ЧАСТЬ ВТОРАЯ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

“. . . а спорили они о чем угодно, начиная с подлинности Священного писания и кончая вопросом, что на самом деле значит надпись “гарантировано"на банках с джемом."

Мюриэл Спарк, Мисс Джин Броди в расцвете лет
(Muriel Spark, The prime of Miss Jean Brodie)

О Г Л А В Л Е Н И Е

ч а с т ь в т о р а я
м а т е м а т и ч е с к а я с т а т и с т и к а

§1. Проблема статистического вывода

Лекция 1

Теория вероятностей создает базу для построения моделей реальных явлений, в основе которых лежат соотношения между частотами появления определенных событий. Располагая вероятностной моделью, мы можем рассчитать вероятности (относительные частоты) этих событий и тем самым оптимизировать свое поведение в условиях неопределенности. Математическая статистика строит модели индуктивного поведения в этих условиях на основе имеющихся вероятностных моделей. Основная проблема состоит в том, чтобы по наблюдениям элементарных исходов (обычно это – значения наблюдаемых случайных величин) дать метод выбора действий, при которых частота ошибок была бы наименьшей. Естественно, эта проблема сопряжена с решением сложных задач на экстремум, но даже в том случае, когда эти задачи не удается решить, теория вероятностей дает метод для расчета средней величины потерь, которые мы будем нести, используя конкретное, выбранное нами правило индуктивного поведения. Таким образом, *математическая статистика есть теория принятия оптимальных решений, когда последствия от действий, предпринимаемых на основе этих решений, носят случайный характер.* Математическая статистика использует методы теории вероятностей для расчета частоты "неправильных" решений или, более общо, для величины средних потерь, которые неизбежно возникают в условиях случайности, как бы мы ни пытались оптимизировать свое поведение в этих условиях.

Приведем два примера, иллюстрирующих задачи математической статистики и, отчасти, методы их решения, с тем чтобы в последующем формализовать общую проблему статистического вывода.

Пример 1.1. *Определение общего содержания серы в дизельном топливе.* Мы снова обращаемся к примеру 7.2 из курса теории вероятностей, где речь шла о важной в экологическом отношении характеристике дизельного топлива – процентном содержании элементарной серы, которая при сжигании и последующем соединении с водой дает серную кислоту. Необходимость использования методов теории вероятностей при аттестации дизельного топлива по этой характеристике была вызвана значительными расхождениями между результатами $x_1, \ldots, x_n$ параллельных и независимых испытаний n проб из партии дизельного топлива. Если даже исключить ошибки эксперимента, связанные с неправильным определением

веса пробы и титрованием, то все равно разброс в параллельных испытаниях будет значительным в силу случайного характера процесса сжигания пробы топлива и выпадения части элементарной серы в золу. Но в таком случае возникает естественный вопрос, что же мы измеряем и что же это за характеристика дизельного топлива, которую мы назвали "общим содержанием серы" ? В практике лабораторных испытаний обычно говорят о *среднем* значении этой характеристики, и дизельное топливо аттестуется величиной $\overline{x} = n^{-1} \sum_1^n x_k$ – арифметическим средним результатов параллельных испытаний. Это и есть то "индуктивное поведение" статистика в условиях случайности, о котором мы говорили в начале лекции, и оправдание разумности такого поведения естественно искать в рамках закона больших чисел.

Действительно, в примере 7.2 мы интерпретировали результат x определения общего содержания серы в одной пробе как результат наблюдения случайной величины X, распределенной по нормальному закону со средним μ и дисперсией σ^2, причем значение (неизвестное экспериментатору) параметра μ являлось математическим выражением той, не совсем понятной для нас характеристики испытуемого топлива, которая называлась "общим содержанием серы" . В рамках этой вероятностной модели естественно трактовать результаты $x_1, \ldots, x_n$ параллельных испытаний n проб дизельного топлива как наблюдения n независимых копий $X_1, \ldots, X_n$ случайной величины X. Термин "копия" в данном случае употребляется для обозначения того факта, что каждая из наблюдаемых случайных величин имеет то же распределение, что и X. Таким образом, постулируется, что $X_1, \ldots, X_n$ независимы и одинаково распределены $\mathcal{N}(\mu, \sigma^2)$, так что в силу закона больших чисел при неограниченном возрастании объема испытаний n

$$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k \underset{P}{\to} \mu.$$

Итак, закон больших чисел гарантирует нам, что при достаточно большом объеме испытаний мы будем близки к истинному значению исследуемой характеристики топлива. Однако на практике в заводских лабораториях обычно сжигаются всего две пробы топлива, и только в исключительных случаях при поверке приборов или тестировании лаборантов делается четыре испытания. Естественно, при $n = 2$ говорить о законе "больших" чисел просто смешно,– следует искать некоторую количественную характеристику последствий от неточной аттестации партии дизель-

ного топлива. Легко понять, что в основу такой характеристики следует положить ошибку $|\,\overline{X} - \mu\,|$ в оценке параметра μ, но, к сожалению, значение μ нам неизвестно, а $\overline{X}$ есть случайная величина, что окончательно делает проблему прогноза ожидаемых ошибок при аттестации конкретной партии топлива неразрешимой. Здесь наблюдается та же ситуация, что и при попытке предсказать сторону монеты, которая выпадет при ее подбрасывании. Точный прогноз невозможен, но методы теории вероятностей позволяют нам рассчитать, как часто мы будем ошибаться в прогнозе при достаточно длительной игре в орлянку. Следовательно, мы должны решить задачу о вычислении вероятности того, что ошибка в оценке μ будет слишком большой – превосходить некоторую предписанную величину Δ. Эта вероятность $P(|\,\overline{X} - \mu\,| > \Delta)$ обычно называется *риском* оценки $\overline{X}$, а вероятность $P(|\,\overline{X} - \mu\,| \le \Delta)$ противоположного события – *надежностью* этой оценки.

Таким образом, риск оценки указывает частоту тех партий дизельного топлива, в паспорте которых общее содержание серы указано с недопустимо большой ошибкой. Зная риск оценки, мы можем вычислить средние затраты на выплату рекламаций по искам потребителей дизельного топлива. Вывести формулу для вычисления риска не представляет особого труда, если обратиться к теореме сложения для нормального распределения (предложение 12.2 курса ТВ). Выборочное среднее $\overline{X}$ есть нормированная на n сумма независимых одинаково распределенных $\mathcal{N}(\mu,\, \sigma^2)$ случайных величин. В силу теоремы сложения эта сумма имеет также нормальное распределение, среднее значение которого равно сумме средних $n\mu$, а дисперсия равна сумме дисперсий $n\sigma^2$. При умножении на $1/n$ среднее умножается на ту же величину, а дисперсия умножается на ее квадрат. Таким образом, $\overline{X} \sim \mathcal{N}(\mu,\, \sigma^2/n)$, надежность оценки

$$P(-\Delta \le \overline{X} - \mu \le \Delta) = \Phi(\Delta\sqrt{n}/\sigma) - \Phi(-\Delta\sqrt{n}/\sigma) = 2\Phi(\Delta\sqrt{n}/\sigma) - 1$$

(напомним, $\Phi(-x) = 1 - \Phi(x)$), а ее риск

$$P\left(|\,\overline{X} - \mu\,| > \Delta\right) = 2\left(1 - \Phi(\Delta\sqrt{n}/\sigma)\right).$$

При вычислении риска оценки необходимо знать величину стандартного отклонения σ. Но значение σ, очевидно, остается постоянным при аттестации различных партий – это параметр, характеризующий точность метода химического анализа топлива, и не имеет отношения к его химическому составу. Естественно, за достаточно короткий срок в лабораториях накапливается большой архивный материал данных испытаний различных партий

топлива, что позволяет оценить значение σ с достаточно высокой точностью. С тем, как это делается, мы познакомимся в одной из ближайших лекций.

Используя формулу риска, мы можем определить минимальный объем испытаний n, гарантирующий предписанную, достаточно малую величину риска. Действительно, если α – заданное ограничение на риск оценки, то разрешая неравенство $2(\Phi(\Delta\sqrt{n}/\sigma) - 1) \leq \alpha$ относительно переменной n, получаем, что требуемый объем испытаний определяется неравенством

$$n \geq \left(\frac{\Phi^{-1}(1 - \alpha/2)\sigma}{\Delta} \right)^2 .$$

Пример 1.2. *Выявление эффекта лечения.* Группа пациентов в количестве 10 человек, обладающих схожими антропометрическими и антропологическими характеристиками, подвергается лечению по некоторой новой методике, подтверждение или опровержение эффективности которой составляет предмет статистического исследования. После лечения дается только качественное заключение о состоянии здоровья каждого пациента, так что результат испытания новой методики можно представить в виде последовательности $x_1, \ldots, x_{10}$, компоненты которой принимают значения 1 (положительный исход лечения) или 0 (отрицательный исход).

Предлагается следующее статистическое правило: новая методика объявляется эффективной, если $x_i = 1$ для всех $i = 1, \ldots, 10$, то есть все пациенты выздоровели. Если же лечение хотя бы одного пациента не привело к положительному исходу, новая методика не рекомендуется к дальнейшему клиническому использованию. Что можно сказать о надежности или, как говорят медики, "достоверности" такого правила индуктивного поведения?

Чтобы ответить на этот вопрос, мы должны построить вероятностную модель проводимых наблюдений. Естественно предполагать, что в силу "однородности" группы пациентов они обладают одинаковой вероятностью p положительного исхода лечения, и если в процессе лечения они не имели возможности излишне тесного общения, то исходы лечений можно представить в виде реализации десяти независимых бинарных случайных величин $X_1, \ldots, X_{10}$, каждая из которых принимает значение 1 с вероятностью p и значение 0 с вероятностью $1-p$. Таким образом, мы пришли к модели испытаний в схеме Бернулли с вероятностью p успешного исхода. Вероятность того, что все 10 исходов были успешными равна p^{10}, и задавая различ-

ные значения p мы можем судить о том, как часто возможны различные результаты апробации нового метода лечения.

Предположим сначала, что новая методика неэффективна. При таком предположении значение p не должно превосходить величины $1/2$, и максимальное значение вероятности события $X_1 = 1, \ldots, X_{10} = 1$ равно $2^{-10} = 1/1024 < 0,001$. Это очень редкое событие, и поэтому предположение о неэффективности новой методики должно быть отвергнуто. При этом вероятность 2^{-10} можно интерпретировать как риск внедрения в медицинскую практику неэффективного метода лечения: *используя предложенное правило выбора между двумя действиями (внедрение или отклонение методики) при испытаниях последующих методик, мы рискуем в среднем не более чем один раз из тысячи внедрить неэффективный метод лечения.*

Интересно заметить, что в предположении "нейтральности" нового метода ($p = 1/2$) вероятность любого исхода $X_1 = x_1, \ldots, X_{10} = x_{10}$ одинакова и равна 2^{-10}, но исход $X_1 = 1, \ldots, X_{10} = 1$ обладает наибольшей вероятностью принятия действительно эффективной методики, ибо

$$p^{\sum_1^{10} x_k}(1-p)^{n-\sum_1^{10} x_k} \leq p^{10},$$

если $p > 1/2$. Столь же просто проверить, что результаты испытаний, в которых лечение только одного пациента окончилось неудачей, имеют вероятность $p^9(1-p)$, и такие 10 результатов $x_1, \ldots, x_{10}$ с одним $x_i = 0$ и другими $x_j = 1$ обладают большей вероятностью, чем исходы с двумя и более количеством неудач, если в действительности $p > 1/2$. Это замечание позволяет нам определить статистическое правило, обладающее наибольшей вероятностью принятия в действительности эффективной методики, но не с таким малым риском, как 2^{-10}.

Дело в том, что в медицинской практике установилась определенная граница риска, равная 0.05, и все события, обладающие меньшей вероятностью, объявляются "редкими" – ими можно пренебречь. В связи с этим позволим себе включить в область принятия новой методики дополнительные исходы с ровно одним неуспехом, и вычислим риск такого статистического правила при $p = 1/2$. Используя известные нам формулы биномиальных вероятностей, находим, что

$$P\left(\sum_1^{10} X_k \geq 9\right) = p^{10} + \mathrm{C}_{10}^1 p^9(1-p),$$

и при $p = 1/2$ эта вероятность равна $2^{-10}(1 + 10) = 11/1024 \approx 0,01$, что по-прежнему достаточно мало по сравнению с 0.05. Следовательно, мы можем включить в область принятия новой методики еще C_{10}^2 результатов испытаний, в которых присутствуют ровно две неудачи. Риск такового статистического правила становится равным

$$P\left(\sum_1^{10} X_k \geq 8\right) = p^{10} + C_{10}^1 p^9(1-p) + C_{10}^2 p^8(1-p)^2,$$

и при $p = 1/2$ эта вероятность равна $2^{-10}(1 + 10 + 45) = 56/1024 \approx 0.05$.

Это как раз соответствует принятой в медицине норме риска статистического правила. Итак, мы рекомендуем новую методику к дальнейшему использованию в клинике, если лечение не более чем двух пациентов из десяти оказалось неудачным, и применение такого правила в испытаниях дальнейших методик может привести к принятию неэффективного метода лечения в среднем в пяти случаях из 100.

Мы рассмотрели две типичных задачи математической статистики – оценка параметров и проверка гипотез. Естественно, круг проблем математической статистики намного шире, но при надлежащей трактовке проблем большинство из них сводится или к задаче оценки параметров, или к задаче выбора одного из нескольких альтернативных высказываний об исследуемом объекте. Опираясь на рассмотренные примеры, мы можем теперь представить достаточно общую схему статистического вывода.

Любое статистическое исследование, проводимое в рамках математической статистики, начинается с описания объекта исследования и формализации *пространства* $\mathcal{D}$ решений d, одно из которых статистик принимает на основе наблюдений независимых копий случайной, возможно векторной, величины X, характеризующей состояние объекта в момент проведения наблюдений. Так, в примере с аттестацией партии дизельного топлива (объект исследования) $\mathcal{D}$ есть интервал (0; 100) (напомним, общее содержание серы измеряется в процентах к весу пробы), а в примере с определением эффективности нового метода лечения (объект исследования) пространство $\mathcal{D}$ состоит из двух точек: d_0 – решение о неэффективности метода (принятие "нулевой" гипотезы) и d_1 – решение о внедрении нового метода в лечебную практику (принятие альтернативной гипотезы).

Наиболее важной и, по-видимому, наиболее сложной частью статистического исследования является этап построения *вероятностной модели*, который состоит в спецификации семейства $\mathcal{P} = \{P_\theta,\, \theta \in \Theta\}$ возможных распределений наблюдаемой случайной величины X. Этот этап связан с достаточно глубоким проникновением в природу исследуемого объекта и метода наблюдений X, – одной математикой здесь, как правило, не обойдешься. Семейство $\mathcal{P}$ индексируется абстрактным *параметром* θ, совокупность значений которого Θ называется *параметрическим пространством*.

В первом примере мы выяснили, что семейство возможных распределений X есть семейство нормальных распределений $\mathcal{N}(\mu,\, \sigma^2)$ с двумерным параметром $\theta = (\mu,\, \sigma)$ и параметрическим пространством $\Theta = \mathrm{R} \times \mathrm{R}^+$. В дальнейшем мы предположили, что значение σ известно, и свели наше параметрическое пространство к эвклидовой прямой: $\Theta = \mathrm{R}$ с $\theta = \mu$. Наконец, поскольку общее содержание серы измеряется в процентах, мы должны окончательно положить $\Theta = (0;\, 100.)$.

Во втором примере мы имели дело с бинарной случайной величиной X, принимающей значение 1 с вероятностью p и значение 0 с вероятностью $1 - p$. Таким образом, вероятностная модель представлялась семейством двухточечных распределений $\mathrm{B}(1,\, p)$ с $\theta = p$ и параметрическим пространством $\Theta = (0;\, 1)$.

Следующий этап статистического исследования состоит в интерпретации решений d в терминах высказываний о соответствующих этому решению значениях параметра θ. Это необходимо сделать, если мы поставили

себе задачу количественного измерения последствий от принятия неверных решений, – в наших примерах риск используемых правил представлял собой функцию от θ. Нетрудно понять, что в первом примере $\mathcal{D} = \Theta$, а во втором примере решению d_0 о неэффективности метода соответствует подмножество параметрического пространства $(0;\, 1/2\,]$, а альтернативному решению d_1 об использовании новой методики соответствует интервал $(1/2;\, 1)$ возможных значений параметра $\theta = p$. Именно таким образом мы сводим конкретные задачи по аттестации партии дизельного топлива и выявлению эффективности нового метода лечения к абстрактным задачам математической статистики – оценке параметра (среднего значения) θ нормального $(\theta,\, \sigma^2)$ распределения и, соответственно, различению двух гипотез $H_0 : \theta \in (0;\, 1/2\,]$ и $H_1 : \theta \in (1/2;\, 1)$ о величине вероятности θ успешного испытания в схеме Бернулли.

Параметрическая интерпретация решений позволяет статистику задать *потери* $L(\theta,\, d)$, которые он несет от принятия решения d, когда θ представляет истинное значение параметра. Среднее значение этих потерь в длинном ряду однотипных статистических исследований с одним и тем же правилом принятия решения определяет величину риска, связанную с принятием неправильных решений. Так, в наших примерах риск определялся вероятностью принятия решения, отстоящего достаточно далеко от того решения, которое соотвествовало истинному значению параметра, и, следовательно, *функция потерь* определялась индикатором некоторого подмножества в $\Theta \times \mathcal{D}$. Это так называемые функции потерь типа 0–1. В первом примере $L(\theta,\, d) = 1$, если $|\, d - \theta\,| > \Delta$, и $L(\theta,\, d) = 0$ в противном случае. Во втором примере $L(\theta,\, d) = 1$, если принималось решение d_1, а $\theta \in (0;\, 1/2\,]$, или принималось d_0, а $\theta \in (1/2;\, 1)$, в остальных точках произведения пространств $\Theta \times \mathcal{D}$ потери $L(\theta,\, d)$ полагались равными нулю. Отметим, что в задаче оценки параметров довольно часто используется квадратичная функция потерь $L(\theta,\, d) = |\, d - \theta\,|^2$.

Каждое из решений d статистик принимает на основе результата $x^{(n)} = x_1, \ldots, x_n$ наблюдений над независимыми копиями $X^{(n)} = (X_1, \ldots, X_n)$ случайной величины X. Строится измеримое отображение $\delta = \delta(\cdot)$ пространства возможных значений $X^{(n)}$ в пространство решений $\mathcal{D}$, с помощью которого принимается решение $d = \delta(x^{(n)})$. Это отображение называется *решающей функцией* или *статистическим правилом*. Так, в первом примере

$\delta(X^{(n)}) = \overline{X}$, а во втором

$$\delta(X^{(n)}) = \begin{cases} d_0, & \text{если } \sum_1^n X_k < 8, \\ d_1, & \text{если } \sum_1^n X_k \geq 8\ . \end{cases}$$

Последствия от использования конкретной решающей функции в длинном ряду однотипных статистических исследований определяются величиной средних потерь $R(\theta;\, \delta) = \mathbf{E}_\theta L(\theta,\, \delta(X^{(n)}))$, которая зависит от θ; функция $R(\theta,\, \delta),\ \theta \in \Theta$, называется *функцией риска.*

Основная проблема математической статистики состоит в построении решающих функций δ, минимизирующих равномерно по $\theta \in \Theta$ функцию риска $R(\theta;\, \delta)$. Мы будем решать эту проблему для задач оценки параметров и проверки гипотез. Естественно, будут также изучаться традиционные, возможно не обладающие оптимальными свойствами, статистические правила, и в этом случае нашей основной задачей будет вычисление их функций риска.

Представленная выше схема статистического вывода весьма далека от общности. Большинство статистических задач имеет дело с наблюдениями одновременно за несколькими объектами, например, новый метод лечения применяется к одной группе пациентов, в то время как другая подвергается лечению традиционным методом, и по данным наблюдений копий двух случайных величин делается вывод о предпочтительности нового метода. Если мы хотим сократить число наблюдений, необходимое для достижения заданной (малой) величины риска, то целесообразно не фиксировать заранее n, а планировать прекращение испытаний после наблюдения каждой копии в зависимости от полученных ранее результатов. Существует большой класс задач управления наблюдениями – оптимального выбора случайной величины, наблюдаемой на каждом шаге статистического эксперимента, а также правила прекращения наблюдений. Все это далеко выходит за рамки тех "кратких начатков" теории статистических выводов, которые будут представлены в нашем семестровом курсе.

Мы завершим этот параграф набором простейших определений и понятий, которые постоянно используются в математической статистике.

Итак, с исследуемым объектом, относительно которого мы должны принять некоторое решение $d \in \mathcal{D}$, соотносится наблюдаемая случайная величина X, распределение которой P_θ известно с точностью до значения параметра θ. Семейство распределений $\mathcal{P} = \{P_\theta,\, \theta \in \Theta\}$, как обычно, назы-

вается *вероятностной моделью*. Пусть $(\mathcal{X}, \mathcal{A})$ – измеримое пространство значений X. В дальнейшем будет всегда предполагаться, что на сигма-алгебре $\mathcal{A}$ существует такая сигма-конечная мера μ, что при любом $\theta \in \Theta$ распределение X можно представить в виде интеграла

$$P_\theta(A) = \mathbf{P}(X \in A) = \int_A f(x \,|\, \theta)\, d\mu(x),\ A \in \mathcal{A},$$

от плотности $f(x \,|\, \theta)$ распределения X по мере μ. В таком случае распределение независимых копий $X^{(n)} = (X_1, \ldots, X_n)$ случайной величины X на произведении $(\mathcal{X}^n, \mathcal{A}^n)$ измеримых пространств $(\mathcal{X}, \mathcal{A})$ определяется функцией плотности

$$f_n(x^{(n)} \,|\, \theta) = \prod_{k=1}^{n} f(x_k \,|\, \theta)$$

по мере $\mu_n = \underbrace{\mu \times \cdots \times \mu}_{n}$, то есть

$$P_{\theta,n}(A_n) = \mathbf{P}(X^{(n)} \in A_n) = \int_{A_n} f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}),\ A_n \in \mathcal{A}^n.$$

Определение 1.1. Вектор $X^{(n)} = (X_1, \ldots, X_n)$ независимых, одинаково распределенных по тому же закону, что и наблюдаемая случайная величина X, случайных величин называется *случайной выборкой объема* n. Измеримое пространство $(\mathcal{X}^n, \mathcal{A}^n)$ значений $X^{(n)}$ называется *выборочным пространством*, а семейство распределений $\mathcal{P}_n = \{P_{\theta,n},\, \theta \in \Theta\}$ на этом пространстве – *статистической структурой* или *статистическим экспериментом*. Вектор $x^{(n)} = (x_1, \ldots, x_n)$ результатов наблюдения случайной выборки $X^{(n)}$ называется *вектором (или совокупностью) выборочных данных*.

Зная распределение выборки, мы можем вычислять риск любого статистического правила δ с помощью n-кратного интеграла

$$R(\theta;\, \delta) = \int_{\mathcal{X}} \ldots \int_{\mathcal{X}} L(\theta,\, \delta(x^{(n)})) f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}).$$

Конечно, если удастся найти распределение G_θ решающей функции δ на измеримом пространстве решений $(\mathcal{D}, \mathcal{D})$, то вычисление риска упрощается:

$$R(\theta;\, \delta) = \int_{\mathcal{D}} L(\theta,\, a)\, dG(a).$$

Так, в первом примере с выборкой из нормального (μ, σ^2) распределения решающей функцией служило выборочное среднее $\overline{X}$. Было показано, что $\overline{X}$ имеет нормальное (μ, σ^2) распределение, и именно это обстоятельство позволило нам найти простое выражение риска статистического правила через функцию распределения стандартного нормального закона. Точно так же во втором примере с выбором из двухточечного распределения $\mathrm{B}(1, p)$ решающая функция была основана на случайной величине $\sum_1^n X_k$, которая имеет распределение Бернулли $\mathrm{B}(n, p)$. Риск нашего решающего правила по выявлению эффективности метода лечения выражался через функцию распределения $\mathrm{B}(n, p)$.

Заметим, что функции от выборочного вектора $X^{(n)}$ играют важную, можно даже сказать самостоятельную, роль в математической статистике.

Определение 1.2. Любое измеримое отображение $T = T(X^{(n)})$ выборочного пространства $(\mathcal{X}^n, \mathcal{A}^n)$ в некоторое измеримое пространство $(\mathcal{T}, \mathcal{B})$ называется *статистикой*.

Существует довольно устоявшийся универсальный набор статистик, постоянно используемых в теории и практике статистического вывода; распределения этих статистик интенсивно изучались на протяжении последних двух столетий. В следующем параграфе мы познакомимся с набором статистик, которые являются выборочными аналогами стандартных характеристик распределения наблюдаемой случайной величины, а также рассмотрим статистики, редуцирующие размерность выборочного вектора до размерности параметрического пространства без потери информации.

§2. Выборочные характеристики. Достаточные статистики

Лекция 3

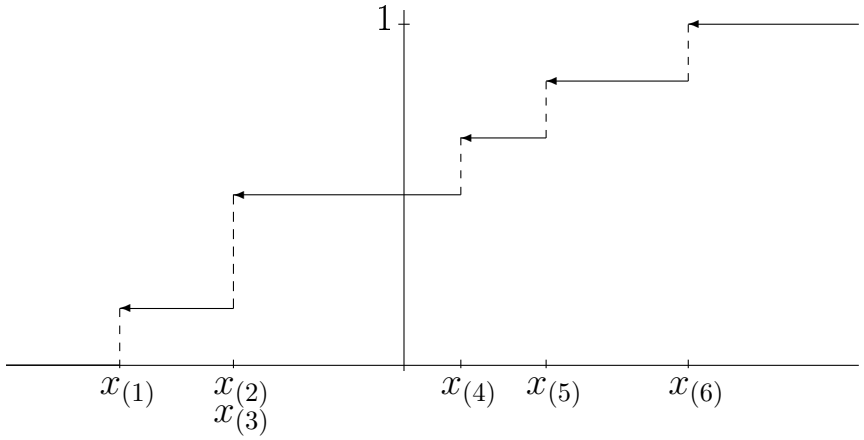
Построение вероятностных моделей в курсе теории вероятностей осуществлялось посредством спецификации функции распределения или функции плотности наблюдаемой случайной величины X. Любая из этих функций однозначно определяет распределение X на сигма-алгебре $\mathcal{A}$ борелевских множеств, порожденной интервалами в пространстве $\mathfrak{X} = \mathrm{R}$ возможных значений X, и с их помощью вычислялись такие характеристики распределения, как среднее, дисперсия, коэффициенты асимметрии и эксцесса, квантили, мода и пр. В прикладной статистике существует традиция, или, можно сказать, обязательное правило, представлять полученные экспериментальные данные с помощью статистик – выборочных аналогов этих функций и характеристик распределения X. Выборочные характеристики являются оценками истинных значений своих прообразов и позволяют судить в общих чертах о характере распределения наблюдаемой случайной величины.

Такая "описательная" статистика обычно начинается с построения *вариационного ряда*: выборочные данные $x_1, \ldots, x_n$ упорядочиваются по возрастанию их значений $x_{(1)} \le \ldots \le x_{(n)}$, и полученный таким образом вектор с неубывающими компонентами служит реализацией случайного вектора $X_{(1)}, \ldots, X_{(n)}$, который, собственно, и следует называть вариационным рядом. Компоненты вариационного ряда называются *порядковыми статистиками*, а $X_{(1)}$ и $X_{(n)}$ – *крайними членами* вариационного ряда. Мы уже сталкивались с порядковыми статистиками, когда изучали структуру пуассоновского процесса и строили вероятностную модель "слабого звена" (распределение Вейбулла).

Упорядоченные данные наносятся на ось абсцисс, и строится ступенчатая функция, возрастающая скачками величины $1/n$ в каждой точке $x_{(1)}, \ldots, x_{(n)}$. Построенная таким образом дискретная функция распределения является реализацией случайной функции

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathrm{I}(X_k < x)$$

($\mathrm{I}(A)$, как обычно, индикатор события A) и называется *эмпирической функцией распределения.*



Таким образом, дискретное эмпирическое распределение приписывает равные вероятности $1/n$ каждой из n компонент выборочного вектора, и при каждом фиксированном $x \in \mathrm{R}$ случайная величина $nF_n(x)$ подчиняется биномиальному распределению $\mathrm{B}(n,\, F(x))$:

$$P(F_n(x) = k/n) = \mathrm{C}_n^k\, F^k(x)(1 - F(x))^{n-k},\ k = 0, 1, \ldots, n.$$

В силу закона больших чисел Бернулли $F_n \underset{P}{\to} F(x)$ при любом $x \in \mathrm{R}$. Более того, теорема Гливенко–Кантелли, утверждение которой

$$D_n = \sup_{x\in\mathrm{R}} |\, F_n(x) - F(x)\,| \underset{P}{\to} 0$$

мы приводим без доказательства, указывает на равномерность этой сходимости на всей числовой оси R.

Мы закончим обсуждение свойств эмпирической функции распределения формулировкой широко известного результата А.Н. Колмогорова:

$$\lim_{n\to\infty} P(\sqrt{n}D_n < x) = \sum_{k=-\infty}^{+\infty} (-1)^k \mathrm{e}^{-k^2x^2}.$$

Полученная им формула для асимптотического ($n \to \infty$) распределения статистики $\sqrt{n}D_n$, характеризующей величину расхождения между теоретическим F и эмпирическим F_n распределениями, используется для построения *критерия согласия* выборочных данных с предположением, что F является истинной функцией распределения, из которого извлекается выборка (гипотезой о том, что F есть функция распределения наблюдаемой случайной величины X).

Итак, мы установили, что эмпирическое распределение сходится по вероятности к истинному (или, как обычно говорят прикладники, теоретическому) распределению, и теперь можем обратиться к вычислению моментных

и квантильных характеристик распределения F_n. Его нецентральные

$$a_k = \int\limits_{\mathrm{R}} x^k dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

и центральные

$$m_k = \int\limits_{\mathrm{R}} (x-a_1)^k dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} (X_i - a_1)^k$$

моменты служат выборочными аналогами соответствующих теоретических моментов $\alpha_k,\ k = 1, 2\ldots,$ и $\mu_k,\ k = 2, 3, \ldots,$ и называются *выборочными моментами.*

Если теоретические моменты существуют, то в силу закона больших чисел выборочные моменты сходятся по вероятности к своим теоретическим прообразам. Среди выборочных моментов особое место занимают моменты первого и второго порядков. Выборочный момент a_1 называется *выборочным средним* и имеет специальное обозначение $\overline{X}$; *выборочная дисперсия* $m_2 = a_2 - \overline{X}^2$ обычно обозначается S^2. Соответствующим образом определяются *выборочный коэффициент асимметрии* $g_1 = m_3/S^3$ и *выборочный коэффициент эксцесса* $g_2 = m_4/S^4 - 3$.

При выборе из m-мерного, $m > 1$, распределения эмпирическое распределение также приписывает массу n^{-1} каждому выборочному (векторному) значению $X_i = (X_{1i}, \ldots, X_{mi}),\ i = 1, \ldots, n$. В соответствии с этим мы можем определить вектор выборочных средних $\overline{X} = (\overline{X}_1, \ldots, \overline{X}_m)$ с компонентами

$$\overline{X}_k = \frac{1}{n}\sum_{i=1}^{n} X_{ki},\ k = 1, \ldots, m,$$

выборочную ковариационную матрицу $S = \| S_{kj} \|$ с элементами

$$S_{kj} = \frac{1}{n}\sum_{i=1}^{n} (X_{ki} - \overline{X}_k)(X_{ji} - \overline{X}_j) = \frac{1}{n}\sum_{i=1}^{n} X_{ki}X_{ji} - \overline{X}_k\overline{X}_j,\ k, j = 1, \ldots, m,$$

и матрицу *выборочных коэффициентов корреляции* $R = \| r_{kj} \|$ с элементами $r_{kj} = S_{kj}/\sqrt{S_{kk}S_{jj}},\ k, j = 1, \ldots, m$. Смешанные моменты более высоких порядков в многомерном случае обычно не вычисляются.

Если выбор происходит из распределения, для которого справедлива теорема сложения (предложение 12.2 курса ТВ), то распределение выборочного среднего устанавливается достаточно просто. В общем же случае

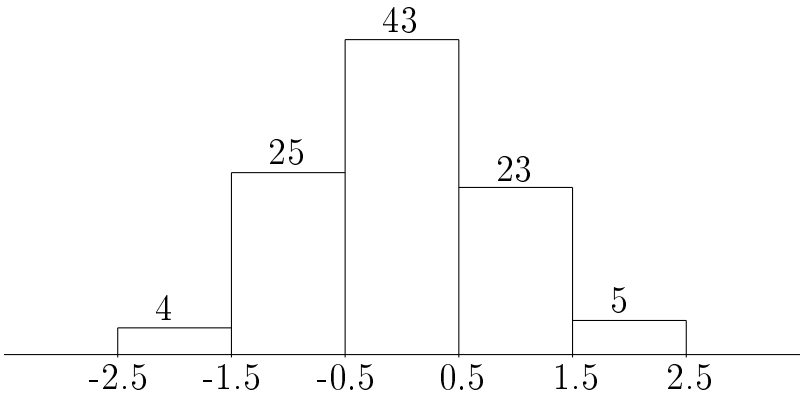
можно только утверждать об асимптотической ($n \to \infty$) нормальности этой статистики при условии существования второго момента у теоретического распределения. Аналогичное утверждение справедливо и для моментов любого k-го порядка, если у $F(x)$ существует момент порядка $2k$.

Обратимся теперь к выборочным аналогам квантилей распределения F наблюдаемой случайной величины X. Напомним, что для непрерывного распределения квантиль порядка p определялась как решение x_p уравнения $F(x) = p$, а в случае дискретного распределения – как наибольшее $x = x_p$ из носителя распределения, при котором $F(x_p) \le p$. Поскольку эмпирическое распределение дискретно, и его функция распределения $F_n(\cdot)$ возрастает скачками в точках, соответствующих компонентам вариационного ряда, то *выборочная квантиль* порядка p полагается равной порядковой статистике $X_{([np])}$, где $[\,x\,]$, как обычно, означает целую часть x. Естественно, для повышения точности оценки истинной квантили x_p можно проводить интерполяцию между статистиками $X_{([np])}$ и $X_{([np]+1)}$. Так, *выборочная медиана*, будучи квантилью порядка $p = 0.5$, обычно определяется как $\left(X_{([n/2])} + X_{([n/2]+1)}\right)/2$. Что же касается оценки моды распределения – точки наибольшего сгущения выборочных данных, то здесь нам придется обратиться к выборочным аналогам функции плотности.

При больших объемах наблюдений выборочные данные обычно подвергаются группировке, при этом индивидуальные выборочные значения не приводятся, а указываются лишь количества наблюдений, попавших в интервалы некоторого разбиения множества $\mathcal{X}$ значений наблюдаемой случайной величины. Поясним процедуру группировки на примере выборки из непрерывного одномерного распределения, когда $\mathcal{X} = \mathrm{R}$.

В декартовой системе координат ось абсцисс разбивается на $r \ge 2$ интервалов $(-\infty,\, a_1]$, $(a_1,\, a_2]$, $\ldots$, $(a_{r-2},\, a_{r-1}]$, $(a_{r-1},\, +\infty)$, причем внутренние интервалы выбираются, как правило, одинаковой длины: $a_i - a_{i-1} = \Delta,\ i = 2, \ldots, r-1$. Выборочные данные сортируются по интервалам разбиения и подсчитываются частоты n_i, $i = 1, \ldots, r$ попадания данных в каждый интервал. Над каждым внутренним интервалом рисуется прямоугольник высоты $n_i/n\Delta$, так что площадь n_i/n каждого прямоугольника с номером $i = 2, \ldots, r-1$ служит реализацией частотной оценки ν_i/n вероятности попадания наблюдаемой случайной величины X в соответствующий интервал. Здесь ν_i – статистика, которую можно записать с помощью индикаторов событий $A_{ij} = \{X_j \in (a_{i-1},\, a_i]\}$, $i = 1, \ldots, r,\ a_0 = -\infty,\, a_r =$

$+\infty$, $j = 1, \ldots, n$, а именно $\nu_i = \sum_{j=1}^{n} \mathrm{I}(A_{ij})$. Полученная таким образом случайная ступенчатая функция, принимающая нулевые значения на крайних интервалах $(-\infty, a_1]$, $(a_{r-1}, +\infty)$ и равная $\nu_i/n\Delta$ на внутренних интервалах с номерами $i = 2, \ldots, r-1$, называется *гистограммной оценкой* f_n функции плотности $f(x)$, $x \in \mathrm{R}$ распределения X, а ее реализация (ν_i заменяются на наблюдаемые частоты n_i, $i = 1, \ldots, r$) – *гистограммой* выборки $x^{(n)}$.



В математической статистике существует ряд теорем, устанавливающих, что при определенных условиях на плотность f, гистограммная оценка $f_n(x) \underset{P}{\to} f(x)$ при любом $x \in \mathrm{R}$, если $n \to \infty$ и одновременно $r \to \infty$, а $\Delta \to 0$ со скоростью, зависящей определенным образом от n и r.

В случае гистограммной оценки функции плотности естественно считать выборочным аналогом (оценкой) моды распределения X середину интервала разбиения, в котором гистограмма принимает наибольшее значение.

Заметим также, что вектор частот $(\nu_1, \ldots, \nu_r)$ имеет мультиномиальное распределение $\mathcal{M}(r, n, \mathbf{p})$ с вероятностями исходов $p_i = F(a_i) - F(a_{i-1})$, $i = 1, \ldots, r$, что позволяет найти распределение оценки $f_n(x)$ при любом $x \in \mathrm{R}$ и построить критерий согласия выборочных данных с гипотезой о виде распределения наблюдаемой случайной величины. Это широко используемый на практике *критерий хи-квадрат*, основанный на статистике (сравните с критерием Колмогорова D_n)

$$X^2 = \sum_{1}^{r} \frac{(\nu_i - np_i)^2}{np_i}\,.$$

Асимптотическое распределение этой статистики мы изучим в параграфе, посвященном статистической проверке гипотез.

Итак, мы рассмотрели основные выборочные аналоги распределения наблюдаемой случайной величины и его основных характеристик. Мы высказали также ряд утверждений о распределении этих статистик, что позволит нам в последующем вычислять последствия от их использования в качестве решающих функций. Для того, чтобы уяснить, насколько важно знать хотя бы среднее значение статистики, претендующей на роль решающей функции, обратимся снова к примеру 1.1 по аттестации партии дизельного топлива, где обсуждалась сопутствующая проблема оценки дисперсии σ^2 наблюдаемой случайной величины $X \sim \mathcal{N}(\mu,\, \sigma^2)$.

Предлагалось оценивать σ^2 по накопленному в лаборатории архиву испытаний аттестуемых партий дизельного топлива, то есть по данным большого числа N выборок $X_1^{(n)}, \ldots, X_N^{(n)}$ малого объема n. Каждая i-ая выборка извлекается из нормального $(\mu_i,\, \sigma^2)$ распределения, причем средние μ_i могут быть различными для разных выборок, $i = 1, \ldots, N$, но дисперсия σ^2 у всех выборок одна и та же. Предлагается следующая оценка σ^2. В каждой выборке вычисляется выборочная дисперсия S_i^2, $i = 1, \ldots, N$, и затем берется их арифметическое среднее: $\hat{\sigma}_N^2 = (1/N)\sum_1^N S_i^2$. Распределение каждой S_i^2 не зависит от μ_i, $i = 1, \ldots, N$, поскольку выборочная дисперсия инвариантна относительно сдвигов $X_k \to X_k + a$. Следовательно, предлагаемая оценка есть нормированная на N сумма независимых, одинаково распределенных случайных величин – копий статистики $S^2 = (1/n)\sum_1^n (X_k - \overline{X})^2$, и в силу закона больших чисел $\hat{\sigma}_N^2 \underset{P}{\to} \mathbf{E}S^2$ при неограниченном возрастании объема N архивных данных. Вычислим это математическое ожидание:

$$\mathbf{E}S^2 = \mathbf{E}\left(\frac{1}{n}\sum_1^n X_k^2 - \overline{X}^2\right) = \mathbf{E}X^2 - \mathbf{E}\overline{X}^2 = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2,$$

поскольку $\alpha_2 = \mathbf{E}X^2 = \mathbf{D}X + \mathbf{E}^2X$, а $\overline{X} \sim \mathcal{N}(\mu,\, \sigma^2/n)$.

Таким образом, предлагаемая оценка обладает значительным смещением при малом объеме n испытаний каждой партии дизельного топлива. Например, в случае $n = 2$ мы занижаем дисперсию в два раза, поскольку $\hat{\sigma}_N^2 \underset{P}{\to} \sigma^2/2$. Естественно, этот дефект легко устраним – достаточно использовать исправленную на смещение оценку $\tilde{\sigma}_N^2 = (n/(n-1))\hat{\sigma}_N^2$.

В завершении этого параграфа мы изучим еще один класс замечательных статистик, используя которые можно редуцировать выборочные данные только к их значениям без потери информации. К сожалению, не все статистические структуры обладают такими статистиками, но, по существу, только в тех структурах, где имеются достаточные статистики, возможно построение оптимального статистического правила, на котором достигается минимум риска.

Идея, состоящая в том, что в определенных случаях для принятия решения без увеличения риска *достаточно* знать только значения некоторых статистик, а не все выборочные данные, не требует введения специальных мер информации, содержащейся в выборочных данных и статистиках, – все становится ясным при рассмотрении следующей простейшей задачи, с которой мы имели дело в самом начале курса теории вероятностей.

Предположим, что мы хотим узнать вероятность наследования доминантного признака в опытах Менделя и располагаем результатами $x_1, \ldots, x_n$ скрещиваний n пар, где, как обычно, каждое x_i есть индикатор наследования признака, $i = 1, \ldots, n$, а совокупность выборочных данных представляет реализацию случайной выборки $X_1, \ldots, X_n$ из двухточечного распределения с функцией плотности $f(x \,|\, \theta) = P_\theta(X = x) = \theta^x(1-\theta)^{1-x}$, отличной от нуля только в точках $x = 0$ и 1. Частотная оценка $\hat{\theta}_n = T/n$ вероятности θ наследования признака определяется статистикой $T = \sum_1^n X_k$, выборочное значение $t = \sum_1^n x_k$ которой соответствует числу потомков в эксперименте, наследовавших доминантный признак. Естественно, возникает вопрос, а нельзя ли извлечь дополнительную информацию о величине параметра θ из номеров $k_1, \ldots, k_t$ выборочных данных, принявших значение 1? Нетрудно понять, что это возможно только в том случае, если *распределение выборочного вектора $X^{(n)}$ при условии, что статистика T приняла фиксированное значение t, зависит от параметра θ*. Действительно, если мы будем наблюдать случайную величину, которая не имеет никакого отношения к интересующему нас параметру, то откуда этой информации взяться? Итак, найдем условное распределение $X^{(n)}$ относительно T.

Используя формулу условной вероятности, получаем, что

$$P_\theta\left(X^{(n)} = x^{(n)} \,|\, T = t\right) = \frac{P_\theta\left(\{X^{(n)} = x^{(n)}\} \bigwedge \{\sum_1^n X_k = t\}\right)}{P_\theta\left(\sum_1^n X_k = t\right)}.$$

Если значения компонент вектора $x^{(n)}$ таковы, что $\sum_1^n x_k \neq t$, то события $X^{(n)} = x^{(n)}$ и $\sum_1^n X_k = t$, очевидно, несовместны, и поэтому в этом случае условная вероятность равна нулю (не зависит от θ). Если же $\sum_1^n x_k = t$, то событие $X^{(n)} = x^{(n)}$ влечет событие $\sum_1^n X_k = t$, и формула для вычисления условной вероятности упрощается:

$$P_\theta\left(X^{(n)} = x^{(n)} \,|\, T = t\right) = \frac{P_\theta\left(X^{(n)} = x^{(n)}\right)}{P_\theta\left(\sum_1^n X_k = t\right)}.$$

Так как

$$P_\theta\left(X^{(n)} = x^{(n)}\right) = f_n(X^{(n)} \,|\, \theta) = \theta^{\sum_1^n x_k}(1-\theta)^{n-\sum_1^n x_k},$$

$$P_\theta\left(\sum_1^n X_k = t\right) = \mathrm{C}_n^t\,\theta^t(1-\theta)^{n-t},$$

то в случае $\sum_1^n x_k = t$ условное распределение выборочного вектора $X^{(n)}$ относительно статистики T имеет вид

$$P_\theta\left(X^{(n)} = x^{(n)} \,|\, T = t\right) = \frac{1}{\mathrm{C}_n^t},$$

и также не зависит от θ.

Итак, наши выкладки показывают, что распределение выборочного вектора на "плоскости" $\sum_1^n X_k = t$ не зависит от θ, и поэтому расположение значений $x_k = 1$ в последовательности $x_1, \ldots, x_n$ при фиксированном количестве таких значений не несет информации о параметре θ.

Определение 2.1. Статистика $T = T(X^{(n)})$ называется *достаточной* для статистической структуры $\mathcal{P}_n = \{P_{\theta,n},\, \theta \in \Theta\}$, если условное распределение выборочного вектора $X^{(n)}$ относительно статистики T не зависит от θ.

В общей теории статистического вывода в рамках более общего определения статистического правила устанавливается замечательный факт: *если статистическая структура обладает достаточной статистикой* T, *то, каково бы ни было статистическое правило* $\delta = \delta(X^{(n)})$, *всегда существует правило* $\delta^* = \delta^*(T)$, *основанное только на* T, *риск которого совпадает с риском правила* δ. Таким образом, построение оптимальных

статистических правил следует начинать с поиска достаточных статистик. Следующая теорема дает критерий существования у статистических структур достаточных статистик и, одновременно, указывает простой способ их нахождения.

Теорема 2.1. *Для того, чтобы $T = T(X^{(n)})$ была достаточной статистикой для статистической структуры, определяемой функцией плотности $f_n(x^{(n)} \,|\, \theta)$, необходимо и достаточно,чтобы эта функция допускала представление*

$$f_n(x^{(n)} \,|\, \theta) = g_\theta \left(T(x^{(n)}) \right) h(x^{(n)}), \tag{1}$$

где функция h не зависит от параметра θ, а функция g зависит от θ и аргумента $x^{(n)}$ только через значения $T(x^{(n)})$ статистики $T = T(X^{(n)})$,

Доказательство теоремы мы проведем только для дискретного распределения наблюдаемой случайной величины, когда функция плотности выборки $f_n(x^{(n)} \,|\, \theta) = P_\theta \left(X^{(n)} = x^{(n)} \right)$. В случае непрерывного распределения схема доказательства та же, но придется делать замену в n-кратном интеграле.

Достаточность. Пусть выполняется факторизационное представление (1); требуется показать, что условное распределение $X^{(n)}$ относительно T не зависит от θ. Как и в только что рассмотренном примере с двухточечным распределением, воспользуемся формулой условной вероятности для вычисления условной плотности $X^{(n)}$ относительно T :

$$P_\theta \left(X^{(n)} = x^{(n)} \,|\, T(X^{(n)}) = t \right) = \frac{P_\theta \left(\{X^{(n)} = x^{(n)}\} \bigwedge \{T(X^{(n)}) = t\} \right)}{P_\theta (T(X^{(n)}) = t)}.$$

События, стоящие в числителе, будут несовместными, если $T(x^{(n)}) \neq t$, и в этом случае условная вероятность равна нулю (не зависит от θ). Если же $T(x^{(n)}) = t$, то первое по порядку событие в числителе влечет второе, и поэтому формула для вычисления условной вероятности упрощается:

$$P_\theta \left(X^{(n)} = x^{(n)} \,|\, T(X^{(n)}) = t \right) = \frac{P_\theta \left(X^{(n)} = x^{(n)} \right)}{P_\theta (T(X^{(n)}) = t)}.$$

Так как $P_\theta(X^{(n)} = x^{(n)}) = f_n(x^{(n)} \,|\, \theta)$, то используя представление (1), получаем, что (напомним, $T(x^{(n)}) = t$)

$$P_\theta \left(X^{(n)} = x^{(n)} \,|\, T(X^{(n)}) = t \right) =$$

$$\frac{g_\theta(T(x^{(n)}))h(x^{(n)})}{\sum\limits_{y^{(n)}:\,T(y^{(n)})=t} g_\theta(T(y^{(n)}))h(y^{(n)})} = \frac{h(x^{(n)})}{\sum\limits_{y^{(n)}:\,T(y^{(n)})=t} h(y^{(n)})}.$$

Таким образом, условное распределение не зависит от θ, и поэтому статистика T достаточна для $\mathcal{P}_n$.

Необходимость. Пусть T – достаточная статистика, так что условное распределение $P_\theta\left(X^{(n)} = x^{(n)} \,|\, T(X^{(n)}) = t\right) = K(x^{(n)}, t)$, где функция K не зависит от θ. Требуется показать, что в этом случае для функции плотности выборки справедливо представление (1).

Имеем

$$f_n(x^{(n)} \,|\, \theta) = P_\theta\left(X^{(n)} = x^{(n)}\right) =$$

$$P_\theta\left(\{X^{(n)} = x^{(n)}\} \bigwedge \{T(X^{(n)}) = T(x^{(n)})\}\right) =$$

$$P_\theta(T(X^{(n)}) = T(x^{(n)})) \cdot P_\theta\left(X^{(n)} = x^{(n)} \,|\, T(X^{(n)}) = T(x^{(n)})\right).$$

Мы получили представление (1) с $g_\theta(T(x^{(n)})) = P_\theta(T(X^{(n)}) = T(x^{(n)}))$ и $h(x^{(n)}) = K(x^{(n)}, T(x^{(n)}))$. Теорема доказана.

Рассмотрим несколько примеров на применения полученного критерия достаточности к статистическим структурам, соответствующим вероятностным моделям из нашего курса теории вероятностей. Начнем с двухточечного распределения (выбор в схеме Бернулли), где мы непосредственными вычислениями условного распределения убедились в достаточности статистики, реализующей число успешных испытаний, – посмотрим, как это делается с помощью представления (1).

1^0. *Двухточечное распределение* $\mathrm{B}(1, \theta)$ имеет функцию плотности

$$f(x \,|\, \theta) = \theta^x (1-\theta)^{1-x},$$

отличную от нуля только в точках $x = 0$ и 1. Параметрическое пространство этого распределения $\Theta = (0;\, 1)$, а функция плотности случайной выборки

$$f_n(x^{(n)} \,|\, \theta) = \theta^{\sum_1^n x_k} (1-\theta)^{n - \sum_1^n x_k}.$$

Представление (1) выполняется с $h(x^{(n)}) \equiv 1$ и $T(x^{(n)}) = \sum\limits_1^n x_k$. Следовательно, $T = \sum\limits_1^n X_k$ – достаточная статистика.

2^0. *Распределение Пуассона* $\mathrm{P}(\theta)$, для которого

$$f(x\,|\,\theta) = \frac{\theta^x\,\mathrm{e}^{-\theta}}{x!},\ x = 0, 1, \ldots, \quad \Theta = \mathrm{R}_+,$$

функция плотности выборки

$$f_n(x^{(n)}\,|\,\theta) = \theta^{\sum_1^n x_k}\,\mathrm{e}^{-n\theta} \,/\, \prod_1^n x_k!.$$

Следовательно, в представлении (1)

$$h(x^{(n)}) = \Big[\prod_1^n x_k!\Big]^{-1},$$

и $T = \sum_1^n X_k$ – достаточная статистика.

3^0. *Показательное распределение* $\mathrm{E}(\theta)$ с

$$f(x\,|\,\theta) = \frac{1}{\theta}\exp\left\{-\frac{x}{\theta}\right\},\ x \geq 0, \quad \Theta = \mathrm{R}_+,$$

и

$$f_n(x^{(n)}\,|\,\theta) = \frac{1}{\theta^n}\exp\left\{-\frac{1}{\theta}\sum_1^n x_k\right\}$$

Также обладает достаточной статистикой $T = \sum_1^n X_k$.

4^0. *Равномерное распределение* $\mathrm{U}(a,\, b)$, функция плотности которого

$$f(x\,|\,\theta) = \frac{\mathrm{I}_{[\,a;\,b\,]}(x)}{b-a}$$

отлична от нуля и постоянна на отрезке $[\,a;\,b\,]$, на что указывает стоящая в числителе индикаторная функция отрезка $[\,a;\,b\,]$. В этом распределении $\theta = (a,\, b)$ – двумерный параметр и параметрическое пространство $\Theta = \{(a,\, b) : (a,\, b) \in \mathrm{R}^2,\, a < b\}$. Статистическая структура определяется функцией плотности

$$f_n(x^{(n)}\,|\,\theta) = \frac{\prod_1^n \mathrm{I}_{[\,a;\,b\,]}\,(x_k)}{(b-a)^n},$$

и поскольку произведение индикаторов, стоящее в числителе этой функции, принимает значение 1 в случае

$$a \leq \min_{1\leq k\leq n} x_k \leq \max_{1\leq k\leq n} x_k \leq b$$

и значение 0 при нарушении этих неравенств, то вектор $T = (\, X_{(1)},\, X_{(n)}\,)$ крайних членов вариационного ряда является достаточной статистикой.

5^0. *Нормальное распределение* $\mathcal{N}(\mu,\, \sigma^2)$. Это распределение обладает двумерным параметром $\theta = (\mu,\, \sigma)$ с областью значений (параметрическим пространством) $\Theta = \mathrm{R} \times \mathrm{R}_+$. Функции плотности наблюдаемой случайной величины X и случайной выборки $X^{(n)}$ определяются соответственно как

$$f(x \,|\, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

и

$$f_n(x^{(n)} \,|\, \theta) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n (x_k - \mu)^2\right\} =$$

$$\frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_1^n x_k^2 - 2\mu\sum_1^n x_k + n\mu^2\right)\right\}.$$

Последнее выражение для плотности $X^{(n)}$ показывает, что двумерная статистика $T = (T_1,\, T_2)$ с $T_1 = \sum_1^n X_k$ $T_2 = \sum_1^n X_k^2$ достаточна для статистической структуры нормального распределения. Кроме того, поскольку $T_1 = n\overline{X}$ и $T_2 = n(S^2 + \overline{X}^2)$, то факторизационное равенство (1) указывает на достаточность статистик $\overline{X}$ и S^2, которые имеют конкретную статистическую интерпретацию и поэтому более удобны для практического использования. Понятно, что это замечание носит общий характер: *любые взаимно однозначные преобразования достаточной статистики наследуют свойство достаточности.*

Отметим также, что в случае известного (фиксированного) σ статистическая структура имеет параметрическое пространство, совпадающее с областью значений параметра μ, и достаточной статистикой будет выборочное среднее $\overline{X}$. Аналогичное утверждение имеет место для статистики S^2 при фиксированном μ.

6^0. *Гамма-распределение* $\mathrm{G}(\lambda,\, a)$ имеет функцию плотности

$$f(x \,|\, \theta) = \frac{1}{a^\lambda \Gamma(\lambda)} x^{\lambda-1} \exp\left\{-\frac{x}{a}\right\},\ x > 0, \quad \theta = (a,\, \lambda),\, a > 0,\, \lambda > 0,$$

так что функция плотности выборочного вектора

$$f(x \,|\, \theta) = \frac{1}{a^{n\lambda}\Gamma^n(\lambda)} \left[\prod_1^n x_k\right]^{\lambda-1} \exp\left\{-\frac{1}{a}\sum_1^n x_k\right\}.$$

Тождество (1) указывает, что достаточной является двумерная статистика $\left(\sum_1^n X_k,\, \prod_1^n X_k\right)$ или более удобная в вычислительном отношении статистика $\left(\sum_1^n X_k,\, \sum_1^n \ln X_k\right)$. Для этого распределения можно сделать то же замечание, что и для нормального: первая компонента достаточной статистики "отвечает" за масштабный параметр a, в то время как вторая соответствует параметру формы λ.

7^0. *Биномиальное распределение* $\mathrm{B}(m,\, p)$. Это дискретное распределение, сосредоточенное в точках $x = 0, 1, \ldots, m$, с функцией плотности

$$f(x \,|\, \theta) = \mathrm{C}_m^x\, p^x (1-p)^{m-x},$$

зависящей от двумерного параметра $\theta = (m,\, p)$, первая компонента m которого может принимать только значения из множества $\mathrm{N} = \{1, 2, \ldots\}$, а вторая компонента $p \in (0;\, 1)$. Функция плотности выборочного вектора

$$f_n(x^{(n)} \,|\, \theta) = \prod_{k=1}^n \mathrm{C}_m^{x_k} \cdot p^{\sum_1^n x_k} (1-p)^{nm - \sum_1^n x_k}.$$

Применение критерия (1) показывает, что для статистической структуры с параметрическим пространством $\Theta = \mathrm{N} \times (0;\, 1)$ достаточной статистикой может быть только весь выборочный вектор $X^{(n)}$, но если $\Theta = (0;\, 1)$ (значение параметра m известно), то $\sum_1^n X_k$ – достаточная статистика.

8^0. *Распределение Коши* $\mathrm{C}(a,\, b)$ имеет функцию плотности выборочного вектора

$$f_n(x^{(n)} \,|\, \theta) = \pi^{-n} b^{-n} \prod_{k=1}^n \left(1 + \left(\frac{x_k - a}{b}\right)^2\right)^{-1},$$

и в силу критерия (1) его статистическая структура обладает только *тривиальной* достаточной статистикой $T = X^{(n)}$.

Мы не будем выписывать статистические структуры многомерных распределений в силу их чрезвычайной громоздкости, но нетрудно установить по аналогии с рассмотренными примерами, что у структуры мультиномиального распределения $\mathcal{M}(m, 1, \mathbf{p})$ с $m \geq 2$ исходами и вектором $\mathbf{p} = (p_1, \ldots, p_m)$ вероятностей соответствующих исходов достаточным будет вектор, состоящий из частот этих исходов в мультиномиальной схеме испытаний, а у структуры многомерного нормального распределения

$\mathcal{N}_m(\mu, \Lambda)$ достаточную статистику образуют вектор выборочных средних и выборочная ковариационная матрица.

На этом завершается вводная часть нашего курса математической статистики. Мы сделали постановку проблемы статистического вывода, провели классификацию основных статистических структур и теперь мы готовы к решению конкретных статистических проблем по оценке параметров распределения наблюдаемой случайной величины и проверке гипотез, касающихся структуры параметрического пространства этого распределения.

§3. Оценка параметров. Метод моментов

Лекция 5

Мы приступаем к решению статистической проблемы оценки неизвестного значения параметра θ, индексирующего семейство $\mathcal{P} = \{P_\theta,\, \theta \in \Theta\}$ возможных распределений наблюдаемой случайной величины X. Будут рассматриваться только конечномерные параметрические пространства $\Theta = \mathrm{R}^k$, $k \geq 1$. Информация о значении θ поступает к нам в виде выборочных данных $x^{(n)} = (x_1, \ldots, x_n)$ – результатов наблюдений n независимых копий $X^{(n)} = (X_1, \ldots, X_n)$ случайной величины X. Напомним, семейство $\mathcal{P}$ мы назвали вероятностной моделью, а случайный вектор $X^{(n)}$ – случайной выборкой объема n.

В этой проблеме, о которой мы несколько раз упоминали в предыдущем параграфе, пространство решений $\mathcal{D}$ совпадает с параметрическим пространством Θ, решающая функция $\delta = \delta(X^{(n)})$ – статистика с областью значений $\mathcal{T} = \Theta$ – называется *оценкой параметра* θ и обычно обозначается $\theta_n,\ \hat{\theta}_n,\ \theta_n^*$ и тому подобное. Функции потерь $L(\theta,\, d)$ в проблеме оценивания обычно выбираются в виде неубывающей функции расстояния $|\, d - \theta\,|$ (в эвклидовой метрике) между значением оценки $d = \hat{\theta}_n(x^{(n)})$ и истинным значением θ оцениваемого параметра.

Основная задача статистической теории оценивания состоит в построении оценки $\theta_n^* = \theta_n^*(X^{(n)})$, *минимизирующей равномерно по* $\theta \in \Theta$ *функцию риска*

$$R(\theta;\, \hat{\theta}_n) = \mathbf{E}_\theta L(\theta,\, \hat{\theta}_n(X^{(n)})).$$

Таким образом, какова бы ни была статистическая оценка $\hat{\theta}_n$, для оценки θ_n^* с равномерно минимальным риском при любом $\theta \in \Theta$ справедливо неравенство $R(\theta;\, \theta_n^*) \leq R(\theta;\, \hat{\theta}_n)$.

Мы рассмотрим одно из решений этой задачи в случае оценки скалярного параметра ($\Theta = \mathrm{R}$) при квадратичной функции потерь $L(\theta, d) = (d-\theta)^2$, но сначала познакомимся с традиционно используемыми в статистической практике методами оценки параметров и изучим распределение этих оценок с целью вычисления их функции риска.

Конечно, далеко не все используемые на практике методы приводят к оптимальным оценкам, иногда бывает трудно найти оценку, обладающую хоть какими-нибудь привлекательными свойствами. Понятно, что считать оценкой любое измеримое отображение выборочного пространства $\mathcal{X}^n$ в параметрическое пространство Θ не совсем разумно, и поэтому мы введем

некоторые условия, которым должна удовлетворять статистика $\hat{\theta}_n$, чтобы претендовать на роль *оценки*. Разрабатывая в дальнейшем методы оценивания и предлагая конкретные оценки, мы всегда будем проверять выполнимость этих условий.

Определение 3.1. Оценка $\hat{\theta}_n$ параметра θ называется *состоятельной*, если $\hat{\theta}_n(X^{(n)}) \underset{P}{\to} \theta$ при любом $\theta \in \Theta$, когда объем выборки $n \to \infty$. Оценка $\hat{\theta}_n$ называется *несмещенной в среднем*, если $\mathbf{E}_\theta \hat{\theta}_n(X^{(n)}) = \theta$, каково бы ни было значение $\theta \in \Theta$.

Напомним, что $\hat{\theta}_n(X^{(n)}) \underset{P}{\to} \theta$ означает, что для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P_\theta \left(\left| \hat{\theta}_n(X^{(n)}) - \theta \right| > \varepsilon \right) = 0,$$

или, что то же,

$$\lim_{n\to\infty} P_\theta \left(\left| \hat{\theta}_n(X^{(n)}) - \theta \right| \leq \varepsilon \right) = 1. \tag{1}$$

Здесь, как обычно, в случае векторного параметра θ запись $|\,\theta_1 - \theta_2\,|$ означает расстояние между точками θ_1 и θ_2 эвклидова пространства Θ.

В предыдущем параграфе мы показали, что выборочные моменты $a_i = (1/n)\sum_1^n X_j^i$ являются состоятельными оценками соответствующих "теоретических" моментов $\alpha_i = \mathbf{E}_\theta X^i$, которые являются функциями оцениваемого параметра: $\alpha_i = \alpha_i(\theta)$, $i = 1, 2, \ldots$. Этот результат указывает нам довольно простой метод построения состоятельных оценок в случае существования у распределения P_θ наблюдаемой случайной величины X момента порядка k, где k – число компонент $\theta_1, \ldots, \theta_k$ оцениваемого параметрического вектора θ.

Приравняем теоретические моменты выборочным и разрешим полученную таким образом систему уравнений $\alpha_i(\theta_1, \ldots, \theta_k) = a_i$, $i = 1, \ldots, k$ относительно переменных $\theta_1, \ldots, \theta_k$. Любое решение $\hat{\theta}_n(\mathbf{a}) = (\hat{\theta}_{1n}(\mathbf{a}), \ldots, \hat{\theta}_{kn}(\mathbf{a}))$, $\mathbf{a} = (a_1, \ldots, a_n)$, этой системы называется *оценкой θ по методу моментов*, и прежде, чем исследовать свойства таких оценок, рассмотрим несколько примеров на применения *метода моментов*.

В курсе теории вероятностей, изучая новые вероятностные модели, мы всегда вычисляли их моментные характеристики. Например, мы знаем, что средние значения двухточечного распределения B(1, θ), распределения Пуассона P(θ) и показательного распределения E(θ) равны θ. Следовательно, выборочное среднее $\overline{X}$ есть оценка по методу моментов параметра θ

любого из этих распределений. Легко видеть, что эта оценка состоятельна и несмещена. Точно так же у нормального распределения $\mathcal{N}(\mu,\, \sigma^2)$ параметр μ означает среднее значение, а σ^2 – дисперсию этого распределения. Следовательно, выборочное среднее $\overline{X}$ и выборочная дисперсия S^2 есть состоятельные оценки соответствующих компонент μ и σ^2 параметрического вектора $\theta = (\mu,\, \sigma^2)$. Исправляя смещение оценки S^2 компоненты σ^2, получаем несмещенную оценку θ. Замечательно то, что все оценки являются достаточными статистиками, и это обстоятельство, как будет видно в дальнейшем, предопределяет их оптимальные свойства. Распределение оценки $\overline{X}$ легко получить, используя теоремы сложения для распределений B, P, E и $\mathcal{N}$, распределение же S^2 при выборе из нормального распределения мы найдем несколько позже.

Рассмотрим теперь примеры, в которых приходится решать систему уравнений, и найденные оценки по методу моментов не являются функциями достаточных статистик.

Пример 3.1. *Оценка параметров биномиального распределения* $\mathrm{B}(m,\, p)$. Проблема состоит в оценке обеих компонент m и p двумерного параметра $\theta = (m,\, p)$. Из курса теории вероятностей нам известно, что среднее значение биномиального распределения равно mp, а дисперсия – $mp(1-p)$. Приравнивая эти теоретические моменты их выборочным аналогам, получаем систему для определения оценок по методу моментов: $mp = \overline{X},\ mp(1-p) = S^2$. Разделив второе уравнение на первое, находим оценку $\hat{p}_n = (\overline{X} - S^2)/\overline{X}$ параметра p, после чего, обращаясь к первому уравнению, находим оценку $\hat{m}_n = \overline{X}^2/(\overline{X} - S^2)$ параметра m. Легко показать, что эти оценки обладают свойством состоятельности (общий метод доказательства таких утверждений смотрите в приведенной ниже теореме 3.1), но при малых n велика вероятность получить отрицательные значения оценок, оценка параметра m, как правило, не будет целым числом, наконец, можно показать, что оценка $p_n^* = \overline{X}$ параметра p будет обладать меньшим квадратичным риском, чем оценка $\hat{p}_n$. Все это, конечно, печально, однако другие методы, приводящие к более точным оценкам, обладают значительными вычислительными трудностями.

Пример 3.2. *Оценка параметров гамма-распределения* $\mathrm{G}(\lambda,\, a)$. У этого двухпараметрического распределения среднее равно λa, а дисперсия – λa^2. Решение системы уравнений $\lambda a = \overline{X},\ \lambda a^2 = S^2$ дает оценки $\hat{a}_n = S^2/\overline{X},\ \hat{\lambda}_n = \overline{X}^2/S^2$, которые, как и в предыдущем примере, не являются

функциями достаточной статистики $\left(\sum_1^n X_k,\, \prod_1^n X_k\right)$, и как показывают не совсем простые вычисления, их риски далеки от возможного минимума. Тем не менее, очевидная вычислительная простота оценок параметров гамма-распределения по методу моментов обеспечивает их популярность в практических применениях.

Изучим теперь асимптотические свойства оценок по методу моментов – установим условия их состоятельности и исследуем поведение их распределений при больших объемах выборок. Для простоты мы ограничимся случаем одномерного параметра θ, оценка которого определяется решением уравнения $\mu(\theta) = \mathbf{E}_\theta X = \overline{X}$, и предположим, что это уравнение имеет единственное решение $\hat{\theta}_n = h(\overline{X})$. Понятно, что $h(\cdot) = \mu^{-1}(\cdot)$, так что $h(\mu(\theta)) \equiv \theta$. О возможности распространения наших результатов на случай векторного θ мы поговорим отдельно.

Теорема 3.1. *Если наблюдаемая случайная величина X имеет конечное среднее значение $\mu = \mu(\theta)$ и функция $h(\cdot)$ непрерывна в области значений выборочного среднего $\overline{X}$, то $\hat{\theta}_n = h(\overline{X})$ является состоятельной оценкой параметра θ по методу моментов.*

Доказательство.. Мы воспользуемся формулой (1) в определении состоятельности оценки и покажем, что для любых $\varepsilon > 0$ и $\alpha > 0$ существует такое $N(\varepsilon,\, \alpha)$, что для всех $n > N(\varepsilon,\, \alpha)$ вероятность

$$P_\theta \left(\left| h(\overline{X}) - \theta \right| \le \varepsilon \right) > 1 - \alpha. \tag{2}$$

Поскольку $h(\mu) = h(\mu(\theta)) = \theta$ и $\overline{X} \underset{P}{\to} \mu$, то нам достаточно показать, что свойство (или определение) непрерывности функции: $h(x) \to h(\mu)$ при $x \to \mu$, остается справедливым при замене обычной сходимости $'' \to ''$ на сходимость по вероятности $'' \underset{P}{\to} ''$, то есть $\overline{X} \underset{P}{\to} \mu$ влечет $h(\overline{X}) \underset{P}{\to} h(\mu)$. Это почти очевидно, поскольку событие, состоящее в попадании в окрестность нуля случайной величины $|\,\overline{X} - \mu\,|$, влечет аналогичное событие для случайной величины $|\,h(\overline{X}) - h(\mu)\,|$, но все же проведем строгое доказательство на языке $''\varepsilon - \delta''$.

Так как $\overline{X} \underset{P}{\to} \mu(\theta)$, а $h(\cdot)$ – непрерывная функция, то найдутся такие $\delta = \delta(\varepsilon,\, \alpha)$ и $N = N(\varepsilon, \alpha)$, что

$$P(\,|\,\overline{X} - \mu(\theta)\,|\, < \delta\,) > 1 - \alpha \tag{3}$$

для всех $n > N$ и *событие* $|\,\overline{X} - \mu(\theta)\,| < \delta$ *повлечет событие* $|\,h(\overline{X}) -$

$h(\mu(\theta))\,| = |\,h(\overline{X}) - \theta\,| \le \varepsilon$. В силу этого неравенство (2) становится следствием неравенства (3). Теорема доказана.

Анализ доказательства показывает, что теорема состоятельности остается справедливой в случае векторного параметра θ, если воспользоваться определением непрерывности векторной функции от векторного аргумента, связав его с расстояниями в эвклидовых пространствах значений функции и ее аргумента.

Обратимся теперь к асимптотическому анализу распределения оценки $\hat{\theta}_n = h(\overline{X})$ при $n \to \infty$. Понятно, что в данном случае употребление термина "оценка" применительно к функции $h(\overline{X})$ ничего особенно не добавляет – речь идет просто об асимптотическом распределении статистики, имеющей вид функции от выборочного среднего.

Теорема 3.2. *Если $X^{(n)}$ – случайная выборка из распределения с конечными средним значением μ и дисперсией σ^2, а функция $h(x)$ обладает ограниченной второй производной $h''(x)$ в некоторой окрестности точки $x = \mu$, то*

$$\lim_{n\to\infty} P\left(\sqrt{n}\left(h(\overline{X}) - h(\mu)\right) < x\right) = \Phi\left(\frac{x}{\sigma|\,h'(\mu)\,|}\right). \tag{4}$$

где $\Phi(\cdot)$ – функция распределения стандартного нормального закона.

Доказательство.. Понятно, что мы должны воспользоваться центральной предельной теоремой (§14 курса ТВ) применительно к статистике $\overline{X} = \frac{1}{n}\sum_1^n X_k$:

$$\lim_{n\to\infty} P\left(\sqrt{n}\left(\overline{X} - \mu\right) < x\right) = \Phi\left(\frac{x}{\sigma}\right). \tag{5}$$

Стандартный прием использования этой теоремы при асимптотическом анализе функций от сумм независимых, одинаково распределенных случайных величин состоит в "линеризации" таких функций с помощью формулы Тейлора. В нашем случае мы разлагаем функцию $h(\cdot)$ в окрестности точки $\overline{X} = \mu$, используя только два члена разложения:

$$h(\overline{X}) = h(\mu) + (\overline{X} - \mu)h'(\mu) + \frac{(\overline{X} - \mu)^2}{2!}h''\left(\mu + \lambda(\overline{X} - \mu)\right),$$

где $0 \le \lambda \le 1$.

Перепишем это разложение в виде

$$\sqrt{n}\left(h(\overline{X}) - h(\mu)\right) = \sqrt{n}(\overline{X} - \mu)h'(\mu) + \frac{\sqrt{n}(\overline{X} - \mu)^2}{2!}h''\left(\mu + \lambda(\overline{X} - \mu)\right),$$

представив тем самым случайную величину $\sqrt{n}\left(h(\overline{X}) - h(\mu)\right)$ (см. формулу (4)) в виде суммы двух случайных величин, первая из которых в силу формулы (5) имеет предельное нормальное распределение, указанное в правой части (4), а вторая сходится по вероятности к нулю. Действительно, $h''\left(\mu + \lambda(\overline{X} - \mu)\right)$ по условию теоремы ограничено с вероятностью сколь угодно близкой к единице, начиная с некоторого n. В силу неравенства Чебышева (предложение 6.2 курса ТВ) для любого $\varepsilon > 0$ вероятность

$$P\left(\sqrt{n}(\overline{X} - \mu)^2 > \varepsilon\right) \leq \frac{\mathbf{E}\sqrt{n}(\overline{X} - \mu)^2}{\varepsilon} = \frac{\sqrt{n}\mathbf{D}\overline{X}}{\varepsilon} = \frac{\sigma^2}{\varepsilon\sqrt{n}} \to 0,$$

и поэтому стоящий перед $h''/2!$ множитель, а с ним и все второе слагаемое, сходятся по вероятности к нулю. Доказательство теоремы завершается ссылкой на предложение 11.1 курса ТВ: если одна последовательность случайных величин имеет невырожденное предельное распределение F, а вторая сходится по вероятности к нулю, то предельное распределение суммы этих последовательностей совпадает с F.

Итак, если $h(\overline{X})$ – оценка $\hat{\theta}_n$ параметра θ по методу моментов, то формула (4) теоремы 3.2 принимает вид

$$\lim_{n\to\infty} P\left(\sqrt{n}\left(\hat{\theta}_n - \theta\right) < x\right) = \Phi\left(\frac{x}{\sigma|\,h'(\mu)\,|}\right).$$

Приведем пример на использование аппроксимации распределения функции от выборочного среднего на практике.

Пример 3.3. *Оценка надежности изделия с показательным распределением долговечности.* При выпуске изделий обычно указывается их гарантийный срок службы t_0; отказ изделия до истечения срока t_0 чреват для поставщика расходами на ремонт или замену изделия. Чтобы планировать расходы на такого рода рекламации со стороны потребителя, поставщик должен знать надежность выпускаемых изделий: $H(t_0) = P(X > t_0)$. Величина $H(t_0)$ указывает среднюю долю изделий, которые не откажут в течении гарантийного срока службы t_0. Чтобы оценить $H(t_0)$ проводятся испытания n изделий, и пусть $x_1, \ldots, x_n$ – наработки на отказ испытуемых изделий, трактуемые как реализации случайной выборки $X^{(n)}$ из распределения $F(\,\cdot\,|\,\theta)$, известного с точностью до значения параметра θ. Наконец, пусть нам известно, что долговечность изделий подчиняется закону "отсутствия последействия" , в силу чего $F(x\,|\,\theta) = 1 - \exp\{-x/\theta\}$ – показательное

распределение. В таком случае проблема состоит в оценке параметрической функции $h(\theta) = \exp\{-t_0/\theta\}$.

Так как $\theta = \mathbf{E}X$ – средняя наработка на отказ, то естественно оценить θ посредством статистики $\overline{X}$ (работает метод моментов), а за оценку надежности взять статистику $h(\overline{X}) = \exp\{-t_0/\overline{X}\}$. Поскольку наибольшую, с экономической точки зрения, опасность представляет завышение надежности, то нас в первую очередь должна интересовать частота грубых превышений, например, на некоторую заданную величину ε. Если положить $\Delta = \varepsilon\sqrt{n}$, то вероятность, указывающая частоту грубых превышений, записывается в виде $R(\theta;\, h(\overline{X})) = P\left(\sqrt{n}\left(h(\overline{X}) - h(\theta)\right) > \Delta\right)$, что позволяет нам непосредственно использовать аппроксимацию (4) при испытаниях достаточно большого числа изделий (насколько большого, это – отдельный вопрос, решить который можно, например, моделируя выборки из показательного распределения с помощью метода Монте-Карло).

Имеем $\sigma^2 = \theta^2,\ h'(\theta) = t_0\theta^{-2}\exp\{-t_0/\theta\}$, и формула (4) дает нам следующую аппроксимацию для риска оценки $h(\overline{X})$:

$$R(\theta;\, h(\overline{X})) \approx 1 - \Phi\left(\frac{\Delta\theta}{t_0}\exp\left\{\frac{t_0}{\theta}\right\}\right).$$

Нетрудно показать, что наибольшее значение риска достигается при $\theta = t_0$ и равно $\Phi(\Delta \cdot \mathrm{e})$.

В том случае, когда оценка имеет вид функции от двух и более выборочных моментов, метод асимптотического анализа ее распределения тот же. Например, пусть $\hat{\theta}_n = h(a_1,\, a_2)$, и функция h удовлетворяет условиям, аналогичным требованиям к h в теореме 3.2. Используя формулу Тейлора, представим h в окрестности точки $(\alpha_1,\, \alpha_2)$ в следующем виде:

$$h(a_1,\, a_2) = h(\alpha_1,\, \alpha_2) + (a_1 - \alpha_1)h'_1(\alpha_1,\, \alpha_2) +$$

$$(a_2 - \alpha_2)h'_2(\alpha_1,\, \alpha_2) + O_p(|\, a - \alpha\,|^2),$$

где $a = (a_1,\, a_2),\ \alpha = (\alpha_1,\, \alpha_2)$. Легко проверяется, что $\sqrt{n}|\, a - \alpha\,|^2 \underset{P}{\to} 0$, так что случайная величина $\sqrt{n}\,(h(a_1,\, a_2) - h(\alpha_1,\, \alpha_2))$ асимптотически нормальна с параметрами, которые выражаются через первые четыре момента наблюдаемой случайной величины X.

§4. Оценка параметров. Метод максимального правдоподобия

Лекция 6

Мы приступаем к изучению более точного метода оценки неизвестного значения параметра. Он превосходит метод моментов и при наличии достаточных статистик дает оптимальные оценки с точки зрения квадратичного риска. Более того, при выполнении определенных условий регулярности этот метод приводит к асимптотически ($n \to \infty$) оптимальным оценкам для широкого класса вероятностных моделей и практически при любых функциях потерь.

Идея метода состоит в математической формализации "разумного" поведения человека в условиях неопределенности. Представим себе ситуацию, что мы ожидаем появления одного из нескольких событий, вероятности которых нам неизвестны и нас интересуют не столько значения этих вероятностей, сколько то событие, которое происходит наиболее часто. Ситуация осложняется тем, что мы располагаем всего одним испытанием, в результате которого произошло некоторое событие A. Конечно, мы примем решение, что A обладает наибольшей вероятностью, и вряд ли можно предложить нечто более разумное, чем такое правило принятия решения.

В этом и состоит *принцип максимального правдоподобия,* который буквально пронизывает всю теорию оптимального статистического вывода. Применение этого принципа к проблеме оценки параметров приводит к следующему статистическому правилу: *если $x^{(n)}$ – результат наблюдения случайной выборки $X^{(n)}$, то за оценку параметра следует брать то его значение, при котором результат $x^{(n)}$ обладает наибольшим правдоподобием.*

Вы спросите, что такое "правдоподобие" результата $x^{(n)}$? Давайте формализуем это понятие.

Если наблюдается дискретная случайная величина, то естественно назвать правдоподобием результата $x^{(n)}$ при фиксированном значении параметра θ вероятность его наблюдения в статистическом эксперименте. Но в дискретном случае эта вероятность совпадает со значением функции плотности в точке $x^{(n)}$: $P_\theta(X^{(n)} = x^{(n)}) = f_n(x^{(n)} \,|\, \theta)$. Следовательно, оценка по методу максимального правдоподобия определяется точкой достижения максимума у функции плотности случайной выборки, то есть

$$\hat{\theta}_n(X^{(n)}) = \arg\max_{\theta\in\Theta} f_n\left(X^{(n)} \,|\, \theta\right). \tag{1}$$

Рассмотрим сразу же простой пример. Пусть $X^{(n)}$ – выборка в схеме Бернулли, и мы оцениваем вероятность θ успешного исхода. В этой модели

$$f(X^{(n)} \,|\, \theta) = \theta^{\sum_1^n X_k} (1-\theta)^{n-\sum_1^n X_k}.$$

Дифференцируя эту функцию по θ и приравнивая производную нулю, находим оценку максимального правдоподобия $\theta = (1/n)\sum_1^n X_k$. Это – давно знакомая нам оценка вероятности успеха в испытаниях Бернулли, которую мы получили с помощью моментов и постоянно использовали при иллюстрации закона больших чисел.

Теперь определим правдоподобие в случае выбора из непрерывного распределения с функцией плотности (по мере Лебега) $f_n(x^{(n)} \,|\, \theta)$, $x^{(n)} \in \mathrm{R}^n$, $\theta \in \Theta$. Пусть $x^{(n)}$ – совокупность выборочных данных, то есть точка в n-мерном выборочном пространстве R^n. Окружим эту точку прямоугольным параллелепипедом малого размера, скажем, $V_\varepsilon = \prod_1^n [\, x_k - \varepsilon/2;\, x_k + \varepsilon/2]$. В силу теоремы о среднем для кратного интеграла вероятность того, что выборочный вектор попадет в этот параллелепипед $P\left(X^{(n)} \in V_\varepsilon\right) \sim f_n(x^{(n)} \,|\, \theta) \cdot \varepsilon^n$, когда $\varepsilon \longrightarrow 0$. Если трактовать эту вероятность, как правдоподобие результата $x^{(n)}$, которое, конечно, зависит от выбора малого ε, мы видим, что проблема максимизации правдоподобия сводится к проблеме отыскания точки достижения максимума по всем $\theta \in \Theta$ у функции плотности f_n. Таким образом, и в случае непрерывного распределения разумно назвать правдоподобием результата $x^{(n)}$ при фиксированном значении параметра θ опять-таки величину функции плотности выборки, то есть $f_n(x^{(n)} \,|\, \theta)$, и определить оценку максимального правдоподобия той же формулой (1).

Рассмотрим пример на построение такой оценки в случае выбора из непрерывного распределения. Пусть наблюдается случайная величина $X \sim \mathcal{N}(\mu, \sigma^2)$, так что функция плотности выборки

$$f_n(x^{(n)} \,|\, \theta) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_1^n (x_k - \mu)^2\right\},$$

где $\theta = (\mu,\, \sigma)$ – двумерный параметр, значение которого нам неизвестно. В соответствии с формулой (1) необходимо отыскать точку достижения максимума функции $f_n(X^{(n)} \,|\, \mu, \theta)$ по переменным $\mu \in \mathrm{R}$ и $\sigma \in \mathrm{R}_+$. Естественно, логарифм этой функции имеет те же точки экстремума, что и сама функция, но логарифмирование упрощает выкладки, поэтому ищем

максимум функции

$$\mathcal{L}(\theta \,|\, X^{(n)}) = \ln f_n(X^{(n)} \,|\, \theta) = -\frac{n}{2}\ln 2\pi - n\ln\sigma - \frac{1}{2\sigma^2}\sum\nolimits_1^n (X_k - \mu)^2.$$

Составляем уравнения, определяющие точки экстремума:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{2\sigma^2}\sum\nolimits_1^n (X_k - \mu) = 0,$$

$$\frac{\partial \mathcal{L}}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum\nolimits_1^n (X_k - \mu)^2 = 0.$$

Из первого уравнения сразу находим оценку параметра μ : $\quad \hat{\mu}_n = \overline{X}$. Подставляя $\overline{X}$ вместо μ во второе уравнение, находим оценку σ : $\quad \hat{\sigma}_n = S$ (выборочное стандартное отклонение). Очевидно, $(\overline{X},\, S)$ – точка максимума. Таким образом, метод максимального правдоподобия приводит к тем же оценкам $\overline{X}$ и S^2 параметров μ и σ^2, что и метод моментов.

Теперь дадим строгое определение правдоподобия и рассмотрим еще несколько примеров, в которых метод максимального правдоподобия дает оценки, отличные от метода моментов.

Определение 4.1. Случайная функция

$$L(\theta \,|\, X^{(n)}) = \prod_{i=1}^{n} f(X_i \,|\, \theta)$$

на параметрическом пространстве Θ называется *функцией правдоподобия*, а значение ее реализации $L(\theta_0 \,|\, x^{(n)})$ при результате наблюдения $X^{(n)} = x^{(n)}$ и фиксированном $\theta = \theta_0$ – *правдоподобием значения* θ_0 *при результате* $x^{(n)}$. Любая точка $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ (статистика) параметрического пространства Θ, доставляющая абсолютный максимум функции правдоподобия, называется *оценкой максимального правдоподобия* параметра θ.

Поскольку функция правдоподобия представляет собой произведение функций от θ, то при отыскании ее максимума методами дифференциального исчисления удобнее иметь дело с логарифмом этой функции. Естественно, точки экстремума у *функции логарифмического правдоподобия*

$$\mathcal{L}(\theta \,|\, X^{(n)}) = \sum_{i=1}^{n} \ln f(X_i \,|\, \theta)$$

те же, что и у функции L, но если функция $L(\cdot \,|\, x^{(n)})$ имеет непрерывные частные производные по компонентам $\theta_1, \ldots, \theta_k$ параметрического вектора θ, то проще дифференцировать $\mathcal{L}$ чем L. В этом случае система уравнений

$$\frac{\partial \mathcal{L}(\theta \,|\, X^{(n)})}{\partial \theta_i} = 0, \;\; i = 1, \ldots, k \tag{2}$$

называется *уравнениями правдоподобия*. Это еще одна разновидность так называемых *оценочных уравнений*, – в предыдущем параграфе мы имели дело с уравнениями метода моментов.

Любое решение системы уравнений (2), доставляющее максимум функции $\mathcal{L}(\cdot \,|X^{(n)})$, может рассматриваться как оценка θ по методу максимального правдоподобия. Мы не будем изучать случаи, когда система (2) имеет несколько решений с возможно одинаковыми значениями функции правдоподобия в этих точках, так что требуются дополнительные априорные знания относительно вероятностной модели, позволяющие выбрать одно из этих решений. Во всех рассмотренных ниже примерах оценка максимального правдоподобия единственна.

Пример 4.1. *Оценка параметра положения равномерного распределения* $\mathrm{U}(0,\, \theta)$. Равномерное на отрезке $[\,0;\, \theta\,]$ распределение имеет функцию плотности $f(x \,|\, \theta) = \theta^{-1}$, если $0 \le x \le \theta$, и $f(x \,|\, \theta) = 0$ вне этого отрезка. Следовательно, функция $L(\theta \,|\, X^{(n)})$ отлична от нуля и равна θ^{-n} только в области $\theta \ge X_{(n)} = \max_{1 \le k \le n} X_k$. Ее максимум по θ достигается в граничной точке $\theta = X_{(n)}$, так что наибольшее значение $X_{(n)}$ выборки $X^{(n)}$ есть оценка максимального правдоподобия параметра θ.

Легко видеть, что оценка θ по методу моментов равна $2\overline{X}$. Эта оценка на порядок хуже оценки максимального правдоподобия с точки зрения квадратичного риска $R(\theta;\, \hat{\theta}_n) = \mathbf{E}_\theta \left(\hat{\theta}_n(X^{(n)}) - \theta \right)^2$. Простые вычисления соответствующих математических ожиданий показывают, что $R(\theta;\, 2\overline{X}) = O(n^{-1})$, в то время как $R(\theta;\, X_{(n)}) = O(n^{-2})$. Данный пример интересен тем, что здесь функция правдоподобия не имеет гладкого максимума, и именно это обстоятельство, как будет видно в дальнейшем, обеспечивает такое различное поведение риска рассматриваемых оценок.

Пример 4.2. *Оценка параметров гамма-распределения* $\mathrm{G}(a,\, \lambda)$. У этого распределения функция плотности

$$f(x \,|\, \theta) = \frac{1}{a^\lambda \Gamma(\lambda)} x^{\lambda - 1} \exp\left\{ -\frac{x}{a} \right\}, \;\; x > 0, \;\; \theta = (a,\, \lambda),$$

отлична от нуля только на положительной полуоси, и логарифмическое правдоподобие

$$\mathcal{L}(a,\,\lambda\,|\,X^{(n)}) = -n\lambda\ln a - n\ln\Gamma(\lambda) + (\lambda-1)\sum_1^n \ln X_k - \frac{1}{a}\sum_1^n X_k.$$

Составляем уравнения правдоподобия:

$$\frac{\partial\mathcal{L}}{\partial a} = -\frac{n\lambda}{a} + \frac{1}{a^2}\sum_1^n X_k = 0,$$

$$\frac{\partial\mathcal{L}}{\partial\lambda} = -n\ln a - n\psi(\lambda) + \sum_1^n \ln X_k = 0,$$

где $\psi(\lambda) = d\ln\Gamma(\lambda)/d\lambda$ – так называемая пси-фукция Эйлера. Исключая из первого уравнения параметр a и подставляя полученный результат во второе, получаем трансцендентное уравнение

$$\ln\lambda - \psi(\lambda) = \ln\overline{X} - \frac{1}{n}\sum_1^n \ln X_k,$$

которое в силу свойства монотонности функции $\ln\lambda - \psi(\lambda)$ имеет единственное решение. При численном решении этого уравнения может оказаться полезной асимптотическая ($\lambda\to\infty$) формула

$$\ln\lambda - \psi(\lambda) = \frac{1}{2\lambda} + \frac{1}{12\lambda^2} + O\left(\frac{1}{\lambda^4}\right).$$

Пример 4.3. *Оценка параметров структурированного среднего при нормальном распределении отклика.* Данная задача весьма часто возникает при калибровке шкалы прибора. Две переменные x и y связаны линейным соотношением $y = a + bx$, и для градуировки значений y на шкале прибора необходимо знать значения параметров a и b этой зависимости. Однако, для каждого стандартного фиксированного значения x прибор замеряет значение y с ошибкой, так что замеры происходят в рамках вероятностной модели $Y = a + bx + \xi$, где ошибка измерения (случайная величина) ξ имеет нормальное распределение с нулевым средним и некоторой дисперсией σ^2, значение которой, как правило, также не известно. Случайная величина Y обычно называется *откликом* на значение *регрессора* x; ее распределение при фиксированном x очевидно нормально $(a + bx,\,\sigma^2)$.

Для оценки a и b производится n измерений отклика $y_1, \ldots, y_n$ при некоторых фиксированных значениях $x_1, \ldots, x_n$ регрессора x, оптимальный выбор которых, обеспечивающий наибольшую точность и надежность калибровки, составляет самостоятельную задачу особой области математической статистики – *планирование регрессионных экспериментов*. Мы будем предполагать, что значения $x_1, \ldots, x_n$ априори фиксированы. В таком случае значения $y_1, \ldots, y_n$ представляют реализации n независимых случайных величин $Y_1, \ldots, Y_n$, и $Y_k \sim \mathcal{N}(a + bx_k,\, \sigma^2),\ k = 1, \ldots, n$. Совместная функция плотности $Y_1, \ldots, Y_n$ равна

$$f_n(y^{(n)}) \,|\, a, b, \sigma) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum\nolimits_1^n (y_k - a - bx_k)^2) \right\},$$

так что логарифмическая функция правдоподобия, необходимая для оценки параметров a, b и σ методом максимального правдоподобия имеет вид

$$\mathcal{L}(a,\, b,\, \sigma \,|\, Y^{(n)}) = -\frac{n}{2} \ln 2\pi - n \ln \sigma - \frac{1}{2\sigma^2} \sum_1^n (Y_k - a - bx_k)^2.$$

Вычисляя производные этой функции по переменным a, b и σ, получаем уравнения правдоподобия

$$\sum_1^n (Y_k - a - bx_k) = 0,$$

$$\sum_1^n x_k (Y_k - a - bx_k) = 0,$$

$$n\sigma^2 - \sum_1^n (Y_k - a - bx_k)^2 = 0.$$

Конечно, это очень простая система уравнений, решение которой не может вызывать каких-либо затруднений, и мы сразу пишем оценки максимального правдоподобия

$$\hat{a}_n = \overline{Y} - \frac{m_{xY}}{s_x^2}\overline{x}, \quad \hat{b}_n = \frac{m_{xY}}{s_x^2}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_1^n \left(Y_k - \hat{a}_n - \hat{b}_n x_k \right)^2,$$

где

$$\overline{x} = \frac{1}{n} \sum_1^n x_k,\ \overline{Y} = \frac{1}{n} \sum_1^n Y_k, \quad s_x = \frac{1}{n} \sum_1^n (x_k - \overline{x})^2,\ S_Y = \frac{1}{n} \sum_1^n (Y_k - \overline{Y})^2,$$

$$m_{xY} = \frac{1}{n}\sum_1^n (x_k - \overline{x})(Y_k - \overline{Y}).$$

Легко видеть, что оценки по методу максимального правдоподобия параметров a и b совпадают с их оценками по *методу наименьших квадратов*. В этом методе "выравнивания" экспериментальных данных оценки ищутся из условия минимизации суммы квадратов *невязок*: $\sum_1^n (Y_k - a - bx_k)^2$, причем под невязкой понимается разность между откликом Y и его "теоретическим" средним значением $a + bx$.

Пример 4.4. *Оценка параметров двумерного нормального распределения: задачи регрессии и прогноза.* Оценка по методу максимального правдоподобия пяти параметров μ_1, μ_2, σ_1^2, σ_2^2, ρ двумерного нормального распределения с функцией плотности

$$f(x, y \,|\, \theta) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\}$$

не представляет особой технической сложности. Эти оценки совпадают с оценками по методу моментов и, таким образом, равны выборочным аналогам тех характеристик двумерного нормального распределения, которые соответствуют указанным пяти параметрам:

$$\hat{\mu}_{1,n} = \overline{X},\ \hat{\mu}_{2,n} = \overline{Y},\quad \hat{\sigma}^2_{1,n} = S_X^2,\ \hat{\sigma}^2_{2,n} = S_Y^2,\quad \hat{\rho}_n = r.$$

Формулы для вычисления выборочных средних $\overline{X}$ и $\overline{Y}$, выборочных дисперсий S_1^2 и S_2^2, а также выборочного коэффициента корреляции r приведены в §2.

Полученные оценки часто используются для оценки параметров линейного прогноза $Y = a + bX$ значений случайной величины Y по результатам наблюдений X. В случае нормального распределения линейный прогноз обладает свойством оптимальности с точки зрения малости средней квадратичной ошибки и совпадает с кривой средней квадратичной регрессии (см. предложение 10.3 курса ТВ)

$$y = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Однако формальная подгонка прогностической кривой с помощью прямой линии используется и вне рамок нормальной модели, и в этом случае оценки

$$\hat{a}_n = \overline{Y} - r\frac{S_2}{S_1}\overline{X}, \quad \hat{b}_n = r\frac{S_2}{S_1}$$

совпадают с оценками по *методу наименьших квадратов*: минимизируется, как и в примере 4.3, сумма квадратов невязок

$$\sum_1^n (Y_k - a - bX_k)^2.$$

Хотя оценки в обоих примерах имеют одинаковый вид, но решаемые в них статистические проблемы весьма различны: в примере 4.3 оценивались параметры некоторой функциональной зависимости с ошибками в наблюдениях отклика, в то время как в примере 4.4 решается задача выявления корреляционной связи и использования этой связи для прогноза.

Лекция 7

Исследуем теперь асимптотические свойства оценок по методу максимального правдоподобия.

Начнем с выяснения достаточных условий состоятельности этих оценок. Такие ограничения на вероятностную модель обычно называются *условиями регулярности*, и в данном случае они имеют следующий вид.

(R1) Параметрическое пространство Θ есть открытый интервал на прямой R.

(R2) Носитель $\mathcal{X}$ распределения P_θ наблюдаемой случайной величины X не зависит от θ, то есть все множества $\mathcal{X} = \{x : \ f(x \,|\, \theta) > 0\}$ можно считать одинаковыми, каково бы ни было $\theta \in \Theta$.

(R3) Распределения P_θ различны при разных θ, то есть при любых $\theta_1 \neq \theta_2, \quad \theta_1, \theta_2 \in \Theta,$ имеет место тождество $\mu\{x : \ x \in \mathcal{X}, f(x \,|\, \theta_1) = f(x \,|\, \theta_2)\} = 0$, где μ – мера, по которой вычисляется плотность $f(x \,|\, \theta)$ распределения P_θ.

Доказательство состоятельности оценок максимального правдоподобия, как и оценок по методу моментов, опирается на закон больших чисел, но

при этом используется следующее достаточно простое, но играющее большую роль в теории вероятностей, неравенство.

Лемма 4.1. (неравенство Йенсена) *Пусть X – случайная величина с конечным математическим ожиданием. Если функция $g(\cdot)$ дважды дифференцируема и выпукла ($g'' > 0$) на некотором интервале, содержащем носитель распределения X, и математическое ожидание $\mathbf{E}\, g(X)$ существует, то справедливо неравенство $\mathbf{E}\, g(X) \geq g(\mathbf{E}X)$, причем знак равенства достигается тогда и только тогда, когда распределение X сосредоточено в одной точке ($X = \text{const.}$).*

Доказательство.. Так как функция g дважды дифференцируема, то справедливо следующее представление Тейлора в окрестности точки $\mu = \mathbf{E}X$:

$$g(X) = g(\mu) + (X-\mu)g'(\mu) + (X-\mu)^2 g''(\mu + \lambda(X-\mu))/2,\ 0 < \lambda < 1.$$

Вычисляя математическое ожидание от обеих частей этого равенства, получаем

$$\mathbf{E}\, g(X) = g(\mathbf{E}X) + \mathbf{E}(X-\mu)^2 g''(\mu + \lambda(X-\mu))/2 \geq g(\mathbf{E}X).$$

Знак равенства возможен только в случае $\mathbf{E}(X-\mu)^2 g''(\mu+\lambda(X-\mu)) = 0$. Но поскольку $g'' > 0$, то последнее равенство с необходимостью влечет $(X-\mu)^2 = 0$, то есть $X = \text{const}$.

Покажем теперь, что справедлива

Теорема 4.1 (состоятельность). *Если функция логарифмического правдоподобия*

$$\mathcal{L}(\theta \,|\, X^{(n)}) = \sum_{k=1}^{n} \ln f(X_k \,|\, \theta) \tag{3}$$

имеет единственный максимум, то при выполнении условий регулярности (R1)–(R3) точка $\hat{\theta}_n$ достижения максимума этой функции (оценка максимального правдоподобия) является состоятельной оценкой параметра θ.

Доказательство.. Покажем, что для любого фиксированного $\theta_0 \in \Theta$ и любого $\varepsilon > 0$ вероятность $P_{\theta_0}\left(|\, \hat{\theta}_n - \theta_0 \,| < \varepsilon \right) \to 1$.

Если θ_0 – истинное значение параметра θ, то в силу условия (R1) θ_0 – внутренняя точка Θ. Тогда сформулированная выше задача состоит в доказательстве следующего утверждения: в некоторой ε-окрестности $(\theta_0 -$

ε; $\theta_0 + \varepsilon$) функция $\mathcal{L}(\,\cdot \,|\, X^{(n)})$ обладает локальным максимумом с вероятностью, стремящейся к единице при $n \to \infty$.

Но если происходит событие

$$A_n = \left\{ \mathcal{L}(\theta_0 \,|\, X^{(n)}) \;>\; \mathcal{L}(\theta_0 \pm \varepsilon \,|\, X^{(n)}) \right\},$$

то внутри этой окрестности имеется точка максимума, и нам остается только показать, что $P_{\theta_0}(A_n) \to 1$, ибо $P_{\theta_0}\left(|\, \hat{\theta}_n - \theta_0 \,| \;<\; \varepsilon \right) \;\geq\; P_{\theta_0}(A_n)$.

Используя условие (R2) и вид функции $\mathcal{L}$ (см. (3)), представим неравенство, определяющее событие A_n, в виде

$$\frac{1}{n} \sum_{k=1}^{n} \ln \frac{f(X_k \,|\, \theta_0 \pm \varepsilon)}{f(X_k \,|\, \theta_0)} \;<\; 0.$$

В силу закона больших чисел Хинчина левая часть этого неравенства сходится по вероятности к

$$\mathbf{E}_{\theta_0} \ln \frac{f(X_k \,|\, \theta_0 \pm \varepsilon)}{f(X_k \,|\, \theta_0)}, \tag{4}$$

и для доказательства утверждения достаточно показать, что это математическое ожидание строго меньше нуля (кстати, докажите сами, что при справедливости условий теоремы математическое ожидание (4) всегда существует, в противном случае закон больших чисел Хинчина не применим).

Так как $g(x) = -\ln x$ – выпуклая функция, то в силу неравенства Йенсена

$$\mathbf{E}_{\theta_0} \ln \frac{f(X \,|\, \theta_0 \pm \varepsilon)}{f(X \,|\, \theta_0)} \;\leq\; \ln \mathbf{E}_{\theta_0} \frac{f(X \,|\, \theta_0 \pm \varepsilon)}{f(X \,|\, \theta_0)} =$$

$$\ln \int\limits_{\mathcal{X}} \frac{f(x \,|\, \theta_0 \pm \varepsilon)}{f(x \,|\, \theta_0)} \cdot f(x \,|\, \theta_0) d\mu(x) = \ln 1 = 0,$$

причем равенство нулю первого члена в этой цепочке неравенств возможно лишь в случае

$$\frac{f(X \,|\, \theta_0 \pm \varepsilon)}{f(X \,|\, \theta_0)} = \text{const.},$$

то есть, поскольку интеграл от плотности равен 1, лишь в случае $f(X \,|\, \theta_0 \pm \varepsilon) = f(X \,|\, \theta_0)$, что невозможно в силу условия (R3). Таким образом, математическое ожидание (4) строго меньше нуля, и состоятельность оценки максимального правдоподобия доказана.

Изучим теперь асимптотическое распределение оценки максимального правдоподобия. Для этого нам потребуется ввести дополнительные условия регулярности.

(R4) Для каждой точки θ_0 параметрического пространства Θ существует некоторая ее окрестность, в которой функция плотности $f(x\,|\theta)$ трижды дифференцируема по параметру θ и

$$\left|\frac{\partial f(x\,|\,\theta)}{\partial\theta}\right| \le H_1(x), \tag{5}$$

$$\left|\frac{\partial^2 f(x\,|\,\theta)}{\partial\theta^2}\right| \le H_2(x), \tag{6}$$

$$\left|\frac{\partial^3 \ln f(x\,|\,\theta)}{\partial\theta^3}\right| \le H_3(x),$$

причем функции H_1 И H_2 интегрируемы по мере μ на носителе $\mathfrak{X}$ распределения X и $\mathbf{E}_\theta H_3(X) < \infty$ в некоторой окрестности каждой точки θ параметрического пространства Θ.

(R5) Функция

$$I(\theta) = \mathbf{E}_\theta\left(\frac{\partial \ln f(X\,|\,\theta)}{\partial\theta}\right)^2 = \int\limits_{\mathfrak{X}}\left(\frac{\partial \ln f(x\,|\,\theta)}{\partial\theta}\right)^2 f(x\,|\,\theta)\,d\mu(x) > 0,$$

каково бы ни было $\theta\in\Theta$.

Естественно, столь громоздкие и, на первый взгляд, странные условия требуют некоторого комментария.

Условие (R4) означает, что соответствующие производные функции плотности равномерно интегрируемы на $\mathfrak{X}$, и поэтому можно выносить производную по θ за знак интеграла.

Условие (R5) требует положительности очень важной, с точки зрения состоятельности статистического вывода, характеристики вероятностной модели: $I(\theta)$ называется *информацией по Фишеру* в точке θ, содержащейся в наблюдении случайной величины X. Если $I(\theta) = 0$, то возникают непреодолимые трудности с принятием корректного решения, соответствующего этой параметрической точке θ. Понятно, что аналогичным образом можно

определить и информацию по Фишеру, содержащуюся в случайной выборке $X^{(n)}$:

$$I_n(\theta) = \mathbf{E}_\theta \left(\frac{\partial \ln f_n(X^{(n)} \,|\, \theta)}{\partial \theta} \right)^2 .$$

Приведем несколько утверждений, касающихся свойств информации по Фишеру.

Лемма 4.2. 1^0. *При выполнении условия (R4) в части (6) для вычисления информации по Фишеру можно использовать формулу*

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

2^0. *При выполнении условия (R4) в части (5) информация по Фишеру обладает свойством аддитивности $I_n(\theta) = nI(\theta)$ – информация, содержащаяся в выборке, равна сумме информаций, содержащихся в наблюдении каждой ее компоненты.*

Доказательство.. 1^0. Условие (R4) в части (6) обеспечивает возможность смены порядка дифференцирования и интегрирования функции плотности, поэтому

$$\mathbf{E}_\theta \frac{\partial^2 \ln f(X \,|\, \theta)}{\partial \theta^2} = \mathbf{E}_\theta \left(\frac{f''_{\theta\theta}(X \,|\, \theta)}{f(X \,|\, \theta)} - \left(\frac{f'_\theta(X \,|\, \theta)}{f(X \,|\, \theta)} \right)^2 \right) =$$

$$\int\limits_{\mathfrak{X}} \frac{f''_{\theta\theta}(x \,|\, \theta)}{f(x \,|\, \theta)} \cdot f(x \,|\, \theta) d\mu(x) - I(\theta) = \frac{d^2}{d\theta^2} \int\limits_{\mathfrak{X}} f(x \,|\, \theta) d\mu(x) - I(\theta) = -I(\theta).$$

2^0. Используя независимость и одинаковую распределенность компонент случайной выборки, получаем, что

$$I_n(\theta) = \mathbf{E}_\theta \left(\frac{\partial \sum_1^n \ln f(X_k \,|\, \theta)}{\partial \theta} \right)^2 =$$

$$\mathbf{E}_\theta \left(\sum_{k=1}^n \left(\frac{\partial \ln f(X_k \,|\, \theta)}{\partial \theta} \right)^2 - \sum_{i \neq j} \frac{\partial \ln f(X_i \,|\, \theta)}{\partial \theta} \cdot \frac{\partial \ln f(X_j \,|\, \theta)}{\partial \theta} \right) =$$

$$\sum_{k=1}^n \mathbf{E}_\theta \left(\frac{\partial \ln f(X_k \,|\, \theta)}{\partial \theta} \right)^2 - \sum_{i \neq j} \mathbf{E}_\theta \frac{\partial \ln f(X_i \,|\, \theta)}{\partial \theta} \cdot \mathbf{E}_\theta \frac{\partial \ln f(X_j \,|\, \theta)}{\partial \theta} = nI(\theta),$$

поскольку, в силу неравенства (5) в условии (R4), математическое ожидание

$$\mathbf{E}_\theta \frac{\partial \ln f(X \,|\, \theta)}{\partial \theta} = \int\limits_{\mathcal{X}} \frac{f'_\theta(x \,|\, \theta)}{f(x \,|\, \theta)} \cdot f(x \,|\, \theta)\, d\mu(x) = \frac{d}{d\theta} \int\limits_{\mathcal{X}} f(x \,|\, \theta)\, d\mu(x) = 0.$$

Теперь приступим к выводу асимптотического распределения оценки максимального правдоподобия скалярного параметра θ.

Теорема 4.2 (асимптотическая нормальность). *При выполнении условий (R1)–(R5) и наличии единственного локального максимума у функции правдоподобия корень $\hat\theta_n = \hat\theta_n(X^{(n)})$ уравнения правдоподобия*

$$\partial \mathcal{L}(\theta \,|\, X^{(n)})/\partial\theta = 0$$

асимптотически ($n \to \infty$) нормален со средним θ и дисперсией $(nI(\theta))^{-1}$, то есть

$$\lim_{n\to\infty} P_\theta \left((\hat\theta_n - \theta)\sqrt{nI(\theta)} \,<\, x \right) = \Phi(x).$$

Доказательство.. Если $\hat\theta_n$ – оценка по методу максимального правдоподобия (корень уравнения правдоподобия), то имеет место тождество $\partial \mathcal{L}(\hat\theta_n \,|\, X^{(n)})/\partial\theta = 0$. Используя условие (R4), разложим его левую часть по формуле Тейлора в окрестности истинного значения θ_0 параметра θ :

$$\partial \mathcal{L}(\hat\theta_n \,|\, X^{(n)})/\partial\theta = \mathcal{L}'(\theta_0 \,|\, X^{(n)}) +$$

$$(\hat\theta_n - \theta_0)\mathcal{L}''(\theta_0 \,|\, X^{(n)}) + (\hat\theta_n - \theta_0)^2 \mathcal{L}'''(\theta_1 \,|\, X^{(n)})/2 = 0,$$

где производные от функции правдоподобия $\mathcal{L}$ вычисляются по параметру θ, а $\theta_1 = \theta_0 + \lambda(\hat\theta_n - \theta_0),\ 0 < \lambda < 1$.

Разрешим полученное уравнение относительно величины $\sqrt{n}(\hat\theta_n - \theta_0)$, которая, согласно утверждению теоремы, должна иметь в пределе при $n \to \infty$ нормальное распределение со средним 0 и дисперсией $[\, I(\theta_0)\,]^{-1}$:

$$\sqrt{n}(\hat\theta_n - \theta_0) = \frac{\mathcal{L}'(\theta_0 \,|\, X^{(n)})/\sqrt{n}}{-\mathcal{L}''(\theta_0 \,|\, X^{(n)})/n - (\hat\theta_n - \theta_0)\mathcal{L}'''(\theta_1 \,|\, X^{(n)})/2n}. \qquad (7)$$

Числитель правой части этого представления

$$\frac{1}{\sqrt{n}}\mathcal{L}'(\theta_0 \,|\, X^{(n)}) = \frac{1}{\sqrt{n}} \sum_1^n \frac{\partial \ln f(X_k \,|\, \theta)}{\partial\theta}$$

есть нормированная на $\sqrt{n}$ сумма независимых, одинаково распределенных случайных величин с нулевыми средними и дисперсиями $I(\theta_0) > 0$ (см. доказательство пункта 2^0 леммы 4.2). Таким образом, в силу центральной предельной теоремы числитель правой части (7) асимптотически нормален с этими параметрами, и для завершения доказательства теоремы достаточно показать, что знаменатель (7) сходится по вероятности к постоянной $I(\theta_0)$, и сослаться на пункт (2) предложения 11.1 (теорема типа Слуцкого) курса ТВ.

В силу закона больших чисел и утверждения 1^0 леммы 4.2 первое слагаемое в знаменателе (7)

$$-\frac{1}{n}\mathcal{L}''(\theta_0 \,|\, X^{(n)}) = -\frac{1}{n}\sum_1^n \frac{\partial^2 \ln f(X_k \,|\, \theta_0)}{\partial\theta^2} \underset{P}{\to} -\mathbf{E}_{\theta_0}\frac{\partial^2 \ln f(X \,|\, \theta_0)}{\partial\theta^2} = I(\theta_0),$$

так что остается показать, что и второе слагаемое сходится по вероятности к нулю.

Так как при выполнении условий (R1)–(R3) оценка максимального правдоподобия состоятельна, то $\hat{\theta}_n - \theta_0 \underset{P}{\to} 0$. Множитель при этой разности

$$\frac{1}{n}\mathcal{L}'''(\theta_1 \,|\, X^{(n)}) = \frac{1}{n}\sum_1^n \frac{\partial^3 \ln f(X_k \,|\, \theta_1)}{\partial\theta^3}$$

в силу условия (R4), начиная с некоторого n по абсолютной величине не превосходит $(1/n)\sum_1^n H_3(X_k)$ (это то n, при котором θ_1 попадает в окрестность точки θ_0 с любой наперед заданной вероятностью $1-\varepsilon$). Применяя к этой сумме закон больших чисел, получаем, что она сходится по вероятности к

$$\mathbf{E}_{\theta_0} H_3(X) < \infty,$$

и поэтому указанный выше сомножитель ограничен с вероятностью единица, а все второе слагаемое в знаменателе правой части (7) сходится по вероятности к нулю.

Доказанная теорема, как будет видно из основного результата следующего параграфа, устанавливает асимптотическую оптимальность оценок максимального правдоподобия с точки зрения квадратичного риска.

§5. Эффективность оценок

Лекция 8

Обсуждая в начале нашего курса общую проблему статистического вывода, мы говорили о главной задаче математической статистики – построении решающих правил $\delta_n = \delta_n(X^{(n)})$, минимизирующих равномерно по всем $\theta \in \Theta$ функцию риска $R(\theta;\, \delta_n)$. К сожалению, без дополнительных ограничений на класс решающих функций эта задача не разрешима. Действительно, рассмотрим проблему оценки параметра θ, в которой пространство решений $\mathcal{D}$ совпадает с параметрическим пространством Θ, а решающая функция $\delta_n = \hat{\theta}_n$ – оценка θ. Возьмем в качестве оценки некоторую фиксированную точку $\theta_0 \in \Theta$, то есть при любом результате $x^{(n)}$ статистического эксперимента будем принимать одно и то же решение $d = \theta_0$. Если функция потерь обладает тем естественным свойством, что $L(\theta,\, \theta) = 0$, каково бы ни было значение $\theta \in \Theta$, то риск такой оценки $R(\theta;\, \theta_0) = L(\theta,\, \theta_0)$ при $\theta = \theta_0$ равен нулю. Таким образом, если мы хотим построить оценку с равномерно минимальным риском в классе всевозможных оценок θ, то мы должны найти оценку θ_n^* с функцией риска $R(\theta,\, \theta_n^*) \equiv 0$, и понятно, что такой оценки не существует. Поэтому мы будем всегда при поиске оптимальных решений указывать класс оценок, в которых ищется оптимальное решение.

Определение 5.1. Оценка $\theta_n^* = \theta_n^*(X^{(n)})$ называется *оптимальной* или *оценкой с равномерно минимальным риском* в классе $\mathcal{K}$ оценок параметра θ, если для любой оценки $\hat{\theta}_n \in \mathcal{K}$ и каждого $\theta \in \Theta$ имеет место неравенство $R(\theta;\, \theta_n^*) \le R(\theta;\, \hat{\theta}_n)$.

Ниже предлагается метод нахождения оптимальных оценок скалярного параметра θ при квадратичной функции потерь в классе несмещенных оценок: $\mathbf{E}_\theta \hat{\theta}_n(X^{(n)}) = \theta$ при любом $\theta \in \Theta \subseteq \mathrm{R}$, но при дополнительных ограничениях на вероятностную модель и соответствующее семейство распределений оценки. Эти ограничения аналогичны тем условиям регулярности, которые мы накладывали на вероятностную модель при изучении асимптотических свойств оценок максимального правдоподобия. Мы покажем, что квадратичный риск любой несмещенной оценки, удовлетворяющей этим условиям, не может быть меньше $[nI(\theta)]^{-1}$ – асимптотической дисперсии оценки максимального правдоподобия (см. теорема 4.2). Следовательно, метод максимального правдоподобия доставляет асимптотическое решение

проблемы оптимальной оценки. Более того, мы покажем, что при наличии достаточных статистик метод максимального правдоподобия может привести и к точному решению проблемы равномерной минимизации функции риска.

Сформулируем условия регулярности, при выполнении которых будет находиться нижняя (достижимая!) граница квадратичного риска оценки.

(B1) Носитель $\mathcal{X}$ распределения P_θ наблюдаемой случайной величины X не зависит от $\theta \in \Theta$ (условие, совпадающее с (R2) в §4).

(B2) Информация по Фишеру $I(\theta)$ строго положительна при любом $\theta \in \Theta$ (условие, совпадающее с (R5) в §4).

(B3) Равенство

$$\int_{\mathcal{X}^n} f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}) = 1$$

можно дифференцировать по θ под знаком интеграла, то есть

$$\int_{\mathcal{X}^n} f_n'(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}) = 0.$$

По аналогии с (R4) в части (5) для этого достаточно потребовать существование такой интегрируемой по мере μ функции $H(x)$, что в некоторой окрестности любой точки $\theta \in \Theta$ выполняется неравенство $|\,\partial f(x\,|\,\theta)/\partial\theta\,| \le H(x),\ x \in \mathcal{X}$.

(B4) Оценка $\hat\theta_n = \hat\theta_n(X^{(n)})$ должна принадлежать классу оценок $\mathcal{K}'$, среднее значение которых

$$\mathbf{E}_\theta\, \hat\theta_n(X^{(n)}) = \int_{\mathcal{X}^n} \hat\theta_n(x^{(n)}) f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)})$$

можно дифференцировать по $\theta \in \Theta$ под знаком интеграла.

Конечно, условие (B4) требует комментария. В "высокой" теории статистического вывода приводятся достаточные условия на семейство распределений $\{P_\theta,\, \theta \in \Theta\}$ наблюдаемой случайной величины X, которые обеспечивают выполнение условия (B4), но формулировка этих условий и, в особенности, доказательство того, что они влекут (B4), настолько технически и концептуально сложны, что могут составить предмет специального курса. Однако все изучаемые нами в курсе ТВ вероятностные модели, за

исключением равномерного распределения, удовлетворяют этим условиям, и поэтому любая оценка их параметров принадлежит классу $\mathcal{K}'$.

Прежде, чем получить основной "технический" результат этого параграфа, вспомним одно замечательное неравенство из курса математического анализа. Это – неравенство Коши–Буняковского, которое в случае интегралов Лебега по вероятностной мере P называется неравенством Шварца. Пусть Y – случайная величина с распределением P и g, h – две интегрируемые с квадратом по мере P функции на области $\mathcal{Y}$ значений Y. Для этих функций имеет место неравенство $(\mathbf{E}\, g(Y)h(Y))^2 \le \mathbf{E}\, g^2(Y) \cdot \mathbf{E}\, h^2(Y)$ или, что то же,

$$\left(\int_{\mathcal{Y}} g(y)h(y)\, dP(y) \right)^2 \le \int_{\mathcal{Y}} g^2(y) dP(y) \cdot \int_{\mathcal{Y}} h^2(y) dP(y),$$

причем знак равенства достигается тогда и только тогда, когда функции g и h линейно зависимы: существуют такие постоянные a и b, что $ag(y) + bh(y) = 0$ для почти всех $y \in \mathcal{Y}$ по мере P.

Теорема 5.1. (неравенство Рао–Крамера) *При выполнении условий (B1)–(B4) для квадратичного риска любой оценки $\hat\theta_n \in \mathcal{K}'$ справедливо неравенство*

$$\mathbf{E}_\theta \left(\hat\theta_n(X^{(n)}) - \theta \right)^2 \ge \mathbf{D}_\theta\, \hat\theta_n(X^{(n)}) \ge \frac{\left[\, d\gamma(\theta)/d\theta \,\right]^2}{nI(\theta)}, \tag{1}$$

где $\gamma(\theta) = \mathbf{E}_\theta\, \hat\theta_n(X^{(n)})$, причем знак равенства между риском и дисперсией оценки $\hat\theta_n$ достигается на несмещенных оценках: $\gamma(\theta) = \theta$, а знак равенства во втором неравенстве (1) имеет место тогда и только тогда, когда существует такая параметрическая функция $C(\theta)$, $\theta \in \Theta$, что

$$\hat\theta_n(X^{(n)}) - \gamma(\theta) = C(\theta) \frac{\partial \mathcal{L}(\theta \,|\, X^{(n)})}{\partial \theta} \tag{2}$$

почти наверное по мере P_θ.

Доказательство.. Продифференцируем обе части равенств

$$\int_{\mathcal{X}^n} f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}) = 1,$$

$$\int_{\mathcal{X}^n} \hat\theta_n(x^{(n)}) f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}) = \gamma(\theta)$$

по параметру θ, занося производные в левых частях под знаки интегралов, что можно сделать благодаря условиям (B3) и (B4). Полученный результат, используя условие (B1), представим в виде

$$\int_{\mathcal{X}^n} \frac{\partial \mathcal{L}(\theta \,|\, x^{(n)})}{\partial \theta} f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}) = 0,$$

$$\int_{\mathcal{X}^n} \hat{\theta}_n(x^{(n)}) \frac{\partial \mathcal{L}(\theta \,|\, x^{(n)})}{\partial \theta} f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}) = \gamma'(\theta).$$

Вычтем из второго равенства первое, умножив его предварительно на $\gamma(\theta)$:

$$\int_{\mathcal{X}^n} \Big(\hat{\theta}_n(x^{(n)}) - \gamma(\theta)\Big) \frac{\partial \mathcal{L}(\theta \,|\, x^{(n)})}{\partial \theta} f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}) = \gamma'(\theta).$$

Применим к левой части полученного равенства неравенство Шварца, полагая $y = x^{(n)}$, $\mathcal{Y} = \mathcal{X}^n$, $g(x^{(n)}) = \hat{\theta}_n(x^{(n)}) - \gamma(\theta)$, $h(x^{(n)}) = \partial \mathcal{L}(\theta \,|\, x^{(n)})/\partial\theta$, $dP(y) = f_n(x^{(n)} \,|\, \theta) d\mu_n(x^{(n)})$. В результате получим неравенство

$$\big(\gamma\prime(\theta)\big)^2 \le \int_{\mathcal{X}^n} \Big(\hat{\theta}_n(x^{(n)}) - \gamma(\theta)\Big)^2 f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}) \cdot$$

$$\int_{\mathcal{X}^n} \left(\frac{\partial \mathcal{L}(\theta \,|\, x^{(n)})}{\partial \theta}\right)^2 f_n(x^{(n)} \,|\, \theta)\, d\mu_n(x^{(n)}), \tag{3}$$

в котором знак равенства достигается тогда и только тогда, когда выполняется соотношение (2).

Мы получили неравенства (1), поскольку первое из них очевидно (на дисперсии достигается минимум всевозможных средних квадратичных уклонений случайной величины от постоянной). Второе неравенство в (1) есть следствие неравенства (3), ибо первый интеграл в правой части (3) равен $\mathbf{D}_\theta \hat{\theta}_n$, а второй интеграл определяет фишеровскую информацию $I_n(\theta)$, содержащуюся в выборке. Наконец, из пункта 2^0 леммы 4.2 следует, что $I_n(\theta) = nI(\theta)$.

Следствие 5.1. *Если $\hat{\theta}_n$ принадлежит подклассу $\mathcal{K} \subseteq \mathcal{K}'$ несмещенных оценок класса $\mathcal{K}'$, то ее квадратичный риск*

$$R(\theta;\, \hat{\theta}_n) = \mathbf{D}_\theta \hat{\theta}_n \;\ge\; [n\, I(\theta)]^{-1}, \tag{4}$$

причем знак равенства тогда и только тогда, когда выполняется равенство (2) с $\gamma(\theta) = \theta$.

Понятно, что это следствие есть частный случай доказанной теоремы. Оно указывает неконструктивный путь к построению несмещенных оценок с равномерно минимальным риском. Достаточно вычислить производную в правой части равенства (2) и затем подбирать статистику $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ и параметрическую функцию $C(\theta)$, для которых имеет место равенство

$$\hat{\theta}_n(X^{(n)}) - \theta = C(\theta)\frac{\partial \mathcal{L}(\theta \,|\, X^{(n)})}{\partial \theta}\,.$$

Обычно это можно сделать в случае статистических структур, обладающих достаточными статистиками, где, в силу теоремы факторизации (теорема 2.1 из §2), функция правдоподобия $L(\theta \,|\, X^{(n)}) = g_\theta(T(X^{(n)}))h(X^{(n)})$, и последнее равенство имеет вид

$$\hat{\theta}_n(X^{(n)}) - \theta = C(\theta)\frac{\partial \ln g_\theta(T(X^{(n)}))}{\partial \theta}\,. \tag{5}$$

Например, для показательного распределения с функцией плотности $f(x \,|\, \theta) = \theta^{-1}\exp\{-x/\theta\},\ x > 0$, функция

$$\ln g_\theta(X^{(n)}) = -n \ln\, \theta - \theta^{-1}\sum\nolimits_1^n X,$$

ее производная

$$\partial \ln g_\theta(T(X^{(n)}))/\partial\theta = -n/\theta + \sum\nolimits_1^n X/\theta^2,$$

и равенство (5) выполняется при $C(\theta) = \theta^2/n$ и $\hat{\theta}_n = \overline{X}$. Таким образом, выборочное среднее $\overline{X}$ есть несмещенная оценка с равномерно минимальным риском для параметра θ показательного распределения. Напомним, что $\overline{X}$ оценка θ как по методу моментов, так и по методу максимального правдоподобия.

Легко понять, что если в (4) достигается знак равенства, то $\hat{\theta}_n$ – оптимальная оценка в классе $\mathcal{K}$, но обратное, вообще говоря, может и не выполняться – мы не располагаем утверждением, что любая оптимальная оценка имеет квадратичный риск, равный $[n\,I(\theta)]^{-1}$. Чтобы подчеркнуть это различие и указать в дальнейшем более конструктивный метод построения оптимальных оценок, введем еще одно определение, рассмотрев более общую задачу несмещенной оценки некоторой параметрической функции $\gamma(\theta)$.

Определение 5.2. Несмещенная оценка $\hat{\gamma}_n = \hat{\gamma}_n(X^{(n)})$ параметрической функции $\gamma(\theta)$ называется *эффективной* в классе $\mathcal{K}'$, если ее квадратичный риск

$$R(\gamma;\, \hat{\gamma}_n) = \mathbf{E}_\theta \left(\hat{\gamma}_n(X^{(n)}) - \gamma(\theta)\right)^2 = \mathbf{D}_\theta\, \hat{\gamma}_n(X^{(n)}) = [\,\gamma'(\theta)\,]^2/n\, I(\theta),$$

то есть (см. теорему 5.1 с $\hat{\theta}_n = \hat{\gamma}_n$) выполняется равенство

$$\hat{\gamma}_n(X^{(n)}) - \gamma(\theta) = C(\theta)\frac{\partial \mathcal{L}(\theta \,|\, X^{(n)})}{\partial \theta}. \tag{6}$$

Оценка $\hat{\gamma}_n$ называется *асимптотически эффективной* в классе $\mathcal{K}'$, если $\mathbf{E}_\theta\, \hat{\gamma}_n(X^{(n)}) \sim \gamma(\theta)$ и $\mathbf{D}_\theta\, \hat{\gamma}_n(X^{(n)}) \sim [\,\gamma'(\theta)\,]^2/n\, I(\theta)$, когда $n \to \infty$.

В силу теоремы 4.2 оценка по методу максимального правдоподобия скалярного параметра θ (в данном случае $\gamma(\theta) = \theta$) является асимптотически эффективной оценкой в классе $\mathcal{K}'$. Покажем, что она дает решение проблемы построения эффективной оценки в классе $\mathcal{K}$.

Пусть $\hat{\theta}_n$ – оценка максимального правдоподобия параметра θ. Определим оценку $\gamma(\hat{\theta}_n)$ параметрической функции $\gamma(\theta)$ с помощью подстановки вместо θ ее оценки $\hat{\theta}_n$.

Теорема 5.2. *Если $\gamma(\hat{\theta}_n)$ есть несмещенная оценка параметрической функции $\gamma(\theta)$ и эффективная в классе $\mathcal{K}'$ оценка γ_n^* параметрической функции $\gamma(\theta)$ существует, то при выполнении условий регулярности (R1)–(R5) и (B3)–(B4) почти наверное $\gamma_n^*(X^{(n)}) = \gamma(\hat{\theta}_n(X^{(n)}))$.*

Доказательство.. Если γ_n^* – эффективная оценка $\gamma(\theta)$, то она удовлетворяет равенству (6):

$$\gamma_n^*(X^{(n)}) - \gamma(\theta) = C(\theta)\partial \mathcal{L}(\theta \,|\, X^{(n)})/\partial\theta, \tag{7}$$

каково бы ни было $\theta \in \Theta$. Но если $\hat{\theta}_n$ – оценка по методу максимального правдоподобия, то $\partial \mathcal{L}(\hat{\theta}_n \,|\, X^{(n)})/\partial\theta = 0$, так что равенство (7) при $\theta = \hat{\theta}_n$ превращается в равенство $\gamma_n^*(X^{(n)}) - \gamma(\hat{\theta}_n) = 0$ почти наверное по вероятности P_θ^n.

Из доказанной теоремы немедленно вытекает, что выборочное среднее $\overline{X}$ есть эффективная (следовательно, и оптимальная) несмещенная оценка параметра θ таких распределений, как двухточечное, биномиальное при известном m, Пуассона, показательное; $\overline{X}$ есть также несмещенная оценка с равномерно минимальным квадратичным риском среднего значения μ нормального $(\mu,\, \sigma^2)$ распределения.

§6. Доверительные интервалы

Лекция 9

Мы рассмотрели несколько методов построения *точечных* оценок для параметров, значения которых определяют распределение наблюдаемой случайной величины. Был получен ряд утверждений о распределении таких оценок, что позволяет судить о надежности оценки при заданной точности, то есть вычислять вероятности событий вида $|\,\hat{\theta}_n(X^{(n)}) - \theta\,| \le \Delta$ при каждом фиксированном значении параметра θ. Поскольку именно значение θ нам неизвестно, то такого рода вычисления зачастую лишены практического смысла – слишком велик размах в надежности оценки $\hat{\theta}_n$ при различных θ, даже в случае, когда мы располагаем некоторой априорной информацией о возможной области значений этого параметра. Поэтому в ряде практических ситуаций пытаются решать обратную задачу: для фиксированной надежности, скажем, $1-\alpha$, где α мало, указать некоторую область значений θ, зависящую, естественно, от выборки $X^{(n)}$, которая с вероятностью, не меньшей $1-\alpha$, накрывает истинное, неизвестное нам значение θ, причем такое надежностное утверждение должно выполняться при любых $\theta \in \Theta$. В таком случае по размерам области, которые определяются выборочными значениями $x^{(n)}$, можно судить о точности такой *интервальной* оценки.

Определение 6.1. Подмножество $\Delta_n = \Delta_n(X^{(n)})$ параметрического пространства Θ называется $(1-\alpha)$-*доверительной областью*, если

$$P_\theta\Big(\Delta_n(X^{(n)}) \ni \theta\Big) \ge 1-\alpha, \tag{1}$$

каково бы ни было значение $\theta \in \Theta$. Заданное (фиксированное) значение $1-\alpha$ называется *доверительным уровнем*, а наименьшее значение левой части неравенства (1) по всем $\theta \in \Theta$ – *доверительным коэффициентом*. В случае $\Theta \subseteq \mathrm{R}$ доверительная область вида $\Delta_n = \big(\underline{\theta}_{\,n}(X^{(n)});\, \overline{\theta}_{\,n}(X^{(n)})\big)$ называется *доверительным интервалом*, в котором различаются *нижний* $\underline{\theta}_{\,n}$ и *верхний* $\overline{\theta}_{\,n}$ *доверительные пределы*. Доверительные интервалы вида $(\underline{\theta}_{\,n};\, \infty)$ и $(-\infty;\, \overline{\theta}_n)$ называются соответственно *нижней* и *верхней доверительными границами*.

Естественно, конфигурация доверительной области выбирается статистиком, сообразуясь с ее геометрической наглядностью и, главное, возможностью гарантировать доверительную вероятность. В случае скалярного

параметра доверительная область обычно выбирается в виде интервала, причем в ряде случаев, например, при оценке надежности или вероятности нежелательного события, в виде одностороннего интервала. В случае многомерного параметра обычно строятся доверительные эллипсоиды или параллелепипеды.

Следует обратить особое внимание на правильную формулировку доверительного утверждения, которая подчеркивается в неравенстве (1) записью $\Delta_n(X^{(n)}) \ni \theta$ вместо обычного $\theta \in \Delta_n(X^{(n)})$. Говорить, что значение параметра θ с вероятностью, не меньшей $1-\alpha$, принадлежит области Δ_n, значит сознательно вводить трудящихся на ниве прикладной статистики в заблуждение. Дело в том, что значение параметра θ в данной вероятностной модели не является случайной величиной, это постоянная, свойственная исследуемому объекту, а постоянная принадлежит какой-либо области только с вероятностью единица или ноль. Вся случайность заключена в самой доверительной области $\Delta_n(X^{(n)})$, и поэтому правильное доверительное утверждение гласит: *область* $\Delta_n(X^{(n)})$ *с вероятностью, не меньшей* $1-\alpha$, *накрывает истинное (неизвестное) значение* θ.

В самом начале нашего курса математической статистики в примере 1.1 с определением содержания общей серы в дизельном топливе мы строили доверительный интервал фиксированной ширины для среднего значения нормального распределения, когда занимались планированием объема испытаний, необходимого для достижения заданной точности и надежности оценки. Рассмотрим еще раз этот пример в свете введенных понятий интервальной оценки параметра.

1^0. Доверительный интервал для среднего значения нормального распределения при известной дисперсии. Итак, в примере 1.1 мы имели дело с выборкой $X^{(n)}$ из нормального $(\mu,\, \sigma^2)$ распределения, причем значение параметра σ нам было известно, так что в качестве неизвестного параметра θ выступало μ. Наша задача состоит в построении такого интервала $\left(\underline{\mu}_n(X^{(n)}),\, \overline{\mu}_n(X^{(n)})\right)$, что $P_\mu\left(\underline{\mu}_n \le \mu \le \overline{\mu}_n\right) \ge 1-\alpha$, при любом значении $\mu \in \mathrm{R}$.

Напомним, что в этом примере оценкой μ служило выборочное среднее $\overline{X}$ – несмещенная оценка μ с минимальным квадратичным риском. Эта линейная оценка обладает замечательным свойством инвариантности: распределение разности $\overline{X}-\mu$ не зависит от μ, и это обстоятельство подска-

зывает нам путь к построению доверительного интервала. Действительно,

$$P\left(\frac{|\,\overline{X}-\mu\,|}{\sigma}\sqrt{n}\,\le\,\lambda\right)=2\Phi(\lambda)-1,$$

и если положить λ равным корню уравнения $2\Phi(\lambda)-1=1-\alpha$, то есть выбрать $\lambda=\lambda_\alpha=\Phi^{-1}(1-\alpha/2)$, то интервал $(\overline{X}-\lambda_\alpha\sigma/\sqrt{n},\ \overline{X}+\lambda_\alpha\sigma/\sqrt{n})$ будет $(1-\alpha)$-доверительным интервалом для среднего значения μ нормального $(\mu,\,\sigma^2)$ распределения при известной дисперсии σ^2.

В этом простейшем примере на построение доверительного интервала ключевым моментом было использование *инвариантной* случайной функции $\hat\theta_n-\theta$ от оценки $\hat\theta_n=\overline{X}$ и параметра $\theta=\mu$. В принципе, именно на подобном выборе *опорной* функции $H(\hat\theta_n,\,\theta)$ с подходящей оценкой $\hat\theta_n$ параметра θ основаны исторически первые методы построения доверительных интервалов и множеств. Опорная функция $H(\,\cdot,\,\cdot\,)$ подбирается таким образом, чтобы она была монотонно возрастающей функцией второго аргумента θ, и при этом вероятность $P_\theta\left(H(\hat\theta_n(X^{(n)}),\,\theta)\le\lambda\right)$ для некоторых значений λ должна оставаться достаточно высокой (близкой к единице), каково бы ни было значение $\theta\in\Theta$. Мы проиллюстрируем этот метод построения доверительных интервалов с помощью подбора инвариантных опорных функций на примере нормального (μ,σ^2) распределения, строя доверительные интервалы для каждого из параметров при известном и неизвестном значениях другого ("мешающего") параметра.

2^0. Верхняя доверительная граница для дисперсии нормального распределения при известном среднем. При выборе из нормального $(\mu,\,\sigma^2)$ распределения с известным средним значением μ метод максимального правдоподобия приводит к несмещенной оценке $\hat\sigma_n^2=n^{-1}\sum_1^n(X_k-\mu)^2$ параметра σ^2. Используя результаты предыдущего параграфа, нетрудно показать, что $\hat\sigma_n^2$ есть несмещенная оценка с равномерно минимальным риском.

Поскольку, в чем мы неоднократно убеждались, $Y_k=(X_k-\mu)/\sigma\ \sim\ \mathcal{N}(0,\,1)$, $k=1,\dots,n$, то естественно рассмотреть в качестве опорной функцию $H(\hat\sigma_n^2,\,\sigma^2)=\sigma^{-2}\sum_1^n(X_k-\mu)^2$. Найдем ее распределение.

Лемма 6.1. *Если $Y_1,\dots,Y_n$ независимы и одинаково распределены по стандартному нормальному закону $\mathcal{N}(0,\,1)$, то $\sum_1^n Y_k^2$ имеет гамма-распределение $\mathrm{G}(n/2,\,2)$.*

Доказательство.. Покажем, что Y^2, где $Y \sim \mathcal{N}(0,\,1)$, имеет гамма-распределение $\mathrm{G}(1/2,\,2)$, после чего просто воспользуемся теоремой сложения для гамма-распределения (см. предложение 12.2, пункт 5^0 курса ТВ). Функция распределения Y^2 вычисляется по формуле $F(x) = P(Y^2 < x) = P(-\sqrt{x} < Y < \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) = 2\Phi(\sqrt{x}) - 1$, так что ее функция плотности

$$f(x) = \frac{d}{dx}\left[\frac{2}{\sqrt{2\pi}}\int\limits_{-\infty}^{\sqrt{x}} \exp\left\{-\frac{t^2}{2}\right\} - 1\right] = \frac{1}{2^{1/2}\Gamma(1/2)}x^{1/2-1}\mathrm{e}^{-x/2},$$

поскольку $\Gamma(1/2) = \sqrt{\pi}$. Мы видим, что это – функция плотности гамма-распределения $\mathrm{G}(1/2,\,2)$ с параметром формы $\lambda = 1/2$ и параметром масштаба $a = 2$, откуда, как было замечено выше, немедленно следует утверждение леммы.

Гамма-распределение $\mathrm{G}(n/2,\,2)$ очень часто используется в различных задачах математической статистики, и оно появилось раньше, чем гамма-распределение $\mathrm{G}(\lambda,\,a)$ общего вида, под названием *хи-квадрат распределение с n степенями свободы*. Функция распределения хи-квадрат обычно обозначается $K_n(x),\ x > 0$, а что касается термина "степени свободы", то его смысл прояснится по мере других применений хи-квадрат распределения.

Теперь мы можем перейти к нашей основной задаче – построению доверительных границ для σ^2. Если обратиться к практической стороне этой проблемы (см. в связи с этим пример 1.1), то легко понять, что статистика должна интересовать только верхняя (а не двусторонняя) граница σ^2, на которую он будет ориентироваться, чтобы обезопасить себя от грубых ошибок при планировании статистического эксперимента. Таким образом, мы должны сформулировать доверительное утверждение в форме $\sigma^2 \le \overline{\sigma}_n^2$. Понятно, что нижняя доверительная граница и двусторонние границы (доверительный интервал), коль скоро они кому-то потребуются, строятся аналогичным образом.

В рамках такой формулировки задачи, мы должны рассмотреть событие

$$A_\lambda = \left\{ H(\hat{\sigma}_n^2,\,\sigma^2) = \frac{\sum_1^n (X_k - \mu)^2}{\sigma^2} \ge \lambda \right\},$$

выбирая λ из условия $P_{\mu,\sigma}(A_\lambda) = 1 - \alpha$. Как мы только что выяснили, эта

вероятность не зависит от μ и σ, и в силу леммы 6.1 постоянная λ определяется квантилью хи-квадрат распределения с n степенями свободы – корнем уравнения $1-K_n(\lambda) = 1-\alpha$. Следовательно, верхняя $(1-\alpha)$-доверительная граница для σ^2 равна $\sum_1^n (X_k-\mu)^2/K_n^{-1}(\alpha)$, где, в соответствии с нашими стандартными обозначениями, $K_n^{-1}(\alpha)$ есть α-квантиль хи-квадрат распределения с n степенями свободы.

Используемые в рассмотренных примерах методы подбора опорных функций, основанные на принципе инвариантности статистик (оценок параметров μ и σ^2) относительно линейных преобразований, позволяют аналогичным образом подбирать такие функции и в случае неизвестных значений мешающего параметра. Так, если рассматривается задача построения доверительных границ для σ^2 при неизвестном μ, то естественно обратиться к оценке $S^2 = n^{-1}\sum_1^n (X_k-\overline{X})^2$ параметра σ^2, замечая, что ее распределение не зависит от μ, поскольку каждая из разностей $X_k - \overline{X}$ инвариантна относительно сдвига, когда X_k заменяется на $X_k - \mu,\ k = 1, \ldots, n$. Если разделить эти разности на σ, то мы получим случайные величины, распределение которых не зависит как от μ, так и от σ, и таким образом мы приходим к инвариантной опорной функции $H(S^2, \sigma^2) = S^2/\sigma^2$. Для вывода распределения этой функции можно обратиться к нормальным $(0, 1)$ случайным величинам $Y_k = (X_k - \mu)/\sigma,\ k = 1, \ldots, n$, поскольку, в чем легко убедиться, $H(S^2, \sigma^2) = n^{-1}\sum_1^n (Y_k - \overline{Y})^2$.

Если обратиться к задаче доверительной интервальной оценки μ при неизвестном значении σ, то здесь инвариантную опорную функцию можно построить, комбинируя ее из опорных функций задач 1^0 и 2^0. Как мы видели при решении этих задач, распределения случайных величин $(\overline{X}-\mu)/\sigma$ и S/σ не зависят от μ и σ, и поэтому в качестве опорной функции при интервальной оценке μ можно использовать опорную функцию, определяемую отношением этих величин, то есть функцию $|\,\overline{X} - \mu\,|/S$.

Однако для построения доверительных интервалов на основе таких функций нам необходимо найти совместное распределение статистик $\overline{X}$ и S^2. Мы получим это распределение в следующей лекции, сформулировав его в виде утверждения, известного в математической статистике как *лемма Фишера.*

Лекция 10

Теорема 6.1. *В случае выбора из нормального* $(\mu,\ \sigma^2)$ *распределения*

статистики $\overline{X}$ *и* S^2 *независимы,* $\overline{X} \sim \mathcal{N}(\mu,\, \sigma^2/n)$, *а* $nS^2/\sigma^2 \sim \chi^2_{n-1}$ *(имеет хи-квадрат распределение с* $n-1$ *степенью свободы).*

Доказательство.. Пусть $Y_1, \ldots, Y_n$ – случайная выборка из стандартного нормального распределения $\mathcal{N}(0,\, 1)$. Покажем, что статистики

$$\overline{Y} = \frac{1}{n}\sum_{k=1}^{n} Y_k \quad \text{и} \quad S_Y = \sum_{k=1}^{n}(Y_k - \overline{Y})^2$$

независимы, $\overline{Y} \sim \mathcal{N}(0,\, 1/n)$, а $S_Y \sim \chi^2_{n-1}$. Тогда утверждение теоремы будет следовать из того факта, что $\sigma Y_k + \mu$ имеют то же распределение, что и X_k, $k = 1, \ldots, n$, и, следовательно, распределение $\overline{X}$ совпадает с распределением $\sigma\overline{Y} + \mu$, а распределение S_Y – с распределением nS^2/σ^2.

Введем случайные величины

$$Z_k = \sum_{i=1}^{n} c_{ki} Y_i,\ k = 1, \ldots, n,$$

которые определяются заданием матрицы $\mathrm{C} = \| c_{ki} \|$ линейных преобразований случайных величин $Y_1, \ldots, Y_n$. Пусть элементы первой строки этой матрицы $c_{11} = \ldots = c_{1n} = 1/\sqrt{n}$, а остальные элементы матрицы C выберем так, чтобы произведение C на транспонированную матрицу C' было единичной матрицей: $\mathrm{CC}' = \mathrm{I}$. Как известно, такой выбор C возможен, и полученная таким образом матрица называется ортонормированной. Случайные величины $Z_1, \ldots, Z_n$ распределены в соответствии с n-мерным нормальным законом, для спецификации которого достаточно найти вектор средних значений этих величин и матрицу их ковариаций.

Средние значения

$$m_k = \mathbf{E}Z_k = \mathbf{E}\sum_{i=1}^{n} c_{ki} Y_k = \sum_{i=1}^{n} c_{ki}\mathbf{E}Y_k = 0,\ k = 1, \ldots, n.$$

Далее, поскольку средние значения равны нулю, ковариации этих случайных величин

$$\operatorname{cov}(Z_k,\, Z_j) = \mathbf{E}(Z_k - m_k)(Z_j - m_j) = \mathbf{E}Z_k Z_j = \mathbf{E}\sum_{i=1}^{n} c_{ki} Y_i \cdot \sum_{i=1}^{n} c_{ji} Y_i =$$

$$\mathbf{E}\sum_{i=1}^{n} c_{ki} c_{ji} Y_i^2 + \mathbf{E}\sum_{i \neq l}^{n} c_{ki} c_{jl} Y_i Y_l.$$

Если занести математические ожидания под знаки сумм и вспомнить, что $Y_1, \ldots, Y_n$ независимы, $\mathbf{E}Y_i = 0$, $\mathbf{E}Y_i^2 = \mathbf{D}Y_i = 1$, а при $i \neq l$ средние значения $\mathbf{E}Y_iY_l = \mathbf{E}Y_i\mathbf{E}Y_l = 0$, то получим, что

$$\operatorname{cov}(Z_k,\, Z_j) = \sum_{i=1}^{n} c_{ki}c_{ji}, \quad k, j = 1, \ldots, n.$$

Поскольку для ортонормированной матрицы последняя сумма равна нулю, если $k \neq j$, и равна единице, если $k = j$, то мы приходим к заключению, что $Z_1, \ldots, Z_n$ независимы и одинаково распределены по стандартному нормальному закону $\mathcal{N}(0,\, 1)$. Таким образом, ортонормированные преобразования случайных величин $Y_1, \ldots, Y_n$ не изменили их совместное распределение.

Теперь представим наши статистики $\overline{Y}$ и S_Y в терминах случайных величин $Z_1, \ldots, Z_n$. Поскольку

$$Z_1 = \sum_{i=1}^{n} c_{1i}Y_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i,$$

то $\overline{Y} = Z_1/\sqrt{n}$. Далее, ортонормированное линейное преобразование сохраняет сумму квадратов компонент преобразуемого вектора, то есть $\sum_1^n Z_k^2 = \sum_1^n Y_k^2$. Следовательно, статистика S_Y в новых переменных приобретает вид

$$\frac{S_Y}{n} = \frac{1}{n}\sum_{1}^{n} Y_k^2 - \overline{Y}^2 = \frac{1}{n}\sum_{1}^{n} Z_k^2 - \frac{Z_1^2}{n} = \frac{1}{n}\sum_{2}^{n} Z_k^2.$$

Итак, распределение $\overline{Y}$ совпадает с распределением $Z_1/\sqrt{n}$, а распределение S_Y – с распределением суммы квадратов $n-1$ независимых в совокупности и независящих от Z_1 нормальных $(0,\, 1)$ случайных величин. Следовательно, $\overline{Y}$ и S_Y независимы, $\overline{Y} \sim \mathcal{N}(0,\, 1/n)$, $S_Y \sim \chi^2_{n-1}$ (см. лемму 6.1), и "лемма Фишера" доказана.

Установив совместное распределение выборочного среднего и выборочной дисперсии в случае выбора из нормального распределения, мы можем приступить к построению доверительных интервалов для каждого из параметров μ и σ при неизвестном значении другого параметра.

3^0. Верхняя доверительная граница для дисперсии нормального распределения при неизвестном среднем. Эта граница на-

ходится наиболее просто, поскольку распределение опорной функции

$$\frac{nS^2}{\sigma^2} = \frac{\sum_1^n (X_k - \overline{X})^2}{\sigma^2} = \sum_{k=1}^n \left(\frac{X_k - \mu}{\sigma} - \frac{1}{n} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} \right)^2 = \sum_1^n (Y_k - \overline{Y})^2 \tag{2}$$

есть хи-квадрат распределение с $n-1$ степенью свободы (см. теорему 6.1). Следовательно, верхняя $(1-\alpha)$-доверительная граница определяется квантилью $\lambda_\alpha = K^{-1}(\alpha)$ хи-квадрат распределения – корнем уравнения

$$P(nS^2/\sigma^2 \geq \lambda) = 1 - K_{n-1}(\lambda) = 1 - \alpha,$$

и доверительное утверждение $\sigma^2 \leq \overline{\sigma}_n^2 = nS^2/K_{n-1}^{-1}(\alpha)$ выполняется с заданной вероятностью $1 - \alpha$.

4^0. Доверительный интервал для среднего значения нормального распределения при неизвестной дисперсии. В этой задаче мы имеем дело с двусторонними доверительными границами (доверительным интервалом), и в соответствии с выбором опорной функции $|\,\overline{X} - \mu\,|/S$, о которой мы говорили перед доказательством теоремы 6.1, нам потребуется знание вероятности события вида $|\,\overline{X} - \mu\,|/S \leq \lambda$.

В начале XIX века английский математик В.Госсет, писавший под псевдонимом "Стьюдент" (Student), нашел распределение случайной величины $T_\nu = \xi\sqrt{\nu}/\sqrt{\chi_\nu^2}$, где $\xi \sim \mathcal{N}(0,\,1)$, а χ_ν^2 – случайная величина, не зависящая от ξ и распределенная по закону хи-квадрат с ν степенями свободы. Естественно, его исследования были связаны с проблемами статистического вывода о среднем значении μ нормального распределения при неизвестной дисперсии, и Стьюдент искал распределение опорной функции (см. (2)) в связи с переходом в записи опорной функции в терминах X_k к Y_k)

$$H = \frac{\overline{X} - \mu}{S}\sqrt{n-1} = \frac{\frac{1}{\sqrt{n}}\sum_1^n Y_k}{\sqrt{\sum_1^n (Y_k - \overline{Y})^2}}\sqrt{n-1},$$

$$Y_k = \frac{X_k - \mu}{\sigma} \sim \mathcal{N}(0,\,1),\ k = 1, \ldots, n,$$

которая отличается от выбранной нами опорной функции только множителем $\sqrt{n-1}$, и поэтому также может быть использована в построении доверительного интервала для μ при неизвестном σ. То, что распределения T_{n-1} и H совпадают, следует из теоремы 6.1: случайная величина в знаменателе

$\xi = \sum_1^n Y_k/\sqrt{n} \sim \mathcal{N}(0,\,1)$ не зависит от $\sum_1^n (Y_k - \overline{Y})^2 \sim \chi^2_{n-1}$, разделив которую на значение степени свободы $n-1$, получаем $\frac{1}{\nu}\chi^2_\nu$ с $\nu = n-1$.

Найдем распределение случайной величины T_ν, которое называется *распределением Стьюдента с ν степенями свободы* или *t-распределением*. Совместная функция плотности независимых случайных величин ξ и $\eta = \chi^2_\nu$ равна

$$f(x,\,y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} y^{\nu/2-1} \exp\left\{-\frac{y}{2}\right\},$$

так что функция распределения случайной величины T_ν

$$\mathrm{S}_\nu(t) = P(\xi\sqrt{\nu/\eta} < t) =$$

$$\iint\limits_{x\sqrt{\nu} < \sqrt{y}t} f(x,\,y)\,dx\,dy = \int\limits_0^\infty dy \int\limits_{-\infty}^{t\sqrt{y/\nu}} f(x,\,y)\,dx.$$

Дифференцируя это выражение по t, находим функцию плотности распределения Стьюдента

$$\mathrm{s}_\nu(t) = \int\limits_0^\infty \sqrt{y/\nu} f(t\sqrt{y/\nu},\,y)\,dy =$$

$$\frac{1}{\sqrt{\pi\nu}2^{(\nu+1)/2}\Gamma(\nu/2)} \int\limits_0^\infty y^{\frac{\nu+1}{2}-1} \exp\left\{-\frac{y}{2}\left(1+\frac{t^2}{\nu}\right)\right\} dt =$$

$$= \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

Вид полученной функции плотности говорит о том, что распределение Стьюдента можно трактовать как обобщение стандартного ($a=0$, $b=1$) распределения Коши $\mathrm{C}(a,\,b)$, которое получается из распределения Стьюдента при числе степеней свободы $\nu = 1$. Это симметричное распределение, и поэтому $\mathrm{S}_\nu(-t) = 1 - \mathrm{S}_\nu(t)$, что позволяет нам довольно просто построить доверительный интервал для μ с помощью квантили распределения $\mathrm{S}_{n-1}(\cdot)$:

$$P(|\,T_{n-1}\,| \le t) = \mathrm{S}_{n-1}(t) - \mathrm{S}_{n-1}(-t) = 2\mathrm{S}_{n-1}(t) - 1 = 1 - \alpha,$$

откуда $t_\alpha = \mathrm{S}_{n-1}^{-1}(1-\alpha/2)$, и $(1-\alpha)$-доверительный интервал для μ определяется пределами $\overline{X} \pm St_\alpha/\sqrt{n-1}$.

Итак, мы построили доверительные пределы для параметров μ и σ^2 нормального распределения. Таблицы нормального, хи-квадрат и стьюдентского распределений, а также квантилей этих распределений, необходимые для численной реализации доверительных оценок, смотрите в книге Большев Л.Н., Смирнов Н.В. Таблицы математической статистики, М.: Наука, 1983, которая в дальнейшем будет цитироваться как ТМС. Еще раз отметим, что возможность доверительной оценки этих параметров определялась, в основном, инвариантностью семейства нормальных распределений относительно линейной группы преобразований. Точно так же мы можем построить доверительные пределы для параметра θ показательного распределения или для параметра масштаба гамма распределения при известном параметре формы; мы вернемся к этим задачам позднее при обсуждении проблемы оптимизации доверительной оценки. Что же касается других распределений, то здесь проблема осложняется отсутствием инвариантных опорных функций и невозможностью получить распределение оценок параметра, для которого строятся доверительные пределы, в явном виде. Тем не менее существует достаточно общий подход к данной проблеме, основанный на асимптотической нормальности распределения оценок по методу моментов или методу максимального правдоподобия.

Пусть $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ – асимптотически нормальная со средним θ и дисперсией $\sigma^2(\theta)/n$ оценка параметра θ (например, при определенных условиях регулярности (см. теорему 4.2) оценка максимального правдоподобия асимптотически нормальна со средним θ и дисперсией $[\,nI(\theta)\,]^{-1}$). Тогда при $n \longrightarrow \infty$ вероятность

$$P_\theta\left(\frac{|\,\hat{\theta}_n - \theta\,|}{\sigma(\theta)}\sqrt{n} \;\le\; \lambda_\alpha\right) \longrightarrow 1-\alpha,$$

при любом $\theta \in \Theta$, если $\lambda_\alpha = \Phi^{-1}(1-\alpha/2)$, и мы получаем *асимптотически* $(1-\alpha)$-доверительное множество

$$\Delta_n(X^{(n)}) = \left\{\theta:\ \frac{|\,\hat{\theta}_n - \theta\,|}{\sigma(\theta)}\sqrt{n} \;\le\; \lambda_\alpha\right\} \bigcap \Theta.$$

Если Δ_n есть интервал на прямой R, то мы решили задачу интервальной оценки параметра θ. Если же это некоторое вычурное и непригодное

к употреблению подмножество R, то можно пойти на дальнейшие упрощения асимптотического утверждения, заменив в определении Δ_n параметрическую функцию $\sigma(\theta)$ на ее оценку $\sigma(\hat{\theta}_n)$. Достаточно потребовать непрерывность функции $\sigma(\theta)$, $\theta \in \Theta$, чтобы, ссылаясь на теорему Слуцкого (предложение 11.1 курса ТВ с $\xi_n \propto \sigma(\hat{\theta}_n) \underset{P}{\to} \sigma(\theta)$), утверждать, что при $n \to \infty$

$$P_\theta \left(\hat{\theta}_n - \frac{\lambda_\alpha \sigma(\hat{\theta}_n)}{\sqrt{n}} \le \theta \le \hat{\theta}_n + \frac{\lambda_\alpha \sigma(\hat{\theta}_n)}{\sqrt{n}} \right) \longrightarrow 1 - \alpha.$$

Проиллюстрируем "работу" этого метода на двух полезных в практическом отношении примерах.

5^0. Асимптотически доверительный интервал для вероятности успеха в испытаниях Бернулли. В схеме испытаний Бернулли – выборе из распределения бинарной случайной величины X, принимающей значение 1 ("успех") с вероятностью p и значение 0 ("неудача") с вероятностью $1 - p$, оптимальной несмещенной оценкой p является выборочное среднее $\overline{X} = n^{-1} \sum_1^n X_k$ или, что то же, относительная частота успешных исходов в n испытаниях. Статистика $n\overline{X}$ имеет биномиальное распределение B(n, p), и это позволяет насчитать таблицы доверительных пределов для p при различных значениях доверительного уровня $1 - \alpha$, объема выборки n и числа успешных исходов $n\overline{x}$ (см., например, ТМС). Что же дает асимптотический подход к построению доверительных интервалов?

Выборочное среднее асимптотически нормально со средним p и дисперсией $p(1-p)/n$. Следовательно, $(1-\alpha)$-доверительная область $\Delta_n = \left\{ p : \ 0 \le p \le 1, \, | \, \overline{X} - p \, | \le \lambda_\alpha \sqrt{p(1-p)/n} \right\}$. Разрешая неравенства в фигурных скобках относительно p, получаем доверительный интервал

$$\frac{n}{n + \lambda_\alpha^2} \left(\overline{X} + \frac{\lambda_\alpha^2}{2n} \pm \lambda_\alpha \sqrt{\frac{\overline{X}(1-\overline{X})}{n} + \frac{\lambda_\alpha^2}{4n^2}} \right),$$

который при больших объемах испытаний n мало отличается от доверительного интервала $\overline{X} \pm \lambda_\alpha \sqrt{\overline{X}(1-\overline{X})/n}$, полученного заменой $\sigma^2(p) = p(1-p)$ на ее оценку $\overline{X}(1-\overline{X})$:

6^0. Асимптотически доверительный интервал для параметра интенсивности распределения Пуассона. Распределение Пуассона $\mathrm{P}(\theta)$ с функцией плотности (по считающей мере) $f(x\,|\,\theta) = P_\theta(X = x) = \theta^x \mathrm{e}^\theta / x!,\ x = 0, 1, 2, \ldots$, индексируется положительным параметром θ, оптимальная несмещенная оценка которого по выборке $X^{(n)}$ объема n, как и в предыдущем примере, определяется выборочным средним $\overline{X}$. Для распределения Пуассона также справедлива теорема сложения: $n\overline{X} \sim \mathrm{P}(n\theta)$, и на основе этого можно построить точные доверительные пределы для θ, таблица которых имеется в упомянутом сборнике ТМС. Но оценка $\overline{X}$ асимптотически нормальна $(\theta,\, \theta/n)$, что позволяет определить асимптотически доверительную область $\Delta_n = \{\theta :\ \ \theta > 0, \,|\,\overline{X} - \theta\,| \le \lambda_\alpha \sqrt{\theta/n}\}$. Решение неравенств в фигурных скобках относительно θ дает асимптотически доверительный интервал

$$\overline{X} + \frac{\lambda_\alpha^2}{2n} \pm \lambda_\alpha \sqrt{\frac{\overline{X}}{n} + \frac{\lambda_\alpha^2}{4n^2}}\,.$$

Наконец, заменяя $\sigma^2(\theta) = \theta$ ее оценкой $\overline{X}$, получаем также асимптотически доверительный, но, как показывают числовые расчеты, менее точный интервал $\overline{X} \pm \lambda_\alpha \sqrt{\overline{X}/n}\,$.

На этом мы заканчиваем изложение простейших методов построения доверительных и асимптотически доверительных интервалов на основе подбора опорных функций. О проблеме оптимального интервального оценивания мы поговорим позднее, изучив теорию оптимальной проверки гипотез – высказываний о возможных значениях параметра θ. Оставшиеся лекции будут посвящены именно этой теории.

§7. Статистическая проверка гипотез (критерии значимости)

Лекция 11

В приложениях математической статистики существует обширный класс задач, в которых требуется проверить истинность некоторого высказывания относительно исследуемого объекта или выбрать одно из альтернативных решений, которое определит дальнейшее поведение статистика по отношению к этому объекту. Например, при аттестации партии дизельного топлива по общему содержанию серы мы должны не только дать точечную оценку данной характеристики топлива, но и принять решение о качестве выпускаемого продукта, которое повлечет за собой одно из следующих действий – или отослать топливо потребителю, или произвести дополнительную очистку топлива от вредных примесей. Точно так же в примере 1.2 мы строили статистическое правило, позволяющее принять одно из двух решений относительно нового лечебного препарата – или признать его эффективным и внедрить в лечебную практику, или запретить его дальнейшее использование. В исследованиях, подобных опытам Менделя, часто надо проверить гипотезу относительно предполагаемого значения вероятности наследования доминантного признака. Селекционер, работающий над получением нового вида пшеницы, должен подкрепить свое заключение о превосходстве нового вида над тем, который уже используется в сельскохозяйственной практике, с помощью сопоставления данных об урожайности этих видов. И так далее, и тому подобное, – вы сами можете привести примеры таких задач по выбору одного из ряда альтернативных решений.

В нашем курсе математической статистики мы рассмотрим задачи, связанные только с выбором одного из двух решений. Пусть мы высказываем некоторое суждение (или предпринимаем действие) об исследуемом объекте, и пусть d_0 – решение об истинности этого суждения, в то время как d_1 – решение о его ложности. Таким образом, пространство решений $\mathcal{D}$ в данной статистической проблеме состоит из точек: $\mathcal{D} = \{d_0, d_1\}$.

Для выбора одного из решений мы наблюдаем случайную выборку $X^{(n)}$ из некоторого распределения P_θ, значение параметра θ которого нам неизвестно. Пусть Θ – область возможных значений θ, которую мы назвали параметрическим пространством. В соответствии с принятой нами в §1 идеологией статистического вывода мы сопоставляем каждому решению $d \in \mathcal{D}$ определенное подмножество Θ_d пространства Θ, то есть интерпретируем каждое решение в терминах высказываний об истинном значении

параметра θ. В нашей статистической проблеме выбора одного из двух решений положим $\Theta_i = \Theta_{d_i},\ i = 0, 1$, и введем ряд понятий и определений, используемых при решении этой проблемы.

Утверждение $H_0 :\ \theta \in \Theta_0$ называется *нулевой гипотезой*, а утверждение $H_1 :\ \theta \in \Theta_1$ – *альтернативной гипотезой* или (коротко) *альтернативой*. Гипотеза H_i называется *простой*, если соответствующее Θ_i состоит из одной точки параметрического пространства Θ; в противном случае H_i называется *сложной* гипотезой; $i = 0, 1$. Так, в примере 1.2 с испытанием нового лечебного препарата параметр θ означал вероятность успешного лечения каждого пациента, и нулевая гипотеза $H_0 :\ \theta = 1/2$ о "нейтральности" препарата есть простая гипотеза, в то время как альтернативная гипотеза $H_1 :\ \theta > 1/2$ об его эффективности – сложная гипотеза.

Правило, по которому принимается или отвергается нулевая гипотеза H_0, называется *критерием*. Иногда добавляется – *критерий согласия* (с нулевой гипотезой), особенно, когда альтернатива H_1 определена не совсем четко и под H_1 подразумевается "все остальное" . В случае полного равноправия гипотез говорят о критерии *различения гипотез*. Критерий определяется заданием особого подмножества S выборочного пространства $\mathcal{X}^n$, которое называется *критической областью*: если выборочные данные $x^{(n)}$ попадают в эту область, то нулевая гипотеза H_0 отклоняется и принимается альтернативное решение – справедлива H_1. Область $A = S^c = \mathcal{X}^n \setminus S$ называется *областью принятия* нулевой гипотезы. Нам будет удобно проводить спецификацию критической области в виде ее индикаторной функции $\varphi = \varphi(X^{(n)})$, которая называется *критической функцией* или, поскольку она определяет статистическое правило проверки гипотезы, просто *критерием*. Итак, функция $\varphi(X^{(n)})$ есть бинарная случайная величина, принимающая значение 1, если произошло событие $X^{(n)} \in S$, и значение 0, если произошло противоположное событие $X^{(n)} \in A$. Понятно, что математическое ожидание $\mathbf{E}\varphi(X^{(n)})$ означает вероятность отклонения гипотезы H_0.

В рассматриваемой статистической проблеме величина риска, связанная с отклонением верной гипотезы, обычно соотносится с функцией потерь типа 1 – 0: потери считаются равными 1, если принята гипотеза H_i, а в действительности $\theta \in \Theta_{1-i},\ i = 0, 1$; если же принята H_i и $\theta \in \Theta_i,\ i = 0, 1$, то потери полагаются равными нулю. Легко видеть, что величина риска при любом значении параметра θ может быть определена с помощью функции $m(\theta) = \mathbf{E}_\theta\varphi(X^{(n)}) = P_\theta(X^{(n)} \in S)$, которая называется *функцией мощности* критерия φ. Эта функция указывает, как часто мы отклоняем нулевую

гипотезу, когда θ – истинное значение параметра, и хорошим следует считать тот критерий, у которого функция $m(\theta)$ принимает близкие к нулю значения в области Θ_0 и близкие к единице – в области Θ_1. В связи с этим вводятся две компоненты функции риска: $\alpha(\theta) = m(\theta)$ при $\theta \in \Theta_0$ и $\beta(\theta) = 1 - m(\theta)$ при $\theta \in \Theta_1$. Функция $\alpha(\theta),\ \theta \in \Theta_0$ называется *вероятностью ошибки первого рода* – она указывает относительную частоту отклонения гипотезы H_0, когда она в действительности верна ($\theta \in \Theta_0$). Функция $\beta(\theta),\ \theta \in \Theta_1$ называется *вероятностью ошибки второго рода* – она указывает относительную частоту принятия гипотезы H_0, когда она ложна (верна альтернативная гипотеза $H_1 :\ \theta \in \Theta_1$). Заметим, что функция мощности $m(\theta)$ в области Θ_1 трактуется как вероятность отклонения гипотезы H_0, когда в действительности выбор идет из распределения с альтернативным значением $\theta \in \Theta_1$, и поэтому часть $m(\theta)$ при $\theta \in \Theta_1$ называется *мощностью* критерия φ.

Легко понять, что при фиксированном объеме наблюдений n невозможно одновременно минимизировать вероятности обеих ошибок, – для уменьшения вероятности ошибки первого рода $\alpha(\theta) = P_\theta(X^{(n)} \in S),\ \theta \in \Theta_0$, необходимо уменьшить критическую область S, что приведет к увеличению области A принятия нулевой гипотезы и, следовательно, к увеличению вероятности ошибки второго рода $\beta(u) = P_u(X^{(n)} \in A),\ u \in \Theta_1$. Здесь возникает такая же ситуация, что и в проблеме построения оценки параметра θ с равномерно минимальным риском, – такие оценки существуют только в определенном классе статистических правил, например, в классе несмещенных оценок. Однако, даже и помимо задачи проверки гипотез с минимальной вероятностью ошибки, и намного раньше создания общей теории наиболее мощных критериев в статистической практике сложился следующий подход к управлению риском критерия.

Предположим, что отклонение гипотезы H_0, когда она в действительности верна, приводит к более тяжким последствиям, чем ее принятие при справедливости альтернативы. В таком случае мы заинтересованы в первую очередь контролировать вероятность ошибки первого рода. С этой целью заранее фиксируется (выбирается) некоторый уровень α, выше которого вероятность ошибки первого рода не допустима, и критическая область S (критерий φ) определяется таким образом, что $\alpha(\theta) \le \alpha$, каково бы ни было $\theta \in \Theta_0$. Это ограничение α на вероятность ошибки первого рода называется *уровнем значимости*, а сам критерий φ, для которого выполняется это ограничение, – *критерием уровня* α. Наибольшее значение вероятности

ошибки первого рода

$$\overline{\alpha} = \sup_{\theta \in \Theta_0} \alpha(\theta)$$

называется *размером* критерия φ, и если $\overline{\alpha} = \alpha$, то говорят о *критерии* φ *размера* α.

В этом выборе ограничения именно на вероятность ошибки первого, а не второго рода проявляется типичная асимметрия в практической ценности гипотезы и альтернативы. Например, если проверяется эффективность нового лекарственного препарата, то нулевой гипотезе должно соответствовать решение о его неэффективности, ибо, отклонив эту гипотезу, когда она верна, мы внедрим в лечебную практику бесполезное или вредное лекарство, что приведет к более тяжким последствиям, чем отклонение в действительности эффективного препарата. Но если мы ищем золото, анализируя состав кернов при бурении предполагаемого месторождения, то естественно принять за нулевую гипотезу утверждение о наличии золота, ибо отклонив ее, когда она верна, мы потеряем намного больше, чем стоимость нескольких дополнительных анализов, удостоверяющих, что золото в разбуренной местности отсутствует.

Следует также обратить особое внимание на общую методологию проверки гипотез, отражаемую в выборе малого значения уровня α. Если наши выборочные данные попадают в область S с исключительно малой вероятностью, то естественно предположить, что то утверждение, которое привело к этому маловероятному событию, не соответствует истине и отклонить его. Поступая таким образом, мы будем терять в действительности верную гипотезу H_0 крайне редко – не более, чем в $100\alpha\%$ случаев.

Простейший метод построения критериев значимости состоит в использовании состоятельных оценок тестируемого параметра θ. Рассмотрим простейший случай: θ – скалярный параметр, вероятностная модель не содержит других (мешающих) параметров и проверяется простая гипотеза $H_0 : \ \theta = \theta_0$ при альтернативе $H_1 : \ \theta \neq \theta_0$, где θ_0 – некоторое, априори фиксированное значение параметра θ (например, в опытах Менделя проверяется гипотеза: вероятность θ наследования доминантного признака равна $\theta_0 = 3/4$). Если $\hat{\theta}_n = \hat{\theta}_n(X^{(n)})$ – состоятельная оценка θ, дисперсия которой стремится к нулю при $n \to \infty$ как $O(1/\sqrt{n})$, то естественно определить критическую область посредством неравенства $\sqrt{n}|\,\hat{\theta}_n(X^{(n)}) - \theta_0\,| > C$. Вероятность ошибки первого рода такого критерия $\alpha(\theta_0,\, C) = P_{\theta_0}(\sqrt{n}|\,\hat{\theta}_n(X^{(n)}) - \theta_0\,| > C)$, и приравнивая эту вероятность заданному уровню значимости α,

находим *критическую константу* $C = C(\alpha)$ как квантиль распределения случайной величины $\sqrt{n}|\,\hat{\theta}_n(X^{(n)}) - \theta_0\,|$; такой выбор C приводит к критерию уровня α. Если $\theta\,(\neq \theta_0)$ – некоторое альтернативное значение параметра, то, в силу состоятельности оценки, $\sqrt{n}|\,\hat{\theta}_n(X^{(n)}) - \theta_0\,| \underset{P}{\to} \infty$, и поэтому вероятность ошибки второго рода $\beta(\theta) = P_\theta(\sqrt{n}|\,\hat{\theta}_n(X^{(n)}) - \theta_0\,| \le C(\alpha)) \to 0$, когда $n \to \infty$. Таким образом, мы получаем критерий заданного уровня α, обладающий к тому же свойством *состоятельности* – его вероятность ошибки второго рода стремится к нулю при неограниченном возрастании объема выборки n.

Сформулируем теперь основную задачу теории статистической проверки гипотез: *требуется найти такой критерий φ уровня α, который равномерно по всем $\theta \in \Theta_1$ максимизирует мощность $m(\theta)$ или, что то же, равномерно по $\theta \in \Theta_1$ минимизирует вероятность ошибки второго рода $\beta(\theta)$*. Мы укажем метод построения таких *равномерно наиболее мощных критериев* заданного уровня α в следующем параграфе, а пока обратимся к иллюстрациям введенных понятий и построению наиболее часто используемых на практике критериев, касающихся проверки гипотез о значениях параметров нормального распределения.

1^0. Проверка гипотезы о величине среднего значения нормального распределения при известной дисперсии. Рассмотрим сначала наиболее часто встречающуюся в практических применениях математической статистики задачу проверки сложной гипотезы $H_0:\ \mu \le \mu_0$ при сложной альтернативе $H_1:\ \mu > \mu_0$ о среднем значении μ нормального $(\mu,\, \sigma^2)$ распределения при известном значении дисперсии σ^2. Выборочное среднее $\overline{X} = n^{-1}\sum_1^n X_k$ есть оптимальная оценка неизвестного значения μ, и поэтому, в соответствии с только что предложенным методом построения состоятельных критериев, рассмотрим критерий, отвергающий нулевую гипотезу H_0, когда $\sqrt{n}(\overline{X} - \mu_0) > C$, или, что то же, $\overline{X} > C$, поскольку значения μ_0 и n фиксированы и известны. Постоянная C должна выбираться по заданному уровню значимости α, ограничивающему максимальное значение вероятности ошибки первого рода.

Так как при выборе из нормального $(\mu,\, \sigma^2)$ распределения статистика $\overline{X} \sim \mathcal{N}(\mu,\, \sigma^2/n)$, то функция мощности этого критерия

$$m(\mu) = P_\mu(\overline{X} > C) = 1 - \Phi\left(\frac{C-\mu}{\sigma}\sqrt{n}\right) = \Phi\left(\frac{\mu - C}{\sigma}\sqrt{n}\right).$$

Легко видеть, что $m(\mu)$ – строго возрастает с ростом μ, так что размер критерия

$$\overline{\alpha} = \max_{\mu \le \mu_0} m(\mu) = m(\mu_0) = 1 - \Phi\left(\frac{C - \mu_0}{\sigma}\sqrt{n}\right).$$

Приравнивая размер критерия уровню значимости α, находим критическое значение $C(\alpha) = \mu_0 + \Phi^{-1}(1-\alpha)\sigma/\sqrt{n}$.

Вероятность ошибки второго рода нашего критерия размера α

$$\beta(\mu) = P_\mu(\overline{X} \le C(\alpha)) = \Phi\left(\frac{C(\alpha) - \mu}{\sigma}\sqrt{n}\right) =$$

$$\Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} + \Phi^{-1}(1-\alpha)\right),\ \mu > \mu_0, \qquad (1)$$

убывает с ростом μ по мере ее отхода от граничного значения μ_0. Наибольшее значение $\beta(\mu)$ достигается в точке $\mu = \mu_0$ и равно $1 - \alpha$. Это значение не зависит от размера выборки n, и поэтому требуются дополнительные соображения при планировании объема наблюдений. Обычно используется метод введения так называемой *зоны безразличия* – интервала (μ_0, μ_1), который выбирается из тех соображений, что при истинном значении $\mu \in (\mu_0, \mu_1)$ принятие нулевой гипотезы H_0 не приводит к слишком тяжелым последствиям. Однако при истинном $\mu \ge \mu_1$ вероятность принятия H_0 должна быть под контролем и не превосходить некоторого предписанного значения β. Это обстоятельство позволяет спланировать объем выборки n, определив его из неравенства $\beta(\mu_1) \le \beta$. Используя формулу (1) для $\beta(\mu)$, находим, что *объем выборки* $n = n(\alpha, \beta, \mu_0, \mu_1)$, *необходимый для различения гипотез* $\mu \le \mu_0$ *и* $\mu \ge \mu_1$ *с заданными ограничениями* α *и* β *на вероятности ошибок первого и второго рода,* равен наименьшему целому n, удовлетворяющему неравенству

$$n \ge \frac{\left[\,\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\,\right]^2}{(\mu_1 - \mu_0)^2}\,\sigma^2.$$

Аналогичным методом строится критерий для проверки простой гипотезы $H_0 :\ \mu = \mu_0$ при сложной альтернативе $H_1 :\ \mu \ne \mu_0$. В этой задаче естественно определить критическую область посредством неравенства $|\,\overline{X} - \mu_0\,| > C$. Функция мощности такого критерия

$$m(\mu) = 1 - \left[\,\Phi\left(\frac{C + \mu_0 - \mu}{\sigma}\sqrt{n}\right) - \Phi\left(\frac{-C + \mu_0 - \mu}{\sigma}\sqrt{n}\right)\,\right]$$

строго убывает при $\mu < \mu_0$, возрастает при $\mu > \mu_0$ и при $\mu = \mu_0$ равна вероятности ошибки первого рода. Таким образом, критическая константа $C = C(\alpha)$ определяется по заданному уровню значимости α из уравнения

$$m(\mu_0) = 1 - \left[\Phi\left(\frac{C}{\sigma}\sqrt{n}\right) - \Phi\left(\frac{-C}{\sigma}\sqrt{n}\right) \right] = 2\left[1 - \Phi\left(\frac{C}{\sigma}\sqrt{n}\right) \right] = 1 - \alpha,$$

откуда $C(\alpha) = \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n}$.

Лекция 12

2^0. Проверка гипотезы о величине дисперсии нормального распределения при неизвестном среднем значении. Это типичная задача контроля за величиной случайной ошибки в параллельных наблюдениях некоторой характеристики исследуемого объекта. Так как превышение характеристики случайной погрешности σ над некоторым номиналом σ_0 в случае, когда мы утверждаем $\sigma \le \sigma_0$, влечет более серьезные последствия, чем неоправданные претензии к слишком большому разбросу в данных, то следует принять за нулевую гипотезу $\sigma > \sigma_0$. Проверка этой гипотезы проводится при естественной альтернативе $H_1 : \ \sigma \le \sigma_0$, причем мы не знаем значения мешающего параметра μ – среднего значения нормального распределения, из которого производится выбор.

Как нам известно, выборочная дисперсия $S^2 = n^{-1}\sum_1^n (X_k - \overline{X})^2$ есть состоятельная оценка σ^2, ее распределение не зависит от μ, а случайная величина nS^2/σ^2 имеет хи-квадрат распределение с $n-1$ степенью свободы. Таким образом, разумно рассмотреть критерий с критической областью $nS^2 < C$. Функция мощности такого критерия

$$m(\sigma) = P_{\mu,\sigma}\left(\frac{nS^2}{\sigma^2} \le \frac{C}{\sigma^2}\right) = K_{n-1}\left(\frac{C}{\sigma^2}\right)$$

монотонно убывает с ростом σ, поэтому наибольшее значение вероятности ошибки первого рода достигается при $\sigma = \sigma_0$, и критическое значение $C(\alpha)$ критерия требуемого размера α определяется из уравнения $K_{n-1}\left(C\sigma_0^{-2}\right) = \alpha$. Итак, $C(\alpha) = \sigma_0^2 K_{n-1}^{-1}(\alpha)$; вероятность ошибки второго рода

$$\beta(\sigma) = P_{\mu,\sigma}\left(nS^2 > C(\alpha)\right) = 1 - K_{n-1}\left(\frac{\sigma_0^2}{\sigma^2}K_{n-1}^{-1}(\alpha)\right), \ \sigma \le \sigma_0,$$

монотонно убывает по мере отхода истинного значения σ от номинала σ_0.

3^0. Проверка гипотезы о величине среднего значения нормального распределения при неизвестной дисперсии (одновыборочный критерий Стьюдента). Вы, наверное, обратили внимание, что при построении критериев значимости мы по существу используем методы построения доверительных множеств? Это, действительно, так – между задачами доверительной оценки и проверки гипотез существует много общего, и, решив одну задачу, мы сразу же получаем решение другой. В конце этого параграфа мы формализуем этот параллелизм, а пока будем использовать его на интуитивном уровне: предлагается использовать для проверки гипотез о среднем значении нормального распределения статистику Стьюдента.

Рассмотрим сначала задачу проверки сложной гипотезы $H_0: \ \mu \le \mu_0$ при сложной альтернативе $H_1: \ \mu > \mu_0$. Так как выборочное среднее $\overline{X}$ есть состоятельная оценка значения μ, то статистика Стьюдента

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n-1}$$

опосредственно, через выборочные данные, характеризует удаленность истинного среднего значения μ от границы μ_0, разделяющей гипотезу и альтернативу. Поэтому предлагается отвергать нулевую гипотезу $\mu \le \mu_0$, если $T > C$, выбирая C, как обычно, по заданному уровню значимости α. Для решения последней задачи необходимо исследовать поведение функции мощности $m(\mu;\, \sigma) = P_{\mu,\sigma}(T > C)$ критерия $T > C$. Если мы покажем, что $m(\mu;\, \sigma)$ есть монотонно возрастающая функция аргумента μ при любом фиксированном значении аргумента σ, то наибольшее значение вероятности ошибки первого рода $\alpha(\mu;\sigma) = m(\mu: \ \sigma), \ \mu \le \mu_0$ при каждом фиксированном σ будет достигаться в точке $\mu = \mu_0$. Следовательно, размер критерия в таком случае будет равен (см. пункт 4^0 предыдущего параграфа) $\overline{\alpha} = m(\mu_0;\, \sigma) = P_{\mu_0,\sigma}(T > C) = 1 - S_{n-1}(C)$, где $S_\nu(\cdot)$ – функция распределения Стьюдента с ν степенями свободы. Таким образом, мы получим свободный от неизвестного значения σ критерий $T > C(\alpha)$ требуемого размера α с критической константой $C(\alpha) = S^{-1}_{n-1}(1-\alpha)$. Это и есть то статистическое правило, которое обычно называется *критерием Стьюдента* или *t-критерием*.

Покажем теперь, что вероятность (функция мощности) $P_{\mu,\sigma}(T > C)$ монотонно возрастает с ростом μ при любых фиксированных значениях σ

и C. С этой целью представим статистику T в следующем виде:

$$T = \frac{\overline{X} - \mu}{S}\sqrt{n-1} + \frac{\mu - \mu_0}{\sigma} \cdot \frac{\sigma}{S}\sqrt{n-1}.$$

Если μ – среднее значение нормального распределения, из которого происходит выбор, то первое слагаемое в этом представлении есть стьюдентовская случайная величина с $n-1$ степенью свободы. Второе слагаемое есть произведение параметрической функции $\Delta(\mu) = (\mu - \mu_0)\sqrt{n-1}/\sigma$ на положительную случайную величину σ/S, распределение которой не зависит от μ и σ. При фиксированном σ функция $\Delta(\mu)$ возрастает с ростом μ и при этом все второе слагаемое возрастает, что влечет увеличение вероятности события перескока статистикой T порога C, то есть вероятности события $T > C$.

Итак, мы построили критерий проверки *односторонней* гипотезы $\mu \le \mu_0$ при односторонней альтернативе $\mu > \mu_0$. Функция мощности этого критерия зависит от μ и σ только через параметрическую функцию $\Delta = (\mu - \mu_0)\sqrt{n-1}/\sigma$, которая называется *параметром нецентральности*. Распределение статистики $T = (\overline{X} - \mu_0)\sqrt{n-1}/S$ при произвольных μ и σ, через которое выражается функция мощности критерия Стьюдента, называется *нецентральным* распределением Стьюдента с $n-1$ степенью свободы; таблицы этого распределения, зависящего от параметра нецентральности Δ, можно найти в ТМС.

Понятно, что построение критерия проверки простой гипотезы $\mu = \mu_0$ при *двусторонней* (сложной) альтернативе $\mu \ne \mu_0$ не вызывает принципиальных затруднений. Это критерий с критической областью $|\,T\,| > C$, где критическая константа $C = C(\alpha) = S^{-1}_{n-1}(1 - \alpha/2)$.

4^0. Сравнение средних значений двух нормальных распределений с общей неизвестной дисперсией (двухвыборочный критерий Стьюдента). Пусть X и Y – независимые случайные величины, причем $X \sim \mathcal{N}(\mu_1, \sigma^2)$, а $Y \sim \mathcal{N}(\mu_2, \sigma^2)$, так что $\mathbf{D}X = \mathbf{D}Y$. По двум независимым выборкам $X^{(n)} = (X_1, \ldots, X_n)$ и $Y^{(m)} = (Y_1, \ldots, Y_m)$ (возможно, разного объема) требуется проверить гипотезу *однородности* $H_0 :\ \mu_1 = \mu_2$ при альтернативе $H_1 :\ \mu_1 > \mu_2$. Типичный пример такой задачи – выявление эффекта нового метода лечения на группе из n пациентов посредством сравнения с контрольной группой из m пациентов, лечение которых проводится по старой методике.

Эта задача является для нас несколько новой, поскольку до сих пор мы

имели дело только с одной выборкой. Тем не менее, она сводится к той, что мы только что рассмотрели в 3^0, с помощью следующих построений.

Рассмотрим сначала разность выборочных средних $\overline{X} - \overline{Y}$. Эта статистика имеет нормальное распределение со средним $\mathbf{E}(\overline{X} - \overline{Y}) = \mu_1 - \mu_2$ и дисперсией $\mathbf{D}(\overline{X} - \overline{Y}) = \mathbf{D}\overline{X} + \mathbf{D}\overline{Y} = \sigma^2(n^{-1} + m^{-1})$. Следовательно, при справедливости нулевой гипотезы $\mu_1 = \mu_2$ случайная величина

$$\xi = \frac{\overline{X} - \overline{Y}}{\sigma}\sqrt{\frac{nm}{n+m}}$$

имеет стандартное нормальное распределение $\mathcal{N}(0, 1)$. Далее, нормированные выборочные дисперсии nS_X^2/σ^2 и mS_Y^2/σ^2 независимы и распределены по закону хи-квадрат с $n-1$ и $m-1$ степенями свободы соответственно. Так как для хи-квадрат распределения, как частного случая гамма-распределения, имеет место теорема сложения, то случайная величина $\eta = \left(nS_X^2 + mS_Y^2\right)/\sigma^2$ имеет хи-квадрат распределение с $n+m-2$ степенями свободы. Таким образом, мы приходим к двухвыборочной статистике Стьюдента

$$T_{n,m} = \frac{\xi}{\sqrt{\eta/(n+m-2)}} = \frac{\overline{X} - \overline{Y}}{\sqrt{nS_X^2 + mS_Y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}},$$

распределение которой при справедливости нулевой гипотезы есть распределение Стьюдента с $n+m-2$ степенями свободы.

Как и в случае одновыборочного критерия Стьюдента в 3^0 нетрудно показать, что при любых фиксированных C и σ функция мощности двухвыборочного критерия Стьюдента $T_{n,m} > C$ есть монотонно возрастающая функция параметра нецентральности $\Delta = (\mu_1 - \mu_2)\sqrt{n+m-2}/\sigma$, так что критическая константа C определяется по заданному уровню значимости из уравнения $P(T_{n,m} > C) = 1 - S_{n+m-2}(C) = \alpha$ и равна квантили распределения Стьюдента: $C(\alpha) = S^{-1}_{n+m-2}(1-\alpha)$. Понятно, что при альтернативе $H_1:\ \mu_1 \neq \mu_2$ критическая константа $C(\alpha) = S^{-1}_{n+m-2}(1-\alpha/2)$.

При использовании этого критерия следует обратить особое внимание на предположение о равенстве дисперсий наблюдаемых случайных величин: $\sigma_X^2 = \sigma_Y^2$. Задача сравнения средних двух нормальных распределений с неравными дисперсиями и с гарантированным ограничением α на вероятность ошибки первого рода называется *проблемой Беренса–Фишера*. Известно лишь асимптотическое решение этой проблемы при больших n и m.

5^0. Сравнение дисперсий двух нормальных распределений при неизвестных средних (критерий Фишера). Независимые выборки $X^{(n)}$ и $Y^{(m)}$ берутся из соответствующих нормальных распределений $\mathcal{N}(\mu_1,\, \sigma_1^2)$ и $\mathcal{N}(\mu_2,\, \sigma_2^2)$, относительно параметров которого проверяется гипотеза $\sigma_1^2 = \sigma_2^2$ при альтернативе $\sigma_1^2 > \sigma_2^2$ с мешающими параметрами μ_1 и μ_2.

В этой задаче естественно рассмотреть критерий, основанный на статистике $F = nS_X^2/mS_Y^2$, которая распределена как

$$\frac{\chi_{n-1}^2}{\chi_{m-1}^2} \cdot \frac{\sigma_1^2}{\sigma_2^2}.$$

Функция мощности критерия $F > C$ (который называется *критерием Фишера* или *F-критерием*)

$$m\left(\frac{\sigma_1^2}{\sigma_2^2}\right) = P_{\mu_1,\mu_2,\sigma_1,\sigma_2}(F > C) = P\left(\frac{\chi_{n-1}^2}{\chi_{m-1}^2} > C \cdot \frac{\sigma_2^2}{\sigma_1^2}\right)$$

есть монотонно возрастающая функция отношения дисперсий σ_1^2/σ_2^2. Для ее вычисления необходимо знать распределение отношения двух независимых случайных величин, распределенных по закону хи-квадрат с $n-1$ и $m-1$ степенями свободы. Это так называемое *распределение Фишера* $F_{n-1,m-1}$, плотность которого

$$f_{n-1,m-1}(x) = \frac{\Gamma\left(\frac{n+m-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{m-1}{2}\right)} \cdot \frac{x^{\frac{n-1}{2}-1}}{(x+1)^{\frac{n+m-2}{2}}},\ x > 0,$$

вычисляется столь же просто, как это мы делали при выводе распределения Стьюдента. Таблицы распределения Фишера можно найти в ТМС. Критическая константа C критерия Фишера заданного размера α определяется как квантиль этого распределения: $C(\alpha) = F_{n-1,m-1}^{-1}(1-\alpha)$.

Мы завершим иллюстрацию методов построения критериев с помощью состоятельных оценок тестируемого параметра примером, в котором не всегда размер критерия совпадает с заданным уровнем значимости.

6^0. Проверка гипотезы о вероятности успеха в испытаниях Бернулли. Рассмотрим задачу проверки гипотезы $p = p_0$ против альтернативы $p < p_0$ о вероятности p успешного исхода в испытаниях Бернулли. Пример такой задачи – проверка гипотезы о вероятности наследования доминантного признака в опытах Менделя, когда альтернативная модель

предписывает этой вероятности меньшее значение. Предлагаемый ниже метод решения позволяет строить критерии проверки такой гипотезы при альтернативах $p > p_0$ или $p \neq p_0$ посредством простой замены неравенства, определяющего критическую область, на обратное или двустороннее.

Итак, если мы располагаем выборкой $X^{(n)}$ из двухточечного распределения $\mathrm{B}(1,\, p)$, то относительная частота успешных испытаний (выборочное среднее) $\overline{X}$ является несмещенной оценкой p с минимальной дисперсией. В соответствии с предложенной выше идеологией проверки гипотез с помощью оценок тестируемого параметра мы должны отвергать гипотезу $p = p_0$ в пользу $p < p_0$, если $\overline{X} - p_0 < C$. Поскольку статистика $T = n\overline{X} = \sum_1^n X_k$ имеет биномиальное распределение $\mathrm{B}(n,\, p)$, а значение p_0 задано, то для вычисления функции мощности удобнее записать критическую область в виде $T < C$. Но статистика T принимает только целочисленные значения $0, 1, \ldots, n$, поэтому бессмысленно рассматривать дробные значения критических констант. Таким образом, мы приходим к наиболее удобной форме записи критической области в виде $T < C$, где C принимает значения $1, 2, \ldots, n$

Функция мощности такого критерия

$$m(p) = P_p\,(T < C) = \sum_{k=0}^{C-1} p^k (1-p)^{n-k},$$

и поскольку проверяется простая гипотеза, то критическая константа C должна определяться по заданному уровню значимости α из неравенства

$$m(p_0) = \sum_{k=0}^{C-1} p_0^k (1-p_0)^{\,n-k} \leq \alpha. \tag{2}$$

Очевидно, что чем больше C, тем больше мощность критерия, и поэтому $C(\alpha)$ следует выбирать как наибольшее целое число, удовлетворяющее неравенству (2). Размер критерия с таким $C(\alpha)$ не обязательно равен α, так что мы можем получить критерий уровня α, но не размера α (в предыдущих примерах с тестовыми статистиками, имеющими распределение непрерывного типа, мы имели критерии размера α). Более того, если p_0 настолько мало, что $(1-p_0)^n > \alpha$, то не существует таких C, при которых имеет место неравенство (2). В таком случае мы должны принимать нулевую гипотезу при любом результате статистического эксперимента, обеспечивая тем самым нулевой размер такого критерия "уровня α".

При больших объемах выборки n можно использовать нормальные аппроксимации биномиального распределения, получая таким образом критерий, размер которого асимптотически ($n \to \infty$) равен α. Статистика T асимптотически нормальна со средним np и дисперсией $np(1-p)$, поэтому неравенство (2) для определения критической константы имеет асимптотический аналог

$$\Phi\left(\frac{C - np_0}{\sqrt{np_0(1-p_0)}}\right) \leq \alpha,$$

откуда $C(\alpha) \approx np_0 - \Phi^{-1}(1-\alpha)/\sqrt{np_0(1-p_0)}$. Легко понять, что такой метод построения критериев асимптотического уровня α применим для любой критической области, в задании которой используется асимптотически нормальная оценка тестируемого параметра (см. пояснения в предыдущем параграфе перед пунктом 5^0).

Этот пример показывает, что в случае дискретных распределений задача построения равномерно наиболее мощных критериев значительно усложняется, поскольку один из двух критериев одного и того же уровня α может иметь большую мощность только потому, что он имеет больший размер. Мы столкнемся с этой проблемой в следующем параграфе, но следует заметить, что современная теория наиболее мощных критериев обходит этот неприятный момент за счет расширения понятия статистического правила, вводя так называемые *рандомизированные* критерии. К сожалению, я не располагаю временем познакомить вас с этим замечательным объектом теории статистического вывода.

Мы закончим этот параграф, как и было обещано, формулировкой *принципа двойственности* между задачами проверки гипотез и доверительного оценивания. Пусть $A(\theta_0) \subset \mathcal{X}^n$ – область принятия некоторого критерия уровня α, тестирующего гипотезу $H_0 : \ \theta = \theta_0$, и пусть $A(\theta_0)$ определена при любом $\theta_0 \in \Theta$. Для каждого результата $x^{(n)}$ наблюдения случайной выборки $X^{(n)}$ введем подмножество $\Delta(x^{(n)})$ параметрического пространства Θ, положив $\Delta(x^{(n)}) = \{\theta : \ x^{(n)} \in A(\theta)\}$. *Тогда* $\Delta(X^{(n)})$ *есть* $(1-\alpha)$-*доверительное множество для параметра* θ, поскольку $P_\theta\left(\Delta(X^{(n)}) \ni \theta\right) = P_\theta\left(X^{(n)} \in A(\theta)\right) \geq 1 - \alpha$.

Например, критерий Стьюдента проверки гипотезы $\mu = \mu_0$ о среднем значении нормального $(\mu,\, \sigma^2)$ распределения с неизвестной дисперсией σ^2

имеет область принятия (см. п. 3^0 данного параграфа)

$$A(\mu_0) = \left\{ X^{(n)} : \ \frac{|\,\overline{X} - \mu_0\,|}{S}\sqrt{n-1} \ \le \ S_{n-1}^{-1}(1-\alpha/2) \right\}.$$

Подставим в это неравенство вместо фиксированного μ_0 параметр μ и разрешим неравенство относительно μ. В результате получим доверительное утверждение (см. п. 4^0 предыдущего параграфа)

$$\overline{X} - St_\alpha/\sqrt{n-1} \ \le \ \mu \ \le \ \overline{X} + St_\alpha/\sqrt{n-1},$$

в котором $t_\alpha = S_{n-1}^{-1}(1-\alpha/2)$.

Вы сами можете сопоставить доверительные интервалы, построенные в §6, с критериями из §7. При этом сопоставлении можно вывести полезное правило, касающееся доверительной оценки скалярного параметра θ. Если имеется состоятельный критерий проверки гипотезы $\theta = \theta_0$ при двусторонней альтернативе $\theta \ne \theta_0$, то его области принятия соответствует двусторонний доверительный интервал. Если же альтернативная гипотеза носит односторонний характер, то при альтернативе $\theta < \theta_0$ мы получаем верхнюю доверительную границу, а при $\theta > \theta_0$ – нижнюю.

Естественно, принцип двойственности применим и к доверительным интервалам, как статистическим правилам проверки гипотез: *гипотеза* $\theta \in \Theta_0$ *отвергается тогда и только тогда, когда* $(1-\alpha)$*-доверительная область принадлежит подмножеству* Θ_1, и такое статистическое правило (критерий) гарантирует заданное ограничение α на вероятность ошибки первого рода.

§8. Равномерно наиболее мощные критерии

Лекция 13

Метод построения критериев заданного уровня α, который равномерно по всем альтернативным значениям параметра θ максимизирует мощность критерия, существенно опирается на следующее, почти очевидное утверждение, которое в теории проверки гипотез обычно называется *леммой Неймана–Пирсона.*

Рассмотрим вероятностную модель, состоящую всего из двух распределений P_0 и P_1, с общим носителем $\mathcal{X}$ и функциями плотности $f_0(x)$ и $f_1(x),\ x \in \mathcal{X}$. По выборке $X^{(n)}$ проверяется простая гипотеза H_0 : выборка взята из распределения P_0 при простой альтернативе H_1 : выборке соответствует распределение P_1. Определим критическую функцию $\varphi^*(X^{(n)})$ как индикаторную функцию критической области

$$L(X^{(n)}) = \prod_{k=1}^{n} \frac{f_1(X_k)}{f_0(X_k)} > C.$$

Статистика L называется *статистикой отношения правдоподобия*, а критерий φ^* – *критерием отношения правдоподобия* или *критерием Неймана–Пирсона.* Критерий φ^* отвергает нулевую гипотезу, если правдоподобие альтернативы $f_{1,n}(X^{(n)}) = \prod_1^n f_1(X_k)$ в C раз превосходит правдоподобие нулевой гипотезы $f_{0,n}(X^{(n)}) = \prod_1^n f_0(X_k)$. Этот критерий обладает следующим замечательным свойством.

Теорема 8.1. *Критерий отношения правдоподобия φ^* является наиболее мощным критерием в классе всех критериев проверки простой гипотезы при простой альтернативе, размер которых не превосходит размера критерия φ^*. Если критерий φ^* имеет размер α, то он обладает наибольшей мощностью в классе всех критериев уровня α.*

Доказательство.. Пусть $\varphi = \varphi(X^{(n)})$ – любой другой критерий, размер которого

$$\mathbf{E}_0\, \varphi(X^{(n)}) \le \mathbf{E}_0\, \varphi^*(X^{(n)}). \tag{1}$$

Требуется показать, что тогда критерий φ^* имеет большую мощность, чем критерий φ, то есть $\mathbf{E}_1\, \varphi^*(X^{(n)}) \ge \mathbf{E}_1\, \varphi(X^{(n)})$.

Рассмотрим интеграл

$$\int\limits_{\mathcal{X}^n} \Big[\, \varphi^*(x^{(n)}) - \varphi(x^{(n)}) \,\Big] \Big[\, f_{1,n}(x^{(n)}) - C f_{0,n}(x^{(n)}) \,\Big]\, d\mu_n(x^{(n)}) =$$

$$\mathbf{E}_1\, \varphi^*(X^{(n)}) - \mathbf{E}_1\, \varphi(X^{(n)}) - C \Big[\, \mathbf{E}_0\, \varphi^*(X^{(n)}) - \mathbf{E}_0\, \varphi(X^{(n)}) \,\Big].$$

Достаточно показать, что этот интеграл неотрицателен, и тогда первое утверждение теоремы будет следовать из неравенства:

$$\mathbf{E}_1\, \varphi^*(X^{(n)}) - \mathbf{E}_1\, \varphi(X^{(n)}) - C \Big[\, \mathbf{E}_0\, \varphi^*(X^{(n)}) - \mathbf{E}_0\, \varphi(X^{(n)}) \,\Big] \;\geq\; 0$$

которое влечет (см. (1))

$$\mathbf{E}_1\, \varphi^*(X^{(n)}) - \mathbf{E}_1\, \varphi(X^{(n)}) \;\geq\; C \Big[\, \mathbf{E}_0\, \varphi^*(X^{(n)}) - \mathbf{E}_0\, \varphi(X^{(n)}) \,\Big] \;\geq\; 0.$$

Покажем, что функции $\varphi^*(x^{(n)}) - \varphi(x^{(n)})$ и $f_{1,n}(x^{(n)}) - C f_{0,n}(x^{(n)})$, произведение которых интегрируется, одновременно положительны или отрицательны при любых $x^{(n)} \in \mathcal{X}^n$. Действительно, если $\varphi^*(x^{(n)}) - \varphi(x^{(n)}) > 0$, то это влечет $\varphi^*(x^{(n)}) = 1$, поскольку критическая функция равна единице, если она не равна нулю. Но, по определению критерия отношения правдоподобия, равенство $\varphi^*(x^{(n)}) = 1$ возможно лишь в случае $f_{1,n}(x^{(n)}) - C f_{0,n}(x^{(n)}) > 0$. Точно также устанавливается, что неравенство $\varphi^*(x^{(n)}) - \varphi(x^{(n)}) < 0$ влечет $f_{1,n}(x^{(n)}) - C f_{0,n}(x^{(n)}) < 0$.

Итак, критерий φ^* наиболее мощен в классе всех критериев, размер которых не превосходит размера φ^*. Если же $\mathbf{E}_0\, \varphi^*(X^{(n)}) = \alpha$, то это утверждение, очевидно, влечет его наибольшую мощность в классе всех критериев уровня α.

Применение этой теоремы к построению равномерно наиболее мощных критериев мы проиллюстрируем на одном частном примере, из которого будет виден общий подход к данной задаче.

Пример 8.1. *Проверка надежности при показательном распределении долговечности.* В примере 3.3 мы рассматривали проблему оценки надежности изделия с показательным распределением долговечности. Напомним, случайная величина X, реализация x которой соответствует промежутку времени от начала работы до момента отказа некоторого изделия, называется долговечностью, и по функции распределения $F(x)$, $x \geq 0$ случайной

величины X можно рассчитать надежность $H(t)$ изделий, соответствующую гарантийному времени t : $H(t) = P(X \geq t) = 1 - F(t)$.

Пусть долговечность X распределена по показательному закону с функцией распределения $F(x \,|\, \theta) = 1 - \exp\{-x/\theta\}$, значение параметра θ которой не известно. Мы должны удостовериться, что надежность выпускаемых изделий достаточно высока: $H(t) \geq P_0$, где P_0 –наименьшая допустимая доля изделий, которые должны прослужить гарантийный срок t.

Это типичная задача проверки гипотез, решение которой начинается с определения нулевой гипотезы H_0. При этом следует помнить, что в статистическом критерии контролируется вероятность отклонения H_0, когда она в действительности верна. В нашей конкретной проблеме спецификация нулевой гипотезы во многом зависит от того, что повлечет за собой отказ изделия. Если мы выпускаем бытовые приборы, то отказ изделия до гарантийного скрока t повлечет издержки на ремонт, которые могут быть незначительными по сравнению со стоимостью изделия. В таком случае естественно выбрать в качестве нулевой гипотезы утверждение о надежности изделий – отклонив эту гипотезу, когда она верна, мы потеряем дорогостоящую продукцию, ремонт которой нам обошелся бы значительно дешевле, чем ее уничтожение или продажа по бросовой цене. Если же отказ изделия приводит к катастрофическим последствиям, например, к гибели людей, то здесь рассуждать нечего, и за нулевую гипотезу следует брать утверждение о "ненадежности" . Отклонив такую гипотезу, когда она в действительности верна, мы столкнемся с неприемлемо большой долей отказов до истечения гарантийного срока, и поэтому риск от принятия "плохих" изделий должен быть контролируем. Остановимся на этом варианте и приступим к построению равномерно наиболее мощного критерия проверки гипотезы "ненадежности" $H_0 : \ H(t) < P_0$ при альтернативе $H_1 : \ H(t) \geq P_0$, когда $H(t) = \exp\{-t/\theta\}$.

В терминах значений параметра θ нулевая гипотеза принимает вид $H_0 : \theta < \theta_0 = -t/\ln\alpha$. Зафиксируем некоторое альтернативное значение $\theta_1 > \theta_0$, и рассмотрим задачу проверки простой гипотезы $H_0' : \ \theta = \theta_0$ при простой альтернативе $H_1' : \ \theta = \theta_1$. Наиболее мощный критерий проверки простой гипотезы при простой альтернативе имеет критическую область вида (см. теорему 8.1)

$$L(X^{(n)}) = \prod_{k=1}^{n} \frac{f_1(X_k)}{f_0(X_k)} = \frac{\theta_0}{\theta_1} \exp\left\{ \left(\frac{1}{\theta_0} - \frac{1}{\theta_1} \right) \sum\nolimits_1^n X_k \right\} > C,$$

где критическая константа C определяется по заданному уровню значимости α из условия $P_{\theta_0}\left(L(X^{(n)}) > C\right) \le \alpha$. Поскольку статистика $T_n = \sum_1^n X_k$ имеет гамма-распределение $\mathrm{G}(n,\, \theta_0)$, то для определения C в последнем неравенстве следует положить знак равенства. Кроме этого, статистика отношения правдоподобия $L(X^{(n)})$ есть монотонная функция статистики T_n, поэтому критическую область $L(X^{(n)}) > C$ можно записать в эквивалентной форме $T_n > C$ и находить новое C из равенства $P_{\theta_0}(T_n > C) = 1 - G_n(C/\theta_0) = \alpha$ (собственно говоря, нам все равно, какое C определять, но на практике, вне сомнения, удобнее иметь дело с критической областью $T_n > C$).

Итак, $C(\alpha) = \theta_0 \cdot G_n^{-1}(1-\alpha)$, где $G_n^{-1}(\cdot)$ – квантиль стандартного гамма-распределения $\mathrm{G}(n,\, 1)$, и критерий $\varphi^*(X^{(n)}) = \mathrm{I}_{\{T_n > C(\alpha)\}}(X^{(n)})$ заданного размера α является наиболее мощным в классе всех критериев уровня α, проверяющих гипотезу H_0' при альтернативе H_1'. Это означает, что для любого другого критерия φ с $\mathbf{E}_{\theta_0}\,\varphi(X^{(n)}) \le \alpha$ выполняется неравенство

$$\mathbf{E}_{\theta_1}\,\varphi(X^{(n)}) \le \mathbf{E}_{\theta_1}\,\varphi^*(X^{(n)}). \tag{2}$$

Но критерий φ^* не зависит от выбора альтернативного значения θ_1 параметра θ – критическая константа $C(\alpha) = \theta_0 \cdot G_n^{-1}(1-\alpha)$! Следовательно, неравенство (2) справедливо при любых $\theta_1 > \theta_0$, и мы приходим к заключению, что *критерий φ^* есть равномерно наиболее мощный критерий в классе всех критериев уровня α, проверяющих простую гипотезу $H_0' :\ \theta = \theta_0$ при сложной альтернативе $H_1 :\ \theta > \theta_0$.*

Далее, функция мощности критерия φ^*, как критерия различения исходных сложных гипотез $H_0 :\ \theta < \theta_0$ и $H_1 :\ \theta \ge \theta_0$, равна $m(\theta) = \mathbf{E}_\theta\,\varphi^*(X^{(n)}) = P_\theta(T_n > C(\alpha)) = 1 - G_n(G_n^{-1}(1-\alpha)\theta_0/\theta),\ \theta > 0$. Это – возрастающая функция θ, поэтому максимум вероятности ошибки первого рода (размер критерия) равен $m(\theta_0) = 1 - G_n\left(G_n^{-1}(1-\alpha)\right) = \alpha$. Таким образом, *критерий φ^* есть критерий размера α проверки гипотезы H_0 при альтернативе H_1, обладающий равномерно наибольшей мощностью в классе всех критериев φ с ограничением $\mathbf{E}_{\theta_0}\,\varphi(X^{(n)}) = \alpha$*. Но в таком случае он будет равномерно наиболее мощным и в более узком классе критериев уровня α, то есть критериев φ, удовлетворяющих ограничению $\mathbf{E}_\theta\,\varphi(X^{(n)}) \le \alpha$ при любом $\theta < \theta_0$.

Более того, нетрудно убедиться, что критерий φ^* обладает минимальной вероятностью ошибки первого рода $\alpha(\theta) = m(\theta),\ \theta \le \theta_0$ в классе всех

критериев уровня α. Для этого достаточно поменять местами нулевую гипотезу и альтернативу и выбрать уровень значимости, равный $1-\alpha$.

В этом примере построение равномерно наиболее мощного критерия стало возможным благодаря особому свойству статистической структуры показательного распределения: *статистика $L(X^{(n)})$ отношения правдоподобия есть монотонная функция статистики $T_n = \sum_1^n X_k$*. Это – частный случай статистических структур, обладающих достаточной статистикой T, ибо в силу теоремы факторизации у таких структур $L(X^{(n)}) = g_{\theta_1}(T)/g_{\theta_0}(T)$ зависит от $X^{(n)}$ только через значения $T(X^{(n)})$. Дополнительное свойство монотонности отношения правдоподобия относительно T обеспечивает существование и возможность конструктивного построения равномерно наиболее мощного критерия, причем критическая область такого критерия обязательно имеет вид $T > C$ или $T < C$. Например, критерий $\sum_1^n X_k > C$ при соответствующем выборе C по заданному уровню значимости α будет равномерно наиболее мощным критерием в классе всех критериев уровня α проверки гипотезы $\theta < \theta_0$ при альтернативе $\theta \geq \theta_0$, когда θ есть среднее значение нормального распределения (дисперсия предполагается известной) или параметр масштаба гамма-распределения (параметр формы известен). Но если θ – параметр таких распределений, как двухточечное или Пуассона, то критерий φ^* с критической областью $\sum_1^n X_k > C$ обладает равномерно наибольшей мощностью только в классе тех критериев, размер которых не больше размера φ^*.

Другие критерии, которые мы рассматривали в предыдущем параграфе, также обладают свойством равномерной наибольшей мощности, и при доказательстве этого также используется лемма Неймана–Пирсона, но методика доказательства совершенно другая и требует разработки методов построения критериев, обладающих свойством инвариантности – независимости от мешающих параметров. Но это уже совсем другая область теории проверки гипотез, поговорить о которой у нас не хватает времени. Я лучше расскажу вам о некоторых дополнительных ухищрениях в практических применениях статистических критериев, которые позволяют с большей степенью наглядности оценить степень согласия проверяемой гипотезы с выборочными данными.

Все рассматриваемые нами критерии заданного уровня α обладают тем свойством, что их критические области можно записать в виде $T(X^{(n)}) > C(\alpha)$, где T – некоторая статистика, характеризующая расхождение вы-

борочных данных с предполагаемыми значениями параметра. Увеличение уровня значимости α приводит к уменьшению $C(\alpha)$, и мы получаем систему вложенных друг в друга критических областей. Это замечательное свойство наших критериев позволяет несколько изменить методологию их практического использования. До сих пор мы фиксировали уровень значимости α, находили по нему критическую константу $C(\alpha)$ и сравнивали ее с выборочным значением $t = T(x^{(n)})$ статистики $T = T(X^{(n)})$. Поступим теперь следующим образом. Получив выборочные данные $x^{(n)}$, вычислим значение $t = T(x^{(n)})$ и рассмотрим критерий $T(X^{(n)}) > t$. Размер такого критерия $\alpha_{\text{кр.}} = P_0\left(T(X^{(n)}) > t\right)$ называется *критическим уровнем значимости*, который трактуется как вероятность получить столь же большие расхождения между выборочными данными и нулевой гипотезой, как и для выборочных данных $x^{(n)}$.

Естественно, мы по-прежнему можем работать с заданным уровнем значимости α, отклоняя нулевую гипотезу, если $\alpha_{\text{кр.}} < \alpha$, и принимая ее в противном случае. Кстати, принимая гипотезу, не следует утверждать, что она верна. На этот счет существует более деликатное выражение: "выборочные данные согласуются с выдвинутой гипотезой," ибо, как говорил один из создателей математической статистики сэр Д.Фишер, "гипотезы не проверяются, а разве лишь отвергаются" . Так вот, в свете этого высказывания более разумно просто сообщать полученный критический уровень значимости, сопровождая его следующим комментарием, который можно считать международным статистическим стандартом. Если $\alpha_{\text{кр.}} \leq 0.01$, то говорят, что расхождение между гипотезой и выборочными данными *высоко значимо*, если $0.01 < \alpha_{\text{кр.}} \leq 0.05$, то просто – *значимо*, если же $0.05 < \alpha_{\text{кр.}} \leq 0.10$ – *почти значимо*, и в случае $\alpha_{\text{кр.}} > 0.10$ – *не значимо*. Заметим также, что в некоторых применениях критериев значимости (особенно, в медицине) $\alpha_{\text{кр.}}$ называют *достоверностью*. Существуют и другие, совершенно фантастические названия $\alpha_{\text{кр.}}$, которые я не буду здесь приводить в силу их крайне неприличного звучания.

Поговорим теперь об оптимальных свойствах доверительных границ, соответствующих равномерно наиболее мощным критериям. Рассмотрим только случай верхней доверительной границы $\overline{\theta}_n = \overline{\theta}_n(X^{(n)})$.

Определение 8.1 Верхняя $(1-\alpha)$-доверительная граница $\overline{\theta}_n$ называется *равномерно наиболее точной*, если она равномерно по всем θ и θ', удовлетворяющим неравенству $\theta' > \theta$, минимизирует вероятность $P_\theta\left(\overline{\theta}_n(X^{(n)}) \geq \theta'\right)$.

Таким образом, в случае равномерно наиболее точной границы $\overline{\theta}_n$ интервал $(-\infty; \overline{\theta}_n\,]$ с заданной вероятностью $1-\alpha$ накрывает истинное значение параметра θ, но он с минимальной вероятностью накрывает любые значения θ, лежащие правее истинного.

Если мы проверяем гипотезу $H:\ \theta = \theta_0$ при альтернативе $K(\theta_0):\ \theta < \theta_0$, и область принятия $A(\theta_0)$ равномерно наиболее мощного критерия размера α обладает тем свойством, что подмножество $\Delta_n(x^{(n)}) = \{\theta:\ x^{(n)} \in A(\theta)\}$ параметрического пространства $\Theta \subseteq \mathrm{R}$ есть интервал $(-\infty:\ \overline{\theta}_n(x^{(n)})\,]$, то $\overline{\theta}_n(X^{(n)})$ есть равномерно наиболее точная верхняя $(1-\alpha)$-доверительная граница. Все объясняется довольно просто: вероятность $P_\theta\,(\overline{\theta}_n(X^{(n)}) \geq \theta')\ = P_\theta\,(X^{(n)} \in A(\theta'))$ есть вероятность ошибки второго рода у критерия проверки гипотезы $H:\ \theta = \theta'$ при альтернативе $K(\theta'):\ \theta < \theta'$. Равномерно наиболее мощный критерий естественно обладает равномерно минимальной вероятностью ошибки второго рода. Все построенные нами в §6 доверительные границы обладают оптимальными свойствами с точки зрения малой вероятности накрытия тех значений параметра, которые не соответствуют истине.

§9. Проверка модельных предположений. Критерии согласия

Лекция 14

Рассмотренные нами методы построения оптимальных решающих функций в проблемах оценки параметров и проверки параметрических гипотез существенно опирались на такие особые свойства вероятностных моделей, как существование достаточных статистик, монотонность отношения правдоподобия относительно некоторой статистики, независимость выборок и прочее. Оценить же последствия от использования конкретных решающих функций (найти функцию риска статистического правила) вообще не представляется возможным без знания вероятностной модели. Отсюда возникает необходимость разработки общих методов тестирования (проверки) предлагаемой вероятностной модели $\mathcal{P} = \{P_\theta,\, \theta \in \Theta\}$ по данным случайной выборки, или нескольких выборок, которые предположительно извлекаются из некоторых распределений семейства $\mathcal{P}$. Значимые расхождения между модельными и эмпирическими распределениями вынуждают статистика пересмотреть посылки, положенные в основу построения вероятностной модели, и тем самым избежать больших потерь от использования заведомо плохих решающих правил (точнее правил, которые оптимальны не для той модели).

Понятно, что речь идет о проверке статистических гипотез без особой спецификации альтернатив к нулевой гипотезе. Статистические правила проверки модельных предположений обычно называются *критериями согласия*, и в математической статистике сложился некоторый традиционный набор таких критериев, обладающих большой универсальностью. Это критерии, с помощью которых можно проверять не только принадлежность распределения наблюдаемой случайной величины к определенному семейству, но и тестировать некоторые более "грубые" черты модели, как то независимость компонент наблюдаемого случайного вектора (векторной случайной величины), возможность объединения нескольких выборок в одну (проверка гипотезы однородности выборок) и множество других предположений, касающихся структуры выборочных данных. Мы познакомимся в этом параграфе с набором универсальных статистических процедур, объединяемых общим названием *критерии хи-квадрат*. Об одном из них мы уже упоминали в §2 в связи с построением гистограммы выборки; это –

1^0. Критерий согласия хи-квадрат. Решается статистическая проблема проверки гипотезы о виде распределения наблюдаемой случайной величины X (возможно, векторной). Начнем с простейшего случая, когда построение вероятностной модели привело к полной спецификации распределения, то есть проблема состоит в проверке простой гипотезы H : распределение X на измеримом пространстве $(\mathcal{X}, \mathcal{A})$ ее значений есть $P(A)$, $A \in \mathcal{A}$.

Построение критерия согласия выборочных данных с распределением P начинается с разбиения пространства $\mathcal{X}$ на $r \geq 2$ частей $A_1, \ldots, A_r$; $\mathcal{X} = \sum_1^r A_i$. Рекомендации по выбору числа r и способу разбиения носят довольно расплывчатый характер, и если не уточнять возможные альтернативы к P, то, как вы сами понимаете, таких рекомендаций не может быть в принципе. Главное, разбиение не должно определяться выборочными значениями, надо стремится к областям одинаковой конфигурации и размера, не следует делать слишком подробное разбиение. Например, если $\mathcal{X} = \mathrm{R}$ (наблюдается действительная случайная величина), то прямая R разбивается на r интервалов вида $(-\infty, a]$, $(a, a+\Delta]$, $(a+\Delta, a+2\Delta]$, $\ldots$, $(a+(r-3)\Delta, a+(r-2)\Delta]$, $(a+(r-2)\Delta, +\infty)$, так что длина внутренних интервалов постоянна и равна Δ. Конечно выбор r зависит от объема выборки n, но даже при исключительно больших n не делается более 15-20 разбиений; этого вполне достаточно, чтобы в гистограмме отразить всю специфику формы тестируемого распределения.

После разбиения $\mathcal{X}$ проводится сортировка выборочных данных по областям разбиений и подсчитываются количества $\nu_1, \ldots, \nu_r$, $\sum_1^r \nu_i = n$, данных, попавших в соответствующие области $A_1, \ldots A_r$. Вычисляются "теоретические" вероятности $p_i = P(A_i)$, $i = 1, \ldots, r$ попадания выборочных данных в эти области и вычисляется значение x^2 тестовой статистики

$$X^2 = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}.$$

Гипотеза H отвергается, если $x^2 > C$, где критическая константа C выбирается по заданному уровню значимости α как наименьшее число, удовлетворяющее неравенству $P(X^2 > C) \leq \alpha$. Естественно, на практике используют критический уровень значимости $\alpha_{\text{кр.}} = P(X^2 > x^2)$, сопровождая его комментариями типа тех, которые были приведены в предыдущем параграфе после введения понятия критического уровня значимости. Однако точное распределение статистики X^2 найти в явном виде не представляется

возможным; предельное распределение X^2 при $n \to \infty$ установил К.Пирсон в самом начале XX века.

Теорема 9.1. *Если число разбиений $r \geq 2$ фиксировано, а объем выборки $n \to \infty$, то распределение X^2 сходится к распределению хи-квадрат с $r-1$ степенью свободы.*

Доказательство.. Очевидно, для вывода предельного распределения X^2 следует в первую очередь обратиться к совместному распределению частот $\nu_1, \ldots, \nu_r$, $\sum_1^n \nu_i = n$. Это мультиномиальное распределение $\mathcal{M}(r, n, p)$, (см. §9 курса ТВ) с функцией плотности

$$f(x_1, \ldots, x_r) = P(\nu_1 = x_1, \ldots, \nu_r = x_r) = \frac{n!}{x_1! \cdots x_r!}\, p_1^{x_1} \cdots p_r^{x_r},$$

сосредоточенное на целочисленной решетке $\sum_1^n x_i = n$. Теорема 9.1 из курса ТВ утверждает, что совместное распределение первых $r-1$ частот $\nu_1, \ldots, \nu_{r-1}$ аппроксимируется $r-1$-мерным нормальным распределением. Естественно, предельное распределение всего вектора частот $\nu_1, \ldots, \nu_r$ при соответствующей нормировке на их средние значения и стандартные отклонения будет вырожденным, ибо $\sum_1^n \nu_i = n$. Вырожденные распределения лучше всего исследовать с помощью характеристических функций, ибо такие распределения можно записать в явном виде, только переходя к системе координат на той гиперповерхности, где сосредоточено такое распределение, и это чрезвычайно усложняет технику асимптотического анализа распределений. Итак, найдем совместную характеристическую функцию $\nu_1, \ldots, \nu_r$.

Вспомним схему мультиномиальных испытаний. Мы наблюдаем выборку $Y_1, \ldots, Y_n$ из распределения случайного вектора $Y = (X_1, \ldots, X_r)$, все компоненты которого, за исключением одной (скажем, X_j), могут принимать только нулевые значения, в то время как $X_j = 1$. Каждая компонента Y_i выборки $Y^{(n)} = (Y_1, \ldots, Y_n)$ есть независимая копия Y, так что $Y_i = (X_{1i}, \ldots, X_{ri})$ и X_{ji} – копия (в смысле одинаковости распределения) X_j, $j = 1, \ldots, r$, $i = 1, \ldots, n$. В таких обозначениях

$$\nu_j = \sum_{i=1}^n X_{ji},\ j = 1, \ldots, r.$$

Если мы найдем характеристическую функцию $\varphi_Y(\mathbf{t})$, $\mathbf{t} = (t_1, \ldots, t_r)$, наблюдаемого вектора Y, то характеристическая функция $\varphi_\nu(\mathbf{t})$ вектора

частот $\nu = (\nu_1, \ldots, \nu_r)$ будет вычисляться по формуле $\varphi_\nu(\mathbf{t}) = \varphi_Y^n(\mathbf{t})$, ибо характеристическая функция суммы независимых случайных величин равна произведению характеристических функций слагаемых (пункт 3^0 теоремы 12.1 курса ТВ). Но характеристическая функция вектора Y (напомним, $\sum_1^r X_j = 1$)

$$\varphi_Y(\mathbf{t}) = \mathbf{E} \exp\left\{ \mathbf{i} \sum_1^r t_j X_j \right\} = \sum_1^r p_j \mathrm{e}^{\mathbf{i} t_j},$$

и поэтому

$$\varphi_\nu(\mathbf{t}) = \left(\sum_1^r p_j \mathrm{e}^{\mathbf{i} t_j} \right)^n .$$

Теперь перейдем к асимптотическому анализу характеристической функции вектора X нормированных частот

$$X_j = \frac{\nu_j - np_j}{\sqrt{np_j}}, \ j = 1, \ldots, r,$$

сумма квадратов компонент которого составляет тестовую статистику X^2 (извините, что использую букву X в новом смысле, но не хочется вводить для обозначения случайных величин новые символы). Характеристическая функция случайного вектора, компоненты которого подвергнуты линейному преобразованию, вычисляется по формуле, аналогичной пункту 2^0 теоремы 12.1:

$$\varphi_X(\mathbf{t}) = \exp\left\{ -\mathbf{i} \sum_1^r t_j \sqrt{np_j} \right\} \left(\sum_1^r p_j \, \exp\left\{ \frac{\mathbf{i}\, t_j}{\sqrt{np_j}} \right\} \right)^n .$$

Разложим логарифм этой функции в ряд Маклорена по степеням $t_1, \ldots, t_r$, как это делалось при доказательстве центральной предельной теоремы:

$$\ln \varphi_X(\mathbf{t}) = -\mathbf{i}\sqrt{n} \sum_1^r t_j \sqrt{p_j} +$$

$$n \ln \left[1 + \frac{\mathbf{i}}{\sqrt{n}} \sum_1^r t_j \sqrt{p_j} - \frac{1}{2n} \sum_1^r t_j^2 + O(n^{-3/2}) \right] =$$

$$= -\frac{1}{2} \sum_1^r t_j^2 + \frac{1}{2} \left(\sum_1^r t_j \sqrt{p_j} \right)^2 + O(n^{-1/2}).$$

Таким образом, характеристическая функция предельного распределения вектора X нормированных частот есть

$$\lim_{n\to\infty}\varphi_X(\mathbf{t})=\exp\left\{-\frac{1}{2}\left[\sum_1^r t_j^2-\left(\sum_1^r t_j\sqrt{p_j}\right)^2\right]\right\}.$$

Это – характеристическая функция r-мерного нормального распределения с нулевыми средними и матрицей ковариаций $\Lambda=\mathrm{I}-\mathbf{pp}'$, где I – единичная матрица, а $\mathbf{p}=(\sqrt{p_1},\ldots,\sqrt{p_r})$ – вектор столбец.

Рассмотрим квадратичную форму

$$Q(\mathbf{t})=\sum_1^r t_j^2-\left(\sum_1^r t_j\sqrt{p_j}\right)^2,$$

коэффициенты которой определяют ковариации компонент вектора $Z=(Z_1,\ldots,Z_r)$, распределенного по нормальному закону. Если произвести ортогональное преобразование $\mathbf{A}$ вектора $\mathbf{t}$, полагая $\mathbf{u}=\mathbf{At}$ и фиксируя последнюю строку матрицы $\mathbf{A}$ таким образом, чтобы в новом векторе $\mathbf{u}=(u_1,\ldots,u_r)$ компонента $u_r=\sum_1^r t_j\sqrt{p_j}$, то мы получим квадратичную форму (вспомните аналогичные ортогональные преобразования нормального вектора при выводе распределения выборочной дисперсии в лемме Фишера)

$$Q(\mathbf{t})=\sum_1^r t_j^2-\left(\sum_1^r t_j\sqrt{p_j}\right)^2=\sum_1^r u_j^2-u_r^2=\sum_1^{r-1}u_j^2.$$

Таким образом, существует ортогональное преобразование $Y=\mathbf{B}Z$ вектора Z, после которого $Y_1,\ldots,Y_{r-1}$ независимы и одинаково нормально распределены со средними, равными нулю, и единичными дисперсиями, а Y_r имеет нулевое среднее и нулевую дисперсию, то есть $Y_r=0$ почти наверное. Все это, конечно, следствие вырожденности нормального распределения вектора Z – оно сосредоточено на гиперплоскости $\sum_1^r Z_j\sqrt{p_j}=0$.

Изучим теперь предельное распределение статистики $X^2=\sum_1^r X_j^2$. Поскольку предельное распределение вектора X совпадает с распределением вектора Z, то предельное распределение статистики X^2 определяется распределением квадратичной формы $\sum_1^r Z_j^2$. Как известно, ортогональные преобразования не меняют суммы квадратов, поэтому $\sum_1^r Z_j^2=\sum_1^r Y_j^2=$

$\sum_1^{r-1} Y_j^2$. Следовательно, предельное распределение статистики X^2 есть распределение суммы квадратов $r-1$ независимых случайных величин, имеющих общее стандартное нормальное распределение. По определению это – хи-квадрат распределение с $r-1$ степенями свободы. Теорема Пирсона доказана.

Лекция 15

Рассмотрим теперь более сложную статистическую проблему, в которой проверяется гипотеза о принадлежности распределения P наблюдаемой случайной величины некоторому параметрическому семейству $\mathcal{P} = \{P_\theta,\, \theta \in \Theta \subseteq \mathrm{R}^s\}$, индексированному s-мерным параметром $\theta = (\theta_1, \ldots, \theta_s)$. В таком случае

$$X^2(\theta) = \sum_{i=1}^{r} \frac{(\nu_i - np_i(\theta))^2}{np_i(\theta)}$$

не может называться статистикой и ее нельзя использовать для проверки сложной гипотезы $H: \; P \in \mathcal{P}$. Естественно воспользоваться какой-либо оценкой $\hat\theta_n = \hat\theta_n(X^{(n)})$ параметра θ и рассмотреть тестовую статистику

$$\hat X^2 = X^2(\hat\theta_n) = \sum_{i=1}^{r} \frac{(\nu_i - np_i(\hat\theta_n))^2}{np_i(\hat\theta_n)}.$$

Понятно, что распределение статистики $\hat X^2$ может зависеть от метода оценки параметра θ. Однако, если определить оценку $\hat\theta_n$ из условия минимума случайной функции $X^2(\theta)$, то, как показал Фишер, при определенных условиях регулярности, которым удовлетворяют все рассмотренные нами в курсе ТВ вероятностные модели, *предельное распределение статистики* $\hat X^2$ *есть хи-квадрат распределение с* $r-s-1$ *степенями свободы*. Если же $\hat\theta_n$ – оценка θ по методу максимального правдоподобия, то предельное распределение $\hat X^2$, также при условиях регулярности типа тех, что обеспечивали асимптотическую нормальность $\hat\theta_n$, имеет функцию распределения $K(x)$, для которой справедлива двусторонняя оценка

$$K_{r-1}(x) \le K(x) \le K_{r-s-1}(x),$$

при любом $x > 0$.

Доказательство этих утверждений достаточно громоздко и мы не будем им заниматься из-за недостатка времени. Идейная сторона проблемы

нам ясна, и коль скоро нам сообщили распределение тестовой статистики, то мы можем использовать его для расчета критического уровня значимости. В случае оценки максимального правдоподобия, когда мы располагаем двусторонней оценкой $\alpha_{\text{кр.}}$, рекомендуется при отклонении гипотезы ориентироваться на $\alpha_{\text{кр.}} = 1 - K_{r-1}(x^2)$ $\left(> 1 - K_{r-s-1}(x^2)\right)$, а в случае ее принятия – на $\alpha_{\text{кр.}} = 1 - K_{r-s-1}(x^2)$ $\left(< 1 - K_{r-1}(x^2)\right)$, чтобы уменьшить риск от принятия неправильного решения.

Критерий хи-квадрат является наиболее универсальным статистическим методом тестирования вероятностной модели, поскольку предельное распределение статистики не зависит от распределения наблюдаемой случайной величины даже в том случае, когда это распределение зависит от параметров, значение которых неизвестно. Критерий Колмогорова $\sqrt{n}D_n = \sqrt{n}\sup_x |F_n(x) - F(x)| > C$, о котором говорилось в начале §2, можно использовать только для проверки простой гипотезы $F(\cdot) = F_0(\cdot)$ о виде функции распределения. Если $F_0(x \mid \theta)$ зависит от параметра θ и в статистику $\sqrt{n}D_n$ вместо $F(x)$ подставляется $F_0\left(x \mid \hat{\theta}_n(X^{(n)})\right)$, то распределение модифицированной таким образом статистики $\sqrt{n}D_n$ зависит как от вида функции F_0, так и от параметра θ. Существует, правда, несколько случаев особой связи между x и θ в записи функции F_0, при наличии которой распределение тестовой статистики не зависит от θ. Это, например, такие функции распределения с параметрами масштаба и сдвига, как нормальное и показательное. Для тестирования таких распределений составляются специальные таблицы критических констант и критических уровней значимости. Следует заметить, что прямое использование критерия Колмогорова с оценками неизвестных значений параметров является наиболее распространенной ошибкой в практических приложениях методов тестирования вероятностных моделей.

Обратимся теперь к проверке гипотез, касающихся не столько вида распределения наблюдаемых случайных величин, сколько их особых свойств, наличие которых позволяет значительно упростить вероятностную модель и добиться ее более четкой спецификации.

2^0. Критерий независимости хи-квадрат (таблицы сопряженности признаков). Следующая задача выявления зависимости между определенными признаками наблюдаемых объектов часто возникает в практических приложениях математической статистики. Предположим, что мы случайно выбрали n особей из некоторой этнической популяции, и хо-

тим выяснить, существует ли зависимость между цветом волос и цветом глаз. Мы различаем $s \geq 2$ уровней первого признака (например, блондин, брюнет, шатен и рыжий) и $r \geq 2$ уровней второго (например, карие, серые, голубые и зеленые). Все n особей разбиваются на sr групп в соответствии с наличием тех или иных уровней каждого признака, и составляется следующая таблица частот особей в каждой группе.

Признаки	1	2	$\cdots$	s	Сумма
1	ν_{11}	ν_{12}	$\cdots$	ν_{1s}	$\nu_{1\cdot}$
2	ν_{21}	ν_{22}	$\cdots$	ν_{2s}	$\nu_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	ν_{r1}	ν_{r2}	$\cdots$	ν_{rs}	$\nu_{r\cdot}$
Сумма	$\nu_{\cdot 1}$	$\nu_{\cdot 2}$	$\cdots$	$\nu_{\cdot s}$	n

Такие таблицы, в которых суммы

$$\nu_{i\cdot} = \sum_{j=1}^{s} \nu_{ij}, \qquad \nu_{\cdot j} = \sum_{i=1}^{r} \nu_{ij},$$

называются *таблицами сопряженности признаков*. Требуется проверить нулевую гипотезу о том, что переменные признаки, по которым построена таблица, независимы. Построим вероятностную модель, соответствующую такого рода табличным данным и составим статистику X^2 для проверки гипотезы независимости.

Пусть p_{ij} – вероятность того, что случайно отобранная особь имеет i-ый уровень по первому признаку и j-ый – по второму, $i = 1, \ldots, r$, $j = 1, \ldots, s$. Гипотеза независимости означает, что $p_{ij} = p_{i\cdot}\, p_{\cdot j}$, где

$$p_{i\cdot} = \sum_{j=1}^{s} p_{ij}, \qquad p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$$

при любых $i = 1, \ldots, r$ и $j = 1, \ldots, s$. Для проверки гипотезы независимости предлагается использовать тестовую статистику

$$X^2 = \sum_{i,j} \frac{\left(\nu_{ij} - n\, p_{i\cdot}\, p_{\cdot j}\right)^2}{n\, p_{i\cdot}\, p_{\cdot j}}\,, \tag{1}$$

в которой суммирование распространяется на все rs групп таблицы сопряженности признаков. Понятно, что X^2 является тестовой статистикой

только в случае известных значений $r+s-2$ параметров $p_{i\cdot}$ и $p_{\cdot j}$, $i=1,\dots,r$, $j=1,\dots,s$ (напомним, $\sum_1^r p_{i\cdot}=\sum_1^s p_{\cdot j}=1$, так что с помощью этих соотношений два из $r+s$ параметров, например, $p_{r\cdot}$ и $p_{\cdot s}$, можно выразить через остальные $r+s-2$ параметров). В этом случае X^2 имеет в пределе ($n\to\infty$) хи-квадрат распределение с $rs-1$ степенями свободы.

Конечно, вся проблема состоит в том, что эти параметры неизвестны. Оказывается, оценки максимального правдоподобия

$$\hat p_{i\cdot}=\frac{\nu_{i\cdot}}{n},\ \hat p_{\cdot j}=\frac{\nu_{\cdot j}}{n},\quad i=1,\dots,r,\ j=1,\dots,s,$$

этих параметров асимптотически эквивалентны оценкам по методу минимума статистики X^2, и поэтому подстановка в правую часть (1) этих оценок приводит к статистике

$$\hat X^2=n\sum_{i,j}\frac{\left(\nu_{ij}-\nu_{i\cdot}\,\nu_{\cdot j}/n\right)^2}{\nu_{i\cdot}\,\nu_{\cdot j}}=n\left(\sum_{i,j}\frac{\nu_{ij}^2}{\nu_{i\cdot}\,\nu_{\cdot j}}-1\right),$$

предельное распределение которой есть хи-квадрат с $rs-(r+s-2)-1=(r-1)(s-1)$ степенями свободы.

Естественно, статистику $\hat X^2$ можно использовать для проверки независимости компонент двумерного вектора $(X,\,Y)$, и при этом таблица сопряженности представляет частотные данные для построения гистограммы двумерной выборки $(X_1,\,Y_1),\dots,(X_n\,Y_n)$. Соответствующим образом нормированная статистика X^2 может служить мерой зависимости признаков (или компонент X и Y случайного вектора).

3^0. Критерий однородности хи-квадрат. Анализируются данные $s\ge 2$ независимых мультиномиальных схем испытаний с одинаковым числом $r\ge 2$ возможных исходов и соответствующими объемами $n_1,\dots,n_s$ наблюдений в каждой схеме. Проверяется гипотеза *однородности*: все схемы испытаний имеют одинаковый вектор вероятностей $\mathbf{p}=(p_1,\dots,p_r)$, $\sum_1^r p_i=1$, появления соответствующих исходов, причем значения компонент вектора $\mathbf{p}$ не известны. Обозначая ν_{ij} частоту появления i-го исхода в j-ом испытании, представим данные наблюдений в виде таблицы, аналогичной таблице сопряженности признаков

исх. \ схем.	1	2	$\cdots$	s	Сумма
1	ν_{11}	ν_{12}	$\cdots$	ν_{1s}	$\nu_{1\cdot}$
2	ν_{21}	ν_{22}	$\cdots$	ν_{2s}	$\nu_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	ν_{r1}	ν_{r2}	$\cdots$	ν_{rs}	$\nu_{r\cdot}$
Сумма	n_1	n_2	$\cdots$	n_s	n

Составим сначала статистику хи-квадрат для случая известного вектора вероятностей $\mathbf{p}$:

$$X^2 = \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(\nu_{ij} - n_j p_i)^2}{n_j p_i}.$$

Внутренняя сумма

$$X_j^2 = \sum_{i=1}^{r} \frac{(\nu_{ij} - n_j p_i)^2}{n_j p_i}$$

представляет статистику хи-квадрат для j-ой схемы мультиномиальных испытаний, и поэтому имеет в пределе ($n_j \to \infty$) хи-квадрат распределение с $r-1$ степенями свободы. Статистика X^2 есть сумма s независимых статистик, каждая из которых имеет предельное хи-квадрат распределение, так что, в силу теоремы сложения, предельное распределение X^2 есть хи-квадрат распределение с $(r-1)s$ степенями свободы.

В случае неизвестных значений вероятностей исходов, которые при справедливости нулевой гипотезы одинаковы для всех схем испытаний, используем их оценки $\hat{p}_i = \nu_{i\cdot}/n,\ i = 1, \ldots, r$, (всего оценивается $r-1$ параметр). Подстановка этих оценок в X^2 дает статистику

$$\hat{X}^2 = n \sum_{i,j} \frac{(\nu_{ij} - n_j \nu_{i\cdot}/n)^2}{n_j \nu_{i\cdot}} = n \left(\sum_{i,j} \frac{\nu_{ij}^2}{n_j \nu_{i\cdot}} - 1 \right),$$

предельное распределение которой есть хи-квадрат с $(r-1)s - (r-1) = (r-1)(s-1)$ степенями свободы. Замечателен тот факт, что мы получили тестовую статистику такого же вида и с тем же предельным распределением, что и при проверке гипотезы независимости признаков.

Естественно, построенный критерий можно использовать для проверки гипотезы однородности распределений, из которых извлекаются $s \geq 2$ выборок. Выборочные данные при этом подвергаются группировке в соответствии с одинаковым для всех выборок разбиением пространства $\mathcal{X}$ на $r \geq 2$ областей.

Л И Т Е Р А Т У Р А

1. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики.– М.: Наука,1983.

2. Боровков А.А. Математическая статистика.– М.: Наука, 1984.

3. Боровков А.А. Теория вероятностей.– М.: Наука, 1986.

4. Козлов М.В., Прохоров А.В. Введение в математическую статистику.– М.: Изд-во МГУ, 1987.

5. Крамер Г. Математические методы статистики.– М.: Мир, 1975.

6. Чистяков В.П. Курс теории вероятностей.– М.: Наука, 1982.

7. Ширяев А.Н. Вероятность.– М.: Наука, 1980.