рис. 1.2.2. График функции $y = F(x)$

Введем событие

$$\Omega_n = \left\{ \max_{1 \leq k \leq R-1} |F_n(a_k) - F(a_k)| < \frac{\varepsilon}{2} \right\}.$$

По уже упомянутой теореме Бернулли, существует $N = N(\varepsilon, \delta)$ такое, что $P\{\Omega_n\} > 1 - \delta$ для всех $n \geq N(\varepsilon, \delta)$. (Другими словами: следствием сходимости в каждой точке является равномерная сходимость на каждом конечном множестве точек).

Теперь покажем, что если произошло событие Ω_n, то при указанном выборе 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 $k = \overline{1, R}$

$$\sup_{x\in\Delta_k} |F_n(x) - F(x)| < \varepsilon. \tag{1.2.1}$$

Поскольку для любой функции $f(\cdot)$

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

для доказательства (1.2.1) 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.

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

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

$$= [F_n(a_k) - F(a_k)] + \frac{1}{R}.$$

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

$$F_n(x) - F(x) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

причем это верно для каждого отрезка Δ_k. Следовательно,

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

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

Для функций с разрывами то же доказательство проходит с некоторыми изменениями. Взамен последовательности $(a_0, a_1, \ldots, a_R)$, рассмотрим конечную 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 $\varepsilon/2$:

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

(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 $P\{x_i \le x\}$ и для $P\{x_i < x\}$).

Как можно построить такую последовательность, показано на рис. 1.2.3.

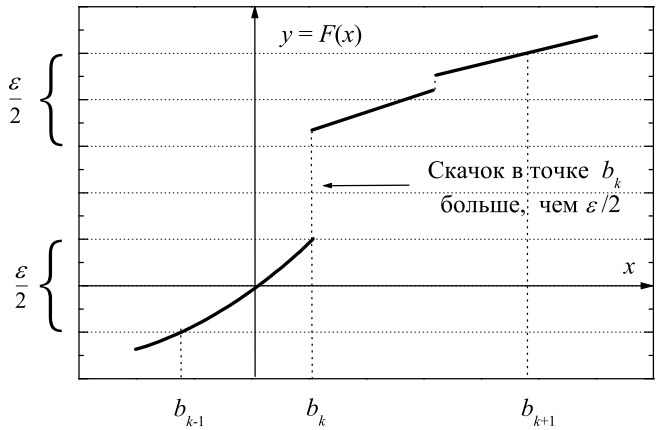


Рис. 1.2.3. Фрагмент графика функции $y = F(x)$. Отмечены несколько точек разбиения b_{k-1}, b_k, b_{k+1}.

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

Событие Ω_n, которое ранее было связано с последовательностью $(a_1, a_2, \ldots, a_{R-1})$, теперь определим так:

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

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

$$P\{\Omega_n\} > 1 - \delta.$$

С этим изменением доказательство проходит так же, как и раньше. (Для крайних отрезков Δ_1 и Δ_R в выкладки входят значения функций $F_n(.)$ и $F(.)$ в точках $a_0 = -\infty$ и $a_R = \infty$. Для этих значений аргумента формально полагаем $F_n(-\infty) = F(-\infty) = 0$, $F_n(\infty) = F(\infty) = 1$). $\square$

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 $\xi_n \to \xi$ (с вероятностью 1 ли, по вероятности ли), 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 в курсе мы будем рассматривать только "слабые" предельные теоремы, утверждающие сходимость по вероятности, даже если известны их усиленные варианты.

§ 3. Глазомерная проверка предположений о типе распределения

Теорема Гливенко позволяет по выборочной функции распределения (объекту наблюдаемому) судить о функции распределения случайных величин, которые образуют эту выборку. Эту истинную функцию распределения иногда — в противоположность выборочной — называют *теоретической* (хотя часто никакой теории за ней не стоит). Так в примере из 1.1, если размеры закле-

пок составляют выборку из нормальной совокупности, то выборочная функция распределения $y = F_n(x)$ (здесь $n = 200$) должна быть близка к некоторой нормальной функции распределения $y = \Phi\Big(\dfrac{x-a}{\sigma}\Big)$, где a и σ — какие-то параметры нормального закона, неизвестные наблюдателю (a — математическое ожидание, σ — стандартное отклонение). К сожалению, невозможно сказать, так ли это, просто глядя на график $y = F_n(x)$ (см. рис. 1.2.1a)). Для статистических проверок обсуждаемого предположения есть различные точные методы, но о них лучше говорить не сейчас. А сейчас стоит познакомиться с хорошим глазомерным способом для проверки нормальности. Для этого надо изобразить зависимости

$$y = F_n(x) \quad \text{и} \quad y = \Phi\Big(\frac{x-a}{\sigma}\Big)$$

в иной системе координат. Взамен переменной y введем новую переменную z, положив $z = \Phi^{-1}(y)$. Здесь $\Phi^{-1}(.)$ обозначает функцию, обратную функции Лапласа $\Phi(.)$. (Так как $\Phi(.)$ — монотонно возрастающая функция, обратная к ней существует). Её также называют *функцией квантилей* стандартного нормального распределения. Её таблицы есть, как правило, во всяком сборнике статистических таблиц. В новой системе координат функция $y = \Phi\Big(\dfrac{x-a}{\sigma}\Big)$ переходит в линейную функцию $z = \dfrac{x-a}{\sigma}$, а её график — в прямую линию. При этом всякое отступление распределения $y = F(x)$ от нормального порождает отклонение графика $z = \Phi^{-1}(F(x))$ от прямой линии. Глаз немедленно эти отклонения замечает. Поскольку $F_n(x) \approx \Phi\Big(\dfrac{x-a}{\sigma}\Big)$, график функции $y = F_n(x)$ на плоскости (x, z), оставаясь ступенчатой ломаной линией, должен подходить близко к этой прямой. (См. график на рис. 1.2.1b). Этот график подтверждает, что числа таблицы 1.1.1 можно считать нормальной, или гауссовской выборкой).

Описанный метод, который позволяет сопоставлять распределения с нормальными с помощью нормальной вероятностной бумаги, может быть полезен и в других обстоятельствах. Применим его, чтобы составить представление о характере сходимости биномиального распределения к нормальному.

Н о р м а л ь н а я а п п р о к с и м а ц и я для биномиальных распределений. На рис. 1.3.1 изображены функции распределения для случайного числа успехов в n испытаниях Бернулли.

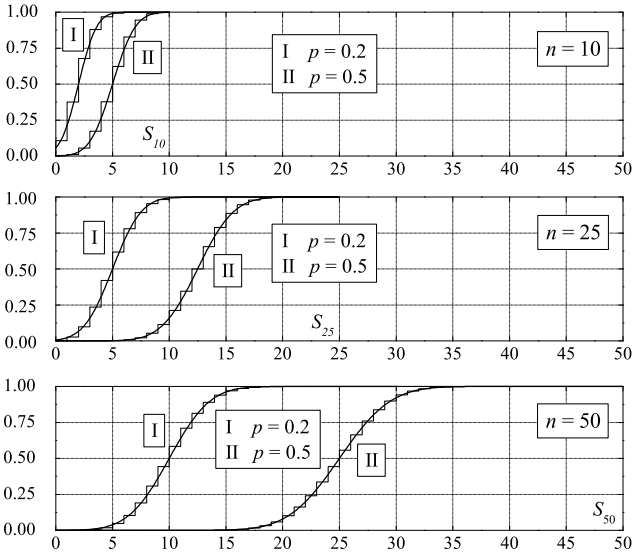


Рис. 1.3.1. Функции распределения для случайного числа успехов в n испытаниях Бернулли при разных n

Число испытаний n принимает значения $n = 10, 25, 50$; для вероятности успеха рассмотрены две возможности: $p = 0,2$ и $p = 0,5$. Согласно теореме Муавра-Лапласа, $P\{S_n < y\} \approx \Phi\Big(\dfrac{y - np}{\sqrt{npq}}\Big)$, и точность приближения увеличивается с ростом n. Видно, что при увеличении n графики функций распределения постепенно приобретают форму, напоминающую функцию распределения нормального закона — с той оговоркой, что функции биномиальных распределений — ступенчатые, а нормальные функции непрерывны. Впрочем, на-глаз трудно судить, велико ли это сходство.

При построении тех же графиков на вероятностной бумаге (рис. 1.3.2) сходство с нормальным распределением становится явным — эти графики "выпрямляются".

Столь же ясным становится и характер отступлений от нормального закона, связанный с дискретностью биномиальных распределений. Ясно видно, например, что для целых y величину

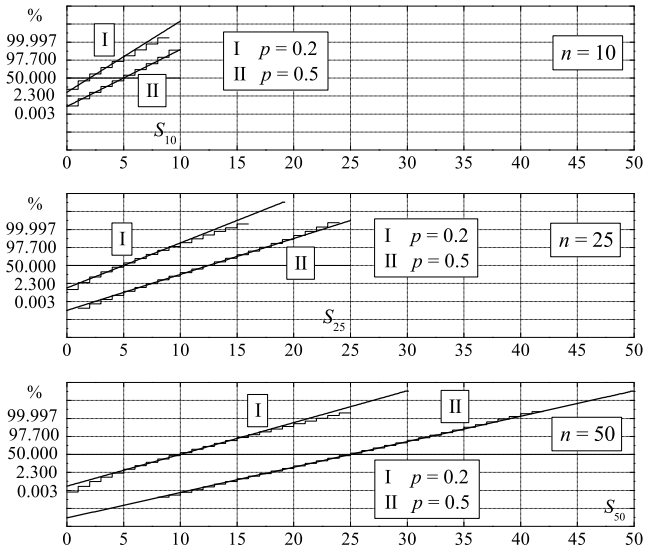


Рис. 1.3.2. Функции распределения для числа успехов в n испытаниях Бернулли на нормальной вероятностной бумаге

$P\{S_n \le y\}$ следует приближать не с помощью $\Phi\Big(\dfrac{y-np}{\sqrt{npq}}\Big)$, а с помощью $\Phi\Big(\dfrac{y+0.5-np}{\sqrt{npq}}\Big)$, ибо разность $\Big|P\{S_n \le y\} - \Phi\Big(\dfrac{y+0.5-np}{\sqrt{npq}}\Big)\Big|$ много меньше, чем $\Big|P\{S_n \le y\} - \Phi\Big(\dfrac{y-np}{\sqrt{npq}}\Big)\Big|$. Последняя величина, в силу локальной предельной теоремы Муавра-Лапласа, имеет порядок $O\left(\dfrac{1}{\sqrt{n}}\right)$, т.е. убывает весьма медленно при увеличении n. Переход от аргумента y к $y+0.5$ при вычислении $P\{S_n \le y\}$ с помощью нормального приближения для случая $p=q=0.5$ улучшает скорость сходимости до $O\left(\dfrac{1}{n}\right)$. Эту поправку в аргументе называют поправкой на непрерывность. Заметим, что по тем же причинам для вычисления $P\{S_n \ge y\}$ надо использовать в качестве приближенного значения $1 - \Phi\Big(\dfrac{y-0.5-np}{\sqrt{npq}}\Big)$.

Прием, который выше был использован для нормального рас-

пределения, имеет весьма общий характер и подходит для всех т.н. масштабно-сдвиговых семейств распределений. Так называют распределения с функциями вида $G\Big(\dfrac{x-a}{\sigma}\Big)$, где $G(.)$ — некоторая заданная функция распределения ("стандартная"), а числа a и $\sigma > 0$ — параметры сдвига и масштаба. Для выявления сходства заданного распределения (например, выборочного) с каким-либо членом указанного масштабно-сдвигового семейства надо рассматривать графики функций $y = F_n(x)$ и $y = G\Big(\dfrac{x-a}{\sigma}\Big)$ на плоскости (x, z), где $z = G^{-1}(y)$.

Рассмотрим как пример показательное распределение. Это распределение сосредоточено на положительной полуоси, его функция $F(x) = 1 - \exp(-x/\theta)$ (для $x \geq 0$) содержит масштабный параметр $\theta > 0$. Описанный прием превращения зависимости в линейную в этом случае удобно применить не к функциям распределения, а к их дополнениям до единицы:

$$R(x) = 1 - F(x), \quad R_n(x) = 1 - F_n(x).$$

Функцию $R(x)$ называют *функцией дожития* (*survival function*). Если $\xi \geq 0$ — случайное время работы (жизни) изделия, то $R(x) = P\{\xi \geq x\}$ есть вероятность того, что изделие прослужит не меньшее время, чем $x > 0$. $R_n(x)$ — выборочная функция дожития. По теореме Гливенко функции $R(x)$ и $R_n(x)$ должны быть близки, когда n велико.

Для показательного распределения $R(x) = \exp(-x/\theta)$ для $x \geq 0$. Если ввести переменную $z = -\ln y$, то график функции $y = \exp(-x/\theta)$ превратится в прямую линию: $z = x/\theta$. Если выборка извлечена из показательного распределения, то график функции $z = -\ln R_n(x)$ тоже должен походить на прямую линию (проходящую через начало координат). Сходство (или несходство) графика с такой прямой при небольшом навыке легко определить на-глаз.

Лекция 2. Начала оценивания

§ 1. Абстрактная статистическая модель

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 практически важных случаях (не всех!) 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. Оценивание: постановка задачи

2.1. В течение нескольких лекций мы будем обсуждать задачу оценивания параметра θ и/или функций от θ. "Оценить" здесь означает "Указать приближенное значение, опираясь на наблюдение X". Решая эту задачу, не будем ограничивать себя единственно имеющимся наблюдением X. Примем во внимание, что в другом, но аналогичном опыте мы можем встретить иное значение X из $\mathcal{X}$. А потому будем искать правило $\delta(\cdot)$, по которому каждое возможное наблюдение X из $\mathcal{X}$ 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). Случайную величину $\delta(X)$ называют *оценкой* θ. Задача статистики: выбрать правило $\delta(\cdot)$ так, чтобы оценить θ как можно лучше (точнее). При этом придется дать понятию сходства (или отличия) какую-либо разумную количественную меру.

2.2. Обратимся к некоторым из рассмотренных ранее примеров и обсудим, какие задачи оценивания и какие оценки там возникают. Для данных из примера раздела 1.1 в качестве статистической модели мы приняли выборку из нормального распределения, возможно, с засорением. Параметры этого нормального закона остаются постоянными, пока оборудование работает исправно. При выходе этих параметров за технически разрешенные границы, производство надо остановить (и произвести его наладку). Для контроля за ходом любого массового производства (в котором сплошная проверка продукции невозможна) надо время от времени брать выборки продукции, по которым оценивать параметры распределения (в данном случае параметры нормального закона).

Итак, пусть $x_1, \ldots, x_n$ — выборка из $N(a, \sigma^2)$. Популярные оценки для a:

Среднее арифметическое: $\bar{x} = \dfrac{1}{n} \sum\limits_{i=1}^{n} x_i$.

Медиана (выборки): $\mathrm{med}(x_1, \ldots, x_n)$. По определению, *медианой* числовой совокупности, называют такое число (скажем, μ), которое делит эту совокупность пополам: число элементов, которые меньше μ, равно числу элементов, которые больше μ. Чтобы точнее (скорее, более операционно) определить медиану конечной совокупности $x_1, \ldots, x_n$, упорядочим ее элементы в порядке возрастания. Дадим элементам этой упорядоченной по-новому совокупности обозначения $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Величины $x_{(1)}, \ldots, x_{(n)}$ называют *порядковыми статистиками*, а всю их совокупность — *вариационным рядом*. Если исходную совокупность $x_1, \ldots, x_n$ составляли случайные величины, случайными величинами являются и порядковые статистики. Для нечетного $n = 2m - 1$ (m — натуральное число) $\mathrm{med}(x_1, \ldots, x_n) = x_{(m)}$. Для $n = 2m$ медианой можно назвать любое число из интервала $(x_{(m)}, x_{(m+1)})$. Для

определенности, в качестве медианы берут его середину:

$$\operatorname{med}(x_1, \ldots, x_n) = \frac{x_{(m)} + x_{(m+1)}}{2}.$$

Усеченное среднее можно определить с помощью порядковых статистик: для $k \leq \left[\frac{n}{2}\right]$

$$\bar{x}_{(k)} = \frac{x_{(k+1)} + \ldots + x_{(n-k)}}{n - 2k}.$$

При вычислении $\bar{x}_{(k)}$ из исходной выборки исключают k самых малых и k самых больших членов; по оставшейся совокупности вычисляют среднее арифметическое. К усеченному среднему как к оценке центра выборки из симметричного распределения прибегают, чтобы исключить влияние крайних элементов наблюденной совокупности. Для засоренной нормальной совокупности этот прием позволяет значительно снизить (исключить) влияние выбросов на величину оценки. Выбросы, если они велики, могут вызвать значительное отклонение среднего арифметического $\bar{x}$ от истинного значения a, интересующего статистика. Медиана, как оценка центра, тоже дает "устойчивую" по отношению к выбросам (robust) оценку центра.

И усеченное среднее, и медиана входят в так называемое семейство L-оценок: линейных комбинаций порядковых статистик вида $\sum\limits_{i=1}^{n} c_i x_{(i)}$, причем $\sum\limits_{i=1}^{n} c_i = 1$. (Слово "оценка" в названии этих статистик надо воспринимать с некоторой осторожностью, поскольку не всегда ясно, что именно оценивает данная L-оценка).

Классической оценкой для дисперсии σ^2 по выборке из $N(a, \sigma^2)$ служит

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Поскольку влияние выбросов на величину s^2 велико, оценивание σ^2 тоже нуждается в поправках в духе тех, что были сделаны выше в отношении $\bar{x}$ как оценки a. Сейчас мы о них говорить не будем.

Линейная регрессия вида

$$y_i = a + bx_i + \varepsilon_i, \quad i = \overline{1, n} \tag{2.2.1}$$

была принята в качестве статистической модели для данных Э. Хаббла и для изменения урожайности зерновых в СССР. (В последнем случае роль фактора x и его значений $x_1, x_2, \ldots$ играли календарные годы $t = 1945, 1946, \ldots$). Оценки для параметров a, b нужны, в частности, чтобы предсказывать будущие значения откликов при будущих значениях фактора. Особенно важен коэффициент наклона b: он показывает, каково влияние x на y.

Рациональные способы оценивания параметров линейной модели сводятся к минимизации (в том или ином смысле) совокупности невязок $y_i - a - bx_i,\ i = \overline{1,n}$. (В этой формуле a, b выступают в качестве переменных. Истинные (и неизвестные) значения этих параметров в формуле (2.2.1) надо было бы обозначить как-то иначе, например, как a_0, b_0 — но так обычно не делают, и мирятся с возникающей двусмысленностью).

Метод наименьших модулей:

$$\sum_{i=1}^{n} |y_i - a - bx_i| \to \min_{a,b}.$$

Метод наименьших квадратов:

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min_{a,b}.$$

Так называемое *M-оценивание*:

$$\sum_{i=1}^{n} \rho(y_i - a - bx_i) \to \min_{a,b}$$

обобщает предыдущие методы. Функцию $\rho(.)$ можно выбирать, руководствуясь различными соображениями. Например, стремясь уменьшить влияние выбросов.

Названные общие методы применимы и к выборке (к оцениванию центра). Метод наименьших квадратов:

$$\sum_{i=1}^{n} (x_i - a)^2 \to \min_{a}$$

дает $\hat{a} = \bar{x}$ (что легко проверить). Метод наименьших модулей:

$$\sum_{i=1}^{n} |x_i - a| \to \min_{a}$$

дает $\hat{a} = \mathrm{med}(x_1, \ldots, x_n)$, что тоже нетрудно подтвердить.

На случай линейной регрессии можно обобщить и $L-$ оценивание (сделав его рекуррентным).

2.3. Вернемся к общей задаче, поставленной в начале этого раздела: по наблюдению X, полученному путем случайного выбора, управляемого распределением вероятностей P_θ, где $\theta \in \Theta$, оценить параметр θ. Что означает: так выбрать измеримую функцию $\delta(.)$, определенную на $\mathcal{X}$, чтобы $\delta(X)$ и θ были близки.

Можно предложить очень много способов, измеряющих близость $\delta(X)$ и θ. Сложилась общая точка зрения: есть функция потерь $L(\theta, d) \geq 0$, 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 случайную величину $L(\theta, \delta(X))$. Например, как в рассмотренных примерах, может быть

$$L(\theta, \delta(X)) = |\theta - \delta(X)|$$

или

$$L(\theta, \delta(X)) = |\theta - \delta(X)|^2 \qquad \text{и т.д.}$$

В каждом отдельном опыте величина потерь случайна. В статистике принято характеризовать статистические правила средними результатами, достигаемыми при многократном применении. По закону больших чисел это:

$$E_\theta L(\theta, \delta(X)).$$

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

$$R(\theta, \delta) = E_\theta L(\theta, \delta(X)).$$

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

$$R(\theta, \delta_1) \leq R(\theta, \delta_2) \tag{2.2.2}$$

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.

О п р е д е л е н и е 2.2.1. Правило $\delta_1(\cdot)$ называют *допустимым*, если нет правила лучшего, чем $\delta_1(.)$, т.е. если не существует правила $\delta_2(\cdot)$, для которого выполняется (2.2.2).

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

О п р е д е л е н и е 2.2.2. *Байесовские правила* — это оптимальные правила в ситуации, когда неизвестный параметр θ получен путем случайного выбора.

В этом случае риск $R(\theta, \delta)$ естественно осреднить 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 случайный вектор называют *байесовским подходом* к статистике. Он имеет как горячих сторонников, так и противников. Мы не будем касаться его в лекциях.

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

О п р е д е л е н и е 2.2.3. Оценка $\delta(\cdot)$ параметра θ (либо функции $\tau(\theta)$) называется *несмещенной*, если $E_\theta \delta(X) = \theta$ (либо $E_\theta \delta(X) = \tau(\theta)$) для всех $\theta \in \Theta$.

Для важной с прикладной точки зрения линейной статистической модели удается найти наилучшие несмещенные 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 дальнейшем 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 — в основном — конечный результат: удается ли получить в конце-концов явные и разумные статистические правила.

Функция риска для несмещенных оценок и квадратичной функции потерь превращается в дисперсию (в матрицу ковариации в векторном случае): $R(\theta,\delta)=E_\theta(\delta(X)-\theta)^2=E_\theta(\delta(X)-E_\theta\delta(X))^2$. 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 каждой оценки (для многомерного параметра — матрицу ковариаций). 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).

§ 3. Неравенство Крамера-Рао для одномерного параметра. (Оно же — неравенство информации, неравенство 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 может меняться от задачи к задаче.

Пусть X — наблюдение (конечномерный вектор), распределение которого зависит от неизвестного параметра θ, причем $\theta\in\Theta\subset R^1$, где Θ — заданное открытое множество. Отдельно будем рассматривать две возможности:

(i) Случайное наблюдение X имеет плотность $p(x,\theta)$ относи-

тельно меры Лебега;

(ii) случайное наблюдение X имеет дискретное распределение; в этом случае $p(x,\theta)$ обозначает вероятность события $X = x$.

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

Пусть $T(X)$ — некоторая статистика, принимающая значения в R^1, для которой существуют математическое ожидание и дисперсия. Пусть $\tau(\theta) := E_\theta T(X)$. Следовательно, $T(X)$ есть несмещенная оценка для $\tau(\theta)$.

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:

$$I(\theta) := E_\theta \left[\frac{\partial}{\partial\theta} \ln p(X,\theta)\right]^2 = E_\theta \lambda^2(X,\theta) \ (= D_\theta \lambda(X,\theta)).$$

Условие

(d) $\qquad 0 < I(\theta) < \infty.$

Т е о р е м а 2.3.1 (неравенство Крамера-Рао). *В перечисленных условиях* (a)–(d)

$$D_\theta T(X) \geq \frac{[\tau'(\theta)]^2}{I(\theta)}. \tag{2.3.1}$$

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

$$D_\theta T(X) \geq \frac{1}{I(\theta)}.$$

Д о к а з а т е л ь с т в о.
1. Заметим, что $E_\theta \lambda(X, \theta) = 0$. Действительно, из (c) мы заключаем, что:

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

Отсюда следует, в частности, что $I(\theta) = D_\theta \lambda(X, \theta)$.
2. Аналогично, из второго равенства (c) мы получаем, что

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

$$= E_\theta [T(X) - \tau(\theta)]\lambda(X, \theta).$$

Последнее равенство — благодаря тому, что $E_\theta \lambda(X, \theta) = 0$.
3. Неравенство Коши (-Буняковского-Римана-Шварца и т.д):

$$(E\xi\eta)^2 \leq E\xi^2 E\eta^2$$

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

$$[\tau'(\theta)]^2 \leq I(\theta) D_\theta T(X).$$

Отсюда и следует указанное в теореме неравенство. □

З а м е ч а н и е 2.3.1. Пусть $X = (X_1, X_2, \ldots, X_n)$ — выборка. Можно говорить о количестве информации, заключенной 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

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

Д о к а з а т е л ь с т в о. Очевидно: правдоподобие θ равно $p\,(X,\theta) = \prod\limits_{i=1}^{n} f(X_i,\theta)$, если через $f(\cdot,\theta)$ обозначить плотность вероятностей отдельных $X_1,\ldots,X_n$.

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

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

Из сказанного можно вывести важное качественное следствие о возможной скорости уменьшения дисперсии несмещенной оценки при возрастании числа независимых наблюдений n: $D_\theta T(X) \geq C/n$, где $C = [i(\theta)]^{-1}$.

З а м е ч а н и е 2.3.2.

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

Этот результат получается простой выкладкой.

§ 4. Экспоненциальные семейства

Случай, когда неравенство Крамера-Рао (2.3.1) выполняется в виде равенства, заслуживает особого рассмотрения. При выводе (2.3.1) мы применили неравенство Коши:

$$(E\xi\eta)^2 \leq E\xi^2 E\eta^2, \tag{2.4.1}$$

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 (2.3.1) (и, следовательно, 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. предположение (а) из § 3) 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, $I_A(.)$ — индикаторная функция множества A. (Заметим, что представление плотности (2.4.4) в указанном виде не единственно).

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

Распределения совокупности n независимых реализаций $(X_1, X_2, \ldots, X_n)$ случайной величины, принадлежащей экспоненциальному семейству (2.4.4), в свою очередь, образуют экспоненциальное семейство с плотностью (вероятностью):

$$p\,(x_1, x_2, \ldots, x_n, \theta) =$$

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

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

Биномиальное распределение.

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

для $x=0,1,\ldots,n;\ \ 0<\theta<1$.

Для параметра θ есть эффективная оценка x/n.

Показательное распределение (и выборка из показательного распределения).

$$p(x,\theta)=\begin{cases}\dfrac{1}{\theta}\exp\left(-\dfrac{x}{\theta}\right), & \text{для } x\geq 0,\ \theta>0;\\ 0, & \text{для } x<0.\end{cases}$$

Для выборки

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

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 функций). Чтобы доказать это, допустим противоположное: для некоторого параметрического семейства есть два равенства вида (2.4.3):

$$T_i(X)=\tau_i(\theta)+a_i(\theta)\lambda(X,\theta),\ \ i=1,2.$$

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

$$T_1(X)=\tau_1(\theta)-\frac{a_1(\theta)}{a_2(\theta)}\tau_2(\theta)+\frac{a_1(\theta)}{a_2(\theta)}T_2(X). \tag{2.4.5}$$

Равенство (2.4.5) возможно, только если:

$$\frac{a_1(\theta)}{a_2(\theta)}=\text{Const},\quad \tau_1(\theta)-\frac{a_1(\theta)}{a_2(\theta)}\tau_2(\theta)=\text{Const}. \tag{2.4.6}$$

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

Из (2.4.6) следует, что все эффективные оценки линейно выражаются одна через другую (см. (2.4.5)), как и соответствующие функции $\tau(\theta)$.

§ 5. Статистические оценки для многомерных параметров

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: $\theta = (\theta_1, \ldots, \theta_r)^T$, $\theta \in \Theta \subset R^r$, 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.

О п р е д е л е н и е 2.5.1. *Случайная матрица* Z есть матрица, элементы z_{ij} которой суть случайные величины, заданные на общем пространстве элементарных исходов, т. е. имеющие совместное распределение вероятностей.

О п р е д е л е н и е 2.5.2. *Математическое ожидание случайной матрицы* $Z = \|z_{ij}\|$ есть

$$EZ = \|Ez_{ij}\|.$$

У т в е р ж д е н и е 2.5.1. Пусть Z — случайная матрица, а неслучайные (постоянные) матрицы A, B, C таковы, что матрица $AZB + C$ существует. (Размерности матриц A, B, Z и C согласованы в том смысле, что указанные действия осуществимы). Тогда:

$$E(AZB + C) = A(EZ)B + C.$$

В частности, если Y — случайный вектор, A — неслучайная матрица и b — неслучайный вектор, то

$$E(AY + b) = A(EY) + b,$$

когда указанные операции (умножения и сложения) осуществимы.

У т в е р ж д е н и е 2.5.2. Пусть Z_1 и Z_2 — две случайные матрицы, определенные на общем для них пространстве элементарных исходов. Пусть их размерности совпадают, так что матрица $Z_1 + Z_2$ существует. Тогда:

$$E(Z_1 + Z_2) = EZ_1 + EZ_2.$$

Утверждения 2.5.1 и 2.5.2 вместе показывают, что операция взятия математического ожидания для случайных матриц обладает привычными для этой операции для случайной величины 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).

О п р е д е л е н и е 5.2.3. *Ковариационная матрица* (она же — матрица ковариаций, дисперсионная матрица и т. п.) векторов X и Y есть

$$\mathrm{Cov}(X, Y) = E(X - EX)(Y - EY)^T.$$

Если $X = (x_1, x_2, \ldots)^T$, $Y = (y_1, y_2, \ldots)^T$, то элемент (i, j) матрицы $\mathrm{Cov}(X, Y)$ есть ковариация случайных величин x_i и y_j:

$$E(x_i - Ex_i)(y_j - Ey_j).$$

Ясно, что:

$$\mathrm{Cov}(X, Y) = EXY^T - (EX)(EY)^T.$$

О п р е д е л е н и е 2.5.4. *Ковариационная матрица случайного вектора* X определяется как:

$$\mathrm{Cov}(X, X) = E(X - EX)(X - EX)^T = EXX^T - (EX)(EX)^T.$$

Диагональные элементы этой матрицы — суть дисперсии случайных величин $x_1, x_2, \ldots$. Обозначение $\mathrm{Cov}(X, X)$ мы будем заменять коротким DX.

У т в е р ж д е н и е 2.5.3. Пусть X — случайный вектор, A — неслучайная (постоянная) матрица, b — неслучайный (постоянный) вектор. Тогда:

$$D(AX + b) = A(DX)A^T,$$

если $AX + b$ существует (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.

У т в е р ж д е н и е 2.5.4.

$$D(a^T X) = a^T (DX)a.$$

5.2. Многомерное неравенство Крамера-Рао

Вернемся к поставленной в начале этого параграфа задаче. Пусть $\varphi(\cdot)$ — некоторая вектор-функция, $\varphi(X)$ — оценка $\tau(\theta)$ (это векторы-столбцы), и пусть $E_\theta \varphi(X) = \tau(\theta)$, где $\tau(\theta) = (\tau_1(\theta), \tau_2(\theta), \ldots, \tau_d(\theta))^T$, $\theta \in \Theta \subset R^r$.

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 z. 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

$$E_\theta \xi = E_\theta \eta = t(\theta),$$

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 $D\xi \le D\eta$ или:

$$z^T [D_\theta \varphi(X)] z \le z^T [D_\theta \psi(X)] z. \tag{2.5.1}$$

Мы можем принять такое о п р е д е л е н и е: $\varphi(X)$ *лучше*, чем $\psi(X)$, если (2.5.1) выполняется для любого вектора $z \in R^d$ (и для некоторых z это неравенство строгое).

По отношению к переменному $z \in R^d$, $z^T[D_\theta\varphi(X)]z$ и $z^T[D_\theta\psi(X)]z$ представляют собой квадратичные формы (неотрицательно определенные). Неравенство (2.5.1), если оно выполняется для всех z, линейная алгебра истолковывает как соотношение между матрицами квадратичных форм. В данном случае, между матрицами ковариаций $D_\theta\varphi(X)$ и $D_\theta\psi(X)$: $D_\theta\varphi(X) \le D_\theta\psi(X)$.

Итак, мы пришли к заключению, что *квадратичным риском* статистики $\varphi(X)$, несмещенно оценивающей $\tau(\theta)$, можно назвать ее матрицу ковариаций: $D_\theta\varphi = E_\theta[\varphi(X) - \tau(\theta)][\varphi(X) - \tau(\theta)]^T$.

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.

Переходим к выводу неравенства. Введем оператор частного дифференцирования по θ, который — в виде исключения! — запишем как строку:

$$\frac{\partial}{\partial\theta} = \left(\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_r}\right).$$

Определим матрицу информации (обобщение количества информации):

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

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

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

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

Покажем, что при принятых в § 3 "условиях регулярности", обобщенных на многомерный случай, справедливо неравенство

$$E_\theta(\varphi(X)-\tau(\theta))(\varphi(X)-\tau(\theta))^T \geq \left(\frac{\partial\tau}{\partial\theta}\right)[I^{-1}(\theta)]\left(\frac{\partial\tau}{\partial\theta}\right)^T. \quad (2.5.2)$$

Д о к а з а т е л ь с т в о. Рассмотрим вектор-строку:

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

Так же, как в пункте 1 из § 3, находим, что

$$E_\theta\lambda(X,\theta)=0. \quad (2.5.3)$$

Дифференцируем по θ тождество

$$\int\limits_A \varphi(x)p\,(x,\theta)\,dx=\tau(\theta),$$

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:

$$E_\theta\varphi(X)\lambda(X,\theta)=\frac{\partial\tau}{\partial\theta}. \quad (2.5.4)$$

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:

$$E_\theta\left[(\varphi-\tau)-\frac{\partial\tau}{\partial\theta}\,I^{-1}(\theta)\lambda^T\right]\left[(\varphi-\tau)-\frac{\partial\tau}{\partial\theta}\,I^{-1}(\theta)\lambda^T\right]^T\geq 0.$$

Левую часть преобразуем:

$$E_\theta(\varphi-\tau)(\varphi-\tau)^T - E_\theta(\varphi-\tau)\left[\frac{\partial\tau}{\partial\theta}\,I^{-1}(\theta)\lambda^T\right]^T -$$

$$-E_\theta\left[\frac{\partial\tau}{\partial\theta}\,I^{-1}(\theta)\lambda^T\right](\varphi-\tau)^T + \qquad (2.5.5)$$

$$+E_\theta\left[\frac{\partial\tau}{\partial\theta}\,I^{-1}(\theta)\lambda^T\right]\left[\frac{\partial\tau}{\partial\theta}\,I^{-1}(\theta)\lambda^T\right]^T \geq 0.$$

Второе слагаемое в (2.5.5):

$$-E_\theta(\varphi-\tau)\lambda\, I^{-1}(\theta)\left(\frac{\partial\tau}{\partial\theta}\right)^T = -\left(\frac{\partial\tau}{\partial\theta}\right)\, I^{-1}(\theta)\left(\frac{\partial\tau}{\partial\theta}\right)^T, \qquad (2.5.6)$$

ибо $E_\theta\varphi\lambda = \dfrac{\partial\tau}{\partial\theta}$ (см. (2.5.4)), $E_\theta\lambda = 0$ (см. (2.5.3)).

Третье слагаемое отличается от второго лишь транспонированием (третье слагаемое — это транспонированное второе). А так как (2.5.6) — симметрично, то третье слагаемое тоже равно (2.5.6).

Наконец, четвертое слагаемое даст:

$$\left(\frac{\partial\tau}{\partial\theta}\right)\, I^{-1}(\theta)\left[E_\theta\lambda^T\lambda\right]\, I^{-1}(\theta)\left(\frac{\partial\tau}{\partial\theta}\right)^T = \left(\frac{\partial\tau}{\partial\theta}\right)\, I^{-1}(\theta)\left(\frac{\partial\tau}{\partial\theta}\right)^T.$$

Приведя в (2.5.5) подобные члены, получим отсюда (2.5.2), что и требовалось. $\square$

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

$$p\,(x,\theta) = \exp\Big[\sum_{j=1}^{n} c_j(\theta)T_j(x) + d(\theta) + S(x)\Big]I_A(x).$$

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

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

Лекция 3. Достаточные статистики и наилучшие несмещенные оценки

§ 1. Условные распределения (элементарная теория)

В этом разделе нам придется оперировать понятиями условной вероятности, условного распределения, условного математического ожидания и т. д. одной случайной величины относительно другой. Чтобы не задерживать развития статистических идей, мы будем поначалу обходиться элементарными формами этих понятий. Они либо уже известны из курса теории вероятностей, либо могут быть получены элементарными средствами (например, предельным переходом). Основательную разработку общего понятия условного математического ожидания и условной вероятности мы на недолгое время отложим.

Начнем с дискретно распределенной случайной величины X. Пусть $P\{X = x\} = p(x)$ для x, принадлежащих носителю распределения (по условию это конечное или счетное множество). Пусть $T = T(X)$ — некоторая функция от X. Обозначим через $P_{X|T}(x, T)$ условную вероятность события $\{X = x\}$ при условии, что $T(X) = T$:

$$P_{X|T}(x,T) = P\{X = x | T(X) = T\} = \begin{cases} \dfrac{p(x)}{\sum\limits_{y:\ T(y)=T} p(y)}, & \text{если } T(x) = T \\ 0, & \text{если } T(x) \neq T. \end{cases}$$

Так определенные условные вероятности образуют условное распределение, ибо при всяком T

$$\sum_{x:\ T(x)=T} P_{X|T}(x,T) = 1.$$

Можно говорить об усреднении по этому распределению как самой случайной величины, так и функций от нее $f(X)$. Эти средние естественно назвать условными математическими ожиданиями X или $f(X)$ при данном T. Их обозначают, соответственно, через $E(X|T)$ и $E(f(X)|T)$:

$$E(X|T) = \sum_{x:\ T(x)=T} x\, P_{X|T}(x,T), \qquad E(f(X)|T) = \sum_{x:\ T(x)=T} f(x)\, P_{X|T}(x,T).$$

Заметим, что условные математические ожидания, так же как условные вероятности $P_{X|T}(x,T)$, являются функциями от случайной величины $T=T(X)$. Тем самым, эти объекты тоже являются *случайными величинами*.

П р и м е р. Пусть $X=(X_1,\dots,X_m)$, где $X_1,\dots,X_m$ суть независимые случайные величины, распределенные по Пуассону с параметрами $\lambda_1,\dots,\lambda_m$, соответственно. Найдем условное распределение X при данном значении $T=X_1+\ldots+X_m$. Для набора $x=(x_1,\dots,x_m)$ целых неотрицательных чисел таких, что $x_1+\ldots+x_m=T$, находим, что

$$P_{X|T}(x,T)=\frac{\prod\limits_{i=1}^{m}P\{X_i=x_i\}}{P\{X_1+\ldots+X_m=x_1+\ldots+x_m\}}.$$

Поскольку сумма независимых пуассоновски распределенных слагаемых $X_1,\dots,X_m$ тоже распределена по Пуассону, но с параметром $\lambda_1+\ldots+\lambda_m$, находим, что

$$P_{X|T}(x,T)=\frac{\lambda_1^{x_1}\dots\lambda_m^{x_m}}{x_1!\dots x_m!}\cdot\frac{(x_1+\ldots+x_m)!}{(\lambda_1+\ldots+\lambda_m)^{x_1+\ldots+x_m}}=$$

$$=\frac{(x_1+\ldots+x_m)!}{x_1!\dots x_m!}\,p_1^{x_1}\dots p_m^{x_m}=\frac{T!}{x_1!\dots x_m!}\,p_1^{x_1}\dots p_m^{x_m},$$

где $p_i=\lambda_i/(\lambda_1+\ldots+\lambda_m)$ для $i=1,\dots,m$. Последнее выражение — это полиномиальная вероятность, т. е. вероятность, что в T испытаниях Бернулли с m различными исходами, вероятности которых суть $p_1,\dots,p_m$, исходы с номерами $1,\dots,m$ произошли $x_1,\dots,x_m$ раз. Следовательно, условное распределение $(X_1,\dots,X_m)$ при данном $T=X_1+\ldots+X_m$ — полиномиальное.

Для непрерывно распределенной случайной величины X её условное распределение при данном значении $T=T(X)$ вводится сложнее. Причина та, что здесь типично, что $P\{T(X)=T\}=0$, равно как и $P\{X=x\}=0$ для всякого x. Поэтому для построения условного распределения в этом случае приходится прибегать к предельным переходам.

§ 2. Распределение вероятностей на поверхности

Предположим, что в пространстве R^d задана вероятностная мера $P(.)$. Предположим, что $P(.)$ имеет плотность, скажем,

$p(x)$, $x \in R^d$, относительно меры Лебега в R^d. Предположим далее, что в пространстве R^d задано гладкое многообразие H. Размерность H обозначим через $d-r$, где $0 < r < d$. Ради краткости и образности будем называть H поверхностью. Мы называем поверхность гладкой, если в каждой точке $x \in H$ существуют касательное и ортогональное к H пространства, которые обозначим через $T(x)$ и $N(x)$, соответственно. Заметим, что мера Лебега в R^d индуцирует на поверхности H некоторую меру, которую будем тоже называть мерой Лебега, или $(d-r)$-мерной лебеговской мерой на H, и обозначать через $s(A)$ — для любого измеримого $A \subset H$. Подобно этому, вероятностная мера $P(.)$ индуцирует на H некоторую вероятностную меру, скажем, $\pi(.)$, к определению которой мы и переходим. В обсуждаемом нами элементарном варианте определение $\pi(.)$ может быть достигнуто предельным переходом. Ради простоты в дальнейшем будем говорить об открытых множествах на H и для произвольного открытого A определим $\pi(A)$. Мы увидим, что мера $\pi(.)$ задается с помощью плотности относительно лебеговской меры $s(.)$, и что эта плотность как функция $x \in H$ лишь множителем (зависящим от H) отличается от $p(x)$.

Для множества $A \subset H$ определим полезное для дальнейшего понятие "поперечного ε-расширения (ε-раздутия)". Пусть $U_\varepsilon(x)$ обозначает r-мерный шар радиуса ε с центром в точке x, лежащий в пространстве $N(x)$.

Поперечным ε-раздутием множества A, $A \subset H$, назовем

$$A^\varepsilon = \bigcup_{x \in A} U_\varepsilon(x).$$

Далее положим по определению

$$\pi(A) = \lim_{\varepsilon \to 0} \frac{P\{A^\varepsilon\}}{P\{H^\varepsilon\}}, \tag{3.2.1}$$

где H^ε — аналогичным образом определенное поперечное ε-раздутие поверхности H. Займемся числителем стоящей в правой части (3.2.1) дроби. (Знаменатель рассматривается аналогично).

Разобьем множество A на непересекающиеся множества Δ_1, $\ldots$, Δ_N. Максимальный из диаметров этих множеств обозначим через $\delta(N)$. В дальнейшем $N \to \infty$ и $\delta(N) \to 0$. Заметим, что поперечные ε-раздутия $\Delta_1^\varepsilon, \ldots, \Delta_N^\varepsilon$ множеств $\Delta_1, \ldots, \Delta_N$ образуют разбиение множества A^ε. Вероятность $P\{A^\varepsilon\}$ представим в виде

интеграла Римана, а последний — как предел интегральных сумм:

$$P\{A^\varepsilon\} = \int\limits_{A^\varepsilon} p\,(x)\,dx = \lim_{\delta(N)\to 0} \sum_{i=1}^{N} p\,(x_i)\Lambda_d(\Delta_i^\varepsilon).$$

Здесь $x_i \in \Delta_i,\ i = \overline{1,N}$; $\Lambda_d(.)$ обозначает d-мерную меру Лебега. Заметим, что $\Lambda_d(\Delta_i^\varepsilon) = \Lambda_r(U_\varepsilon)s(\Delta_i)[1+\varepsilon\delta o(1)]$ при $\varepsilon \to 0,\ \delta \to 0$. Здесь U_ε — шар радиуса ε в пространстве R^r. Поэтому

$$\sum_{i=1}^{N} p\,(x_i)\Lambda_d(\Delta_i^\varepsilon) = \Lambda_r(U_\varepsilon) \sum_{i=1}^{N} p\,(x_i)s(\Delta_i)[1+\varepsilon\delta o(1)].$$

Последняя интегральная сумма при измельчении разбиения сходится к $\int\limits_A p\,(x)s(dx)$.

Аналогичные преобразования знаменателя (3.2.1) приводят к появлению и там множителя $\Lambda_r(U_\varepsilon)$ — интегральной суммы, распространенной по всей поверхности H (с теми оговорками, которые необходимы для толкования интеграла по, возможно, некомпактной (бесконечной) области). После сокращения общего для числителя и знаменателя множителя $\Lambda_r(U_\varepsilon)$ мы заключаем, что

$$\pi(A) = \int\limits_A \frac{p\,(x)}{\int\limits_H p\,(y)s(dy)}\, s(dx).$$

Таким образом, распределение $P\,(.)$ во всем пространстве R^d, имеющее плотность $p\,(.)$, индуцирует на поверхности H некоторое распределение $\pi(.)$, также имеющее плотность относительно лебеговской меры на поверхности H, и эта плотность равна

$$\frac{p\,(x)}{\int\limits_H p\,(y)s(dy)} \quad \text{для} \quad x \in H.$$

Доказательство было проведено для гладких поверхностей, но, разумеется остается справедливым и для кусочно-гладких H.

С п е ц и а л ь н ы й с л у ч а й: условное распределение. Предположим, что распределение $P\,(.)$ в пространстве R^d задано случайной величиной X, а поверхность H — это одно из множеств уровня случайной функции $T = T(X):\ H = \{x:\ T(x) = t\}$. В

этом случае распределение $\pi(.)$ на поверхности H, рассмотренное выше, называют *условным распределением* случайной переменной X при заданном значении случайной величины $T(X)$, а плотность этого распределения — *условной плотностью* X при заданном значении $T(X)$. Иногда для этой условной плотности принимают обозначение

$$p_{X|T}(x,T) = \frac{p\,(x)}{\int\limits_{T(y)=T} p\,(y)s(dy)} \text{ для } x \text{ таких, что } T(x)=T. \quad (3.2.2)$$

Главный качественный вывод: условная плотность пропорциональна исходной плотности.

П р и м е р. Пусть $X=(X_1,\ldots,X_m)$, где $X_1,\ldots,X_m$ суть независимые случайные величины, распределенные равномерно на отрезке $[0,a]$, где $a>0$. Найти условное распределение X при данном значении $T(X)=\max(X_1,\ldots,X_m)$.

По условию, m-мерная случайная величина X имеет равномерное распределение на m-мерном кубе

$$Q(a)=\{\ x=(x_1,\ldots,x_m):\ 0\le x_1\le a,\ldots,0\le x_m\le a\ \}.$$

Для $0\le T\le a$ множество уровня $T(X)=T$ — это та часть поверхности куба $Q(T)$, что лежит в положительном октанте. Мера $s(.)$ на этой поверхности — это обычная $(m-1)$-мерная мера Лебега. Согласно (3.2.2), условная плотность X при данном T постоянна на поверхности уровня $T(x_1,\ldots,x_m)=T$. Поэтому условное распределение X при данном T — равномерное (на указанной поверхности).

З а д а ч а. Найдите условное распределение случайной величины X_1 при заданном значении T.

§ 3. 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.

О п р е д е л е н и е 3.3.1. 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).

Дискретный случай. Когда распределение X дискретно, понятие условного распределения X вводится элементарно:

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

$$= \begin{cases} \dfrac{P_\theta\{X = x\}}{P_\theta\{T(X) = t\}}, & \text{если} \quad T(X) = t; \\ 0, & \text{если} \quad T(X) \neq t. \end{cases}$$

Пример: испытания Бернулли. Пусть $X = (X_1, X_2, \ldots, X_n)$ — 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 (X_i — число успехов в испытании с номером i), T — общее число успехов в n испытаниях. Элементарная выкладка показывает, что в этом примере (где $x = (x_1, x_2, \ldots, x_n)$ — заданная последовательность нулей и единиц):

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

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

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_θ обладает плотностью 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 лебеговской меры 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).

Д о с т а т о ч н ы е р а з б и е н и я. 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). Достаточная статистика $T = T(X)$ разбивает выборочное пространство $\mathcal{X}$ на множества уровня $\{x : T(x) = \mathrm{Const}\}$. Условные распределения X на элементах этих разбиений одинаковы для всех распределений θ, когда $\theta \in \Theta$.

П р и м е р. Пусть $X = (X_1, \ldots, X_n)$ — выборка из показательного распределения, где плотность отдельного наблюдения X_i равна:

$$f(u,\theta) = \begin{cases} \dfrac{1}{\theta}\exp\Big(-\dfrac{u}{\theta}\Big) & \text{для } \ u \geq 0; \\ 0 & \text{для } \ u < 0. \end{cases}$$

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:

$$\prod_{i=1}^{n} f(u_i,\theta) = \begin{cases} \Big(\dfrac{1}{\theta}\Big)^n \exp\left(-\dfrac{\sum\limits_{i=1}^{n} u_i}{\theta}\right) & \text{для } \ u_1 \geq 0, \ldots, u_n \geq 0; \\ 0 & \text{в противном случае.} \end{cases}$$

Условная плотность X при фиксированном T равна (в точке u

такой, что $\sum\limits_{i=1}^{n} u_i = T$, $u_1 \geq 0, \ldots, u_n \geq 0$):

$$\frac{\left(\frac{1}{\theta}\right)^n \exp\left(-\frac{\sum\limits_{i=1}^{n} u_i}{\theta}\right)}{\int\limits_{\{y:\,\sum y_i = T, y_i \geq 0\}} \left(\frac{1}{\theta}\right)^n \exp\left(-\frac{\sum\limits_{i=1}^{n} y_i}{\theta}\right) dy} =$$

$$= \frac{\left(\frac{1}{\theta}\right)^n \exp\left(-\frac{T}{\theta}\right)}{\left(\frac{1}{\theta}\right)^n \exp\left(-\frac{T}{\theta}\right) \int\limits_{\{y:\sum y_i = T, y_i \geq 0\}} dy} = \mathrm{Const}(T).$$

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 u. Это означает, что указанное условное распределение X — равномерное (на каждом множестве уровня).

Выкладки, которые мы проделали в двух рассмотренных примерах, по существу повторяются при доказательстве следующей теоремы.

Т е о р е м а (факторизации). *Статистика $T = T(X)$ достаточна для параметра θ, $\theta \in \Theta$ т. и т. т., когда существуют функции $g(t, \theta)$ и $h(x)$ такие, что*

$$p\,(x, \theta) = g(T(x), \theta) h(x) \tag{3.3.1}$$

при всех $\theta \in \Theta$.

З а м е ч а н и е. Величина $p\,(x, \theta)$ обозначает либо плотность наблюдения X в точке x, если модель непрерывна, либо вероятность точки x, если модель дискретна.

Доказательство проведем для дискретного случая. Для непрерывного случая оно, по существу, повторяется.

Д о к а з а т е л ь с т в о (дискретный случай).

($\Leftarrow$) Если выполнено (3.3.1), то $T = T(X)$ — достаточная статистика для θ.

Надо показать, что условное распределение X при данном значении $T(X)$ 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:

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

$$= \frac{P_\theta\{X = x\}}{P_\theta\{T(X) = t\}} = \frac{g(T(x),\theta)h(x)}{g(t,\theta) \sum\limits_{y:\, T(y)=t} h(y)} = \frac{h(x)}{\sum\limits_{T(x)=t} h(x)}.$$

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 (3.3.1) доказана.

($\Rightarrow$) Если T — достаточная статистика, то (3.3.1) выполнено.

Если T — достаточна для $\theta \in \Theta$, то для таких x, что $T(x) = t$ и для всех $\theta \in \Theta$:

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

— условная вероятность не зависит от θ, обозначим ее через $h(x)$. Подробнее:

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

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. $\square$

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

П р и м е р (применения теоремы факторизации для нахождения достаточной статистики).

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 конечномерным (принадлежит n-мерному пространству), $X = (X_1, \dots, X_n)^T$ — вектор-столбец. Его координаты считаем независимыми случайными величинами, распределенными по нормальному закону, причем $DX_i = \sigma^2$, $i = \overline{1,n}$. Значение σ^2 не известно. Относительно EX предположим, что EX, будучи неизвестным, принадлежит заданному линейному подпространству L, $L \subset R^n$. Если обозначить $EX = l$, $E(X - EX)(X - EX)^T = D_\theta X = \sigma^2 I$ (I — единичная матрица), то $X \sim N(l, \sigma^2 I)$, причем $l \in L$, L — задано.

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 $M \subset R^n$; $L^\perp$ обозначает ортогональное дополнение L до R^n, т.е. $R^n = L \oplus L^\perp$. Для доказательства достаточно указать правдоподобие (l, σ^2) и затем его преобразовать:

$$p\,(X,\theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - l_i)^2\right\} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}|X-l|^2\right\} =$$

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

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 правдоподобие (l, σ^2) равно

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

Мы видим, что правдоподобие зависит от статистик $\mathrm{proj}_L X$ и $|\mathrm{proj}_{L^\perp} X|$, 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).

§ 4. Линейная регрессия

Задача *линейной регрессии* — одна из частных форм задачи 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 $t_1, t_2, \ldots, t_n$ переменной t (называемой часто фактором), были произведены измерения переменной x (называемой откликом). Они дали значения $x_1, x_2, \ldots, x_n$. При этом $x_i = f(t_i) + \varepsilon_i$, где $\varepsilon_1, \ldots, \varepsilon_n$ — какие-то, сопровождающие измерения, ошибки. Основное предположение состоит в том, что мы считаем упомянутые $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ независимыми случайными величинами. Менее важные предположения: ε_i распределены одинаково и распределены по нормальному закону $N(0, \sigma^2)$. Предположение $E\varepsilon_i = 0$ 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 $\varphi_1, \varphi_2, \ldots$): существуют параметры $\theta_1, \ldots, \theta_m$ такие, что

$$f(t) = \theta_1\varphi_1(t) + \theta_2\varphi_2(t) + \cdots + \theta_m\varphi_m(t).$$

В этом случае вектор $X = (x_1, x_2, \ldots, x_n)^T$ представляется в виде линейной комбинации векторов:

$$\Phi_j = (\varphi_j(t_1), \varphi_j(t_2), \ldots, \varphi_j(t_n))^T,\ j = \overline{1, m}$$

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 EX, в данном случае порождено векторами $\Phi_1, \Phi_2, \ldots, \Phi_m$.

§ 5. Нормальная выборка

Рассмотрим выборку $x_1, x_2, \ldots, x_n$ из нормальной совокупности $N(a, \sigma^2)$, где параметры $a \in R^1$, $\sigma^2 \in (0, \infty)$ неизвестны. Теорема факторизации помогает найти достаточные статистики для (a, σ^2). Выпишем функцию правдоподобия этой модели (пользуясь независимостью гауссовских случайных величин $x_1, x_2, \ldots, x_n$) и преобразуем ее:

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\Big\{ -\frac{(x_i - a)^2}{2\sigma^2} \Big\} =$$

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

Поскольку правдоподобие зависит от переменных $x_1, x_2, \ldots, x_n$ 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 $\{T(X) = \mathrm{Const}\}$, 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 пара $(\bar{x}, s^2)$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Легко видеть, что $(\bar{x}, s^2)$ взаимно однозначно связана с $(\sum\limits_{i=1}^{n} x_i, \sum\limits_{i=1}^{n} x_i^2)$. О преимуществах, которые дает статистика $(\bar{x}, s^2)$ перед другими статистиками для (a, σ^2), мы подробнее будем говорить

позже. Сейчас отметим лишь то, что $\bar{x}$ и s^2 несмещенно оценивают a и σ^2:

$$E\bar{x} = a, \quad Es^2 = \sigma^2.$$

Важность этих свойств будет ясна уже на следующей лекции. Заметим, что эти соотношения справедливы для любой, не только гауссовской, выборки (если Dx_i^2 существуют).

Выборка из $N(a, \sigma^2)$ является частным случаем линейной модели. Рассмотрим вектор $X = (x_1, x_2, \ldots, x_n)^T$. 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 $DX = \sigma^2 I$. Таким образом, предпосылки линейной модели соблюдены. Достаточные статистики общей линейной модели в данном случае суть:

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

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

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:

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

Рассмотрим вектор

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

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

$$S_{n-2} = \Big\{y : \ y \in R^n, \ \sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} y_i^2 = 1\Big\}.$$

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

- вектор Y как случайный элемент не зависит от $\bar{x}, s^2$;
- (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 выборочный коэффициент асимметрии $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$ и выборочный коэффициент эксцесса $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4-3$ статистически независимы от $(\bar{x}, s^2)$. А их распределения не зависят от неизвестных параметров a, σ^2 и могут быть вычислены и табулированы.

Упомянутые статистики могут служить для проверки нормальности имеющейся выборки. Если знать, что данная выборка — нормальная, её можно исследовать очень детально. (Из дальнейших лекций будет видно, как). Для общей линейной гауссовской модели утверждение о равномерном распределении случайного вектора $(X-\mathrm{proj}_L X)/|\mathrm{proj}_{L^\perp} X|$ (на единичной сфере размерности $n-r$, где $r=\dim L$) и его статистической независимости от пары $(\mathrm{proj}_L X, |\mathrm{proj}_{L^\perp} X|^2)$ доказывается аналогично. Аналогично мы можем составить коэффициенты асимметрии и эксцесса, и тоже использовать их для проверки нормальности распределения X в линейной модели.

Лекция 4. Наилучшие несмещенные оценки

Под этим названием обычно разумеют 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 наилучшие 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 несмещенных оценках удается продвинуть (а часто и полностью решить), если для неизвестного параметра существует достаточная статистика. Несмещенное 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 лекции. А сейчас, чтобы завершить тему наилучшего несмещенного оценивания, мы ограничимся неформальным толкованием этого понятия. А также укажем некоторые его свойства, необходимые для упомянутой цели.

§ 1. Условные математические ожидания

Пусть случайные величины X и Y заданы на одном вероятностном пространстве. (Содержательно это означает, что значе-

ния переменных X, Y получены в одном эксперименте). Понятие условного математического ожидания X при данном значении Y — далее $E(X|Y)$ — можно ввести элементарными средствами, если при 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 $E(X|Y)$. Поскольку Y — случайная величина, $E(X|Y)$ тоже является случайной величиной.

Если совместное распределение (X, Y) имеет плотность $p(x, y)$ (либо дискретно), то формулу для $E(X|Y)$ 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

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

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: $E(X|Y)$ 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) $E(X_1 + X_2|Y) = E(X_1|Y) + E(X_2|Y)$.

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

2) $E(kX|Y) = kE(X|Y)$,
   где k — постоянный (неслучайный) множитель.

3) $E[f(Y)X|Y] = f(Y)E(X|Y)$, где $f(Y)$ — функция Y.
Это свойство тоже естественно, ибо при фиксированном значении Y случайная величина $f(Y)$ 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 $E|X|$ (в первом пункте должны существовать $E|X_1|$ и $E|X_2|$).

Наиболее важным является свойство

4) $E[E(X|Y)] = EX$.

§ 2. Улучшение несмещенных оценок

Вернемся к обсуждавшейся задаче о несмещенных оценках с минимальной дисперсией. В её 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.:

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

причем $E_\theta|d(X)|$ существует.

Пусть $T(X)$ — достаточная статистика для параметра θ. Рассмотрим условное математическое ожидание $d(X)$ при данном T:

$$\varphi(T) = E(d(X)|T).$$

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

Т е о р е м а (Blackwell-Rao, 1947-1949). *При указанных выше условиях*

(a) $E_\theta\varphi(T) = \tau(\theta)$,

(b) $D_\theta\varphi(T) \le D_\theta d(X)$.

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

Доказательство. Утверждение (a) выполняется в силу свойства условных математических ожиданий $EE(X|Y) = EX$:

$$E_\theta E[d(X)|T] = E_\theta d(X) = \tau(\theta).$$

Доказательство свойства (b) проведем сначала для одномерных φ, d и τ; многомерный случай рассмотрим ниже.

Одномерный случай.

$$D_\theta d(X) = E_\theta[d(X) - \tau(\theta)]^2 = E_\theta[(d(X) - \varphi(T)) + (\varphi(T) - \tau(\theta))]^2 =$$

$$E_\theta(d - \varphi)^2 + E_\theta(\varphi - \tau)^2 + 2E_\theta(d - \varphi)(\varphi - \tau) = E_\theta(d - \varphi)^2 + D_\theta\varphi,$$

поскольку

$$E_\theta(d-\varphi)(\varphi-\tau) = E_\theta E_\theta[(d-\varphi)(\varphi-\tau)|T] = E_\theta(\varphi - \tau)E_\theta[(d - \varphi)|T] = 0,$$

ибо $E_\theta[(d(X) - \varphi(T))|T] = E(d|T) - E(\varphi|T) = \varphi - \varphi = 0$.

(Последнее равенство — с вероятностью 1 для каждого распределения P_θ). Равенство в (b) достигается, если (и только если)

$$E_\theta[d(X) - \varphi(T)]^2 = 0 \qquad \text{при всех } \theta.$$

Это возможно, если (и только если) $d(X) = \varphi(T)$ с вероятностью 1 для всех P_θ распределений.

Многомерный случай. Пусть $d(X)$, $\tau(\theta)$ принимают значения в R^p, записываем их в виде столбцов, $D_\theta d < \infty$. Пусть $z \in R^p$, z — произвольный вектор. Рассмотрим скалярные величины:

$$\xi = \xi(X) := z^T d(X),$$

$$\eta = \eta(T) := E[\xi(X)|T] = z^T E[d(X)|T] = z^T\varphi(T),$$

$$t = t(\theta) := z^T\tau(\theta).$$

Ясно, что $E_\theta\xi(X) = t(\theta) = E_\theta\eta(T)$. По одномерной теореме Блеквелла-Рао

$$D_\theta\eta(T) \le D_\theta\xi(X).$$

Откуда

$$D_\theta(z^T\varphi) \le D_\theta(z^T d),$$

или

$$z^T(D_\theta\varphi)z \le z^T(D_\theta d)z,$$

или

$$D_\theta\varphi \le D_\theta d.$$

Равенство — если

$$P_\theta\{\eta(T) = \xi(X)\} = 1,$$

или

$$P_\theta\{z^T(\varphi(T) - d(X)) = 0\} = 1$$

для всех $\theta \in \Theta$ и для всех $z \in R^p$. $\square$

§ 3. 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.

О п р е д е л е н и е 4.3.1. Достаточная статистика $T = T(X)$ называется *полной*, если уравнение относительно функции $f(\cdot)$

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

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, $\theta \in \Theta$).

Т е о р е м а (Леман, Шеффе, 1955). *Если $T = T(X)$ — полная достаточная статистика и $\varphi = \varphi(T)$ — несмещенная оценка $\tau(\theta)$, $\theta \in \Theta$, то $\varphi(T(X))$ — наилучшая несмещенная оценка $\tau(\theta)$.*

Д о к а з а т е л ь с т в о. Достаточно доказать единственность такой оценки φ. Предположим, что существует другая (отличная от $\varphi(T)$) несмещенная оценка $\psi(T)$, так что

$$E_\theta\psi(T) = E_\theta\varphi(T) = \tau(\theta) \qquad \text{для всех } \theta \in \Theta.$$

В этом случае $E_\theta[\psi(T) - \varphi(T)] = 0$ для всех $\theta \in \Theta$. Поскольку статистика T — полная, отсюда следует, что

$$\psi(T) - \varphi(T) = 0 \ \text{ почти наверное, для всех } \ \theta \in \Theta.$$

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

П р и м е р 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 биноминальным:

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

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

$$E_\theta f(S_n) = 0. \tag{4.3.1}$$

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 (4.3.1) имеет вид:

$$\sum_{m=0}^{n} C_n^m f(m)\theta^m(1-\theta)^{n-m} = 0 \ \text{ для всех } \ \theta \in (0,1). \tag{4.3.2}$$

Введем переменную $z = \dfrac{\theta}{1-\theta}$. Очевидно, что $z \in (0,\infty)$ и пробегает это множество, когда θ пробегает множество (0, 1). Сократив (4.3.1) на множитель $(1-\theta)^n$, получаем для последовательности $f(0), f(1), \ldots, f(n)$, — т. е. для функции $f(\cdot)$ — уравнение:

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

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 (4.3.1) имеет лишь тривиальное решение, т. е. статистика S_n — полная.

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

Пример 2. Выборка из показательного распределения. Пусть $x_1, x_2, \ldots, x_n$ — выборка из распределения с плотностью

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

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:

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

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 т. н. "стандартное" гамма-распределение, с плотностью

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

где n может принимать натуральные значения).

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

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

или

$$\int_0^\infty f(x) \frac{1}{(n-1)!} \left(\frac{x}{\theta}\right)^{n-1} \frac{1}{\theta} e^{-\frac{x}{\theta}} dx = 0 \qquad \text{для } \theta > 0.$$

Введем новую переменную $t = \dfrac{1}{\theta}$. После сокращений получим уравнение

$$\int_0^\infty f(x)\, x^{n-1} e^{-tx}\, dx = 0 \qquad \text{для всех } \ t > 0.$$

Левая часть этого уравнения — это преобразование Лапласа функции $x^{n-1} f(x)$. Оно тождественно (относительно t) равно нулю только для $f(\cdot) = 0$. Отсюда следует, что статистика T — полная.

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

$$p\,(x, \theta) = \Big\{ \exp\Big[\sum_{j=1}^{k} c_j(\theta) T_j(x) + d(\theta) + S(x) \Big] \Big\} I_A(x). \qquad (4.3.3)$$

По теореме факторизации $T(X) = (T_1(X), T_2(X), \ldots, T_k(X))$ есть достаточная статистика для θ, $\theta \in \Theta$.

Т е о р е м а. *Если область значений векторной функции $(c_1(\theta), c_2(\theta), \ldots, c_k(\theta))$, которую она заполняет, когда θ пробегает параметрическое множество Θ, содержит какое-либо открытое множество, то статистика T — полная.* (*Семейство распределений с плотностями* (4.3.3) — *полное*).

Доказательство этой теоремы не приводим. Оно может быть основано на свойствах преобразований Лапласа (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 $(\bar{x}, s^2)$, достаточной для параметров нормального распределения $N(a, \sigma^2)$ в случае выборки из этого распределения.

П р и м е р 3. 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, σ^2).

Лекции 5-6. Условные математические ожидания и условная вероятность

§ 1. Определения и простейшие свойства

1.1. Напоминания: вероятностное пространство и случайные величины (А.Н. Колмогоров, 1933)

- *Вероятностной моделью*, или *вероятностным пространством* называют набор $(\Omega, \mathcal{A}, P)$.

  Ω — это множество точек ω; $\mathcal{A}$ — σ-алгебра подмножеств из Ω; P — вероятностная мера на $\mathcal{A}$.

- Множество Ω называют *пространством элементарных исходов* (или элементарных событий).

- Множества из $\mathcal{A}$ называют *исходами* или *событиями*.

- Множество $A \in \Omega$ 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

$$\{\omega : \xi(\omega) \in B\} \tag{5.1.1}$$

  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 (5.1.1), когда B пробегает множество $\mathcal{B}$.

1.2. Производная Радона-Никодима (1930)

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

$$\Omega = \bigcup_{i=1}^{\infty} A_i, \quad \text{причем} \quad \mu(A_i) < \infty, \ i = 1, 2, \ldots.$$

Т е о р е м а 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*

$$\lambda(A) = \int\limits_A f(\omega)\mu(d\omega)$$

для всякого $A \in \mathcal{F}$. 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:

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

1.3. 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 $E(X|Y)$.

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 $E(X|Y)$.

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

Представим X в виде

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

где $X^+ \geq 0$, $X^- \geq 0$. Определим $E(X^+|\mathcal{G})$ и $E(X^-|\mathcal{G})$, и затем положим по определению:

$$E(X|\mathcal{G}) = E(X^+|\mathcal{G}) - E(X^-|\mathcal{G}), \tag{5.1.2}$$

если хотя бы одно из этих условных математических ожиданий конечно. Таким образом, $E(X|\mathcal{G})$ может принимать значения $+\infty$ или $-\infty$. (Такую возможность имеет и EX при этом способе определения). Впрочем, можно ограничиться случаем, когда $E|X| < \infty$. В силу (5.1.2) надо определить $E(X|\mathcal{G})$ для $X \geq 0$.

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

$$Q(A) = \int\limits_A X(\omega)\, P(d\omega). \tag{5.1.3}$$

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

$$Q(A) = \int\limits_A f(\omega)\, P(d\omega). \tag{5.1.4}$$

Функцию $f(\omega)$ из (5.1.4) назовем *условным математическим ожиданием* X (здесь $X \geq 0$) *относительно* σ*-алгебры* $\mathcal{G}$, т. е.:

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

Определив $E(X^+|\mathcal{G})$ и $E(X^-|\mathcal{G})$, по формуле (5.1.2) определим $E(X|\mathcal{G})$ для произвольной случайной величины X.

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

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

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

1.4. Некоторые свойства $E(X|\mathcal{G})$

- $$\int\limits_A E(X|\mathcal{G})P\,(d\omega) = \int\limits_A XP\,(d\omega) \quad \text{для всякого} \quad A \in \mathcal{G}. \tag{5.1.5}$$

  Это свойство — всего лишь другая запись определения (5.1.4). Заметим различие между X и $E(X|\mathcal{G})$: случайная величина X, вообще говоря, не измерима относительно $\mathcal{G}$ (она измерима относительно более "богатой" σ-алгебры $\mathcal{A}$, $\mathcal{G} \subset \mathcal{A}$).

- $$EE(X|\mathcal{G}) = EX. \tag{5.1.6}$$

  Для доказательства надо положить $A = \Omega$ в (5.1.5). Тогда:

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

  что и требовалось.

- $$E(aX + bY|\mathcal{G}) = aE(X|\mathcal{G}) + bE(Y|\mathcal{G}) \tag{5.1.7}$$

  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 $A \in \mathcal{G}$

  $$\int\limits_A E(aX + bY|\mathcal{G})\,dP = \int\limits_A [aE(X|\mathcal{G}) + bE(Y|\mathcal{G})]\,dP \tag{5.1.8}$$

  и что $aE(X|\mathcal{G}) + bE(Y|\mathcal{G})$ измеримо относительно $\mathcal{G}$. Последнее, впрочем, очевидно. Преобразуем левую часть (5.1.8):

  $$\int\limits_A E(aX + bY|\mathcal{G})\,dP = \int\limits_A (aX + bY)\,dP =$$

$$a\int\limits_A E(X|\mathcal{G})dP + b\int\limits_A E(Y|\mathcal{G})\,dP = \int\limits_A [aE(X|\mathcal{G}) + bE(Y|\mathcal{G})]\,dP,$$

что и требовалось.

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

  В частности,

  $$E(X|\mathcal{A}_X) = X. \tag{5.1.9}$$

- Если X и Y независимы, то

  $$E(X|Y) = EX. \tag{5.1.10}$$

  Для доказательства достаточно проверить, что для любого $A \in \mathcal{A}_Y$:

  $$\int\limits_A E(X|Y)\,dP = \int\limits_A (EX)\,dP. \tag{5.1.11}$$

  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 (5.1.11), заметив предварительно, что правая часть (5.1.11) равна $(EX)P\{A\}$. Имеем:

  $$\int\limits_A E(X|Y)\,dP = \int\limits_A X\,dP = E(XI_A) = (EX)(EI_A) = (EX)P\{A\}$$

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

- У с л о в н ы е в е р о я т н о с т и.

  Как мы только что вспомнили,

  $$P\{A\} = EI_A.$$

  По аналогии с этим равенством, условную вероятность события A относительно σ-алгебры $\mathcal{G}$ определим как

  $$P\{A|\mathcal{G}\} = E(I_A|\mathcal{G}). \tag{5.1.12}$$

  Соответственно этому, условная вероятность события A относительно случайной величины Y (при данном Y) есть

  $$P\{A|Y\} := P\{A|\mathcal{A}_\mathcal{Y}\}. \tag{5.1.13).}$$

- 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

$$P_X\{B|\mathcal{G}\} := P\{X \in B|\mathcal{G}\}(\omega),\ B \in \mathcal{B}. \tag{5.1.14}$$

  Не следует забывать, что (5.1.14) — это случайная величина, определенная с точностью до множества меры нуль. Можно показать, что существует такой вариант её определения, что (5.1.14), как функция B, $B \in \mathcal{B}$ с вероятностью 1 образует на $\mathcal{B}$ (случайную) вероятностную меру. В этом случае

$$E(X|\mathcal{G})(\omega) = \int X(\omega')P_X(d\omega'|\mathcal{G})(\omega) \quad \text{п. н.} \tag{5.1.15}$$

  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).

§ 2. Простые случайные величины

В этом параграфе мы рассмотрим $E(X|Y)$ 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

$$Y = \sum_j y_j I(D_j), \tag{5.2.1}$$

где $I(D) = I_D(\omega)$ — индикаторная функция множества D. (По удобствам обозначения $I(D)$ предпочтительнее, 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 (5.2.1) образует разбиение пространства Ω: $D_j \bigcap D_i = \oslash$, 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 $Y = Y(\omega)$, $D_j = \{\omega : Y(\omega) = y_j\}$).

Далее мы будем рассматривать σ-алгебры, порожденные конечными (или счетными) разбиениями. Пусть $\mathcal{G}$ — такая σ-алгебра. Порождающее её разбиение обозначим, как и выше, через $D_1, D_2, \ldots$. Пусть X — простая случайная величина. Тогда для $E(X|\mathcal{G})$ можно дать элементарное определение.

Начнем с определения условной вероятности. Положим по определению для всякого $A \in \mathcal{A}$

$$P\{A|\mathcal{G}\} = P\{A|\mathcal{G}\}(\omega) = \sum_j P\{A|D_j\} I(D_j). \tag{5.2.2}$$

Ясно, что $P\{A|\mathcal{G}\}$ есть измеримая относительно $\mathcal{G}$ случайная величина. Главное свойство условной вероятности (5.2.2):

$$EP\{A|\mathcal{G}\} = P\{A\}. \tag{5.2.3}$$

Доказательство очевидно:

$$EP\{A|\mathcal{G}\} = \sum_j P\{A|D_j\} EI(D_j) = \sum_j P\{A|D_j\} P\{D_j\} = P\{A\}.$$

Пусть

$$X = \sum_i x_i I(A_i). \tag{5.2.4}$$

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

$$E(X|\mathcal{G}) = \sum_i x_i P\{A_i|\mathcal{G}\}. \tag{5.2.5}$$

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

$$EE(X|\mathcal{G}) = EX. \tag{5.2.6}$$

Доказательство (5.2.6) очевидно:

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

Покажем, что определение (5.2.5) совпадает с общим определением математического ожидания из параграфа 1. Для этого достаточно проверить, что для любого $B \in \mathcal{G}$:

$$\int\limits_B E(X|\mathcal{G})\,dP = \int\limits_B X\,dP. \tag{5.2.7}$$

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

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

Поэтому (5.2.7) достаточно доказать для множеств D_k, $k = 1, 2, \ldots$. Итак, положив $B = D_k$, преобразуем левую часть (5.2.7), используя (5.2.5) и (5.2.2):

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

$$= \sum_i x_i \sum_j P\{A_i|D_j\}EI(D_k)I(D_j) =$$

$$= \sum_i x_i P\{A_i|D_k\}P\{D_k\} = \sum_i x_i P\{A_i D_k\}.$$

Преобразование правой части (5.2.7) дает тот же результат:

$$\int\limits_{D_k} X\,dP = EI(D_k)\sum_i x_i I(A_i) = \sum_i x_i EI(A_i D_k) = \sum_i x_i P\{A_i D_k\},$$

что и требовалось.

О п р е д е л е н и е 5.2.1. Условное математическое ожидание как усреднение. *Усреднением* набора чисел $x_1, \ldots, x_n$ с весами $p_1 \geq 0, \ldots, p_n \geq 0$, $\sum\limits_{i=1}^{n} p_i = 1$ называют $\sum\limits_{i=1}^{n} x_i p_i$. (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 $E(X|\mathcal{G})$, суть усреднения значений X. Действительно,

$$E(X|\mathcal{G}) = \sum_i x_i P\{A_i|\mathcal{G}\} = \sum_i x_i \sum_j P\{A_i|D_j\} I(D_j) =$$

$$= \sum_j \Big[\sum_i x_i P\{A_i|D_j\} \Big] I(D_j).$$

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

$$y_j = \sum_i x_i P\{A_i|D_j\}.$$

Отметим, что $P\{A_i|D_j\} \geq 0$, 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. Некоторые дальнейшие свойства условных математических ожиданий

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

$$E[\varphi(Y)X|Y] \overset{\text{п.н.}}{=} \varphi(Y)E(X|Y). \tag{5.3.1}$$

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

$$E(XY|\mathcal{G}) \overset{\text{п.н.}}{=} YE(X|\mathcal{G}) \tag{5.3.2}$$

при условии, что эти математические ожидания существуют.

Доказательство этого равенства начнем с простых случайных величин.

3.1. Доказательство (5.3.2) для случая простых случайных величин

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

$$E(XY|\mathcal{G}) \overset{\text{п.н.}}{=} YE(X|\mathcal{G}). \tag{5.3.3}$$

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

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

Чтобы получить (5.3.3), достаточно показать, что

$$E\big[I(B)X|\mathcal{G}\big] = I(B)E(X|\mathcal{G}),$$

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

$$\int\limits_A I(B)E(X|\mathcal{G})\,dP = \int\limits_A I(B)X\,dP. \tag{5.3.4}$$

Преобразуя левую часть, докажем тем самым, (5.3.4):

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

что и требовалось. □

3.2. Общий случай

Пусть Y измерима относительно $\mathcal{G}$, $E|X| < \infty$, $E|Y| < \infty$, $E|XY| < \infty$. Тогда

$$E(XY|\mathcal{G}) \overset{\text{п.н.}}{=} YE(X|\mathcal{G}). \tag{5.3.5}$$

Доказательство. Основывается на пункте 3.1 и обобщенной теореме Лебега о мажорированной сходимости, которая будет дана в пункте 3.3.

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

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

в силу (5.3.3). По упомянутой теореме

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

Кроме того,

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

Это доказывает (5.3.5). □

С л е д с т в и е.

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

3.3. Лемма

Л е м м а 5.3.1. *Пусть* $|\alpha_n| \le \eta$, $E\eta < \infty$ *и* $\alpha_n \overset{\text{п.н}}{\longrightarrow} \alpha$ *при* $n \to \infty$. *Тогда*

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

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

Сравним с т е о р е м о й Л е б е г а (о мажорированной сходимости): *Пусть* $|\xi_n| \le \eta$, $E\eta < \infty$ *и* $\xi_n \overset{\text{п.н.}}{\longrightarrow} \xi$ *при* $n \to \infty$. *Тогда*

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

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

Д о к а з а т е л ь с т в о л е м м ы 5.3.1. Положим

$$\xi_n := \sup_{m:m\ge n} |\alpha_m - \alpha|.$$

Ясно, что $\xi_n \ge |\alpha_n - \alpha|$. Так как $\alpha_n \overset{\text{п.н.}}{\longrightarrow} \alpha$, то $\xi_n \downarrow 0$ п. н. Теперь

$$|E(\alpha_n|\mathcal{G}) - E(\alpha|\mathcal{G})| = |E[(\alpha_n - \alpha)|\mathcal{G}]| \le E(|\alpha_n - \alpha|\,\big|\,\mathcal{G}) \le E(\xi_n|\mathcal{G}). \quad (5.3.6)$$

Докажем, что $E(\xi_n|\mathcal{G}) \xrightarrow{\text{п.н.}} 0$.

Заметим, что $0 \le E(\xi_{n+1}|\mathcal{G}) \le E(\xi_n|\mathcal{G})$ п. н. Поэтому существует предел (почти наверное):

$$h := \lim_{n\to\infty} E(\xi_n|\mathcal{G}) \ge 0.$$

Далее: $0 \le \int\limits_\Omega h dP \le \int\limits_\Omega E(\xi_n|\mathcal{G})\, dP = \int\limits_\Omega \xi_n dP = E\xi_n \to 0$. Последнее заключение есть следствие цитированной теоремы Лебега, ибо

$$0 \le \xi_n \le 2\beta,\ E\beta < \infty,\ \xi_n \xrightarrow{\text{п.н.}} 0.$$

Получили, что $\int\limits_\Omega h\, dP = 0$. Т.к. $h \ge 0$, то $h = 0$ п. н. Следовательно:

$$E(\xi_n|\mathcal{G}) \xrightarrow{\text{п.н.}} 0.$$

Это и доказывает лемму. $\square$

3.4. σ-аддитивность условной вероятности $P\{A|\mathcal{G}\}$

Пусть $A = \sum\limits_i A_i$, причем $A_i \bigcap A_j = \oslash$, если $i \ne j$. Тогда

$$P\{A|\mathcal{G}\} \overset{\text{п.н.}}{=} \sum_i P\{A_i|\mathcal{G}\}. \tag{5.3.7}$$

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.

3.5. Условная дисперсия

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

$$D(X|\mathcal{G}) = E\{[X - E(X|\mathcal{G})]^2|\mathcal{G}\}. \tag{5.3.8}$$

З а д а ч а. Покажите, что

$$DX = ED(X|\mathcal{G}) + DE(X|\mathcal{G}) \tag{5.3.9}$$

(при условии, что DX существует).

3.6. Наилучший квадратичный прогноз

(Формулируется в виде з а д а ч и).

З а д а ч а. 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 $g(\xi)$:

$$E\big(\eta - f(\xi)\big)^2 \le E\big(\eta - g(\xi)\big)^2.$$

О т в е т: $f(\xi) = E(\eta|\xi)$.

З а д а ч а. Обобщите этот результат для случайных векторов $\eta \in R^p$.

§ 4. Пример вычисления $E(X|Y)$

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

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

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

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

Вычислим $E(\xi|\eta)$.

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. Её составляют (пересеченные с Ω) прямые произведения борелевских множеств, лежащих на прямой $x = y$, и прямой $\{(x, y) : x+y = 0\}$. Очевидно, что ξ не измерима относительно A_η и η не измерима относительно A_ξ.

По определению, $E(\xi|\eta)$ — такая измеримая относительно A_η функция $f(x, y)$, для которой

$$\int\limits_{(x,y)\in A} f(x, y)\, dP \;=\; \int\limits_{(x,y)\in A} x\, dP \qquad \text{для любого } \ A \in A_\eta. \tag{5.4.1}$$

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 (5.4.1) достаточно рассматривать только множества A вида:

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

где z — произвольно.

При таком выборе $f(x,y)$ и A условие (5.4.1) примет вид:

$$\int\limits_{\{x+y\le z,\,(x,y)\in\,\Omega\}} g\,(x+y)\;dP = \int\limits_{\{x+y\le z,\,(x,y)\in\,\Omega\}} x\;dx\,dy. \tag{5.4.2}$$

В интегралах (5.4.2) 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 (5.4.2) представим в виде повторных. Для простоты возьмем $z \in [0,1]$. (Случай $z \in [1,2]$ легко сводится к рассматриваемому). Получим уравнение для $g(\cdot)$:

$$\frac{1}{2}\int\limits_0^z \Big(g(u)\int\limits_{-u}^{u} dv\Big)du = \frac{1}{2}\int\limits_0^z \Big(\int\limits_{-u}^{u}\frac{u+v}{2}dv\Big)du. \tag{5.4.3}$$

Отсюда $\displaystyle\int\limits_0^z u\,g(u) = \int\limits_0^z \frac{1}{2}u^2 du$, или $g(z) = \dfrac{z}{2}$.

Таким образом, здесь $E(\xi|\eta) = \eta/2$, или

$$E(x|x+y) = \frac{x+y}{2}. \tag{5.4.4}$$

Заметим, что при вычислении $E(X|X+Y)$, если X и Y независимы и одинаково распределены (как в рассмотренном выше примере), можно обойтись практически без вычислений, если вспомнить некоторые из перечисленных ранее свойств условных математических ожиданий. Во-первых, в силу симметрии,

$$E(X|X+Y) = E(Y|X+Y).$$

Затем $X+Y = E(X+Y|X+Y) = E(X|X+Y) + E(Y|X+Y)$.

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

Лекция 7. Линейная гауссовская модель

В абстрактной форме эта статистическая модель о (векторном) наблюдении X, $X \in R^n$, X — вектор-столбец, $X = (X_1, X_2, \ldots, X_n)^T$.

Предположим, что X — случайный вектор, распределенный по нормальному закону, причем математическое ожидание X, т. е. вектор EX, принадлежит заданному линейному подпространству L, $L \subset R^n$; матрица ковариаций вектора X равна $\sigma^2 I$ (скалярная матрица). Вектор $l := EX$ 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.

§ 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 $T = (\mathrm{proj}_L X, |\mathrm{proj}_{L^\perp} X|^2)$. Примем без доказательства тот факт, что эта статистика T — полная. Поэтому наилучшая (имеющая наименьшую матрицу ковариаций) несмещенная оценка параметра θ должна быть функцией достаточной статистики (такая оценка единственна).

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

- Операцию усреднения (вычисления математического ожидания) и проектирования X можно поменять местами (проектирование X на подпространство — линейная операция; 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 $\mathrm{proj}_{L^\perp} X$.

1.1. Несколько вспомогательных определений

О п р е д е л е н и е 7.1.1. *Распределение хи-квадрат.* Пусть $\eta_1, \eta_2, \ldots, \eta_r$ суть независимые случайные величины, распределен-

ные каждая по стандартному нормальному закону $N(0,1)$. Случайной величиной *хи-квадрат с r степенями свободы* называют

$$\chi^2(r) := \eta_1^2 + \eta_2^2 + \cdots + \eta_r^2.$$

Распределение случайной величины $\chi^2(r)$ для любого r (r — натуральное) 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 r-мерного вектора $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_r) \sim N_r(0, I)$, 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.

О п р е д е л е н и е 7.1.2. *Нецентральное распределение хи-квадрат.* Пусть $\vec{a} = (a_1, a_2, \ldots, a_r)$ — заданный вектор. Рассмотрим случайную величину

$$\chi^2(r, \Delta) := (\eta_1 + a_1)^2 + (\eta_2 + a_2)^2 + \cdots + (\eta_r + a_r)^2.$$

Здесь $\Delta = a_1^2 + a_2^2 + \cdots + a_r^2$. Из следствия леммы 7.1.1 (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 $\chi^2(r, \Delta)$. Величину $\Delta = \sum\limits_{i=1}^{r} a_i^2$ называют *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 Δ, $\Delta > 0$, если r фиксировано. Иными словами, если $0 \le \Delta_1 \le \Delta_2$, то для любого $x > 0$

$$P\{\chi^2(r, \Delta_1) > x\} \le P\{\chi^2(r, \Delta_2) > x\}.$$

(О доказательстве скажем позже).

Графики функций распределения $F(x) = P\{\chi^2(r, \Delta) \le x\}$ при разных $\Delta > 0$ выглядят так:

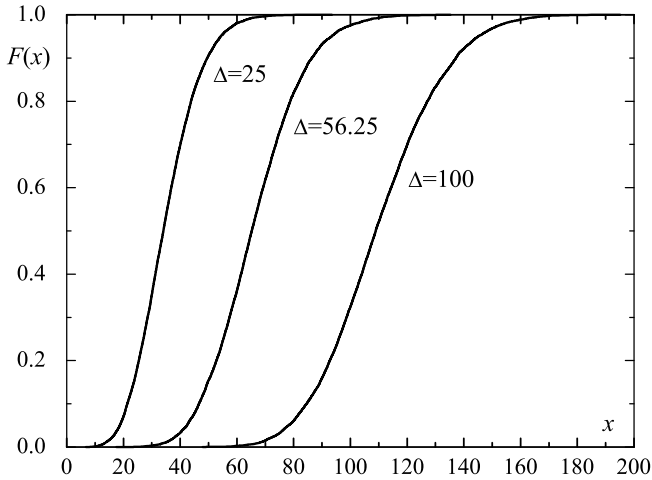


Рис. 7.1.1. Графики функций распределения $y = P\{\chi^2(10, \Delta) \le x\}$ при разных Δ

1.2. Две леммы о круговых нормальных распределениях

Л е м м а 7.1.1. *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*).

Д о к а з а т е л ь с т в о. Для доказательства достаточно вычислить матрицу ковариации вектора $Y = CX$. Поскольку для любой матрицы A матрица ковариаций вектора AX есть

$$D(AX) = A(DX)A^T,$$

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

$$DY = C(DX)C^T = C(\sigma^2 I)C^T = \sigma^2 I,$$

что и требовалось. □

С л е д с т в и е. *Пусть* $\eta_1, \eta_2, \ldots, \eta_r$ *суть независимые* $N(0,1)$. *Тогда*:

$$(\eta_1 + a_1)^2 + (\eta_2 + a_2)^2 + \cdots + (\eta_r + a_r)^2 \stackrel{d}{=} (\eta_1 + \sqrt{\Delta})^2 + \eta_2^2 + \cdots + \eta_r^2,$$

где $\Delta = a_1^2 + a_2^2 + \cdots + a_r^2$.

Это утверждение доказывает правильность употребления выражения $\chi^2(r, \Delta)$ для распределения квадрата длины вектора $\vec{\eta} + \vec{a}$. Здесь $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_r)$.

Д о к а з а т е л ь с т в о. Доказательство основывается на том, что вектор $\vec{\eta} + \vec{a}$ можно ортогональным преобразованием (скажем, C) перевести в вектор с координатами:

$$(\tilde{\eta}_1 + \sqrt{\Delta}, \tilde{\eta}_2, \ldots, \tilde{\eta}_r)^T,$$

где $\tilde{\eta} = C\eta$. При ортогональных преобразованиях длина вектора не меняется; распределение $C\eta$, так же как и распределение η, есть $N(0, I)$.

Л е м м а 7.1.2. *Пусть* $L_1, L_2, \ldots$ — *попарно ортогональные подпространства, прямая сумма которых составляет всё пространство* R^n:

$$L_1 \oplus L_1 \oplus \ldots = R^n.$$

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* $\mathrm{proj}_{L_1} X$, $\mathrm{proj}_{L_2} X$, ... *независимы* (*в совокупности*) *и распределены нормально, причем*

$$E\,\mathrm{proj}_{L_i} X = \mathrm{proj}_{L_i} l,\ i = 1, 2, \ldots;$$

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

Д о к а з а т е л ь с т в о. Рассмотрим в R^n новый ортонормированный базис, который строим, объединяя ортонормированные базисы подпространств $L_1, L_2, \ldots$.

Ради определенности введем соответствующие обозначения:

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

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

. . . . . .

Рассмотрим координаты вектора $X = (X_1, \ldots, X_n)$ в базисе $\{f\}$. Обозначим их через $Y_1, Y_2, \ldots, Y_n$.

Как известно, с помощью матрицы перехода от одного базиса к другому — обозначим эту матрицу через C — векторы-столбцы $Y = (Y_1, Y_2, \ldots, Y_n)$ и X связаны соотношением $Y = CX$. Заметим, что C — ортогональная матрица, и поэтому $Y \sim N(Cl, \sigma^2 I)$. Следовательно, случайные величины $Y_1, Y_2, \ldots, Y_n$ независимы, распределены нормально и имеют одну и ту же дисперсию σ^2. Заметим, что

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

$$\mathrm{proj}_{L_2} X = Y_{r_1+1} f_{r_1+1} + \ldots + Y_{r_1+r_2} f_{r_1+r_2}, \quad \text{и т. д.}$$

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

Для доказательства (b) заметим, что

$$|\mathrm{proj}_{L_1} X|^2 = Y_1^2 + \ldots + Y_{r_1}^2 =$$

$$\sigma^2 \Big[\Big(\frac{1}{\sigma} Y_1 \Big)^2 + \Big(\frac{1}{\sigma} Y_2 \Big)^2 + \cdots + \Big(\frac{1}{\sigma} Y_{r_1} \Big)^2 \Big] = \sigma^2 \chi^2(r_1, \Delta_1),$$

ибо $\Big(\frac{1}{\sigma} Y_1 \Big)^2, \ldots, \Big(\frac{1}{\sigma} Y_{r_1} \Big)^2$ — суть независимые случайные величины с общей дисперсией.

Параметр нецентральности — это квадрат длины математического ожидания вектора $\frac{1}{\sigma} (Y_1, Y_2, \ldots, Y_{r_1})^T$. По сказанному выше,

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

Лемма 7.1.2 доказана. □

1.3. 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 $|\mathrm{proj}_{L^\perp} X|^2$. Согласно лемме 7.1.2,

$$|\mathrm{proj}_{L^\perp} X|^2 = \sigma^2 \chi^2(n - r, \Delta), \quad \text{где} \quad n - r = \dim L^\perp = n - \dim L.$$

Параметр нецентральности Δ здесь равен

$$\Delta = \frac{1}{\sigma^2}|\mathrm{proj}_{L^\perp} EX|^2 = 0,$$

ибо $EX \in L$ по условиям модели, так что $\mathrm{proj}_{L^\perp} EX = 0$.

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

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

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

Отметим также, что в силу леммы 7.1.2, $\mathrm{proj}_L X$ и $\mathrm{proj}_{L^\perp} X$ статистически независимы (как случайные векторы).

З а м е ч а н и е о вычислении $\mathrm{proj}_L X$ и $\mathrm{proj}_{L^\perp} X$. По определению, *проекцией* точки (вектора) X на множество L называют такую точку множества L, на которой достигается минимум расстояния:

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

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

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

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

1.4. Простой пример линейной гауссовской модели

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

$$X = (X_1, X_2, \ldots, X_n)^T, \quad \text{где} \quad X_i \sim N(a, \sigma^2).$$

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 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ и $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Эта же пара $(\bar{X}, s^2)$ и служит достаточной статистикой для (a, σ^2). Статистики $\bar{X}$ и s^2 независимы, $\bar{X} \sim N(a, \frac{1}{n}\sigma^2)$, $(n-1)s^2 \sim \sigma^2\chi^2(n-1)$.

§ 2. 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.

2.1. Однофакторная гауссовская модель

Некий фактор может принимать несколько различных значений, называемых уровнями: $A_1, A_2, \ldots, A_r$. При каждом значении A_j, $j = \overline{1,r}$ производится несколько (скажем n_j) 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.

С т а т и с т и ч е с к а я м о д е л ь:

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

где $a_1, a_2, \ldots, a_r$ — некие числа (обычно неизвестные экспериментатору), ε_{ij} — суть 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 $n_1 + n_2 + \cdots + n_r$), 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 EX),

порождено r векторами вида:

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

Оценки параметров $a_1, a_2, \ldots, a_r$ и σ^2 мы получим в этой модели, применяя общие результаты. Здесь $a_j^* = \dfrac{1}{n_j}\sum\limits_{i=1}^{n_j} x_{ij}$ для $j = \overline{1,r}$; $s^2 = \dfrac{1}{\sum\limits_{j=1}^{r}(n_j-1)}\sum\limits_{j=1}^{r}\sum\limits_{i=1}^{n_j}(x_{ij}-a_j^*)^2$. Статистики $a_1^*, a_2^*, \ldots, a_r^*, s^2$ независимы.

2.2. 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, $A_1, A_2, \ldots, A_r$ — уровни фактора A. Фактор B принимает уровни $B_1, B_2, \ldots, B_s$.

Планы эксперимента в этой схеме могут быть более разнообразны, чем в однофакторной модели. В данном случае план опыта указывает, какое количество независимых повторений n_{ij} надо произвести для комбинации A_i и B_j уровней факторов A и B, $i = \overline{1,r}$, $j = \overline{1,s}$. Наиболее простой и популярный план: $n_{ij} = 1$. (Специальное выражение: "одно наблюдение в клетке"). Результаты опыта можно записать таблицей

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

С т а т и с т и ч е с к а я м о д е л ь (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 пред-

ставление 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 разности $a_i - a_{i'}$, $b_j - b_{j'}$ 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:

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

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 этих данных естественную структуру матрицы (прямоугольной, размера $r \times s$) и положить $X = \|x_{ij}, \ i = \overline{1,r}, \ j = \overline{1,s}\|$.

Матрицы данного размера образуют линейное пространство размерности rs. Подпространство L, которому принадлежит EX, имеет размерность $r+s-1$. Оно порождено $r+s$ матрицами особого вида. Каждая из таких матриц имеет либо строку, либо столбец из единиц; прочие их элементы равны нулю. Симметрии ради (и

не изменяя L), к перечисленным матрицам можно присоединить матрицу, сплошь состоящую из единиц.

Оценки параметров μ, $\vec{\alpha}$, $\vec{\beta}$ получают, проецируя случайный вектор X на подпространство L, т. е. действуя по методу наименьших квадратов. Иначе говоря, решая экстремальную задачу:

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

при условиях

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

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

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

(Точка замещает индекс, по которому произведено осреднение отклика). В этих обозначениях наилучшие несмещенные оценки параметров суть:

$$\mu^*=x_{\cdot\cdot},\quad \alpha_i^*=x_{i\cdot}-x_{\cdot\cdot},\quad \beta_j^*=x_{\cdot j}-x_{\cdot\cdot},$$

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

При этом $(r-1)(s-1)s^2\sim\sigma^2\chi^2\big((r-1)(s-1)\big)$. Указанные выше оценки можно получить как прямым решением приведенной выше экстремальной задачи, так и на основе тождества:

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

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

$$+(x_{\cdot j}-x_{\cdot\cdot}-\beta_j)^2+(x_{\cdot\cdot}-\mu)^2\Big],$$

которое верно, если $\sum\limits_{i=1}^{r}\alpha_i=0,\quad \sum\limits_{j=1}^{s}\beta_j=0.$

§ 3. 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 её канонической форме, когда вектор наблюдений X и его ожидаемое значение $l = EX$ — это n-мерные векторы-столбцы.

Пусть векторы (столбцы) $F_1, F_2, \ldots, F_r$ порождают подпространство L. Эта совокупность векторов может быть как линейно-независимой (базис L), так и нет.

Так как $l \in L$, то $l = \theta_1 F_1 + \theta_2 F_2 + \cdots + \theta_r F_r$ при некоторых коэффициентах $\theta_1, \theta_2, \ldots, \theta_r \in R^1$. Это представление l можно записать в матричной форме. Для этого введем матрицу F (размера $n \times r$), столбцами которой служат векторы $F_1, F_2, \ldots, F_r$:

$$F := \|F_1, F_2, \cdots, F_r\|.$$

Определим r-мерный вектор-столбец θ, положив $\theta = (\theta_1, \theta_2, \ldots, \theta_r)^T$. Теперь вектор l можно представить короче:

$$l = F\theta.$$

Исходная линейная модель представима в виде

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

где $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T \sim N(0, \sigma^2 I)$, $\theta \in R^r$, 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:

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

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 $\dfrac{\partial}{\partial\theta}$ строкой:

$$\frac{\partial}{\partial\theta} = \left(\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_r}\right).$$

При таком соглашении $Q(\theta + d\theta) = Q(\theta) + \dfrac{\partial Q}{\partial\theta}\, d\theta + o(d\theta)$. Далее,

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

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 $F_1, F_2, \ldots, F_r$ — линейно независимая. В этом (и только в этом) случае матрица $F^T F$ невырождена. В этом случае

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

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:

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

Оценкой (несмещенной, наилучшей) для 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 $\hat\theta_1, \hat\theta_2 \ldots, \hat\theta_r$. Это облегчает интерпретацию результатов.

Лекция 8. Доверительное (интервальное) оценивание

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 самое точность приближения у нас не было.

Между тем, точность приближения — это общенаучное понятие, так же как и способ её 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 этих примерах для неизвестной величины a мы указываем её приближенное значение x, причем $|x - a| < \varepsilon$ для некоторого определенного $\varepsilon > 0$. Здесь ε — гарантированная точность приближения $x \approx a$.

В задачах статистического оценивания мы получаем аналогичное приближенное равенство $\hat{\theta}(X) \approx \theta$. (Либо $\hat{\theta}(X) \approx \tau(\theta)$, 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 с большой вероятностью. Рассмотрим на примере нормальной выборки, как реализуются эти соображения.

§ 1. Нормальное распределение $N(a, \sigma^2)$: доверительный интервал для a

Пусть $x_1, \ldots, x_n$ суть независимые измерения некоторой величины a, причем $x_i \sim N(a, \sigma^2)$ для $i = \overline{1, n}$. Оценкой для a может служить $\bar{x} = \dfrac{1}{n} \sum\limits_{i=1}^{n} x_i$, так что $\bar{x} \approx a$. Как можно судить о точности этого приближения, то есть о $|\bar{x} - a|$? С какой вероятностью для данного $\varepsilon > 0$ выполняется неравенство $|\bar{x} - a| < \varepsilon$? 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 $\sqrt{n}\ \dfrac{\bar{x} - a}{\sigma} \sim N(0, 1)$. Зададимся какой-либо (обычно близкой к 1) вероятностью, для удобства обозначив ее через $1 - 2\alpha$. Здесь α задано, $0 < \alpha < \dfrac{1}{2}$. Пусть $z_{1-\alpha}$ обозначает $(1 - \alpha)$-квантиль стандартного нормального распределения. Иными словами, $\Phi(z_{1-\alpha}) = 1 - \alpha$, где $\Phi(\cdot)$ — функция стандартного нормального распределения, или функция Лапласа. Ввиду симметрии (относительно нуля) $z_{1-\alpha} = -z_\alpha$, где $\Phi(z_\alpha) = \alpha$. Поэтому для $\sqrt{n}\ \dfrac{\bar{x} - a}{\sigma}$ справедливо утверждение

$$P\Big\{\Big|\sqrt{n}\ \frac{\bar{x} - a}{\sigma}\Big| < z_{1-\alpha}\Big\} = 1 - 2\alpha, \tag{8.1.1}$$

или

$$P\Big\{|\bar{x} - a| < \frac{\sigma}{\sqrt{n}}\ z_{1-\alpha}\Big\} = 1 - 2\alpha. \tag{8.1.2}$$

Мы можем сказать, что с вероятностью $1 - 2\alpha$ точность приближения $\bar{x} \approx a$ не хуже, чем $\dfrac{\sigma}{\sqrt{n}}\ z_{1-\alpha}$.

Соотношение (8.1.1) можно преобразовать далее, и написать, что

$$P\Big\{\bar{x} - \frac{\sigma}{\sqrt{n}}\ z_{1-\alpha} < a < \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}\Big\} = 1 - 2\alpha. \tag{8.1.3}$$

Отсюда следует, что интервал (случайный)

$$\Big(\bar{x} - \frac{\sigma}{\sqrt{n}}\ z_{1-\alpha},\ \bar{x} + \frac{\sigma}{\sqrt{n}}\ z_{1-\alpha}\Big) \tag{8.1.4}$$

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 $\bar{x}$ и a, следует вычислить доверительные интервалы для нескольких доверительных вероятностей, скажем, для 0.50, 0.90, 0.95 и 0.99. Рис. 8.1.1, на котором выделены эти доверительные интервалы, дает нам наглядное представление о точности статистического приближения $\bar{x} \approx a$.

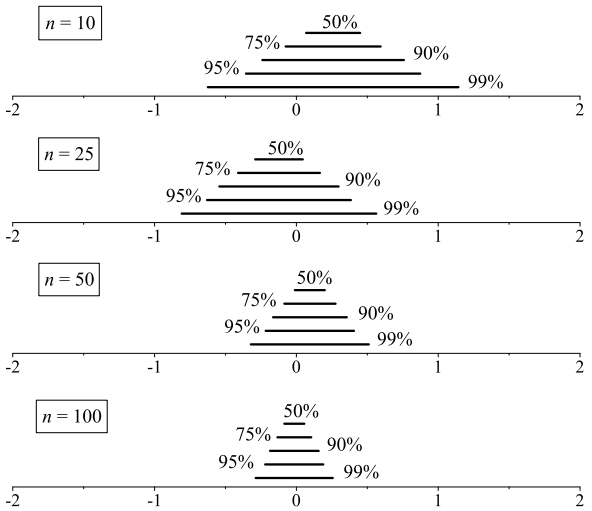


Рис. 8.1.1. Доверительные интервалы для a по модельным выборкам из $N(0, 1)$ объемов $n = 10, 25, 50, 100$ для доверительных вероятностей 0.50, 0.75, 0.90, 0.95, 0.99

Отметим некоторые очевидные, но важные свойства полученных доверительных интервалов.

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

- Интервалы тем шире, чем больше квантиль $z_{1-\alpha}$, которая, в свою очередь, возрастает при приближении $1-\alpha$ к 1. (Эта 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 может быть самое это суждение.

- Наконец, на точность приближения $\bar{x} \approx a$ 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. Достичь этого трудно. Поэтому на практике большие выборки встречаются не часто.

§ 2. Распределение Стьюдента

Предыдущие результаты верны, но бесполезны, когда σ не известно, что чаще всего на практике и бывает. Естественная мысль: заменить неизвестное σ его оценкой s, где $s^2 = \dfrac{1}{n-1}\sum\limits_{i=1}^{n}(x_i-\bar{x})^2$, и рассмотреть случайную величину

$$t = \sqrt{n}\,\frac{\bar{x}-a}{s}. \tag{8.2.1}$$

Её называют *отношением Стьюдента* (*Student's ratio* — стьюдентовская дробь, стьюдентовское отношение). Легко видеть, что распределение (8.2.1) не зависит от неизвестных параметров нормальной выборки (a, σ^2) и совпадает с распределением отношения стандартной нормальной величины $N(0,1)$ к случайной величине $\sqrt{\dfrac{1}{n-1}\,\chi^2(n-1)}$, причем эти случайные величины независимы (см. лекцию о распределении $\bar{x}, s^2$).

Распределение случайной величины (8.2.1) называют *распределением Стьюдента с* $(n-1)$ *степенями свободы.*

Приведем общее определение.

О п р е д е л е н и е 8.2.1. 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

$$t = t(m, \mu) = \frac{\xi_0 + \mu}{\sqrt{\frac{1}{m}(\xi_1^2 + \cdots + \xi_m^2)}}, \tag{8.2.2}$$

где $\mu \in R^1$ — 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 $\mu = 0$, распределение случайной величины $t(m) = t(m, 0)$ называют *центральным распределением Стьюдента*. Эпитет "центральное" обычно 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 приведем формулу плотности для $t(m)$:

$$\frac{1}{\sqrt{m}\, B\left(\frac{1}{2}, \frac{m}{2}\right)} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}. \tag{8.2.3}$$

Из (8.2.3), а также и из (8.2.2) следует, что распределение Стьюдента с одной степенью свободы (при $m = 1$) совпадает с распределением Коши.

Отметим важное для дальнейшего с в о й с т в о распределений Стьюдента: *при каждом m семейство $t(m, \mu)$ стохастически упорядочено* (*стохастически монотонно возрастает*) *относительно μ. Это означает, что для любого $x \in R^1$*

$$P\{t(m, \mu_1) > x\} < P\{t(m, \mu_2) > x\}, \quad \text{если} \quad \mu_1 < \mu_2. \tag{8.2.4}$$

Д о к а з а т е л ь с т в о почти очевидно: из (8.2.2) следует, что

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

$$= EP\{\,\xi_0 + \mu_1 > x\sqrt{\frac{1}{m}\,\chi^2(m)}\;\Big|\;\chi^2(m)\}.$$

Для завершения доказательства остается заметить, что для любого $z \in R^1$

$$P\{\xi_0 + \mu_1 > z\} < P\{\xi_0 + \mu_2 > z\},$$

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 $N(a,\sigma^2)$). Её решение теперь почти не отличается от рассмотренного в первом параграфе. 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

$$P\{t(n-1) < t_{1-\alpha}\} = 1-\alpha. \tag{8.2.5}$$

Ввиду симметрии распределения Стьюдента

$$P\Big\{\Big|\sqrt{n}\,\frac{\bar{x}-a}{s}\Big| < t_{1-\alpha}\Big\} = 1-2\alpha. \tag{8.2.6}$$

Преобразуя (8.2.6), получаем оценку точности для приближения $\bar{x} \approx a$:

$$P\{|\bar{x}-a| < \frac{s}{\sqrt{n}}\,t_{1-\alpha}\} = 1-2\alpha \tag{8.2.7}$$

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

$$P\{\bar{x} - \frac{s}{\sqrt{n}}\,t_{1-\alpha} < a < \bar{x} + \frac{s}{\sqrt{n}}\,t_{1-\alpha}\} = 1-2\alpha. \tag{8.2.8}$$

Все сделанные в § 1 замечания о свойствах доверительного интервала (8.1.4), остаются верными и для (8.2.8). Равно как и рекомендации не ограничиваться каким-либо одним доверительным

интервалом (и какой-либо одной доверительной вероятностью), а вычислять несколько таких интервалов — для нескольких коэффициентов доверия.

Тем же приемом можно выводить для a и другие доверительные утверждения. Например, *доверительные пределы* (границы сверху или снизу).

Выбираем доверительную вероятность $1-\alpha$. Если мы хотим получить для a границу снизу, берем α-квантиль $t_\alpha = t_\alpha(n-1)$; для границы сверху берем $(1-\alpha)$-квантиль $t_{1-\alpha}$. (Заметим, что из-за симметрии $t_\alpha = -t_{1-\alpha}$). Далее заметим, что для (8.2.1) выполняется соотношение

$$P\Big\{t_\alpha < \sqrt{n}\,\frac{\bar{x}-a}{s}\Big\} = 1-\alpha.$$

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

$$P\Big\{a < \bar{x} + \frac{s}{\sqrt{n}}\,t_{1-\alpha}\Big\} = 1-\alpha, \tag{8.2.9}$$

так что $\bar{x} + \dfrac{s}{\sqrt{n}}\,t_{1-\alpha}$ — это верхняя доверительная граница для a, с коэффициентом доверия $1-\alpha$. Нижняя $(1-\alpha)$-доверительная граница для a, равная $\bar{x} - \dfrac{s}{\sqrt{n}}\,t_{1-\alpha}$, получается аналогично.

Пересечение двух полученных доверительных областей дает для a уже известный доверительный интервал (8.2.8), с доверительной вероятностью $1-2\alpha$.

§ 3. Центральные величины

Обсудим в общем виде тот прием, который мы применяли в параграфах 1 и 2. Пусть распределение наблюдения X определяется неизвестным параметром θ, $\theta \in \Theta$. 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 $\sqrt{n}\,\dfrac{\bar{x}-a}{s}$). $G(X,\theta)$ 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:

$$P\{g_\alpha < G(X,\theta)\} = 1-\alpha. \tag{8.3.1}$$

(Точнее было бы в этом равенстве употребить символ P_θ для распределения вероятностей, зависящих от $\theta, \theta \in \Theta$. Но поскольку (8.3.1) 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

$$S_{1-\alpha}(X) := \{\theta : \ g_\alpha < G(X, \theta), \ \theta \in \Theta\}. \tag{8.3.2}$$

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 (8.3.2), но использовать всё 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:

$$\text{если} \ \ 0 < \alpha_1 < \alpha_2 < 1, \ \ \text{то} \quad S_{1-\alpha_1}(X) \supset S_{1-\alpha_2}(X). \tag{8.3.3}$$

Когда θ — одномерный параметр, достаточным условием для (8.3.3) служит монотонность $G(X, \theta)$ по переменной θ (при каждом фиксированном X). Точнее: для (8.3.3) нужно, чтобы $G(X, \theta)$ монотонно убывала по θ. В этом случае $S_{1-\alpha}(X)$ — полупрямая (точнее, это пересечение Θ с полупрямой); его правый конец — это верхняя $(1 - \alpha)$-доверительная граница для θ, $\theta \in \Theta$.

Другая система доверительных областей возникает из аналогичного (8.3.1) соотношения

$$P\{G(X, \theta) < g_{1-\alpha}\} = 1 - \alpha. \tag{8.3.4}$$

Действуя как выше, т. е. решая неравенство относительно θ, получим для θ доверительную область

$$T_{1-\alpha}(X) = \{\theta : G(X, \theta) < g_{1-\alpha}, \ \theta \in \Theta\}.$$

В оговоренном выше одномерном монотонном случае множество $T_{1-\alpha}(X)$ — это полупрямая (пересеченная с Θ). Её левый конец для θ дает $(1 - \alpha)$-доверительную границу снизу. Пересечение областей $S_{1-\alpha}(X) \bigcap T_{1-\alpha}(X)$ дает для θ доверительную область (как правило, ограниченную) с доверительной вероятностью $1 - \ 2\alpha$.

§ 4. Приближенные доверительные границы для вероятности успеха в испытаниях Бернулли

В этой задаче 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, $\theta \in (0,1)$; пусть S_n — число успехов, случившееся в n проведенных испытаниях Бернулли. По теореме Муавра-Лапласа, случайная величина

$$\frac{S_n - n\,\theta}{\sqrt{n\,\theta(1-\theta)}} \xrightarrow{d} N(0,1) \qquad \text{при } n \to \infty.$$

Как обычно, мы заключаем из этой теоремы, что "для достаточно больших n" и $z \in R^1$

$$P\Big\{\Big|\frac{S_n - n\,\theta}{\sqrt{n\,\theta(1-\theta)}}\Big| < z\Big\} \approx \Phi(z) - \Phi(-z).$$

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

$$P\Big\{\Big|\frac{S_n - n\,\theta}{\sqrt{n\,\theta(1-\theta)}}\Big| < z_{1-\alpha}\Big\} \approx 1-2\alpha. \tag{8.4.1}$$

Неравенство в (8.4.1) надо разрешить относительно θ, $\theta \in (0,1)$. После тождественных преобразований получим для этого неравенства эквивалентную форму

$$(S_n - n\,\theta)^2 - n\,\theta(1-\theta)z_{1-\alpha}^2 < 0. \tag{8.4.2}$$

Левая часть — квадратный трехчлен относительно θ, причем коэффициент при θ^2 положителен. Поэтому решение (8.4.2) имеет вид

$$\underline{\theta}(S_n) < \theta < \overline{\theta}\,(S_n), \tag{8.4.3}$$

где $\underline{\theta}(S_n)$, $\overline{\theta}\,(S_n)$ — суть корни квадратного трехчлена в (8.4.2). Здесь

$$\underline{\theta}(S_n),\ \overline{\theta}\,(S_n) = \frac{S_n + \dfrac{z_{1-\alpha}^2}{2} \mp z_{1-\alpha}\sqrt{\dfrac{S_n(n-S_n)}{n} + \dfrac{z_{1-\alpha}^2}{4}}}{n + z_{1-\alpha}^2}.$$

Выражение (8.4.3) дает для θ доверительный интервал, доверительная вероятность которого приближенно равна $1-2\alpha$.

§ 5. Регрессионная модель

Метод центральной величины пригоден для того чтобы строить доверительные области для параметров гауссовских линейных моделей. Рассмотрим регрессионную модель

$$X = F\theta + \varepsilon, \tag{8.5.1}$$

где X — наблюдаемый n-мерный вектор (столбец); $\theta = (\theta_1, \ldots, \theta_m)^T$ — неизвестный параметр, $\theta \in R^m$; F — заданная $n \times m$ матрица, $F = \|F_1, \ldots, F_m\|$; все её 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

$$\hat{\theta} = (F^T F)^{-1} F^T X. \tag{8.5.2}$$

Из теории гауссовских линейных моделей (точнее, из леммы об ортогональных разложениях) вытекает, что $|X - F\hat{\theta}|^2$ и $\hat{\theta}$ статистически независимы, причем

$$|X - F\hat{\theta}|^2 = \sigma^2 \chi^2(n-m), \tag{8.5.3}$$

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

Для построения центральной величины нам понадобится изложенная ниже лемма, а также еще одно семейство распределений.

Л е м м а 8.5.1. *Пусть* $\xi \sim N_p(a, A)$, *причем* A^{-1} *существует. Тогда*

$$\xi^T A^{-1} \xi = \chi^2(p, \Delta),$$

где параметр нецентральности $\Delta = a^T A^{-1} a$.

Д о к а з а т е л ь с т в о. 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 $A \geq 0$ 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

$$|\eta|^2 = \chi^2(p, \Delta), \quad \text{где} \quad \Delta = |Ba|^2 = (Ba)^T Ba = a^T A^{-1} a.$$

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 (8.5.2). Получим, что

$$(\hat{\theta} - \theta)^T (F^T F)(\hat{\theta} - \theta) = \sigma^2 \chi^2(m). \tag{8.5.4}$$

Э ф - р а с п р е д е л е н и е. Называемое также распределением Снедекора, распределением Фишера, распределением дисперсионного отношения Фишера и т. д.

О п р е д е л е н и е 8.5.1. Пусть случайные величины X_1 и X_2 независимы и распределены по законам хи-квадрат:

$$X_1 = \chi^2(m_1, \Delta), \quad X_2 = \chi^2(m_2, 0).$$

Случайная величина

$$F = F(m_1, m_2, \Delta) = \frac{\frac{1}{m_1} X_1}{\frac{1}{m_2} X_2} \tag{8.5.5}$$

называется *F-отношением* (эф-отношением, дисперсионным отношением Фишера). Распределение (8.5.5) называют *нецентральным эф-распределением* с m_1 и m_2 степенями свободы и параметром нецентральности Δ. Если $\Delta = 0$, распределение называют

центральным. Эпитет "центральное" часто опускают и говорят просто об 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 (8.5.5) 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 $F(m_1, m_2, 0)$:

$$\frac{\left(\frac{m_1}{m_2}\right)^{\frac{m_1}{2}} x^{\frac{m_1}{2}-1}}{B\left(\frac{m_1}{2}, \frac{m_2}{2}\right)\left(1+\frac{m_1}{m_2}\,x\right)^{\frac{m_1+m_2}{2}}} \quad \text{для} \quad x \geq 0.$$

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, которым была доказана упорядоченность семейства $\chi^2(m, \Delta)$ по Δ (m — любое).

Вернемся к доверительному оцениванию θ в модели (8.5.1). Из двух независимых случайных величин (8.5.3) и (8.5.4) составим эф-отношение

$$F(m, n-m) = \frac{\frac{1}{m}(\hat\theta - \theta)^T (F^T F)(\hat\theta - \theta)}{\frac{1}{n-m}|X - F\hat\theta|^2}. \tag{8.5.6}$$

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

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

— это несмещенная оценка для σ^2.

Теперь видно, что $(1-\alpha)$-доверительное множество для θ, заданное неравенством

$$\{\theta:\ (\hat{\theta}-\theta)^T(F^TF)(\hat{\theta}-\theta) < m\,s^2F_{1-\alpha}(m,n-m)\}, \tag{8.5.7}$$

представляет собой внутреннюю часть (случайного) эллипсоида с центром в точке $\hat{\theta}$. Эта область (внутренность эллипсоида) накрывает неизвестное θ (точку $\theta \in R^m$) с вероятностью $1-\alpha$.

Можно указать доверительные интервалы и для отдельных параметров θ_i, $\theta = (\theta_1, \ldots, \theta_m)^T$. Из (8.5.2) 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 $(F^TF)^{-1} = \|a_{ij}\|$. С учетом (8.2.2) (и независимости $\hat{\theta}_i$ и s^2) можно утверждать, что случайная величина

$$t := \frac{\hat{\theta}_i - \theta}{s\sqrt{a_{ii}}} \tag{8.5.8}$$

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 параграфе 2.

Отметим, что если матрица F^TF 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 (8.5.8). 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 Шеффе (или Тьюки).

Лекция 9. Проверка статистических гипотез

§ 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 $\mathcal{P}_0 \bigcap \mathcal{P}_1 = \oslash$, $\mathcal{P}_1 \bigcup \mathcal{P}_0 = \mathcal{P}$. (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, $\Theta_1 \subset \Theta$ и $\Theta_0 \bigcap \Theta_1 = \oslash$.

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 наблюденному значению X и в результате примем 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 её 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 I 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. Аргументом служит распределение вероятностей P на $\mathcal{X}$, $P \in \mathcal{P}$.

О п р е д е л е н и е 9.1.1. *Мощностью* $\beta(P)$ критерия S называют

$$\beta(P, S) = \beta(P) = P\{X \in S\},$$

т.е. вероятность события $\{X \in S\}$, когда случайный выбор X, $X \in \mathcal{X}$, происходит согласно распределению вероятностей P. (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 $\beta(P) \leq \alpha$ для всех $P \in \mathcal{P}_0$. Поскольку каждый критерий уровня α есть одновременно и критерий уровня α', если $\alpha < \alpha'$ то полезно определить для критерия его минимальный уровень значимости

$$\sup_{P \in \mathcal{P}_0} \beta(P).$$

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 $\sup\limits_{\theta \in \Theta_0} P_\theta\{X \in S\}$.

Перечислим еще раз желательные свойства любого статистического критерия, предназначенного для проверки статистической гипотезы $P \in \mathcal{P}_0$:

- малый размер;
- быстрое возрастание функции мощности (при удалении распределения P от гипотетического множества распределений $\mathcal{P}_0$).

§ 2. Пример реальной проверки статистической гипотезы

2.1. 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 одинакова жизнеспособность потомства (например, всхожесть семян) 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 фенотипов A и a среди гибридов второго поколения 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, № 1, стр.38–42. (См. также: А.Н. Колмогоров, Теория вероятностей и математическая статистика. – М.: "Наука", 1986 – 535с. и В.Н. Тутубалин (1992), Теория вероятностей и случайных процессов. – М.: изд-во МГУ, 1992 – 400 с., часть 2, глава 3, § 1).

Работа Колмогорова основывается на экспериментальных данных Н.И. Ермолаевой: "Еще раз о гороховых законах"; Яровизация (1939), № 2 (23). Н.И. Ермолаева экспериментировала с томатами. В её опытах результаты разделялись по семействам. Например, семейство составляли все растения, выросшие в одном ящике. Семейства мы занумеруем индексом i, $i = 1 \dots N$; N — их общее число. Чистые линии, которые подвергались скрещиванию (гибридизации), отличались внешне: одни имели гладкие, а другие — морщинистые листья.

Пусть μ_i, $i = 1 \dots N$, обозначают частоты фенотипа a в каждой из N серий, а n_i обозначает число растений в серии. С вероятностной точки зрения μ_i — это число "успехов" в n_i испытаниях Бернулли, если назвать успехом появление фенотипа a. При гипотезе (т. е. если закон Менделя верен) $p = \frac{1}{4}$; в противном случае $p \neq \frac{1}{4}$.

Если численности n_i не слишком малы (порядка нескольких десятков), то по теореме Муавра Лапласа и при справедливости законов Менделя нормированные частоты (где $p = \frac{1}{4}$)

$$\xi_i = \frac{\mu_i - n_i p}{\sqrt{n_i p\,(1-p)}}, \quad \text{а точнее,} \quad \xi_i = \frac{\mu_i - \frac{1}{4}\,n_i}{\sqrt{\frac{3}{16}\,n_i}}$$

имеют (приближенно) распределение $N(0,1)$. Поэтому на совокупность $\xi_1, \xi_2, \dots, \xi_n$ можно смотреть как на выборку (объема N) из $N(0,1)$. Все это — если верен закон Менделя.

2.2. Естественная мысль — сопоставить выборочную функцию $F_N(x)$, построенную по выборке $\xi_1, \ldots, \xi_N$ 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

$$D_N = \sup_x |F_N(x) - \Phi(x)| \tag{9.2.1}$$

при больших N должна быть малой, если верны законы Менделя, ибо в этом случае $D_N \xrightarrow{P} 0$ 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 $D_N \xrightarrow{P} c > 0$, если закон Менделя неверен.

Этих соображений, однако, недостаточно для точных статистических выводов. Надо привлечь следующую теорему.

Т е о р е м а К о л м о г о р о в а (1933). *Пусть $F_n(x)$ — функция распределения выборки объема n, извлеченной из распределения с непрерывной функцией $F(x)$. Пусть*

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

Тогда при $n \to \infty$ равномерно по $z > 0$

$$P\{\sqrt{n} D_n < z\} \to K(z), \quad \text{где} \quad K(z) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 z^2}.$$

Функцию $K(z)$ часто называют *функцией Колмогорова*.

Из этой теоремы следует, что при справедливости закона Менделя статистика $\sqrt{N}\, D_N = \sup_x \sqrt{N}\, |F_N(x) - \Phi(x)|$ при больших N подчиняется распределению Колмогорова. В случае же его нарушения $\sqrt{N}\, D_N \xrightarrow{P} \infty$ при $N \to \infty$.

Это значит, что для конечных значений N статистика $\sqrt{N} D_N$ должна принимать большие значения, если гипотеза неверна.

Таким образом, статистика $\sqrt{N} D_N$ различно ведет себя при гипотезе и при ее нарушении (при альтернативе). Именно это позволяет по наблюдаемой величине $\sqrt{N} D_N$ сделать вывод о том, что же действует на самом деле: гипотеза или альтернатива.

2.3. В данном случае естественно следующее решающее правило: отвергать гипотезу о том, что выборка извлечена из распределения с функцией $F(\cdot)$, если статистика $\sqrt{N}D_N$ 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 C_ε, такое, что $K(C_\varepsilon) = 1 - \varepsilon$.
- Если наблюдаемое значение $\sqrt{N}D_N$ превосходит C_ε, мы проверяемую гипотезу отвергаем (как говорят — на уровне ε). В данном случае — это гипотеза (закон) Менделя.

2.4. Судить о том, совместимо ли наблюденное в опыте значение статистики $\sqrt{N}D_N$ с проверяемой гипотезой можно и иначе. Как было сказано, против гипотезы (закона Менделя) говорят большие значения $\sqrt{N}D_N$, и тем сильнее, чем наблюденное значение больше.

Рассмотрим вероятность того, что в независимом повторении проведенного опыта мы получим такое же или даже большее значение статистики $\sqrt{N}D_N$, 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 $1 - K(\sqrt{N}D_N)$.

Вернемся к опытам Ермолаевой. Всего было две выборки: $N = 98$ и $N = 123$. В обеих выборках наблюденные значения 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.

Лекция 10. Статистические критерии

§ 1. Оптимальный критерий Неймана-Пирсона (J. Neyman, S. Pearson, 1933)

Вводные о п р е д е л е н и я. Статистический критерий S для проверки гипотезы $H_0 : P \in \Theta_0$ против альтернативы $H_1 : P \in \Theta_1$ 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

$$\beta(P, R) \le \beta(P, S) \qquad \text{для всех } P \in \mathcal{P}_1. \tag{10.1.1}$$

Критерий S называют *оптимальным критерием* уровня α, если для любого другого критерия R уровня α выполняется соотношение (10.1.1). 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 $H_1 : P \in \mathcal{P}_1$ 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 либо 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 множества вида

$$S_\lambda = \{x : \ f_1(x) - \lambda f_0(x) \ge 0\} \ \text{ для } \ \lambda > 0 \tag{10.1.2}$$

как критерии 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 поверки H_0 против H_1 по наблюдению $X \in \mathcal{X}$. Предположим, что для некоторого $\lambda > 0$

$$P_0\{X \in R\} \le P_0\{X \in S_\lambda\}. \tag{10.1.3}$$

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 (10.1.2) с тем же уровнем значимости. Тогда в (10.1.3) стоит равенство). Тогда

$$\text{(a)} \qquad P_1\{X \in R\} \le P_1\{X \in S_\lambda\}, \tag{10.1.4}$$

$$\text{(b)} \qquad P_0\{X \in S_\lambda\} \le P_1\{X \in S_\lambda\}.$$

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 мощность 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 (10.1.2) называют *критериями Неймана-Пирсона*, а сформулированное выше утверждение об оптимальности критериев (10.1.2) — *леммой* (*теоремой*) *Неймана-Пирсона*.

Доказательства для распределений, имеющих плотности, и для дискретных распределений проходят одинаково — с той разницей, что интегралы заменяются суммами. Поэтому достаточно рассмотреть что-либо одно, для определенности — плотности. Для простоты предположим, что $\mathcal{X}$ — это конечномерное арифметическое пространство, f_0 и f_1 — плотности относительно меры Лебега.

Записи будут компактными, если вместе с критериями R и S_λ рассмотреть их индикаторные функции $I_R(x)$ и $I_{S_\lambda}(x)$:

$$I_R(x) = \begin{cases} 1, & \text{для } x \in R, \\ 0, & \text{для } x \notin R, \end{cases} \qquad I_{S_\lambda}(x) = \begin{cases} 1, & \text{для } x \in S_\lambda, \\ 0, & \text{для } x \notin S_\lambda. \end{cases}$$

С помощью I_R, I_{S_λ} 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, $P_0\{X \in R\} = E_0 I_R(X)$, а предложение (10.1.3) имеет вид

$$E_0 I_R(X) \le E_0 I_{S_\lambda}(X). \tag{10.1.5}$$

Доказательство утверждения (a). Легко проверить, что справедливо неравенство

$$I_R(x)[f_1(x) - \lambda f_0(x)] \le I_{S_\lambda}(x)[f_1(x) - \lambda f_0(x)]. \tag{10.1.6}$$

Действительно, если $f_1(x) - \lambda f_0(x) > 0$, то $I_{S_\lambda}(x) = 1$ и (10.1.6) превращается в очевидное утверждение $I_R(x) \le 1$. Если же $f_1(x) - \lambda f_0(x) < 0$, то $I_{S_\lambda}(x) = 0$, и потому правая часть (10.1.6) обращается в нуль, а левая часть (10.1.6) при этом не положительна, так что (10.1.6) верно и в этом случае.

Интегрируем (10.1.6) по всему пространству. Результат запишем в виде математических ожиданий.

$$E_1 I_R(X) - \lambda E_0 I_R(X) \le E_1 I_{S_\lambda}(X) - \lambda E_0 I_{S_\lambda}(X),$$

или

$$E_1 I_{S_\lambda}(X) - E_1 I_R(X) \ge \lambda [E_0 I_{S_\lambda}(X) - E_0 I_R(X)]. \tag{10.1.7}$$

В силу (10.1.5) и $\lambda > 0$, правая часть (10.1.7) неотрицательна, что и доказывает (a). $\square$

Доказательство утверждения (b). Для доказательства утверждения (b) надо порознь рассмотреть для λ, определяющего S_λ в (10.1.2), две возможности: $\lambda \ge 1$ и $\lambda < 1$.

- Допустим, что $\lambda \ge 1$. Тогда из (10.1.2) следует, что $f_1(x) \ge f_0(x)$ для $x \in S_\lambda$. Поэтому

  $$P_0\{X \in S_\lambda\} = \int I_{S_\lambda}(x) f_0(x) dx \le \int I_{S_\lambda}(x) f_1(x) dx = P_1\{X \in S_\lambda\},$$

  что и требуется.

- Допустим, что $\lambda < 1$. Рассмотрим множество

  $$\overline{S}_\lambda = \{x : \ f_1(x) \le \lambda f_0(x)\}.$$

Его индикатор есть $1 - I_{S_\lambda}(x)$. При $\lambda < 1$ получаем, что $f_1(x) \le f_0(x)$ для $x \in \overline{S}_\lambda$. Поэтому

$$P_1\{X \in \overline{S}_\lambda\} = \int [1 - I_{S_\lambda}(x)] f_1(x) dx \le \int [1 - I_{S_\lambda}(x)] f_0(x) dx = P_0\{X \in \overline{S}_\lambda\}.$$

Отсюда следует, что при $\lambda < 1$

$$1 - P_1\{X \in S_\lambda\} \le 1 - P_0\{X \in S_\lambda\}.$$

Это доказывает (b) и в этом случае. $\square$

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

$$P_0\{X \in S_\lambda\} = \varepsilon. \tag{10.1.8}$$

В случае плотности это означает, что мы должны решить относительно λ уравнение

$$\int\limits_{\{x: f_1(x) \ge \lambda f_0(x)\}} f_0(x)\, dx \;=\; \varepsilon.$$

В типичном случае решение существует (и единственно). Для дискретно распределенных наблюдений X уравнение (10.1.8) разрешимо не для всех $\varepsilon > 0$. В таком случае — в поисках оптимального критерия уровня ε — либо останавливаются на критерии вида (10.1.2) с меньшим уровнем, чем назначенный ε (с меньшей вероятностью ошибки I рода, увеличивая тем самым вероятность ошибки II рода), либо изменяют выбор уровня значимости так, чтобы (10.1.8) стало разрешимо. Последнее правильнее, ибо назначение уровня значимости — решение в немалой степени произвольное.

§ 2. Равномерно наиболее мощные критерии

Определение равномерно наиболее мощных критериев дано в начале § 1. 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.

П р и м е р. Проверка односторонних гипотез против односторонних альтернатив в схеме Бернулли.

Пусть проведено n (n задано) испытаний Бернулли. Пусть θ, $\theta \in (0,1)$ — 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

$$H_0 : \ \theta \le \theta^0$$

против альтернативы

$$H_1 : \ \theta > \theta^0,$$

где θ^0 — задано, $\theta^0 \in (0,1)$.

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 $(\theta^0, 1)$. Следовательно,

$$0 < a \le \theta^0 < b < 1. \tag{10.2.1}$$

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

$$S_\lambda = \Big\{ x : \ \frac{f_1(x)}{f_0(x)} \ge \lambda \Big\}, \quad \lambda > 0$$

или

$$S_\lambda = \Big\{x:\ \Big(\frac{b}{a}\cdot\frac{1-a}{1-b}\Big)^{T_n(x)}\cdot\Big(\frac{1-b}{1-a}\Big)^{n} \geq \lambda\Big\},\quad \lambda>0. \qquad (10.2.2)$$

Мы уже отмечали, что критерии Неймана-Пирсона образуют семейство оптимальных критериев. Из этого семейства потом выбирают критерий заданного уровня значимости. Сейчас семейство (10.2.2) параметризовано параметром λ, $\lambda>0$. Любая другая параметризация этого семейства не будет хуже.

В частности, семейству (10.2.2) можно дать форму

$$\Big\{x:\ \Big(\frac{b}{1-b}\cdot\frac{1-a}{a}\Big)^{T_n(x)} \geq \lambda'\Big\},\quad \lambda'>0,$$

где $\lambda' = \lambda\Big(\dfrac{1-a}{1-b}\Big)^{n}$. Впрочем, связь между новым параметром λ' и старым параметром λ не важна. При дальнейших изменениях параметризации мы такие связи отмечать не будем. Ввиду (10.2.1)

$$\frac{b}{1-b}\cdot\frac{1-a}{a}>1.$$

Поэтому (10.2.2) можно еще упростить:

$$\{x:\ T_n(x)\geq t\},\quad t>0. \qquad (10.2.3)$$

Отметим главную особенность (10.2.3) как статистического критерия: его вид не зависит от конкретных $a\leq\theta^0$, $b>\theta^0$. Этот критерий — общий для всех $a\in(0,\theta^0]$, $b\in(\theta^0,1)$. Это означает, что критерий (10.2.3) в рассматриваемой задаче является равномерно наиболее мощным. $\square$

Статистическое правило теперь таково. Отвергать гипотезу $H_0:\ \theta\leq\theta^0$ против альтернативы $H_1:\ \theta>\theta^0$, если произошло событие

$$T_n(X)\geq t, \qquad (10.2.4)$$

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. Это характерно для всякого разумного критерия в тех статистических моделях, где существуют достаточные статистики.

Остается определить критическое значение t в (10.2.4). Для этого зададимся некоторым уровнем значимости ε. Для t должно выполняться условие

$$P_\theta\{T_n(X) \geq t\} \leq \varepsilon \quad \text{для всех } \theta \leq \theta^0.$$

Из утверждения (b) леммы Неймана-Пирсона следует, что

$$\sup_{\theta \leq \theta^0} P_\theta\{T_n(X) \geq t\} = P_{\theta^0}\{T_n(X) \geq t\}.$$

Поэтому условие для выбора t упрощается:

$$P_{\theta^0}\{T_n(X) \geq t\} \leq \varepsilon. \tag{10.2.5}$$

Ради достижения наибольшей мощности против альтернативы $\theta > \theta^0$ в качестве критического значения следует взять наибольшее t, удовлетворяющее (10.2.5). Выбор t при заданных θ и n помогают осуществить таблицы для вероятности

$$P_\theta\{T_n(X) \geq t\} = \sum_{k=t}^{n} C_n^k \, \theta^k (1-\theta)^{n-k}$$

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 статистики T_n обозначить как T_n(набл.), сохранив за T_n смысл случайной переменной, то P-значением T_n(набл.) служит

$$P_{\theta^0}\{T_n \geq T_n(\text{набл.})\}. \tag{10.2.6}$$

Сопоставляя это выражение с (10.2.5), видим, что P-значение — это наименьший уровень значимости, на котором еще можно опровергнуть гипотезу H_0 по правилу (10.2.5).

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, то 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 $H_0 : \theta \leq \theta^0$ 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, 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, очень велика.

Существование равномерно наиболее мощных критериев (для проверки односторонних гипотез против односторонних альтернатив) типично 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

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

Биноминальные 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

- $T_n \geq t$, либо $T_n \leq t$.

Лекция 11. Проверка линейных гипотез (в линейных гауссовских моделях)

§ 1. Примеры линейных гипотез

1.1. Выбор степени многочлена

В задачах регрессии y по x функциональный вид зависимости ожидаемого значения отклика $E(y|x)$, как функции x, бывает известен далеко не всегда. В таких случаях аппроксимирующее выражение для $E(y|x)$ подбирают эмпирически. Часто для приближения $E(y|x)$ используют многочлены от x.

Пусть, для простоты, x — скалярная переменная. Предположим, что:

$$E(y|x) = a_0 + a_1 x + \cdots + a_p x^p \tag{11.1.1}$$

для некоторой степени $p \geq 0$ и некоторых коэффициентов a_0, a_1,..., a_p. Далее предположим, что при некоторых заданных значениях $x_1, x_2, \ldots, x_n$ фактора x проведены независимые измерения $y_1, y_2, \ldots, y_n$ отклика y, так что

$$y_i = a_0 + a_1 x_i + \cdots + a_p x_i^p + \varepsilon_i, \quad i = \overline{1,n}, \tag{11.1.2}$$

где $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ суть независимые случайные величины (случайные ошибки). Мы предположим, что $\varepsilon_i \sim N(0, \sigma^2)$, $i = \overline{1,n}$, причем дисперсия ошибки σ^2 не известна.

Выбор степени p аппроксимирующего многочлена в формуле (11.1.1) всегда представляет определенную проблему. Эту степень надо выбрать так, чтобы погрешность в (11.1.1) (она же — систематическая ошибка в (11.1.2)) не влияла на статистические выводы о $E(y|x)$, которые мы сумеем сделать по наблюдениям (x_i, y_i), $\overline{1,n}$. Понятно, что чем ниже эта степень, тем легче интерпретировать результаты опытов. На практике эта степень редко превышает три.

Особенно часто приходится отвечать на вопрос: можно ли для аппроксимации $E(y|x)$ обойтись многочленом первой степени, т. е. простой линейной регрессией, или же надо обратиться к параболической регрессии, т. е. к многочлену степени два?

Статистически проблема выглядит так.

Предположим, что наблюдения y удовлетворяют статистической модели

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 + \varepsilon_i, \quad i = \overline{1,n}, \tag{11.1.3}$$

где $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ суть независимые $N(0, \sigma^2)$, a_0, a_1, a_2 — неизвестные коэффициенты. По наблюдениям (11.1.3) надо проверить гипотезу

$$H_0 : \ a_2 = 0 \tag{11.1.4}$$

против альтернативы

$$H_1 : \ a_2 \neq 0.$$

Гипотеза H_0 (11.1.4) состоит в том, что зависимость отклика от фактора можно передать моделью

$$y_i = a_0 + a_1 x_i + \varepsilon_i, \quad i = \overline{1,n}, \tag{11.1.5}$$

при тех же, что и выше, предположениях об ошибках $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.

1.2. Однофакторный дисперсионный анализ

Пусть наблюдаются $k \geq 2$ независимых выборок, объемы которых обозначим через $n_1, n_2, \ldots, n_k$. Элементы выборки с номером j, $j = \overline{1,k}$, обозначим через x_{ij}, i меняется от 1 до n_j. Предположим, что

$$x_{ij} = a_j + \varepsilon_{ij}, \tag{11.1.6}$$

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 (11.1.6) 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 худших и т. п.) способов обработки. Целесообразно, однако, прежде задаться вопросом: дают ли наши данные основания для такого выбора? По-видимому, нет, если с наблюдениями (11.1.6) совместима гипотеза

$$H_0 : \ a_1 = a_2 = \cdots = a_k. \tag{11.1.7}$$

Легко видеть, что гипотеза (11.1.4) в модели (11.1.3) и гипотеза (11.1.7) в модели (11.1.6) являются частными формами общей линейной гипотезы в линейной модели, как она формулируется в следующем параграфе.

§ 2. Общая линейная гипотеза

Мы говорим, что в отношении наблюдения X (X — элемент линейного пространства, в наших рассмотрениях $X \in R^n$) 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 $E\xi = 0$, $D\xi = \sigma^2 I$, причем σ^2 неизвестно. (Такая форма матрицы ковариаций ξ означает, что компоненты вектора X независимы и имеют одинаковые дисперсии).

- Линейная гипотеза $H_0 : l \in L_0$, где L_0 — заданное линейное подпространство, причем $L_0 \subset L$. Альтернативой к H_0 выступает отрицание H_0 в рамках линейной модели:
- Альтернатива $H_1 : l \notin L_0$, но при этом $l \in L$.

Линейную гипотезу можно рассматривать как частный случай общей параметрической гипотезы о распределении наблюдения X. Предположим, что случайная величина X имеет плотность $f(x, \theta)$, где θ — неизвестный параметр, $\theta \in \Theta$. (Плотность относительно некоторой меры. В нашем случае — относительно меры Лебега в 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 $\tilde{\theta}$ обозначает аналогичную оценку, но вычисленную в предположении, что $\theta \in \Theta_0$.
- Критические события теперь имеют вид

$$S_\lambda = \Big\{ X : \ \frac{f(X, \hat{\theta})}{f(X, \tilde{\theta})} \geq \lambda \Big\}. \tag{11.2.1}$$

Параметр λ, как обычно, выбирают по заданному уровню значимости ε из условия $P\{S_\lambda | H_0\} \leq \varepsilon$.

Критерий (11.2.1) называют *критерием отношения правдоподобий.* В рассматриваемой нами линейной модели оценки $\hat{\theta}$, $\tilde{\theta}$ (для пары (l, σ^2)) нам известны, и вскоре мы к ним обратимся. В общей задаче в качестве $f(x, \hat{\theta})$ и $f(x, \tilde{\theta})$ обычно берут

$$f(X, \hat{\theta}) = \max_{\theta \in \Theta} f(X, \theta), \quad f(X, \tilde{\theta}) = \max_{\theta \in \Theta_0} f(X, \theta).$$

Получаемые по такому правилу оценки θ

$$\hat{\theta} = \arg\max_{\theta \in \Theta} f(X, \theta) \quad \text{и} \quad \tilde{\theta} = \arg\max_{\theta \in \Theta_0} f(X, \theta)$$

называют *оценками наибольшего правдоподобия* (при условиях $\theta \in \Theta$ и $\theta \in \Theta_0$). Эти оценки мы будем изучать в лекции 14. Критерий отношения правдоподобий в этом случае имеет такие критические события

$$S_\lambda = \Big\{ X : \ \frac{\max\limits_{\theta \in \Theta} f(X, \theta)}{\max\limits_{\theta \in \Theta_0} f(X, \theta)} > \lambda \Big\}.$$

Само выражение $f(X, \theta)$, рассматриваемое как функция θ, называют *правдоподобием* θ. Отсюда и названия: *оценки наибольшего правдоподобия*, *критерий отношения правдоподобий*. Свойства оценок наибольшего правдоподобия мы еще будем изучать позже.

§ 3. Применение критерия отношения правдоподобий к проверке линейных гипотез

Применим критерий отношения правдоподобий к проверке линейных гипотез. В рассматриваемой гауссовской модели правдоподобие есть

$$f(X, \theta) = \left(\frac{1}{\sqrt{2\pi}} \right)^n (\sigma^2)^{-n/2} \exp \Big\{ -\frac{1}{2\sigma^2} |X - l|^2 \Big\}. \tag{11.3.1}$$

При условии, что $l \in L$ оценки $\hat{l},\ \hat{\sigma^2}$ суть

$$\hat{l} = \mathrm{proj}_L X,\quad \hat{\sigma^2} = \frac{1}{n-m}|\mathrm{proj}_{L^\perp} X|^2 = \frac{1}{n-m}|X - \mathrm{proj}_L X|^2, \quad (11.3.2)$$

где $m = \dim L$.

При условии, что $l \in L_0$ оценки $\tilde{l},\ \tilde{\sigma^2}$ суть

$$\tilde{l} = \mathrm{proj}_{L_0} X,\quad \tilde{\sigma^2} = \frac{1}{n-m_0}|\mathrm{proj}_{L_0^\perp} X|^2 = \frac{1}{n-m_0}|X - \mathrm{proj}_{L_0} X|^2, \quad (11.3.3)$$

где $m_0 = \dim L_0$

В обоих случаях показатель экспоненты $-\dfrac{|X-l|^2}{2\sigma^2}$ 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 это $-(n-m_0)/2$.

Поэтому семейство критических событий (11.2.1) для проверки гипотезы H_0 имеет вид

$$\left\{ X : \frac{|X - \mathrm{proj}_{L_0} X|^2}{|X - \mathrm{proj}_L X|^2} > \lambda \right\}. \quad (11.3.4)$$

(Параметр λ в (11.3.4) не тождественен параметру λ в (11.2.1); несмотря на это мы употребили для них один и тот же символ. Как уже отмечалось, нам важно, чтобы множество критических событий было как-либо параметризовано, но связь между различными возможными параметризациями не важна. Поэтому соотношение между параметрами в (11.2.1) и (11.3.4) мы можем оставить без внимания).

Ради дальнейшего упрощения (11.3.4), введем в рассмотрение еще одно линейное подпространство: ортогональное дополнение L_0 до L. Обозначим его через L_1. Итак, $L_1 \perp L_0$, $L_0 + L_1 = L$. Теперь R^n представимо в виде суммы трех попарно ортогональны подпространств L_0, L_1 и $L^\perp$. (Как обычно, $L^\perp$ обозначает ортогональное дополнение L до всего пространства R^n):

$$R^n = L_0 + L_1 + L^\perp.$$

В связи с этим для X действует разложение

$$X = \mathrm{proj}_{L_0} X + \mathrm{proj}_{L_1} X + \mathrm{proj}_{L^\perp} X,$$

причем

$$|X - \mathrm{proj}_{L_0} X|^2 = |\mathrm{proj}_{L_1} X|^2 + |\mathrm{proj}_{L^\perp} X|^2. \tag{11.3.5}$$

В силу (11.3.5) критерию отношения правдоподобий (11.3.4) можно дать вид:

$$F := \frac{\frac{1}{m_1}|\mathrm{proj}_{L_1} X|^2}{\frac{1}{n-m}|\mathrm{proj}_{L^\perp} X|^2} \geq \lambda, \tag{11.3.6}$$

с учетом замечаний к (11.3.4).

Вспомним, что оценкой для σ^2 при условии, что $l \in L$, служит

$$\frac{1}{n-m}|\mathrm{proj}_{L^\perp} X|^2. \tag{11.3.7}$$

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

$$\frac{1}{m_1}|\mathrm{proj}_{L_1} X|^2. \tag{11.3.8}$$

Если гипотеза H_0 неверна, оценка σ^2 (11.3.8) приобретает смещение — тем большее, чем больше $|\mathrm{proj}_{L_1} l|^2$. (Но о смещении — чуть позже, когда будем говорить о распределениях (11.3.7) и (11.3.8)). Поэтому критериальная статистика в (11.3.6) — 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 (11.3.6) при гипотезе.

Обсудим распределение статистики из (11.3.6) при гипотезе и при альтернативе. Лемма об ортогональном разложении из лекции 6 говорит, что

$$|\mathrm{proj}_{L^\perp} X|^2 \stackrel{d}{=} \sigma^2 \chi^2(n-m),$$

$$|\mathrm{proj}_{L_1} X|^2 \stackrel{d}{=} \sigma^2 \chi^2(m_1, \Delta),$$

где параметр нецентральности $\Delta = \frac{1}{\sigma^2}|\mathrm{proj}_{L_1} EX|^2$.

Если верна гипотеза H_0, то $\Delta = 0$. Следовательно, критериальная статистика из (11.3.6) распределена как $F(m_1, n-m, \Delta)$:

$$\frac{\frac{1}{m_1}|\mathrm{proj}_{L_1} X|^2}{\frac{1}{n-m}|\mathrm{proj}_{L^\perp} X|^2} \stackrel{d}{=} F(m_1, n-m, \Delta). \tag{11.3.9}$$

(Соотношение (11.3.9) объясняет, между прочим, и принятое для эф-отношения название дисперсионного отношения Фишера).

Примечательно, что при гипотезе H_0 статистика (11.3.9) распределена свободно (от влияния неизвестных параметров $l \in L_0$ и σ^2). (Это свойство получено нами сверх ожиданий. Ничто в наших выкладках того не обещало). Поэтому выбор критического значения λ в (11.3.6) очень упрощается: для этого надо (с помощью таблиц распределения, например) решить уравнение

$$P\{F(m_1, n-m) \geq \lambda\} = \varepsilon.$$

В качестве критического значения (для проверки H_0 на уровне ε) в (11.3.6) 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 (11.3.9) может быть выражение

$$\frac{\frac{1}{m-m_0}|\mathrm{proj}_L X - \mathrm{proj}_{L_0} X|^2}{\frac{1}{n-m}|X - \mathrm{proj}_L X|^2}. \tag{11.3.10}$$

Итак, получили статистическое правило:

- Отвергаем гипотезу H_0 на уровне ε, если статистика (11.3.9) или (11.3.10) превосходит $F_{1-\varepsilon}(m_1, n-m)$.

Из свойств эф-отношения следует, что мощность этого критерия монотонно возрастает вместе с ростом параметра нецентральности $\Delta = \frac{1}{\sigma^2}|\mathrm{proj}_{L_1} EX|^2$.

§ 4. Пример: две нормальные выборки

Рассмотрим две независимые нормальные выборки $x_1, x_2, \ldots, x_m$, где $x_i \sim N(a, \sigma^2)$, и $y_1, y_2, \ldots, y_n$, где $y_i \sim N(b, \sigma^2)$, параметры a, b и σ^2 неизвестны. Подлежащая проверке гипотеза

$$H_0 : \ a = b. \tag{11.4.1}$$

Альтернатива к ней $\quad H_1 : \ a \neq b.$

В $(n+m)$-мерном пространстве рассмотрим векторы

$$e_1 = (\underbrace{1, \ldots, 1}_{m}, \underbrace{0, \ldots, 0}_{n})^T, \quad e_2 = (\underbrace{0, \ldots, 0}_{m}, \underbrace{1, \ldots, 1}_{n})^T,$$

$$Z = (x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n)^T, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_m, \varepsilon_{m+1}, \ldots, \varepsilon_{m+n})^T,$$

где $\varepsilon_1, \varepsilon_2, \ldots$ суть независимые $N(0, \sigma^2)$.

Вектор Z можно представить в виде $Z = a\, e_1 + b\, e_2 + \varepsilon$. Ясно, что Z следует линейной гауссовской модели, причем $EZ \in L(e_1, e_2)$, где $L(e_1, e_2)$ обозначает (двумерное) линейное подпространство с базисом e_1, e_2. При гипотезе H_0 вектор EZ лежит в одномерном линейном подпространстве L_0, порожденном единственным вектором $e_1 + e_2$. Для проверки H_0 против H_1 с помощью статистики (11.3.10) надо вычислить $|\mathrm{proj}_L Z - \mathrm{proj}_{L_0} Z|^2$ и $|Z - \mathrm{proj}_L Z|^2$. Будем использовать обозначения

$$\bar{x} = \tfrac{1}{m} \sum_{i=1}^{m} x_i, \quad s_x^2 = \tfrac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2,$$

$$\bar{y} = \tfrac{1}{n} \sum_{j=1}^{n} y_j, \quad s_y^2 = \tfrac{1}{n-1} \sum_{j=1}^{n} (y_j - \bar{y})^2.$$

Легко видеть, что

$$\mathrm{proj}_L Z = \bar{x} e_1 + \bar{y} e_2, \quad \mathrm{proj}_{L_0} Z = \big(\tfrac{m}{m+n}\, \bar{x} + \tfrac{n}{m+n}\, \bar{y}\big)(e_1 + e_2).$$

Отсюда

$$|Z - \mathrm{proj}_L Z|^2 = (m-1)s_x^2 + (n-1)s_y^2, \quad |\mathrm{proj}_L Z - \mathrm{proj}_{L_0} Z|^2 = \tfrac{mn}{m+n}(\bar{x} - \bar{y})^2.$$

В этих обозначениях статистика (11.3.10) и последующее статистическое правило таковы:

- Отвергать $H_0 : \ a = b$ на уровне ε, если

$$\frac{mn(m+n-2)}{m+n}\cdot\frac{(\bar{x}-\bar{y})^2}{\sum\limits_{i=1}^{m}(x_i-\bar{x})^2+\sum\limits_{j=1}^{n}(y_j-\bar{y})^2} > F_{1-\varepsilon}(1,m+n-2). \quad (11.4.2)$$

Обычно вместо эф-статистики (11.4.2) рассматривают статистику Стьюдента t, причем $t^2 = F$:

$$t = \frac{\sqrt{\frac{mn}{m+n}}(\bar{x}-\bar{y})}{\sqrt{\frac{1}{m+n-2}[(m-1)s_x^2+(n-1)s_y^2]}}. \quad (11.4.3)$$

При гипотезе H_0 статистика (11.4.3) распределена по Стьюденту, с $m+n-2$ степенями свободы. С помощью (11.4.3) можно отдельно проверять H_0 против односторонних альтернатив: против правосторонней $H^+ : a > b$ или левосторонней $H^- : a < b$.

§ 5. Заключение

Теория гауссовских линейных моделей составляет классическую главу математической статистики, её 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 распределения наблюдений (или ошибок) определенно не гауссовские. 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. (О влиянии выбросов будем говорить позже, в лекции 15).

Поэтому для приложений нужны и другие статистические методы. Об одном из них, не опирающемся на какую-либо параметрическую форму распределений (и поэтому называемом непараметрическим), простом математически и достаточно универсальном, будем рассказывать далее.

Лекция 12. Ранговые методы: критерий ранговых сумм (Wilcoxon)

§ 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 подругому).

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 $1, 2, \ldots, n$, если n — численность исходной совокупности.

Пусть теперь исходная совокупность $X = (x_1, x_2, \ldots, x_n)$ — выборка из некоторого непрерывного распределения. С вероятностью 1 эта выборка не имеет одинаковых элементов. Рассмотрим ранги величин $x_1, x_2, \ldots, x_n$. Для определенности, при упорядочении в порядке возрастания. Обозначим их через $R(x_1)$, $R(x_2)$,...,$R(x_n)$. Пусть $(r_1, r_2, \ldots, r_n)$ — произвольная перестановка чисел $(1, 2, \ldots, n)$. Основное свойство случайных рангов:

$$P\{\vec{R}(X) = \vec{r}\} := P\{R(x_1) = r_1, \ldots, R(x_n) = r_n\} = \frac{1}{n!}.$$

Это значит, что для независимых одинаково распределенных непрерывных случайных величин распределение их рангов — равномерное и не зависит от исходного распределения. Поэтому с помощью рангов его изучать нельзя. Но для более сложных статистических моделей ранговые методы могут быть очень полезны, так как они приводят к выводам, не зависящим от исходных распределений (и от того, известны ли эти распределения). Это свойство — независимость выводов от распределения — часто называют *свободой от распределения*. Ранговые методы — один из примеров свободных от распределения методов. Другое название для таких методов — *непараметрические*. Название дано, чтобы

противопоставить их статистическим методам, разрабатываемым специально для того или иного параметрического семейства распределений — нормального, например. Обсуждая гауссовские линейные модели, мы занимались именно параметрическим исследованием.

На прошлой лекции мы обсудили задачу о двух выборках, которые могут отличаться сдвигом, когда эти выборки гауссовские. Теперь покажем, как ту же задачу для выборок из произвольного непрерывного распределения можно решать с помощью ранговых методов.

§ 2. Сравнение двух выборок, могущих отличаться сдвигом

- Пусть $X = (x_1, x_2, \ldots, x_m)$ — выборка из распределения с функцией распределения $F(u) = P\{x_i \le u\}$.
- Пусть $Y = (y_1, \ldots, y_n)$ — независимая от X выборка из распределения с функцией распределения $F(u - \theta)$.
- Здесь $\theta \in R^1$ — параметр сдвига, $F(\cdot)$ — некоторая непрерывная функция, неизвестная наблюдателю.

В этой обстановке надо

- Проверить гипотезу $H:\ \theta = 0$ против лево- либо правосторонних альтернатив $H^-:\ \theta < 0,\ \ H^+:\ \theta > 0$;
- Построить доверительные интервалы и доверительные пределы для θ;
- Указать точечную оценку θ.

Всё это возможно с помощью ранговых средств.

Р а н г о в ы й м е т о д (проверки гипотезы H). Рассмотрим объединенную совокупность (X, Y): $x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n$.

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 $(S_1, S_2, \ldots, S_n)$ 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 $1/[N(N-1)\dots(N-n+1)]$. В частности, $P\{R(y_j) = s\} = \dfrac{1}{N}$ для любого $s = 1, 2, \dots, N$.

Чтобы понять, каково распределение рангов игреков (S_1, S_2, …, S_n) при альтернативах H^- или H^+, представим выборку из Y как продолжение выборки из X, но "со сдвигом":

$$y_1 = \theta + x_{m+1}, \dots, y_n = \theta + x_{m+n}.$$

Здесь $x_{m+1}, x_2, \dots, x_{m+n}$ — независимые (в совокупности) и независящие от $x_1, x_2, \dots, x_n$ случайные величины, имеющие ту же (что $x_1, x_2, \dots, x_n$) функцию распределения $F(\cdot)$.

Теперь ясно, что:

- Если $H^+ :\ \theta > 0$ (альтернатива H^+), то $P\{y_j > x_i\} > \dfrac{1}{2}$.
- Если же $H^- :\ \theta < 0$, то верно противоположное неравенство $P\{x_i > y_j\} > \dfrac{1}{2}$.

Поэтому при $H^+ :\ \theta > 0$ для рангов игреков, т. е. для случайных величин $(S_1, S_2, \dots, S_n)$, более вероятны значения из правой части ряда $1, 2, \dots, N$, чем из левой.

При $H^- :\ \theta < 0$ — наоборот, для $(S_1, S_2, \dots, S_n)$ более вероятны малые числа из $1, 2, \dots, N$.

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 вычислены и сведены в таблицы (табулированы).

При альтернативе H^+ для $W_{m,n}$ становятся более вероятными большие значения: для $z > 0$

$$P\{W_{m,n} \geq z \,|H^+\} > P\{W_{m,n} \geq z \,|H\}.$$

При H^- справедливо противоположное неравенство:

$$P\{W_{m,n} \leq z \,|H^-\} > P\{W_{m,n} \leq z \,|H\}.$$

Взяв во внимание эти различия в статистическом поведении $W_{m,n}$ при гипотезе и альтернативах, можно предложить правила проверки H против H^- либо H^+.

П р а в и л о проверки H против H^+

- Выбираем уровень значимости $\varepsilon > 0$.
- 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:

  $$P\{W_{m,n} \geq w(\varepsilon, m, n)|H\} = \varepsilon.$$

  (Лучше выбрать ε так, чтобы это уравнение имело решение — из-за дискретности распределения $W_{m,n}$ это возможно только для некоторых значений ε).
- Отвергаем H в пользу H^+ на уровне ε, если наблюденное значение $W_{m,n}$ равно или превосходит $w(\varepsilon, m, n)$, т. е. если

  $$\text{набл. } W_{m,n} \geq w(\varepsilon, m, n).$$

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:

- отвергать 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 $E_0 W_{m,n}$. (Индексом ноль отмечаем распределения, соответствующие $\theta = 0$). Легко видеть, что

$$E_0 W_{m,n} = \frac{n(m+n+1)}{2}.$$

Можно показать, что функции мощности приведенных выше критериев возрастают по мере удаления значения θ от $\theta = 0$ в нужном направлении.

§ 3. Связь доверительного оценивания и проверки гипотез

Пусть X — наблюдение, P_θ — распределение X, θ — неизвестный параметр, $\theta \in \Theta$. Предположим, что для проверки гипотезы $H_t : \theta = t$ мы располагаем статистическим критерием, уровень которого $\leq \varepsilon$. Пусть $\delta(X, t)$ — индикаторная функция критерия. (Отвергаем $H_t : \theta = t$, если $\delta(X, t) = 1$).

Те значения параметра $t \in \Theta$, для которых гипотеза $H_t : \theta = t$ не отвергается по наблюдению X на выбранном уровне, образуют множество $C(X) = \{t : \delta(X, t) = 0,\ t \in \Theta\}$, которое является доверительным для неизвестного значения θ с доверительной вероятностью $\geq 1 - \varepsilon$. Т. е. доверительное множество образуют те значения параметра, которые совместимы с наблюдением X (точнее, с X совместимы распределения вероятностей). Легко видеть, что

$$P_\theta\{\theta \in C(X)\} \geq 1 - \varepsilon,$$

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 $\geq 1 - \varepsilon$.

П р и м е р. Оценка (доверительная) сдвига одной гауссовской выборки относительно другой. Пусть $X = (x_1, x_2, \ldots, x_m)$ — выборка из $N(a, \sigma^2)$, $Y = (y_1, y_2, \ldots, y_n)$ — 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

$$F = \frac{mn}{m+n} \cdot \frac{(\bar{x} - \bar{y})^2}{s^2},$$

которая при гипотезе $H_0 : a = b$ следует эф-распределению с $(1, m + n - 2)$ степенями свободы. Здесь

$$s^2 = \frac{1}{m+n-2}\Big[\sum_{i=1}^{m}(x_i - \bar{x})^2 + \sum_{j=1}^{n}(y_j - \bar{y})^2\Big].$$

Рассмотрим гипотезу $H_t : \theta = t,\ t$ — задано. Мы сведем задачу к предыдущей, если выборку Y преобразуем в $Z = (z_1, z_2, \ldots, z_n)$,

где $z_j = y_j - t,\ \ j = \overline{1,n}$. Критериальная статистика для проверки $H_t:\ \theta = t$ теперь равна:

$$\frac{mn}{m+n} \cdot \frac{(\bar{x} - \bar{y} - t)^2}{s^2}.$$

(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 отвергать H_t, если $\sqrt{\dfrac{mn}{m+n}} \cdot \dfrac{|\bar{x} - \bar{y} - t|}{s} < t_{1-\varepsilon/2}$,

где $t_{1-\varepsilon/2}$ — это $(1-\varepsilon/2)$- квантиль распределения Стьюдента с $(m+n-2)$ степенями свободы. Решая это неравенство относительно t, получим для θ доверительный интервал

$$\left\{\bar{y} - \bar{x} - s\sqrt{\frac{mn}{m+n}}\, t_{1-\varepsilon/2} < \theta < \bar{y} - \bar{x} + s\sqrt{\frac{mn}{m+n}}\, t_{1-\varepsilon/2}\right\}.$$

§ 4. Доверительное оценивание сдвига

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 об однородности выборок

$$x_1, x_2, \ldots, x_m \text{ и } y_1 - t, y_2 - t, \ldots, y_n - t, \tag{12.4.1}$$

для произвольных $t \in R^1$.

Обозначим статистику Уилкоксона для (12.4.1) через $W_{m,n}(t)$:

$$W_{m,n}(t) = \sum_{j=1}^{n} R(y_j - t),$$

где $R(y_1–t), \ldots, R(y_n–t)$ — ранги случайных величин $y_1–t, \ldots, y_n–t$ в объединенной совокупности $x_1, \ldots, x_m, y_1 - t, \ldots, y_n - t$. Теперь доверительное множество для неизвестного истинного значения

параметра сдвига θ (доверительная вероятность которого равна $1-2\alpha$) есть

$$\{t:\ nN - w(\alpha,m,n) < W_{m,n}(t) < w(\alpha,m,n)\}. \qquad (12.4.2)$$

Остается решить это неравенство относительно t, т. е. дать явный вид этому доверительному множеству.

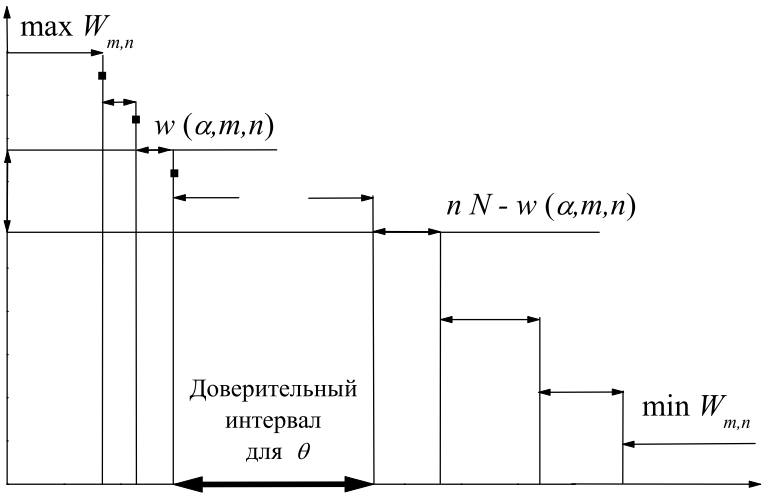


Рис. 12.4.1. График функции $y = W_{m,n}(t)$, $t \in R^1$

Рассмотрим статистику $W_{m,n}(t)$ как функцию переменного $t \in R^1$. 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 $\max W_{m,n} = nN - \dfrac{n(n-1)}{2} = \dfrac{1}{2}\,n(n+2m+1)$. При $t \to \infty$ по противоположным соотношениям между $y_j - t$ и x_i находим

$$W_{m,n}(t) = 1 + 2 + \ldots + n = \frac{n(n+1)}{2} = \min W_{m,n}.$$

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

$y_j - x_i$ ($i = \overline{1,m};\ j = \overline{1,n}$). (Для контроля: $\max W_{m,n} + \min W_{m,n} = 2E_0 W_{m,n}$, $\max W_{m,n} - \min W_{m,n} = mn$, т. е. разница между наибольшим и наименьшим значениями $W_{m,n}(t)$ равна количеству единичных скачков).

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 её графика следует, что доверительное множество (12.4.2) есть интервал; его концами служат некоторые элементы из множества $\{y_j - x_i,\ i = \overline{1,m};\ j = \overline{1,n}\}$, которые нетрудно указать точно. Для этого сказанное множество нужно упорядочить, и затем выбрать порядковые статистики с нужными номерами. (Из рисунка 12.4.1 видно, какие это номера).

§ 5. Точечная оценка сдвига

Статистика $W_{m,n}(t)$ количественно выражает степень согласия (однородности) двух выборок: $x_1, x_2, \ldots, x_n$ и $y_1 - t, y_2 - t, \ldots, y_n - t$. Чем более отклоняется $W_{m,n}(t)$ от $E_0 W_{m,n}$ (от ожидаемого значения $W_{m,n}$ при полной однородности), тем больше (сильнее) различаются выборки. Обратно: две выборки тем ближе к однородным (если мерить с помощью статистики Уилкоксона), чем ближе $W_{m,n}(t)$ к $E_0 W_{m,n}$. Отсюда вытекает предложение: выбрать в качестве точечной оценки неизвестного сдвига θ величину $\hat{\theta}$ такую, что $W_{m,n}(\hat{\theta}) = E_0 W_{m,n}$, т. е. в качестве оценки θ взять решение уравнения $W_{m,n}(t) = \dfrac{n(m+n+1)}{2}$. Понятно, что $\hat{\theta} = \operatorname{med}(x_i - y_j,\ i = \overline{1,m};\ j = \overline{1,n})$, т. е. медиана совокупности, состоящей из mn разностей вида $y_j - x_i$ для $i = 1, \ldots, m;\ j = 1, \ldots, n$. Эту оценку называют *медианой Ходжеса-Лемана.*

§ 6. Совпадения

Если наблюдаемые случайные величины распределены непрерывно, то среди их реализаций не может быть одинаковых (вероятность совпадений равна нулю). На практике, однако, совпадения встречаются нередко хотя бы из-за округлений при записи результатов. Первая трудность, которая при этом возникает —

как назначить ранги? Принятый способ — так называемые средние ранги. Его проще пояснить на примере. Пусть $x_1 = 3$, $x_2 = 1$, $x_3 = 2$, $x_4 = 2$, $x_5 = 4$. Упорядоченные по возрастанию, эти числа дают вариационный ряд $1, 2, 2, 3, 4$. Число 2 в нем встречается дважды, занимая второе и третье места. Будь x_3 и x_4 различны, они получили бы ранги 2 и 3. Сейчас каждому из них дают средний ранг $\frac{2+3}{2} = 2.5$. Получают $R_1 = 4, R_2 = 1, R_3 = 2.5, R_4 = 2.5, R_5 = 5$.

Назначив ранги (если надо — средние), действуют по описанным правилам. Следует помнить, что при наличии совпадений статистические характеристики (уровни значимости, доверительные вероятности и т. д.) не являются точными: совпадения указывают, что базовое предположение о непрерывности основного распределения не выполняется. Все ранговые выводы становятся приближенными, и чем выше доля совпадений, тем менее надежны эти выводы. Впрочем, если доля совпадений невелика, то невелики и возможные ошибки, и ранговыми методами можно пользоваться без опасений. К сожалению, границу здесь провести трудно.

§ 7. Другие ранговые правила

Ранговые (шире: непараметрические) методы можно прилагать к решению многих задач. В частности, так можно исследовать не только задачу о двух выборках, как мы сделали это выше, но и другие линейные модели. Большое достоинство, что случайные ошибки при этом могут иметь произвольное распределение (непрерывное). Как пример, обратимся к уже упоминавшейся (лекция 11) однофакторной модели, где $x_{ij} = a_j + \varepsilon_{ij}$, $j = \overline{1,k}$, $i = \overline{1,n_j}$. Здесь x_{ij} — наблюдения, $a_1, \ldots, a_k$ — неизвестные постоянные, ε_{ij} — независимые одинаково распределенные случайные ошибки. Для этой модели приведем и гауссовское, и непараметрическое правила для проверки гипотезы

$$H_0 : \ a_1 = \ldots = a_k.$$

Гауссовская постановка: случайные ошибки ε_{ij} независимы и распределены нормально $N(0, \sigma^2)$, причем дисперсия σ^2 не известна. Упомянутую гипотезу можно проверить, сравнивая две оценки для σ^2, общую и при гипотезе. Описанный в лекции 11 общий ме-

тод приводит к критериальной статистике

$$F = \frac{\dfrac{1}{k-1}\sum\limits_{j=1}^{k} n_j (x_{\cdot j} - \bar{x})^2}{\dfrac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j} (x_{ij} - x_{\cdot j})^2}.$$

Здесь $N = \sum\limits_{j=1}^{k} n_j$ — общее число наблюдений, $x_{\cdot j} = \frac{1}{n_j}\sum\limits_{i=1}^{n_j} x_{ij}$ — среднее по столбцу j, $\bar{x} = \sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j} x_{ij}/N$ — генеральное среднее.

При гипотезе H_0 эта статистика следует эф-распределению с $(k-1, N-k)$ степенями свободы. Гипотезу H_0 отвергают на уровне значимости α, если F превосходит $(1-\alpha)$-квантиль сказанного эф-распределения. (Проще говоря, если вычисленное F оказывается неправдоподобно большим).

Непараметрическая постановка: случайные ошибки ε_{ij} независимы и одинаково распределены, причем это распределение непрерывное. Ранговый метод начинается с перехода от наблюдений x_{ij} к их рангам r_{ij} (в объединенной совокупности). Ранговая критериальная статистика (одна из возможных) для проверки той же гипотезы H_0 выглядит сходно с числителем эф-отношения: она равна $G = \frac{12}{N(N+1)}\sum\limits_{j=1}^{k} n_j (r_{\cdot j} - \bar{r})^2$, где $r_{\cdot j} = \frac{1}{n_j}\sum\limits_{i=1}^{n_j} r_{ij}$, $\bar{r} = \frac{1}{N}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j} r_{ij}$. В данном случае $\bar{r} = \frac{N+1}{2}$. Множитель $\frac{12}{N(N+1)}$ поставлен ради удобного перехода к пределу: при гипотезе H_0: $G \xrightarrow{d} \chi^2(k-1)$, когда $n_1, \ldots, n_k \to \infty$.

При гипотезе H_0, когда все возможные размещения рангов имеют одинаковые вероятности, статистика G распределена свободно. Поэтому её распределение для каждого набора $(n_1, \ldots, n_k)$ и k может быть вычислено. Реально таблицы распределений для G составлены лишь для немногих таких наборов. По счастью, упомянутая аппроксимация распределением хи-квадрат неплохо действует даже для относительно небольших численностей $n_1, \ldots, n_k$. Более детальный рассказ здесь неуместен. Для практического применения ранговых методов рекомендую обратиться к М. Холлендер, Д. Вулф, "Непараметрические методы статистики". – М.: Финансы и статистика, 1983. – 518 с.

Лекция 13. Асимптотическая нормальность статистики ранговых сумм Уилкоксона

§ 1. Формулировки теорем

Мы продолжаем обсуждение статистики ранговых сумм $W_{m,n}$. Нам уже известно главное: для однородных независимых выборок распределение статистики $W_{m,n}$ одно и то же для всех непрерывных генеральных совокупностей (т. е. для любого непрерывного распределения случайных наблюдений). Поэтому для проверки с помощью $W_{m,n}$ гипотезы об однородности двух выборок можно для всех них использовать одни и те же таблицы распределения этой статистики.

В разных сборниках таблиц (и пакетах статистических программ) сведения о распределении $W_{m,n}$ могут быть даны поразному. Удобно для приложений, если указана, например,

$$P_0\{W_{m,n} \geq x\} \tag{13.1.1}$$

для первых натуральных m, n и различных x. Так сделано в книге М. Холлендер, Д. Вулф "Непараметрические методы статистики". – М.: Финансы и статистика, 1983. – 518 с., таблица А.5. Через $P_0\{.\}$ здесь обозначена вероятность при гипотезе, т. е. когда обе независимые выборки извлечены из общего непрерывного распределения. Но как бы далеко по m, n ни были рассчитаны таблицы, неизбежен вопрос о вероятностях (13.1.1) для численностей выборок m, n за их пределами. В этой лекции будет показано, что при больших m, n статистика $W_{m,n}$ распределена приближенно нормально, и не только при гипотезе. По современному обыкновению этот результат формулируют в виде предельной теоремы.

Т е о р е м а 13.1.1. *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* $m, n \to \infty$

$$\frac{W_{m,n} - EW_{m,n}}{\sqrt{DW_{m,n}}} \xrightarrow{d} N(0,1).$$

Говоря точнее, мы будем доказывать асимптотическую нормальность другой статистики — так называемой *статистики Манна-Уитни* (Mann-Whitney)

$$H_{m,n} = \sum_{i=1}^{m} \sum_{j=1}^{n} I(x_i < y_j),$$

где $I(.)$ — индикаторная функция события. Легко видеть, что статистики $W_{m,n}$ и $H_{m,n}$ связаны соотношением

$$W_{m,n} = H_{m,n} + \frac{n(n+1)}{2}.$$

Для доказательства достаточно заметить, что

$$R(y_j) = \sum_{i=1}^{m} I(x_i < y_j) + \sum_{k=1}^{n} I(y_k < y_j) + 1.$$

Мы докажем следующую теорему:

Т е о р е м а 13.1.2. *В условиях теоремы* 13.1.1

$$\frac{H_{m,n} - EH_{m,n}}{\sqrt{DH_{m,n}}} \xrightarrow{d} N(0,1).$$

Очевидно, что теорема 13.1.1 следует из теоремы 13.1.2, и обратно. На практике теоремы 13.1.1 и 13.1.2 используют как основания для приближенных вычислений, скажем:

$$P_0\{W_{m,n} \le x\} \approx \Phi\Big(\frac{x - E_0 W_{m,n}}{\sqrt{D_0 W_{m,n}}}\Big).$$

Для целых x, как обычно, более точное приближение обеспечивает поправка на непрерывность:

$$P_0\{W_{m,n} \le x\} \approx \Phi\Big(\frac{x + 0.5 - E_0 W_{m,n}}{\sqrt{D_0 W_{m,n}}}\Big),$$

$$P_0\{W_{m,n} \ge x\} \approx 1 - \Phi\Big(\frac{x - 0.5 - E_0 W_{m,n}}{\sqrt{D_0 W_{m,n}}}\Big).$$

Аналогичные формулы верны и для $H_{m,n}$. Как легко видеть,

$$E_0 W_{m,n} = \frac{n(m+n+1)}{2}, \quad E_0 H_{m,n} = \frac{mn}{2}.$$

В дальнейшем будет показано, что

$$D_0 H_{m,n} = D_0 W_{m,n} = \frac{mn(m+n+1)}{12}.$$

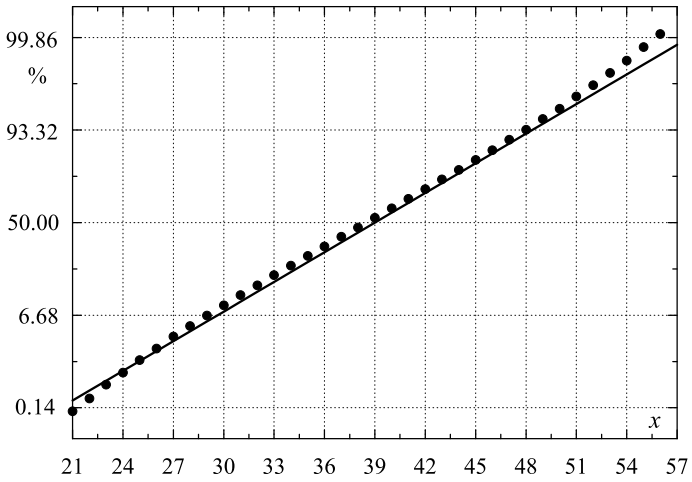


Рис. 13.1.1. График функции $y = P\{W_{m,n} \le x\}$ при $m = n = 6$ на нормальной бумаге

Отметим важное свойство этой аппроксимации: она дает удовлетворительную точность даже для малых выборок. Чтобы убедиться, достаточно взглянуть на рис. 13.1.1, где на нормальной бумаге изображены графики функций распределения $W_{m,n}$ и соответствующего нормального распределения $N(E_0 W_{m,n}, D_0 W_{m,n})$.

Это свойство нормальной аппроксимации позволяет составлять сборники статистических таблиц так, что для m, n за их пределами нормальная аппроксимация для $W_{m,n}$ дает достаточную для практики точность (которую авторы обычно указывают).

Вернемся к теореме 13.1.2. Теорема 13.1.2 — это частный случай общей теоремы об асимптотическом поведении так называемых *U-статистик* (*U-statistics*). В данном случае $H_{m,n}$ — это двухвыборочная U-статистика

$$U_{m,n} := \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i, y_j)$$

с ядром $f(x, y) = I(x < y)$. Мы докажем следующую теорему:

Т е о р е м а 13.1.3. *Пусть* $(x_1, \ldots, x_m)$ *и* $(y_1, \ldots, y_n)$ — *две*

независимые выборки; пусть функция $f(x,y)$ *такова, что*

$$Df^2(x_1,y_1)<\infty,\ D\big(E[f(x_1,y_1)|x_1]\big)^2>0,\ D\big(E[f(x_1,y_1)|y_1]\big)^2>0.$$

Тогда, при $m,n\to\infty$ $\quad \dfrac{U_{m,n}-EU_{m,n}}{\sqrt{DU_{m,n}}}\xrightarrow{d}N(0,1).$

План действий таков:

- Доказать теорему 13.1.3.
- Затем вывести из нее теорему 13.1.2, ограничиваясь случаем однородных выборок, поскольку этот случай для нас более важен, и, поскольку в этом случае легко вычислить $E_0H_{m,n}$ и $D_0H_{m,n}$.
- Теорема 13.1.1 из теоремы 13.1.2 вытекает непосредственно (к тому же $DW_{m,n}=DH_{m,n}$).

По ходу доказательства теоремы 13.1.3 нам будет необходима так называемая

Т е о р е м а С л у ц к о г о. *Пусть случайная последовательность* $\{\xi_n,\ n>1\}$ *по распределению сходится к случайной величине* ξ; *пусть случайная последовательность* $\{\eta_n,\ n>1\}$ *сходится по вероятности к постоянной величине* C.

Тогда, при $n\to\infty$

(a) $\quad \xi_n+\eta_n\xrightarrow{d}\xi+C,$

(b) $\quad \xi_n\eta_n\xrightarrow{d}C\xi.$

§ 2. Доказательство теоремы 13.1.3

Для простоты предположим, что $Ef(x_i,y_j)=0$. Тогда $EU_{m,n}=0$. Введем случайные величины $\alpha(x_1)$ и $\beta(y_1)$:

$$\alpha(x_1)=E[f(x_1,y_1)|x_1],\qquad \beta(y_1)=E[f(x_1,y_1)|y_1].$$

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 $\Delta_{m,n} = \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n}[f(x_i,y_j) - \alpha(x_i) - \beta(y_j)]$.

Далее дробь $U_{m,n}/\sqrt{DU_{m,n}}$, предельное поведение которой есть предмет теоремы 13.1.3, когда $EU_{m,n} = 0$, представляем в виде:

$$\frac{U_{m,n}}{\sqrt{DU_{m,n}}} = \underbrace{\frac{n\sum\limits_{i=1}^{m}\alpha(x_i) + m\sum\limits_{j=1}^{n}\beta(y_j)}{\sqrt{D[n\sum\limits_{i=1}^{m}\alpha(x_i) + m\sum\limits_{j=1}^{n}\beta(y_j)]}}}_{\xi_{m,n}} \times$$

$$\times \underbrace{\sqrt{\frac{D[n\sum\limits_{i=1}^{m}\alpha(x_i) + m\sum\limits_{j=1}^{n}\beta(y_j)]}{DU_{m,n}}}}_{C_{m,n}} + \underbrace{\frac{\Delta_{m,n}}{\sqrt{DU_{m,n}}}}_{\eta_{m,n}}$$

или, коротко:

$$\frac{U_{m,n}}{\sqrt{DU_{m,n}}} = \xi_{m,n}\, C_{m,n} + \eta_{m,n}.$$

Для доказательства теоремы 13.1.3 достаточно показать, что

(a) $C_{m,n} \to 1$,

(b) $\xi_{m,n} \xrightarrow{d} N(0,1)$,

(c) $\eta_{m,n} \xrightarrow{P} 0$.

Затем применить теорему Слуцкого.

§ 3. Вычисление дисперсии U-статистик

Ключевую роль играет вычисление дисперсии U-статистики. Поэтому мы выделяем это в отдельный параграф.

Так как $Ef(x_i,y_j) = 0$, то

$$DU_{m,n} = EU_{m,n}^2 = \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{n}\sum_{j'=1}^{n} Ef(x_i,y_j)f(x_{i'},y_{j'}).$$

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. $Ef(x_i, y_j) = 0$, а случайные величины $f(x_i, y_j)$ и $f(x_{i'}, y_{j'})$ независимы, если индексы различны.

2) $\sum_2 = mn(n-1)Ef(x_1, y_1)f(x_1, y_2) = mn(n-1)D\alpha(x_1)$, т. к.

$$Ef(x_1, y_1)f(x_1, y_2) = EE[f(x_1, y_1)f(x_1, y_2)|x_1] =$$
$$= E\{E[f(x_1, y_1)|x_1]E[f(x_1, y_2)|x_1]\} = E\alpha(x_1)\alpha(x_1) = D\alpha(x_1),$$

ибо $E\alpha(x_1) = 0$.

3) $\sum_3 = mn(m-1)D\beta(y_1)$ по аналогичной причине.

4) $\sum_4 = mnE[f(x_1, y_1)]^2 = mnDf(x_1, y_1)$.

Поэтому

$$DU_{m,n} = mn(n-1)D\alpha(x_1) + nm(m-1)D\beta(y_1) + mnDf(x_1, y_1). \quad (13.3.1)$$

§ 4. Доказательство вспомогательных утверждений из параграфа 2

- Утверждение (a) очевидно, ибо

$$D[n\sum_{i=1}^{m}\alpha(x_i) + m\sum_{j=1}^{n}\beta(y_j)] = n^2mD\alpha(x_1) + m^2nD\beta(y_1).$$

- Утверждение (b) есть одна из форм центральной предельной теоремы. Её легко доказать методом характеристических функций, по аналогии с доказательством центральной предельной теоремы для суммы независимых одинаково распределенных случайных величин.

- Утверждение (c). Заметим, что $\Delta_{m,n}$ — это U-статистика с ядром $\tilde{f}(x,y)=f(x,y)-\alpha(x)-\beta(y)$, причем $E\tilde{f}(x_1,y_1)=0$. Выражение для $D\Delta_{m,n}$ дает формула (13.3.1), в которой $f,\ \alpha,\ \beta$ надо заменить на $\tilde{f},\ \tilde{\alpha},\ \tilde{\beta}$, где

$$\tilde{\alpha}(x_1) = E[\tilde{f}(x_1,y_1)|x_1],$$

$$\tilde{\beta}(y_1) = E[\tilde{f}(x_1,y_1)|y_1].$$

Легко видеть, что $\tilde{\alpha}(x_1)=0$, $\tilde{\beta}(y_1)=0$.
Поэтому $D\Delta_{m,n} = mnD\tilde{f}$. Теперь утверждение (c) есть следствие неравенства Чебышева для $\eta_{m,n}$, ибо

$$E\eta_{m,n}=0 \quad \text{и} \quad D\eta_{m,n} = \frac{D\Delta_{m,n}}{mn[nD\alpha + mD\beta + \text{Const}]} \to 0.$$

§ 5. Доказательство теоремы Слуцкого

Напомним о п р е д е л е н и е: случайная последовательность ξ_n, $n=1,2,\ldots$ *слабо*, или *по распределению* сходится к случайной величине ξ, если для любой непрерывной и ограниченной функции $f(.)$

$$Ef(\xi_n) \to Ef(\xi) \quad \text{при} \quad n\to\infty. \tag{13.5.1}$$

Ограничимся доказательством утверждения (a), т. к. (b) доказывается аналогично. Надо показать, что для любой непрерывной ограниченной функции $f(\cdot)$ при $n\to\infty$

$$E[f(\xi_n+\eta_n) - f(\xi+C)] \to 0. \tag{13.5.2}$$

Д о к а з а т е л ь с т в о. Заметим, что для любого $\varepsilon>0$ существует такое число $A>0$, что $P\{|\xi|>A\}<\varepsilon$. Так как $\xi_n \xrightarrow{d} \xi$, то для достаточно больших n

$$P\{|\xi_n|>A\} < 2\varepsilon. \tag{13.5.3}$$

Так как $\eta_n \xrightarrow{P} C$, то для указанного выше ε и любого фиксированного $\delta>0$ для достаточно больших n

$$P\{|\eta_n - C|>\delta\} < \varepsilon. \tag{13.5.4}$$

Окончательно δ мы выберем ниже, а пока положим $\delta\le 1$. Поскольку

$$E[f(\xi_n+\eta_n) - f(\xi+C)] =$$

$$= E[f(\xi_n + \eta_n) - f(\xi_n + C)] + E[f(\xi_n + C) - f(\xi + C)], \quad (13.5.5)$$

то для (13.5.2) достаточно показать, что каждое из слагаемых в (13.5.5) для достаточно больших n становится меньше любого наперед заданного числа.

Сходимость к нулю второго слагаемого очевидна: в качестве $f(x)$ в (13.5.1) надо взять $f(x + C)$, чтобы из $\xi_n \xrightarrow{d} \xi$ заключить, что $\xi_n + C \xrightarrow{d} \xi + C$.

Первое слагаемое представим в виде

$$E\{[f(\xi_n + \eta_n) - f(\xi_n + C)][I(|\xi_n| \le A) + I(|\xi_n| > A)] \times$$

$$\times [I(|\eta_n - C| \le \delta) + I(|\eta_n - C| > \delta)]\} = \quad (13.5.6)$$

$$= E\left\{[f(\xi_n + \eta_n) - f(\xi_n + C)]I(|\xi_n| \le A)I(|\eta_n - C| \le \delta)\right\} + R_n.$$

Через R_n обозначена сумма трех математических ожиданий, которые получаются при раскрытии скобок в левой части (13.5.6). В каждом из этих математических ожиданий есть либо сомножитель $I(|\xi_n| > A)$, либо $I(|\eta_n - C| > \delta)$, либо оба. Поэтому каждое из этих математических ожиданий при достаточно больших n можно оценить сверху числом $4\varepsilon \max\limits_x |f(x)|$, а их сумму R_n — числом $12\varepsilon \max\limits_x |f(x)|$. Например,

$$|E\{[f(\xi_n + \eta_n) - f(\xi_n + C)]I(|\xi_n| > A)I(|\eta_n - C| \le \delta)\}| \le$$

$$\le 2 \max_x |f(x)| I(|\xi_n| > A) \le 4\varepsilon \max_x |f(x)|.$$

Итак, для достаточно больших n

$$|R_n| \le 12\varepsilon \max_x |f(x)|. \quad (13.5.7)$$

Обратимся к главному слагаемому в (13.5.5) и заметим, что в нем $|\xi_n| \le A$, $|\eta_n - C| \le \delta$. Поэтому значения $|\xi_n + \eta_n - C| \le A + \delta \le A + 1$. Следовательно, случайная величина $\xi_n + \eta_n$ принадлежат компакту $K = [C - A - 1, C + A + 1]$. Так как функция $f(\cdot)$ из (13.5.2) непрерывна, то на компакте K она равномерно непрерывна. Это значит, что для выбранного выше $\varepsilon > 0$ существует число $\delta > 0$ такое, что $|f(u) - f(v)| < \varepsilon$, если $|u - v| < \delta$ и $u, v \in K$. Число δ зависит от ε и K, и именно таким мы выберем δ в (13.5.4). В обсуждаемом главном слагаемом из (13.5.5)

$$|(\xi_n + \eta_n) - (\xi_n + C)| < \delta, \quad \xi_n + \eta_n \in K, \ \xi_n + C \in K.$$

Поэтому $|f(\xi_n+\eta_n)-f(\xi_n+C)|<\varepsilon$.

В итоге получаем с учетом (13.5.7), что для достаточно больших n

$$\big|E\big[f(\xi_n+\eta_n)-f(\xi_n+C)\big]\big|<\varepsilon+12\varepsilon\max_x f(x).$$

За счет выбора $\varepsilon>0$ эта оценка может быть сделана сколь угодно малой. Теорема доказана. $\square$

§ 6. Применение теоремы 13.1.1 для вычислений критических значений

Теорема 13.1.1 бывает полезна для вычисления критических значений статистики $W_{m,n}$ при больших m,n. Чтобы воспользоваться теоремой 13.1.1, надо вычислить $D_0W_{m,n}$ (дисперсию при гипотезе, т. е. для однородных выборок $(x_1,\dots,x_m)$ и $(y_1,\dots,y_n)$). Мы вычислим $D_0H_{m,n}$, которая равна $D_0W_{m,n}$.

Воспользуемся результатом § 3, положив

$$f(x_1,y_1)=I(x_1<y_1)-EI(x_1<y_1)=I(x_1<y_1)-P\{x_1<y_1\}.$$

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

$$\alpha(x_i)=E\{[I(x_i<y_j)-1/2]|x_i\}=P\{x_i<y_j|x_i\}-1/2=$$
$$=1-P\{y_j<x_i|x_i\}-1/2=1/2-F(x_i).$$

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 рис. 13.6.1. Ясно, что $P\{F(X)<z\}=z$ для $z\in(0,1)$.

Получили, что при гипотезе (однородности) $\alpha(x_i)=1/2-U_i$, $\beta(y_j)=V_j-1/2$, где $U_1,\dots,U_m$, $V_1,\dots,V_n$ суть независимые случайные величины, равномерно распределенные на [0, 1]. Очевидно, что

$$E[\alpha(x_i)]^2=DU_i=\frac{1}{12},\quad E[\beta(y_j)]^2=DV_j=\frac{1}{12}.$$

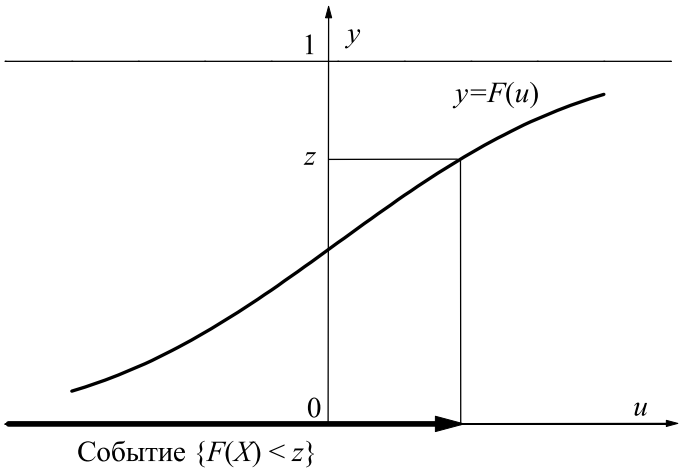


Рис. 13.6.1.

Поэтому

$$D_0H_{m,n}{=}D_0W_{m,n}{=}\frac{mn(n-1)}{12}+\frac{m(m-1)n}{12}+\frac{mn}{4}{=}\frac{mn(m+n+1)}{12},$$

ибо $D_0f(x_1,y_1){=}D_0I(x_1<y_1){=}P\{x_1<y_1\}\big(1-P\{x_1<y_1\}\big){=}1/4$.

Итак, для непрерывных однородных выборок теорема 13.1.1 дает:

$$W^*_{m,n}=\frac{W_{m,n}-n(n+m+1)/2}{\sqrt{mn(m+n+1)/12}}\xrightarrow{d}N(0,1)$$

при $m,\ n\to\infty$.

Пользоваться этим нормальным приближением (для вероятностей, не слишком близких к 0 или 1) можно при $m,n\geq 10$. Центральная предельная теорема не дает нам оценок для скорости сходимости. Сказанное правило подтверждается сравнением точного распределения $W_{m,n}$ для, скажем $m=10,\ n=10$, которое можно найти в таблицах, и его нормальной аппроксимации. (И убежденностью в том, что для бóльших m, n аппроксимация будет еще лучше).

Лекция 14. Метод наибольшего правдоподобия

§ 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 $\{X = x\}$. Другая частая возможность: $p(x, \theta)$ — это плотность относительно меры Лебега.

О п р е д е л е н и е 14.1.1. *Правдоподобием* значения параметра θ называют (случайную величину) $p(X, \theta)$.

О п р е д е л е н и е 14.1.2. То значение параметра θ, для которого правдоподобие принимает наибольшее значение, называют *оценкой наибольшего правдоподобия* (параметра θ):

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} p(X, \theta). \tag{14.1.1}$$

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 $\theta^0 \in \Theta$. В этом случае упомянутое выше правдоподобие $p(X, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$.

Относительно оценки (14.1.1) 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. Этот результат позволит нам указать для неизвестного параметра θ^0 асимптотические доверительные интервалы.

О п р е д е л е н и е 14.1.3. Оценка $t = t(X)$ параметра θ называется *состоятельной*, если $t(X) \xrightarrow{P} \theta^0$ при $n \to \infty$.

О п р е д е л е н и е 14.1.4 (способ выражения). Говорят, что случайная величина ξ_n (на самом деле — последовательность случайных величин $\xi_n,\ n = 1, 2, \ldots$) распределена *асимптотически нормально* с параметрами a_n, σ_n^2, если

$$\frac{\xi_n - a_n}{\sigma_n} \xrightarrow{d} N(0,1) \ \text{ при } \ n \to \infty. \tag{14.1.2}$$

При этом a_n называют *асимптотическим математическим ожиданием* ξ_n (асимптотическим средним), а σ_n^2 — *асимптотической дисперсией* ξ_n.

Следует отметить, что математическое ожидание ξ_n не только может не совпадать с a_n, но и вообще не существовать. То же относится и к дисперсии ξ_n. Наконец, из приведенного определения видно, что последовательность (a_n, σ_n), $n = 1, 2, \ldots$ определена не однозначно.

Для (14.1.2) употребительна и более выразительная запись

$$\xi_n \overset{\text{ac.}}{\sim} N(a_n, \sigma_n^2).$$

В этом разделе мы встретимся с асимптотически нормальными оценками $\hat{\theta}_n$ параметра θ, для которых

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2(\theta)) \ \text{ при } n \to \infty, \ \text{ или } \ \hat{\theta}_n \overset{\text{ac.}}{\sim} N\left(\theta, \frac{\sigma^2(\theta)}{n}\right).$$

§ 2. Состоятельность оценок наибольшего правдоподобия

2.1. Лемма (вариант т. н. неравенства теории информации)

Л е м м а 14.2.1. *Пусть $f(\cdot)$, $g(\cdot)$ — две плотности вероятности. Тогда*

$$\int f(x) \ln f(x) dx \geq \int f(x) \ln g(x) dx, \tag{14.2.1}$$

причем равенство возможно, только если $f = g$ почти всюду.

С о г л а ш е н и я:

- Для интегралов допускается значение $-\infty$.

- Будем считать, что $\int\limits_A f(x)\ln g(x)dx = 0$, если $f(x) = 0$ для $x \in A$, вне зависимости от значений $g(\cdot)$.

Д о к а з а т е л ь с т в о. Достаточно показать, что

$$\int f(x)\ln\frac{g(x)}{f(x)}dx \le 0.$$

Заметим, что $\ln(1+x) \le x$ для $x \ge -1$. (См. на рис. 14.2.1 графики функций $y = x$ и $y = \ln(1+x)$).

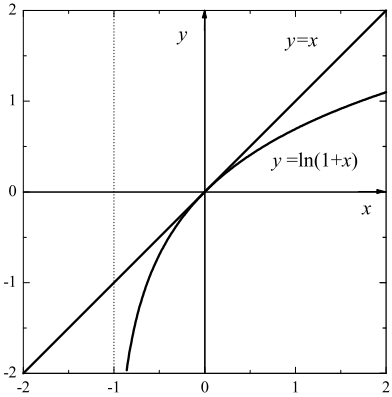


Рис. 14.2.1. Графики функций $y = x$ и $y = \ln(1+x)$

Рассмотрим множество $A = \{x : f(x) > 0\}$. Для $x \in A$:

$$\ln\frac{g(x)}{f(x)} \equiv \ln\left[1 + \left(\frac{g(x)}{f(x)} - 1\right)\right] \le \frac{g(x)}{f(x)} - 1.$$

Умножив обе части неравенства на $f(\cdot)$, интегрируем:

$$\int f(x)\ln\frac{g(x)}{f(x)}dx = \int\limits_A f(x)\ln\frac{g(x)}{f(x)}dx \le$$

$$\le \int\limits_A [g(x) - f(x)]dx = \int\limits_A g(x)dx - \int\limits_A f(x)dx \le 0,$$

т. к. $\int\limits_A g(x)dx \le \int g(x)dx = 1$. Ч. т. д. $\square$.

2.2. 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 $\prod\limits_{i=1}^{n} f(x_i, \theta)$, а оценка наибольшего правдоподобия (14.1.1) есть

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f(x_i, \theta)$$

или

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \log f(x_i, \theta) \right]. \tag{14.2.2}$$

(Точка экстремума не изменяется при переходе от функции к её логарифму и при умножении на положительное число).

В силу закона больших чисел при $n \to \infty$

$$\frac{1}{n} \sum_{i=1}^{n} \log f(x_i, \theta) \xrightarrow{P} E_0 \log f(x_1, \theta), \tag{14.2.3}$$

где E_0 означает усреднение по плотности $f(x, \theta^0)$, где θ^0 — истинное значение θ. Поэтому естественно ожидать, что

$$\arg\max_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \log f(x_i, \theta) \right] \xrightarrow{P} \arg\max_{\theta \in \Theta} E_0 \log f(x_1, \theta).$$

Согласно лемме 14.2.1, справедливо (14.2.1). Это неравенство для $g(x) = f(x, \theta)$, $f(x) = f(x, \theta^0)$ дает:

$$E_0 \log f(x_1, \theta) = \int [\log f(x, \theta)] f(x, \theta^0) dx \leq \int [\log f(x, \theta^0)] f(x, \theta^0) dx.$$

Следовательно, $\arg\max\limits_{\theta \in \Theta} E_0 \log f(x_1, \theta) = \theta^0$.

Доказательство сходимости $\hat{\theta}_n \xrightarrow{P} \theta^0$ надо проводить, учитывая свойства $E_0 \log f(x_1, \theta)$ как функции θ, $\theta \in \Theta$. Если эта функция и $f(x, \theta)$ непрерывны по θ, обычно удается такой план:

- Показать, что сходимость в (14.2.3) равномерна по θ на компакте, содержащем θ^0.

- В этом случае можно утверждать, что существует последовательность локальных экстремумов функции $\hat{\theta}_n$ из (14.1.1) или (14.2.2), по вероятности сходящаяся к θ^0:

$$\hat{\theta}_n \xrightarrow{P} \theta^0 \quad \text{при} \quad n \to \infty. \tag{14.2.4}$$

2.3. Доказательство сходимости $\hat{\theta}_n \xrightarrow{P} \theta^0$ для одномерного случая

В одномерном случае доказательство проще. Предположим, что $\log f(x, \theta)$ при всяком x непрерывно зависит от $\theta \in \Theta$, где Θ — открытое множество, $\Theta \subset R^1$.

Чтобы доказать (14.2.4), мы покажем, что (локальный) экстремум функции

$$\frac{1}{n} \sum_{i=1}^{n} \log f(x_i, \theta)$$

лежит со сколь угодно близкой к 1 вероятностью — при достаточно больших n — внутри интервала $(\theta^0 - h,\ \theta^0 + h)$, где h — произвольное число.

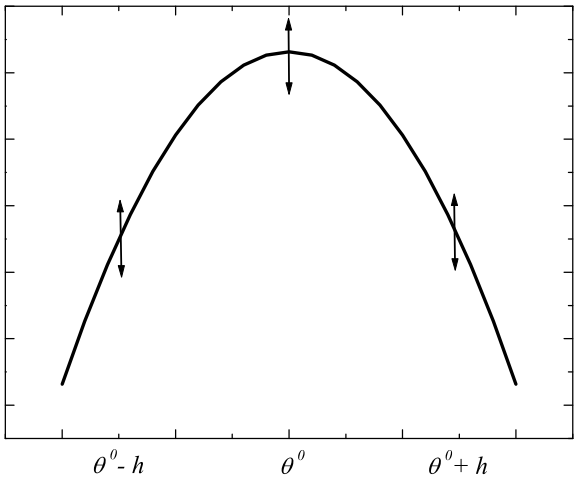


Рис. 14.2.2. График функции $y = E_0 \log f(x_1, \theta)$

Так как

$$E_0 \log f(x_1, \theta^0) > E_0 \log f(x_1, \theta^0 \pm h),$$

то можно подобрать такое $\varepsilon > 0$, что

$$E_0 \log f(x_1, \theta^0) - \varepsilon > E_0 \log f(x_1, \theta^0 \pm h) + \varepsilon.$$

Для произвольного фиксированного $\delta > 0$, в силу упомянутого закона больших чисел (14.2.3), для достаточно больших n выполняются неравенства:

$$P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} \log f(x_i, \theta^0) - E_0 \log f(x_1, \theta^0)\Big| < \varepsilon\Big\} > 1 - \delta,$$

$$P\Big\{\Big|\frac{1}{n}\sum_{i=1}^{n} \log f(x_i, \theta^0 \pm h) - E_0 \log f(x_1, \theta^0 \pm h)\Big| < \varepsilon\Big\} > 1 - \delta.$$

Поэтому

$$P\Big\{\frac{1}{n}\sum_{i=1}^{n} \log f(x_i, \theta^0) > \frac{1}{n}\sum_{i=1}^{n} \log f(x_i, \theta^0 \pm h)\Big\} > 1 - 3\delta.$$

Поэтому (при достаточно больших n) экстремум (локальный) функции правдоподобия из (14.1.1) лежит в сколь угодно узкой окрестности точки θ^0. Поэтому последовательность этих локальных экстремумов сходится (по вероятности) к θ^0, что и требовалось доказать. $\square$

З а д а ч а. Исследуйте на максимум функцию $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \log f(x_i, \theta)$ и сопоставьте её с $E_0 \log f(x_1, \theta)$ для нормального семейства $N(a, \sigma^2)$ и для семейства равномерных распределений на отрезке $[\,0, \theta\,]$, где $\theta \in (0, \infty)$.

§ 3. Асимптотическая нормальность оценок наибольшего правдоподобия (по выборке из регулярного семейства)

3.1. Одномерный случай

Пусть $X = (x_1, \ldots, x_n)$ — выборка из распределения с плотностью (вероятностью) $p\,(x, \theta)$, $\theta \in \Theta \subset R^1$. (После того, как мы

закончим исследование одномерного параметра θ, мы обсудим, какие изменения надо сделать, когда $\theta \in \Theta \subset R^r$). Множество Θ будем считать открытым.

В рассматриваемом случае оценка наибольшего правдоподобия есть решение уравнения правдоподобия

$$\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log p\,(x_i,\theta)=0. \tag{14.3.1}$$

Считая, что $p\,(x,\theta)$ трижды дифференцируема по θ, предположим, что существует функция $M(x)$ такая, что

- для любого $\theta\in\Theta$: $\left|\dfrac{\partial^3}{\partial\theta^3}\log p\,(x,\theta)\right|<M(x)$,
- причем $E_\theta M(x_1)<\infty$ для всех $\theta\in\Theta$.
- $0<i(\theta^{\,0})<\infty$, где $i(\theta^{\,0})=E_0\left[\dfrac{\partial}{\partial\theta}\log p\,(x_1,\theta^{\,0})\right]^2$.

В дальнейшем, ради краткости будем писать

$$l(x,\theta)=\frac{\partial}{\partial\theta}\log p\,(x,\theta).$$

Исследование уравнения правдоподобия. Введем новую переменную τ, положив $\theta=\theta^{\,0}+\dfrac{\tau}{\sqrt{n}}$. Теперь уравнение правдоподобия (14.3.1) имеет вид

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}l(x_i,\theta^{\,0}+\frac{\tau}{\sqrt{n}})=0. \tag{14.3.2}$$

Разлагаем левую часть (14.3.2) по формуле Тейлора в точке $\theta^{\,0}$. Получаем:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}l(x_i,\theta^{\,0})+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}l'_\theta(x_i,\theta^{\,0})\frac{\tau}{\sqrt{n}}+\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}l''_{\theta\theta}(x_i,\tilde\theta_n)\left(\frac{\tau}{\sqrt{n}}\right)^2, \tag{14.3.3}$$

где $\tilde\theta_n$ — некая промежуточная точка между $\theta^{\,0}$ и θ.

Заметим, что если мы ограничим область изменения переменной τ произвольным компактом, т. е. предположим, 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

$$\frac{1}{n} \sum_{i=1}^{n} M(x_i) \xrightarrow{P} E_\theta M(x_1).$$

Сопоставим решение уравнения (14.3.2), левая часть которого представлена в форме (14.3.3), и решение уравнения

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} l(x_i, \theta^0) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l'_\theta(x_i, \theta^0) \frac{\tau}{\sqrt{n}} = 0. \tag{14.3.4}$$

(Левая часть как в (14.3.3), но без третьего слагаемого).

Решение (14.3.4) очевидно:

$$\tau_n^* = \frac{\dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} l(x_i, \theta^0)}{-\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} l'_\theta(x_i, \theta^0)}. \tag{14.3.5}$$

При этом легко увидеть, что при $n \to \infty$

$$\tau_n^* \xrightarrow{d} N\big(0, [\, i(\theta^0)]^{-1}\big). \tag{14.3.6}$$

Здесь $i(\theta^0)$ — количество информации (по Фишеру) о θ, содержащейся в одном наблюдении x_1.

Действительно, числитель (14.3.5) есть сумма независимых случайных величин $\dfrac{1}{\sqrt{n}} \dfrac{\partial}{\partial\theta} \displaystyle\sum_{i=1}^{n} \log p\,(x_i, \theta^0)$, $i = \overline{1, n}$. При обсуждении неравенств Крамера-Рао мы отметили, что

$$E_\theta \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \log p\,(x_i, \theta) = 0 \quad \text{для} \quad \theta \in \Theta,$$

и что

$$E_\theta \left[\frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log p\,(x_i, \theta) \right]^2 = n\, i(\theta).$$

Поэтому, по центральной предельной теореме, числитель (14.3.5) по распределению сходится к $N(0, i(\theta^0))$, когда $n \to \infty$.

Знаменатель (14.3.5) по закону больших чисел сходится (по вероятности) к $-E_\theta l'_\theta(x_1, \theta)$, где $\theta{=}\theta^0$. Мы (при упомянутых выше обсуждениях) отмечали, что

$$E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log p\,(x_1, \theta) \right] = -i(\theta).$$

Поэтому по теореме Слуцкого выполнено (14.3.6).

Остается убедиться, что решение уравнения (14.3.2) асимптотически эквивалентно решению уравнения (14.3.4) — эквивалентно в том смысле, что при $n \to \infty$ разность между решениями стремится к нулю (по вероятности).

Мы уже отмечали, что левые части (14.3.2) и (14.3.4) отличаются бесконечно мало (и притом равномерно по τ), когда $|\tau| < \text{Const}$ — меньше произвольной постоянной.

Рассмотрим график левой части (14.3.4) как функцию от τ: $y = \psi_n(\tau)$. Для достаточно больших n график левой части (14.3.2), 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 (14.3.2) найдется решение $\hat{\tau}_n$, такое, что $\hat{\tau}_n - \tau_n^* \xrightarrow{P} 0$ — при том дополнительном условии, что уравнение (14.3.4) имеет решение, принадлежащее компакту $\{\tau : |\tau| < C\}$.

Остается сделать последнее замечание, чтобы завершить исследование (14.3.2). Так как τ_n^* (решение (14.3.4)) асимптотически нормально, можно выбрать упомянутый выше компакт $\{\tau : |\tau| < C\}$ так, чтобы для произвольно выбранного $\delta > 0$

$$P\{|\tau_n^*| < C\} > 1 - \delta \tag{14.3.7}$$

для достаточно больших n.

Таким образом, со сколь угодно близкой к 1 вероятностью, уравнение (14.3.2) имеет корень на выбранном компакте (для достаточно больших n), и этот корень сколь угодно близок к τ_n^*.

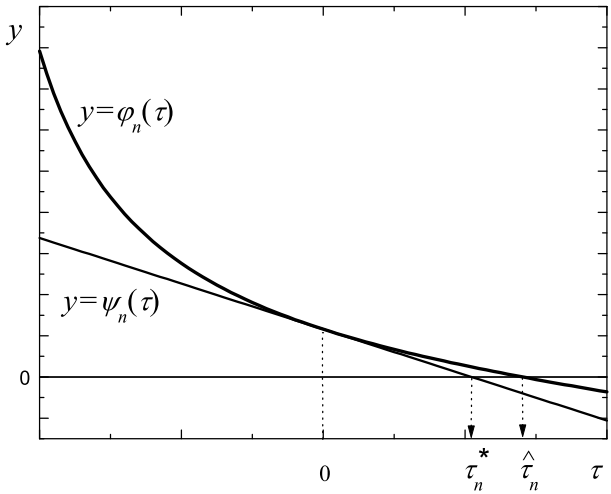


Рис. 14.3.1. Графики функций $y = \varphi_n(\tau)$ и $y = \psi_n(\tau)$

Поэтому корень $\hat{\tau}_n$ распределен так же, как и τ_n^*, т. е. при $n \to \infty$

$$\hat{\tau}_n \xrightarrow{d} N\big(0, \big[i(\theta^0)\big]^{-1}\big). \tag{14.3.8}$$

Вернемся к переменной $\theta = \theta^0 + \dfrac{\tau}{\sqrt{n}}$, $\hat{\theta}_n = \theta^0 + \dfrac{\hat{\tau}}{\sqrt{n}}$, $\hat{\tau}_n = \sqrt{n}(\hat{\theta}_n - \theta^0)$. Утверждение (14.3.8) означает, что при $n \to \infty$

$$\sqrt{n}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N\big(0, [i(\theta^0)]^{-1}\big). \tag{14.3.9}$$

Следовательно, мы доказали, что (при наложенных выше условиях на $p\,(x, \theta)$) существует решение (точнее, последовательность решений) уравнения правдоподобия (14.3.1) $\hat{\theta}_n$, сходящееся к θ^0 и распределенное асимптотически нормально с параметрами θ^0 и $\big[\, n\, i(\theta^0)\,\big]^{-1}$.

3.2. Многомерный случай

Для многомерного параметра θ всё исследование проходит также, как и для одномерного, с очевидными изменениями.

- Уравнение правдоподобия (14.1.1) теперь — векторное (т. е. (14.1.1) 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 $i(\theta)$ занимает матрица информации $\mathcal{J}(\theta)$.

Окончательный результат принимает вид

$$\hat{\theta}_n \xrightarrow{d} N(\theta^0, n^{-1}\mathcal{J}^{-1}(\theta^0)). \tag{14.3.10}$$

3.3. Асимптотически эффективные оценки

Несмещенные оценки параметра, для которых неравенство Крамера-Рао обращается в равенство, ранее были названы *эффективными*. Для эффективных оценок дисперсия (матрица ковариаций в многопараметрическом случае) равна обратному количеству информации (обратной матрице информации).

Обратим внимание на сходство с эффективностью результатов (14.3.9) и (14.3.10): для оценки наибольшего правдоподобия (в условиях регулярности) асимптотическое математическое ожидание совпадает с оцениваемым параметром, а её асимптотическая дисперсия — с обратным количеством информации. В связи с этим сходством принято следующее определение.

О п р е д е л е н и е 14.3.1. Асимптотически нормальную оценку $\hat{\theta}_n$ (не обязательно оценку наибольшего правдоподобия) параметра θ называют *асимптотически эффективной*, если её асимптотические параметры можно выбрать так, чтобы выполнялось (14.3.9) (или (14.3.10) — в многопараметрическом случае).

И хотя неверно утверждение, что асимптотическая дисперсия не может быть меньше, чем обратная информация (аналог неравенства Крамера-Рао), всё же асимптотически эффективные оценки можно считать наилучшими. (Исследований на эту тему было очень много. Окончательные формулировки, на мой взгляд, еще не найдены. Об асимптотической эффективности см., например, Э. Леман, "Теория точечного оценивания": Пер. с англ.– М.: Наука, 1991 – гл. 6).

§ 4. Одношаговые оценки

Результаты § 3 говорят о существовании асимптотически эффективных оценок наибольшего правдоподобия $\hat{\theta}_n$, но не говорят, как выделить такую оценку среди решений уравнений правдоподобия, если решение не единственно. Это серьезный недостаток. Затем само решение уравнения правдоподобия может оказаться трудной задачей. Предлагаемый ниже метод улучшения оценки позволяет обойти оба эти затруднения. Сохраним в силе все сделанные в § 3 предположения.

Предположим, что для параметра θ (для простоты, одномерного) мы располагаем $\sqrt{n}$-состоятельной оценкой, скажем, оценкой $\theta_n^{(1)}$. (Оценку $\theta_n^{(1)}$ называют *$\sqrt{n}$-состоятельной*, если случайная величина $\sqrt{n}(\theta_n^{(1)} - \theta)$ ограничена по вероятности). Найти такую оценку обычно не составляет труда. В этом случае мы можем, не решая уравнения правдоподобия, а используя метод Ньютона, получить для параметра θ новую оценку, скажем, $\theta_n^{(2)}$, эквивалентную асимптотически эффективной оценке $\hat{\theta}_n$.

Ради краткости на время положим $\varphi(\theta) = \dfrac{1}{n}\sum_{i=1}^{n} l(x_i, \theta)$ и запишем уравнение правдоподобия в форме

$$\varphi(\theta) = 0. \tag{14.4.1}$$

Так как и $\theta_n^{(1)}$, и $\hat{\theta}_n$ сходятся к θ, то $\theta_n^{(1)}$ близко к $\hat{\theta}_n$, корню уравнения (14.4.1). При этом

$$\theta_n^{(1)} - \hat{\theta}_n = O_P\left(\frac{1}{\sqrt{n}}\right). \tag{14.4.2}$$

Мы можем рассматривать $\theta_n^{(1)}$ как приближенное решение уравнения (14.4.1). Метод Ньютона позволяет получить новое приближенное решение $\theta_n^{(2)}$, лучшее, чем $\theta_n^{(1)}$ (см. рис. 14.4.1). Идея метода — замещение функции $y = \varphi(\theta)$ ее касательной (в точке $\theta = \theta_n^{(1)}$)

$$y = \varphi(\theta_n^{(1)}) + \varphi'(\theta_n^{(1)})(\theta - \theta_n^{(1)})$$

и затем решения линейного уравнения

$$\varphi(\theta_n^{(1)}) + \varphi'(\theta_n^{(1)})(\theta - \theta_n^{(1)}) = 0.$$

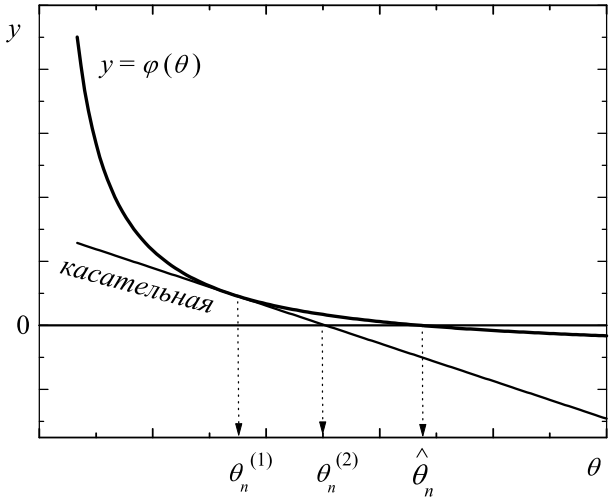


Рис. 14.4.1. Функция $y = \varphi(\theta)$ и касательная к ней в точке $\theta_n^{(1)}$

Это дает $\theta_n^{(2)} = \theta_n^{(1)} - \dfrac{\varphi(\theta_n^{(1)})}{\varphi'(\theta_n^{(1)})}$. Легко видеть, что

$$\theta_n^{(2)} - \hat{\theta}_n = \frac{(\theta_n^{(1)} - \hat{\theta}_n)\varphi'(\theta_n^{(1)}) - \varphi(\theta_n^{(1)})}{\varphi'(\theta_n^{(1)})}. \tag{14.4.3}$$

Разлагая по формуле Тейлора числитель этого выражения, получим, что

$$\theta_n^{(2)} - \hat{\theta}_n = (\theta_n^{(1)} - \hat{\theta}_n)^2 \frac{\varphi''(\tilde{\theta}) - \frac{1}{2}\varphi''(\tilde{\tilde{\theta}})}{\varphi'(\theta_n^{(1)})}, \tag{14.4.4}$$

где $\tilde{\theta}$, $\tilde{\tilde{\theta}}$ — некоторые значения, промежуточные между $\theta_n^{(1)}$ и $\hat{\theta}_n$.

Принятые в § 3 условия регулярности обеспечивают ограниченность по вероятности второго сомножителя в (14.4.4). В силу (14.4.2) получаем $\theta_n^{(2)} - \hat{\theta}_n = O_P\left(\dfrac{1}{n}\right)$, что и означает эквивалентность одношаговой оценки $\theta_n^{(2)}$ и асимптотически эффективной оценки $\hat{\theta}_n$.

Лекция 15. Устойчивые оценки

§ 1. Функция влияния

Начнем с примера. Пусть $x_1, \ldots, x_n$ — выборка из $N(a, \sigma^2)$. По этой выборке мы хотим оценить a. Как известно, в определенном смысле наилучшей оценкой a является выборочное среднее $\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$: среди всех несмещенных оценок параметра a оценка $\bar{x}$ имеет наименьшую дисперсию. В этом отношении $\bar{x}$ превосходит, например, выборочную медиану $\mu_n = \mathrm{med}(x_1, \ldots, x_n)$, тоже несмещенную оценку a.

Теперь предположим, что к упомянутой выборке добавилось некое постороннее число z. Такие посторонние, ошибочные данные нередко присутствуют в массивах наблюдений, засоряя их. Вспомним хотя бы выборку из книги А. Хальда (200 измерений заклепок), которую я, как пример, приводил на первой лекции: в русском издании одно из двухсот чисел приведено с опечаткой. Из-за этого оно далеко отступило от основного массива. Такие грубо ошибочные данные, не являющиеся результатом случайного выбора из интересующей нас генеральной совокупности, часто называют "*выбросами*".

Посмотрим на упомянутых примерах, как присутствие выбросов сказывается на качестве оценки — в данном случае на оценивании неизвестного a. Основой для оценивания теперь служит совокупность $x_1, \ldots, x_n, z$. Ясно, что выборочное среднее в этих условиях равно:

$$\frac{n}{n+1}\,\bar{x} + \frac{1}{n+1}\,z.$$

Видно, что эта оценка может оказаться весьма далекой от a, если число $|z|$ достаточно велико. В этом примере и при этом способе оценивания даже единичный выброс (единичное наблюдение) может сколь угодно сильно повлиять на результат. В этом смысле среднее арифметическое не является устойчивой оценкой a (центра распределения).

Иные свойства у выборочной медианы. Пусть, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n-1)} \le x_{(n)}$ — вариационный ряд и, для определенности, $n = 2m$. Тогда $\mu_n = \dfrac{x_{(m)} + x_{(m+1)}}{2}$. Предположим, что $z > x_{(m+1)}$. Тогда выборочная медиана μ_{n+1} "засоренной" совокупности равна $\mathrm{med}(x_1, \ldots, x_n, z) = x_{(m+1)}$. Видно, что разница между μ_{n+1}

и μ_n мала, как бы ни было велико z. При больших n эта разница пренебрежимо мала, ибо $\mu_{n+1} - \mu_n = O_P\left(\frac{1}{n}\right)$, в то время как точность статистического оценивания по порядку величины не выше, чем $1/\sqrt{n}$. Мы получили, что выборочная медиана как оценка центра симметричного распределения устойчива по отношению в выбросам. По-английски устойчивые оценки называют *robust*. Иногда их так называют и по-русски: *робастные*.

Рассмотрим теперь вероятностное — в противоположность выборочному — воплощение той же идеи. Для простоты будем рассматривать распределения на прямой. В этом случае удобным средством для описания распределения служит его функция распределения, скажем $F(.)$. Предположим, что нас интересует некоторая характеристика этого распределения, т. е. некоторый функционал $T(.)$. Для распределения F это $T(F)$. Далее предположим, что к распределению F в некоторой доле λ, $0 \le \lambda \le 1$ "примешано" распределение вероятностей, как и ранее, сосредоточенное в некоторой точке z. Пусть Δ_z обозначает функцию этого вырожденного распределения:

$$\Delta_z(x) = 0 \quad \text{для } x \le z, \quad \Delta_z(x) = 1 \quad \text{для } \quad x > z.$$

Функция распределения описанной смеси распределений равна

$$(1-\lambda)F(x) + \lambda\Delta_z(x). \tag{15.1.1}$$

Вопрос: как скажется присутствие засорения на функционале $T(.)$? На распределении (15.1.1) функционал $T(.)$ принимает значение $T[(1-\lambda)F+\lambda\Delta_z]$. Отнесенное к λ влияние на $T(.)$ сказанного засорения равно

$$\lambda^{-1}\{T[(1-\lambda)F + \lambda\Delta_z] - T(F)\}.$$

Это влияние интересует нас, в первую очередь, при малых λ. Отсюда вытекает следующее

О п р е д е л е н и е 15.1.1. *Функцией влияния* (влияния на функционал T в точке F засорения z) называют предел:

$$h(z) = \lim_{\lambda \to +0} \lambda^{-1}\{T[(1-\lambda)F + \lambda\Delta_z] - T[F]\}, \tag{15.1.2}$$

если этот предел существует.

Для функции влияния (15.1.2) употребляют и более сложное обозначение $IF(z;T,F)$, расшифровывающее смысл этого понятия (выше в определении 15.1.1 приведен в скобках; IF — начальные буквы *Influence Function*).

Вернемся к нормальному распределению $N(a,\sigma^2)$, выборки из которого мы рассматривали выше. Параметр a можно представить с помощью функционалов "математическое ожидание" и "медиана распределения". Первый — это

$$T(F)=\int x\,dF(x),$$

второй — это $T(F)=t$, где t — решение уравнения $F(t)=\dfrac{1}{2}$.

Для функционала $\int x\,dF(x)$ вычисление функции влияния несложно: формула (15.1.2) сразу дает:

$$h(z)=z-\int x\,dF(x). \tag{15.1.3}$$

Мы видим, что функция влияния неограниченно возрастает (убывает) с ростом (убыванием) z, что и отражает неустойчивость этого функционала к засорениям.

Функция влияния для медианы требует несколько бóльших вычислений. Обозначим через t_λ медиану распределения (15.1.1). При этом медиана распределения F получает обозначение t_0, а искомая функция влияния

$$h(z)=\frac{\partial t_\lambda}{\partial\lambda}\Big|_{\lambda=0}.$$

Так как t_λ как функция от λ определяется неявно с помощью уравнения

$$t_\lambda:\quad (1-\lambda)F(t)+\lambda\Delta_z(t)=\frac{1}{2}, \tag{15.1.4}$$

то и производную её необходимо искать как производную неявной функции.

Заметим, что в точке $z=t_0$ функция влияния $h(z)=0$. Поэтому далее рассмотрим случай $z\neq t_0$. Обратим внимание, что для фиксированного z и достаточно малых λ

$$\Delta_z(t_\lambda)=\frac{1}{2}[\,\mathrm{sign}(t_0-z)+1\,].$$

Подставим $t = t_\lambda$ в уравнение (15.1.4). Получим тождество, которое при достаточно малых $\lambda > 0$ имеет вид:

$$(1-\lambda)F(t_\lambda) + \frac{\lambda}{2}[\,\mathrm{sign}(t_0 - z) + 1\,] = \frac{1}{2}.$$

Вычислив производную по λ и положив затем $\lambda = 0$, получим

$$-F(t_\lambda)\Big|_{\lambda=0} + (1-\lambda)F'(t_\lambda)\frac{\partial t_\lambda}{\partial \lambda}\Big|_{\lambda=0} + \frac{1}{2}\,\mathrm{sign}(t_0 - z) + \frac{1}{2} = 0.$$

Отсюда

$$h(z) = \frac{\mathrm{sign}(z - t_0)}{2F'(t_0)}, \tag{15.1.5}$$

если $F'(t_0) > 0$. Формула верна и для $z = t_0$.

Мы видим, что как функция z функция влияния ограничена. Это означает уже знакомое нам свойство медианы: этот функционал устойчив к засорениям.

§ 2. М-оценки

Мы убедились, что если есть опасность засорения наблюдений, то лучше пользоваться устойчивыми оценками. Как такие оценки получить? До сих пор мы могли только, уже имея некоторую оценку, определить, устойчива она или нет, но не предложить устойчивую оценку. Да и вообще, нам пока что известен только один универсальный метод оценивания — метод наибольшего правдоподобия. (К слову: каковы функции влияния для оценок наибольшего правдоподобия, мы скоро узнаем). Сейчас мы введем метод оценивания, при котором функцию влияния оценки можно указать заранее (по крайней мере с точностью до постоянного множителя). Этот метод называется М-*оцениванием*; оценки, полученные по этому методу — М-*оценками*. Для М-оценивания упомянутый метод наибольшего правдоподобия является частным случаем.

Рассмотрим этот метод на примере выборки $x_1, \ldots, x_n$ из распределения с функцией $F(x, \theta)$ и плотностью (вероятностью) $f(x, \theta)$, где $\theta \in \Theta$, Θ — заданное открытое множество, $\Theta \subset R^1$. Истинное значение параметра обозначим через θ^0, $\theta^0 \in \Theta$.

М-оценки введем по аналогии с оценками наибольшего правдоподобия. Вспомним, что оценку наибольшего правдоподобия можно рассматривать как решение уравнения правдоподобия (точнее,

как одно из его решений)

$$\frac{1}{n}\sum_{i=1}^{n} l(x_i,\theta)=0, \quad \text{где} \quad l(x,\theta)=\frac{\partial}{\partial\theta}\log f(x,\theta).$$

Эту оценку можно рассматривать как значение на выборочной функции F_n функционала T, где $T(F)=t$ — решение уравнения

$$\int l(x,t)\,dF(x)=0. \tag{15.2.1}$$

Для распределения $F(x,\theta^0)$ функционал $T(F)$ есть решение уравнения

$$\int l(x,t\,)dF(x,\theta^0)=0.$$

Мы уже знаем, что $t=\theta^0$ является его решением, каково бы ни было истинное значение $\theta^0\in\Theta$. Иначе это можно сказать так:

$$\int l(x,\theta)\,dF(x,\theta)=0 \ \text{ для всякого } \ \theta\in\Theta. \tag{15.2.2}$$

Уравнение (15.2.1) служит образцом, по которому вводят М-оценки: с помощью некоторой функции $\psi(x,t)$ определяют функционал $T(F)=t$ как решение уравнения

$$\int \psi(x,t)\,dF(x)=0. \tag{15.2.3}$$

Если мы хотим, чтобы М-оценка оценивала именно параметр θ, то при выборе функции $\psi(x,t)$, по сходству с (15.2.2), надо потребовать, чтобы

$$\int \psi(x,\theta)\,dF(x,\theta)=0 \ \text{ для всех } \ \theta\in\Theta. \tag{15.2.4}$$

В таком случае среди решений уравнения

$$\int \psi(x,t)\,dF(x,\theta^0)=0$$

будет и $t=\theta^0$. Итак, выбор функции $\psi(x,t)$ должен быть согласован с $F(x,\theta)$.

Для введенной выше выборки из $F(x,\theta^0)$ М-оценка — это решение уравнения

$$\sum_{i=1}^{n}\psi(x_i,\theta)=0.$$

Ф у н к ц и я в л и я н и я для М-оценки. Функцию засоренного распределения (15.1.1) обозначим через $F_\lambda(x)$,

$$F_\lambda(x)=(1-\lambda)F(x,\theta)+\lambda\Delta_z(x).$$

Значение М-функционала $T(.)$ на F_λ, т. е. $T(F_\lambda)=\theta_\lambda$, определим как решение уравнения

$$t:\quad\int\psi(x,t)\,dF_\lambda(x)=0.$$

В этих обозначениях функция влияния $h(z)=\dfrac{\partial\theta_\lambda}{\partial\theta}\Big|_{\lambda=0}$. Как уже было однажды, θ_λ определена как неявная функция. Упомянутую производную по λ получим, дифференцируя по λ тождество

$$\int\psi(x,\theta_\lambda)\,dF_\lambda(x)=0$$

и переходя затем к пределу при $\lambda\to 0$. Производная по λ:

$$\int\frac{\partial}{\partial\lambda}\psi(x,\theta_\lambda)\frac{\partial\theta_\lambda}{\partial\lambda}\,dF_\lambda(x)+\int\psi(x,\theta_\lambda)\frac{\partial}{\partial\lambda}\,[\,dF_\lambda(x),\,]=0.$$

Положив $\lambda=0$ и заметив, что $\theta_\lambda\Big|_{\lambda=0}=\theta,\ \ F_\lambda(x)\Big|_{\lambda=0}=F$, получим

$$h(z)\int\psi_t'(x,\theta)\,dF(x,\theta)-\int\psi(x,\theta)\,dF(x,\theta)+\int\psi(x,\theta)\,d\,\Delta_z(x)=0.$$

В силу (15.2.4) получаем отсюда, что:

$$h(z)=\frac{\psi(z,\theta)}{-\int\psi_t'(x,\theta)\,dF(x,\theta)}.\qquad(15.2.5)$$

Для оценки наибольшего правдоподобия (15.2.5) дает:

$$h(z)=\frac{\dfrac{\partial}{\partial\theta}\log f(z,\theta)}{i(\theta)},\qquad(15.2.6)$$

ибо здесь

$$-\int \psi_t'(x,\theta)\,dF(x,\theta) = -\int \left[\frac{\partial^2}{\partial\theta^2}\log f(x,\theta)\right] dF(x,\theta) = i(\theta).$$

Важное свойство М-оценки: её функция влияния лишь постоянным множителем отличается от $\psi(x,t)$. М-оценка окажется устойчивой, если функция $\psi(x,t)$ будет ограниченной по x.

§ 3. Асимптотическое распределение $T(F_n)$ — наводящие соображения

Главным для нас в этом разделе будет связь между выборочным функционалом $T(F_n)$ и его вероятностным аналогом $T(F)$. Ясно, что если $T(.)$ — непрерывный функционал, то $T(F_n) \xrightarrow{P} T(F)$, так как $F_n \xrightarrow{P} F$ при $n \to \infty$. Наша цель — найти асимптотическое распределение $T(F_n) - T(F)$. С использованием функций влияния мы приведем правдоподобные соображения в пользу того, что при $n \to \infty$

$$\sqrt{n}\big(T(F_n) - T(F)\big) \xrightarrow{d} N(0,\sigma^2),$$

где $\sigma^2 = \operatorname{Var} IF(x_1;T,F)$.

Начнем со связи функций влияния с дифференцируемостью функционалов. Предположим, что засорение распределения F сосредоточено теперь в нескольких точках $z_1, z_2, \ldots, z_p$ с весами $k_1 > 0,\ k_2 > 0,\ \ldots,\ k_p > 0$, причем $k_1 + k_2 + \ldots + k_p = 1$. Инфинитезимальное влияние на $T(F)$ этого засорения равно:

$$\lim_{\lambda\to 0} \lambda^{-1}\{T[(1-\lambda)F + \lambda\sum_{i=1}^{p} k_i\Delta_{z_i}] - T(F)\} = \sum_{i=1}^{p} k_i h(z_i) = \int h(x)dK(x),$$

если положить $K(x) = \sum\limits_{i=1}^{p} k_i\Delta_{z_i}(x)$, $K(.)$ — функция распределения засорений. Предположим, что аналогичная формула справедлива для произвольного распределения засорений:

$$\lim_{\lambda\to 0} \lambda^{-1}\{T[(1-\lambda)F + \lambda K] - T(F)\} = \int h(x)\,dK(x). \qquad (15.3.1)$$

Линейный функционал (15.3.1) называют *слабым дифференциалом (дифференциалом Гато)* функционала $T(.)$ в точке F. Заметим, что

$$\int h(x)\,dF(x) = 0. \tag{15.3.2}$$

Чтобы в этом убедиться, достаточно в (15.3.1) положить $K = F$. Предположим теперь, что функционал $T(.)$ дифференцируем в каком-либо смысле. В таком случае, если G и F — две функции распределения, то:

$$T(G) = T\big(F + (G - F)\big) = T(F) + dT + R,$$

где dT — дифференциал, R — остаток, т. е. переменная величина, стремящаяся к нулю быстрее, чем $(G - F)$. Заметим, что если функционал $T(.)$ дифференцируем и dT существует, то dT совпадает с дифференциалом Гато. Так что для дифференцируемого функционала $T(.)$

$$T(G) = T(F) + \int h(x)\,d(G - F) + R,$$

или:

$$T(G) = T(F) + \int h(x)\,dG(x) + R, \tag{15.3.3}$$

поскольку $\int h(x)\,dF(x) = 0$, как уже было отмечено. Формула (15.3.3) приобретает точный смысл, если снабдить пространство функций распределения какой-либо нормой и предположить, что по отношению к этой норме функционал $T(.)$ дифференцируем (в сильном смысле, или по Фреше). Этот путь оказывается продуктивным лишь для немногих функционалов из тех, которые представляют интерес для математической статистики. Поэтому соотношение (15.3.3) обычно истолковывают, как наводящее, причем применяют его лишь для сопоставления $T(F)$ и $T(F_n)$. Формально применим (15.3.3), положив $G = F_n$. Получим, что

$$T(F_n) = T(F) + \int h(x)\,dF_n(x) + R_n.$$

В данном случае $F_n \xrightarrow{P} F$ со скоростью $1/\sqrt{n}$ при $n \to \infty$. Поэтому можно ожидать, что $R_n = o_P\Big(\dfrac{1}{\sqrt{n}}\Big)$. Далее заметим, что

$$\int h(x)\,dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} h(x_i).$$

Случайные величины $h(x_1), h(x_2), \ldots, h(x_n)$ независимы и одинаково распределены, причем $Eh(x_i) = 0$ в силу (15.3.2). Предположим, что

$$0 < Dh(x_1) = \mathrm{Var}\, IF(x_1, T, F) < \infty$$

и обозначим эту дисперсию через σ^2. Тогда по центральной предельной теореме

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(x_i) \xrightarrow{d} N(0, \sigma^2).$$

Отсюда заключаем, что при $n \to \infty$

$$\sqrt{n}\big(T(F_n) - T(F)\big) \xrightarrow{d} N(0, \sigma^2). \tag{15.3.4}$$

Для доказательства (15.3.4) достаточно показать, что:

$$\sqrt{n}\Big(T(F_n) - T(F) - \frac{1}{n} \sum_{i=1}^{n} h(x_i)\Big) \xrightarrow{P} 0. \tag{15.3.5}$$

Обычно для конкретных задач утверждение (15.3.5) удается доказать без обращения к концепции дифференцирования функционалов в сильном смысле. Так, между прочим, мы и действовали применительно к оценкам максимального правдоподобия.

Специально отметим замечательное свойство (15.3.4): во всех известных случаях оценивания эта формула дает верный результат. Для примера найдем с помощью (15.3.4) асимптотическое выражение для выборочной медианы μ_n (по выборке объема n из распределения с функцией $F(.)$, причем для медианы μ этого распределения выполнено $F'(\mu) > 0$). Функция влияния для медианы известна:

$$h(z) = \frac{\mathrm{sign}(z - \mu)}{2F'(\mu)}.$$

Формула (15.3.4) утверждает, что:

$$\sqrt{n}(\mu_n - \mu) \xrightarrow{d} N\Big(0, \frac{1}{4[F'(\mu)]^2}\Big),$$

так как $D\,\mathrm{sign}(x_1 - \mu) = 1$. Это — правильный результат. Доказать его можно разными способами, но мы этого делать здесь не будем.

Лекция 16. Критерии согласия типа Пирсона-Фишера

§ 1. Теорема К. Пирсона

Упомянутые критерии относятся к независимым испытаниям с несколькими исходами и к гипотезам об их вероятностях. Рассмотрим независимые испытания с m ($m \geq 2$) исходами. Обозначим исходы через $A_1, A_2, \ldots, A_m$. Вероятности этих исходов неизменны во всех испытаниях. Обозначим эти вероятности через $p_1, p_2, \ldots, p_m$, 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) $\mu_1, \mu_2, \ldots, \mu_m$ исходов $A_1, A_2, \ldots, A_m$; при этом $\sum\limits_{i=1}^{m} \mu_i = n$. Теоремы, которые мы обсудим, касаются проверок гипотез о $\vec{p} = (p_1, \ldots, p_m)^T$ по частотам $\vec{\mu} = (\mu_1, \ldots, \mu_m)^T$.

Начнем с первого критерия такого рода, установленного К. Пирсоном (Karl Pearson) в 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:

$$H_0 : \vec{p} = \vec{p}^{\,0} \quad \text{или, подробнее,} \quad H_0 : p_1 = p_1^0,\ p_2 = p_2^0,\ \ldots,\ p_m = p_m^0,$$

где $p_1^0,\ p_2^0,\ \ldots,\ p_m^0$ — заданные положительные вероятности, $\sum\limits_{i=1}^{m} p_i^0 = 1$. Альтернативой к H_0 служит ее отрицание

$$\overline{H}_0 : \vec{p} \neq \vec{p}^{\,0}.$$

Правило Пирсона имеет асимптотический характер и может корректно применяться при численностях испытаний n "достаточно больших" (что это означает — обсудим позже).

П р а в и л о К. Пирсона. *Отвергнуть* $H_0 : \vec{p} = \vec{p}^{\,0}$ *на (приближенном) уровне* $\varepsilon > 0$, *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-ε)-квантиль распределения хи-квадрат с $(m-1)$ степенью свободы. Вопрос о том, какие численности n достаточно велики для того чтобы можно было обращаться к этому правилу, довольно темен, несмотря на долгую его историю. Осторожная (консервативная) рекомендация: должны выполняться соотношения $np_i^0 \geq 5$ 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:

$$X_n^2 \xrightarrow{P} \infty \quad \text{при} \quad n \to \infty.$$

Асимптотическое поведение X_n^2 при гипотезе $\vec{p} = \vec{p}^{\,0}$:

Т е о р е м а 16.1.1. (Karl Pearson, 1900г. — примерно). *Случайная величина*

$$\sum_{i=1}^{m} \frac{(\mu_i - np_i^0)^2}{np_i^0} \xrightarrow{d} \chi^2(m-1) \quad \textit{при} \quad n \to \infty.$$

(*Случайная величина X_n^2 при $n \to \infty$ сходится по распределению к случайной величине 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 $H_0 : \vec{p} = \vec{p}^{\,0}$.

§ 2. Доказательство теоремы Карла Пирсона

Т е о р е м а 16.2.1. (Многомерная теорема Муавра-Лапласа). *В описанной выше схеме испытаний Бернулли с m исходами*

$$\sqrt{n}\left(\frac{1}{n}\,\vec{\mu} - \vec{p}\right) \xrightarrow{d} N(0, \mathcal{P} - \vec{p}\vec{p}^{\,T}), \quad \textit{при } n \to \infty,$$

где $\mathcal{P} = \operatorname{diag}(p_1, \ldots, p_m)$ *—диагональная матрица.*

Д о к а з а т е л ь с т в о. 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). В многомерном случае этот прием не оправдан.

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 $n \to \infty$

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\vec{x}_j - E\vec{x}_j) \xrightarrow{d} N(0, \Sigma),$$

где

$$\Sigma = E\vec{x}_j\vec{x}_j^{\,T} - (E\vec{x}_j)(E\vec{x}_j)^T.$$

Очевидный подсчет дает $\quad E\vec{x}_j = \vec{p}, \quad D\vec{x}_j = \mathcal{P} - \vec{p}\vec{p}^{\,T}.$ $\square$

Заметим, что матрица $\mathcal{P} - \vec{p}\vec{p}^{\,T}$ вырождена. Её ранг равен $(m-1)$. Если бы не это обстоятельство, предельное распределение хи-квадрат для нормы вектора

$$\xi_n \xrightarrow{d} N(0, B)$$

мы могли бы получить немедленно. Ибо очевидно, что

$$\xi_n^T B^{-1} \xi_n \xrightarrow{d} \chi^2(m).$$

Д о к а з а т е л ь с т в о теоремы Карла Пирсона. Введем в рассмотрение вектор

$$\xi_n := \sqrt{n}\,\mathcal{P}^{-1/2}\Big(\frac{1}{n}\,\vec{\mu} - \vec{p}\Big).$$

Легко видеть, что при $n \to \infty$

$$\xi_n \xrightarrow{d} N(0, I - zz^T),$$

где I — единичная матрица, $z = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_m})^T$.

Введем ортогональную матрицу V, первая строка которой есть $(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_m})$, 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:

$$D(V\xi_n) = V(D\xi_n)V^T = V(I - zz^T)V^T =$$

$$VV^T - (Vz)(Vz)^T = I - \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{pmatrix},$$

ибо $Vz = (1, 0, \ldots, 0)^T$. Теперь

$$|\xi_n|^2 = \sum_{i=1}^{m} \left(\frac{1}{\sqrt{p_i}} \sqrt{n} \left(\frac{1}{n} \mu_i - p_i \right) \right)^2 = \sum_{i=1}^{m} \frac{(\mu_i - np_i)^2}{np_i},$$

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

$$(0, \eta_2, \ldots, \eta_m)^T,$$

где $\eta_2, \ldots, \eta_m$ суть независимые стандартные гауссовские случайные величины $N(0, 1)$. По определению,

$$\eta_2^2 + \ldots + \eta_m^2 = \chi^2(m-1). \quad \square$$

§ 3. Сложные гипотезы

Здесь мы рассмотрим гипотезы о $\vec{p}$ вида

$$H : \vec{p} \in Q,$$

где Q — некоторое заданное гладкое многообразие, принадлежащее симплексу $\{\vec{p} : \sum\limits_{i=1}^{m} p_i,\ p_1 \geq 0, \ldots, p_m \geq 0\}$. "Гладкое" здесь означает, что в каждой точке $\vec{p} \in Q$ существует касательное линейное многообразие. Размерность $\vec{p}$ обозначим через r.

Т е о р е м а 16.3.1. (J. Neyman, E. Pearson, 1928). *При* $n \to \infty$

$$\min_{\vec{p} \in Q} \sum_{i=1}^{m} \frac{(\mu_i - np_i)^2}{np_i} \xrightarrow{d} \chi^2(m-r-1). \tag{16.3.1}$$

Заметим, что при вычислении статистики из (16.3.1) обычно находят и то значение $\vec{p} \in Q$, при котором достигается минимум в (16.3.1). 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 $\mu_1, \ldots, \mu_m$. (Либо иная оценка, но с теми же асимптотическими свойствами, что и $\hat{\theta}_n$). Тогда справедлива

Т е о р е м а 16.3.2. *При* $n \to \infty$

$$\sum_{i=1}^{m} \frac{\bigl(\mu_i - np_i(\hat{\theta})\bigr)^2}{np_i(\hat{\theta})} \xrightarrow{d} \chi^2(m-r-1). \qquad (16.3.2)$$

Эти теоремы и другие, подобные, часто связывают с именем Р. Фишера (R.A. Fisher). Фишер действительно 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 § 4.

А сейчас, чтобы окончить, сформулируем правило проверки $H : \vec{p} \in Q$, основанное на приведенных выше теоремах. А также на том факте, что статистики (16.3.1) или (16.3.2) 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 (16.3.1) или (16.3.2) превосходит $\chi^2_{1-\varepsilon}(m-r-1)$ — $(1-\varepsilon)$-квантиль распределения хи-квадрат с $(m-r-1)$ степенями свободы.

Это правило применимо для "достаточно больших n". Осторожная (консервативная) практическая рекомендация: $\mu_i \geq 5$. (Впрочем, разные авторы говорят несколько различное на эту тему).

§ 4. Таблицы сопряженности

Предположим, что каждый объект некоторой (бесконечной) совокупности может быть классифицирован по двум признакам A и B. Признак A при этом имеет 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 значений признаков A и B, где $i = \overline{1,r}$; $j = \overline{1,s}$.

Пусть p_{ij} обозначает вероятность того, что наудачу взятый объект обладает комбинацией признаков A_iB_j. Пусть μ_{ij} — это 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 для всех $i=\overline{1,r};\ j=\overline{1,s}$

$$p_{ij} \equiv P\{A_iB_j\} = P\{A_i\}\,P\{B_j\}.$$

Вероятность появления A_i и вероятность появления B_j обозначим через $p_{i\cdot}$ и $p_{\cdot j}$ соответственно. При этом

$$p_{i\cdot} = \sum_{j=1}^{s} p_{ij}, \quad p_{\cdot j} = \sum_{i=1}^{r} p_{ij}.$$

Гипотеза о 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 $\vec{p}=\|p_{ij}\|$ размера $(r\times s)$) принадлежит $(r+s-2)$-мерному многообразию. (Размерность 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 параграфа 2. Для этого найдем оценки наибольшего правдоподобия для $p_{i\cdot}$ и $p_{\cdot j}$, и затем применим теорему 16.3.2.

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

$$\mathrm{Const}\prod_{i=1}^{r}(p_{i\cdot})^{\mu_{i\cdot}}\prod_{j=1}^{s}(p_{\cdot j})^{\mu_{\cdot j}},$$

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:

$$\hat{p}_{i\cdot}=\frac{\mu_{i\cdot}}{n},\quad \hat{p}_{\cdot j}=\frac{\mu_{\cdot j}}{n}\quad \text{для}\quad i=\overline{1,r},\ j=\overline{1,s}.$$

Статистика X_n^2 из теоремы 16.3.2 здесь

$$X_n^2=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\Big(\mu_{ij}-n\,\dfrac{\mu_{i\cdot}}{n}\cdot\dfrac{\mu_{\cdot j}}{n}\Big)^2}{n\,\dfrac{\mu_{i\cdot}}{n}\cdot\dfrac{\mu_{\cdot j}}{n}}.$$

При гипотезе о независимости признаков

$$X_n^2\xrightarrow{\ d\ }\chi^2\big((r-1)(s-1)\big),$$

ибо $rs-(r+s-2)-1=(r-1)(s-1)$.

Гипотезу о независимости признаков следует отвергать, если наблюденное (вычисленное) значение статистики X_n^2 слишком велико. Точнее: гипотезу о независимости признаков A и B следует отвергнуть на уровне ε, если статистика X_n^2 превосходит $(1-\varepsilon)$-квантиль распределения хи-квадрат с $(r-1)(s-1)$ степенями свободы.

З а д а ч а. Согласно гипотезе Ф. Бернштейна, наличие у людей четырех групп крови O (I группа), A (II группа), B (III группа) и AB (IV группа) вызвано тремя генами A, B и O, причем A и B доминируют над O. Если индивидуум имеет генную пару OO, его кровь относится к группе O; генные пары AO и AA приводят к группе крови A; пары BO и BB — к группе B. Наконец, генная пара AB приводит к группе AB. Из популяции случайно выбрана большая группа испытуемых и у каждого определена группа крови. Так получены частоты для групп крови O, A, B и AB в выборке. Как по этим частотам можно проверить генную гипотезу Бернштейна?

Список литературы

[1] Беляев Ю.К., Носко В.П. Основные понятия и задачи математической статистики. – М.: изд-во МГУ, 1998.

[2] Бикел П., Доксам К. Математическая статистика. Вып. 1 и 2: пер. с англ. – М.: Финансы и статистика, 1983. – 278 с. и 254 с.

[3] Ивченко Г.Н., Медведев Ю.И. Математическая статистика: Учебное пособие для втузов. – М.: Высш. шк., 1992. – 304 с.

[4] Чибисов Д.М., Пагурова В.И. Задачи по математической статистике. – М.: изд-во МГУ, 1990.

[5] Ширяев А.Н. Вероятность. – М.: Наука, 1989. –576 с.

Дополнительная литература

[1] Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. – М.: Наука, 1983. – 416 с.

[2] Боровков А.А. Математическая статистика. – Новосибирск.: Наука, изд-во Института математики, 1997. – 772 с.

[3] Вероятность и математическая статистика: Энциклопедия. /Гл. ред. Ю.В. Прохоров – М.: Большая Российская энциклопедия, 1999. – 910 с.

[4] Леман Э. Теория точечного оценивания: Пер. с англ. Ю.В. Прохорова. – М.: Наука, 1991. – 448 с.

[5] Рао С.Р. Линейные статистические методы и их применения: пер. с англ. – М.: Наука, 1968. – 548 с.

[6] Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность в статистике. Подход на основе функций влияния: Пер. с англ. под ред. В.М. Золотарева. – М.: Мир, 1989. – 512 с.

Тюрин Юрий Николаевич

Математическая статистика.
Записки лекций.
М.: Изд-во ЦПИ механико-математического
факультета МГУ, 2003. – 192 с.

Подписано в печать . .2003 г.
Формат 60 × 90 Объем 12 п.л.
Заказ Тираж 600 экз.

Издательство Центра прикладных исследований при
механико-математическом факультете МГУ.
Лицензия на издательскую деятельность ИД B04059
от 20.02.2001

Отпечатано с оригинал-макета на типографском
оборудовании механико-математического факультета
и Франко-русского центра им. А.М. Ляпунова.