Оглавление

1. Параметрические гипотезы

Понятие статистической гипотезы

В самой общей постановке задача проверки статистических гипотез формулируется следующим образом. Пусть задан набор наблюдений $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $1 \leqslant n < \infty$. Предположим, что результат x, полученный в эксперименте, есть реализация n-мерной с.в. $\xi = (\xi_1, \ldots, \xi_n)$. Статистическая гипотеза – это предположение о виде распределения с.в. ξ. Решить задачу проверки статистической гипотезы означает сформулировать правило, которое по результату наблюдений x будет давать ответ на вопрос, верна гипотеза или нет.

С математической точки зрения решение задачи проверки статистической гипотезы сводится к разбиению множества R возможных реализаций выборки на два подмножества: те реализации, по которым мы будем делать вывод, что гипотеза верна, и те реализации, которые свидетельствуют против гипотезы. Понятно, что вследствие случайного характера наблюдений в общем случае любое решающее правило будет давать ошибочные решения, например, согласно правилу мы делаем вывод, что гипотеза верна, а на самом деле распределение с.в. ξ отличается от гипотетического. Таким образом, при построении решающего правила необходимо контролировать вероятности ошибочных решений.

Более простыми являются постановка и решение задач проверки параметрических гипотез. В таких задачах, как и при оценивании параметров, мы предполагаем, что распределение с.в. ξ задаётся известной априори функцией правдоподобия $L^{(n)}(x, \theta)$, $x \in \mathbb{R}^n$, $\theta \in \Theta$, с неизвестным параметром θ. В случаях, когда размерность выборки несущественна для наших рассуждений, мы пишем $L(x, \theta)$ вместо $L^{(n)}(x, \theta)$. Гипотеза заключается в том, что $\theta \in \Theta$ удовлетворяет каким-либо условиям.

В непараметрических задачах проверки гипотез предположение формулируется относительно самого́ вида распределения.

Начнём с самой простой задачи: выбора между двумя конкретными значениями параметра θ.

1.1. Простая гипотеза и простая альтернатива

Пусть в функции правдоподобия $L(x,\theta)$, $x \in \mathbb{R}^n$, $\theta \in \Theta$, множество Θ состоит всего из двух значений параметра, $\Theta = \{\theta_H, \theta_K\}$.

Наша гипотеза H заключается в том, что параметр распределения имеет значение θ_H; предположение о том, что параметр равен θ_K, назовём альтернативой K. Итак,

$$H\colon\ \theta = \theta_H, \qquad K\colon\ \theta = \theta_K. \tag{1.1}$$

Такие гипотеза и альтернатива называются простыми. Введём множество

$$\mathsf{R} = \{x \in \mathbb{R}^n\colon L(x,\theta_H) + L(x,\theta_K) > 0\} \tag{1.2}$$

реализаций с.в. ξ, возможных или при гипотезе, или при альтернативе.

Для решения задачи проверки гипотезы H при альтернативе K разобъём множество R на два непересекающихся подмножества, $\mathsf{R} = \mathsf{D}_H + \mathsf{D}_K$, одно из которых ($\mathsf{D}_H$) отвечает тем реализациям, которые свидетельствуют в пользу гипотезы и против альтернативы, а другое (D_K) отвечает тем реализациям, которые свидетельствуют в пользу альтернативы и против гипотезы. Такое разбиение множества реализаций называется *статистическим критерием*. Всюду далее мы пишем D вместо D_K и называем D *критическим множеством*, а множество D_H обозначаем как $\mathsf{R}\backslash\mathsf{D}_K$ и называем *множеством принятия гипотезы*. Альтернативным образом критерий можно задать функцией

$$\varphi\colon \mathsf{R} \to \mathbb{R}, \qquad \varphi(x) = \begin{cases} 1, & \text{если } x \in \mathsf{D}, \\ 0, & \text{если } x \in \mathsf{R}\backslash\mathsf{D}. \end{cases} \tag{1.3}$$

Она называется *критической функцией*. Видно, что критическая функция – это индикатор критического множества.

При проверке простой гипотезы против простой альтернативы после задания критерия легко найти вероятности ошибочных решений. Понятно, что в данном случае, имея наблюдение x, мы можем совершать ошибки двух типов:

1) согласно критерию мы обязаны отклонить гипотезу или, что то же самое, принять альтернативу (т. е. реализация $\xi = x$ попала в множество D), но на самом деле гипотеза верна (т. е. $\xi \sim L(\,\cdot\,, \theta_H)$);

2) согласно критерию мы обязаны принять гипотезу или, что то же самое, отклонить альтернативу (т. е. реализация x с.в. ξ попала в множество $\mathsf{R}\backslash\mathsf{D}$), но на самом деле гипотеза неверна, верна альтернатива (т. е. $\xi \sim L(\,\cdot\,, \theta_K)$).

Тогда вероятности первого и второго ошибочных решений рассчитываются соответственно по формулам[1)]

$$\begin{aligned} \alpha_1 &\stackrel{\text{def}}{=} P_H(\xi \in \mathsf{D}) = \int_{\mathsf{D}} L(x, \theta_H)\, dx, \\ \alpha_2 &\stackrel{\text{def}}{=} P_K(\xi \in \mathsf{R}\backslash\mathsf{D}) = \int_{\mathsf{R}\backslash\mathsf{D}} L(x, \theta_K)\, dx. \end{aligned} \tag{1.4}$$

Эти вероятности называются *ошибками первого и второго рода*. Очевидно, что

$$\int_{\mathsf{D}} L(x, \theta_H)\, dx + \int_{\mathsf{R}\backslash\mathsf{D}} L(x, \theta_H)\, dx = \int_{\mathsf{R}} L(x, \theta_H)\, dx = 1,$$

и аналогично для интегралов по распределению $L(\,\cdot\,, \theta_K)$. Ошибку первого рода часто называют *уровнем значимости статистического критерия*. Величина

$$1 - \alpha_1 = P_H(\xi \in \mathsf{R}\backslash\mathsf{D}) = \int_{\mathsf{R}\backslash\mathsf{D}} L(x, \theta_H)\, dx$$

равна вероятности принять гипотезу, когда гипотеза и в самом деле верна. Эта вероятность называется *уровнем доверия статистического критерия*. Далее во избежание путаницы мы будем использовать для величины α_1 только термин «ошибка первого рода». Для $1 - \alpha_2$ также вводится специальный термин: величина

$$\beta \stackrel{\text{def}}{=} P_K(\xi \in \mathsf{D}) = \int_{\mathsf{D}} L(x, \theta_K)\, dx \tag{1.5}$$

[1)] Здесь и далее мы считаем, что распределение выборки абсолютно непрерывно и записываем вероятность как интеграл от функции правдоподобия (плотности вероятности). В случае дискретного распределения интеграл, конечно, нужно заменить на сумму. Кроме того, мы используем краткие обозначения P_H вместо P_{θ_H} и P_K вместо P_{θ_K}.

называется *мощностью критерия*. Мощность критерия равна вероятности отклонить гипотезу, когда гипотеза неверна.

Мы видим, что любой статистический критерий проверки простой гипотезы $\theta = \theta_H$ против простой альтернативы $\theta = \theta_K$ характеризуется двумя параметрами: ошибкой первого рода (у которой мы далее не пишем нижний индекс и иногда называем её просто ошибкой критерия) и мощностью критерия:

$$\begin{aligned} \alpha &\stackrel{\text{def}}{=} P_H(\xi \in \mathsf{D}) = \int_{\mathsf{D}} L(x, \theta_H)\, dx = \int_{\mathsf{R}} \varphi(x) L(x, \theta_H)\, dx, \\ \beta &\stackrel{\text{def}}{=} P_K(\xi \in \mathsf{D}) = \int_{\mathsf{D}} L(x, \theta_K)\, dx = \int_{\mathsf{R}} \varphi(x) L(x, \theta_K)\, dx, \end{aligned} \tag{1.6}$$

где мы воспользовались определением (1.3) критической функции. Ошибка второго рода равна $1-\beta$. Наша цель – построить статистический критерий, т. е. выбрать множество D, так, чтобы величина α была как можно меньше, а величина β – как можно больше. Но понятно, что α и β связаны друг с другом, и почти всегда попытка уменьшить α приводит к уменьшению β, а увеличение β – к увеличению α.

Задача проверки простой гипотезы против простой альтернативы ставится следующим образом: пусть задан некоторый приемлемый размер ошибки $\alpha_0 \in (0, 1)$; понятно, что нас интересуют малые α_0, обычно выбирают α_0 порядка нескольких процентов. Введём множество критических функций вида (1.3), для которых ошибка первого рода не превышает α_0:

$$\mathrm{Cr} = \mathrm{Cr}(\alpha_0) = \left\{ \varphi(\,\cdot\,) \colon \int_{\mathsf{R}} \varphi(x) L(x, \theta_H)\, dx \leqslant \alpha_0 \right\}, \tag{1.7}$$

(см. правое выражение для α в (1.6)). Требуется найти критическую функцию $\varphi^*(\,\cdot\,) \in \mathrm{Cr}$, такую что

$$\beta^* = \int_{\mathsf{R}} \varphi^*(x) L(x, \theta_K)\, dx = \max_{\varphi \in \mathrm{Cr}} \int_{\mathsf{R}} \varphi(x) L(x, \theta_K)\, dx. \tag{1.8}$$

Введя критическое множество $\mathsf{D}^* \subset \mathsf{R}$, представим критическую функцию как

$$\varphi^*(x) = \begin{cases} 1, & \text{если } \ x \in \mathsf{D}^*, \\ 0, & \text{если } \ x \in \mathsf{R} \backslash \mathsf{D}^*. \end{cases} \tag{1.9}$$

Этот критерий обладает максимальной мощностью в классе всех критериев с ошибкой первого рода, не превосходящей заданной величины $\alpha_0 \in (0,1)$. Он так и называется: *наиболее мощный критерий* (НМК).

Понятно, что найти НМК означает найти множество $\mathsf{D}^* \subset \mathsf{R}$, такое что для критической функции (1.9) выполнено (1.8).

Ответ на вопрос, как выглядит НМК, дает следующее утверждение, которое представляет собой упрощённый вариант *фундаментальной леммы Неймана–Пирсона.*

Теорема 1. *Решение задачи* (1.8) *в классе* (1.7) *имеет вид*

$$\varphi^*(x) = \begin{cases} 1, & \text{если } L(x,\theta_K) > C \cdot L(x,\theta_H), \\ 0, & \text{если } L(x,\theta_K) \leqslant C \cdot L(x,\theta_H), \end{cases} \qquad x \in \mathsf{R}, \tag{1.10}$$

где постоянная $C \geqslant 0$ *находится из уравнения*

$$\int_{\mathsf{R}} \varphi^*(x) L(x,\theta_H)\, dx = \alpha_0. \tag{1.11}$$

Упрощённый характер этого утверждения связан с тем, что мы считаем уравнение (1.11) разрешимым (для заданного α_0 или для любого $\alpha_0 \in (0,1)$). Ниже мы приведём пример, когда это уравнение не имеет решений, и докажем лемму Неймана–Пирсона в общем случае.

Доказательство. Рассмотрим любую другую критическую функцию φ из множества (1.7) с критическим множеством D. Понятно, что с учётом вида критической функции

$$\int_{\mathsf{R}} \varphi(x) L(x,\theta)\, dx = \int_{\mathsf{D}} L(x,\theta)\, dx, \qquad \theta = \theta_H \text{ или } \theta = \theta_K.$$

Введём обозначения для мощностей двух критериев – критерия, заданного в (1.10), и произвольного из множества Cr:

$$\beta^* = \int_{\mathsf{D}^*} L(x,\theta_K)\, dx, \qquad \beta = \int_{\mathsf{D}} L(x,\theta_K)\, dx.$$

Оценим разность мощностей

$$\Delta\beta = \beta^* - \beta = \int_{\mathsf{D}^*} L(x,\theta_K)\, dx - \int_{\mathsf{D}} L(x,\theta_K)\, dx.$$

Очевидно, что $\mathsf{D}^* = (\mathsf{D}^* \cap \mathsf{D}) + (\mathsf{D}^* \backslash \mathsf{D})$ и $\mathsf{D} = (\mathsf{D} \cap \mathsf{D}^*) + (\mathsf{D} \backslash \mathsf{D}^*)$, и в $\Delta\beta$ интегралы по общей части $\mathsf{D}^* \cap \mathsf{D}$ сокращаются,

$$\Delta\beta = \int_{\mathsf{D}^* \backslash \mathsf{D}} L(x, \theta_K)\, dx - \int_{\mathsf{D} \backslash \mathsf{D}^*} L(x, \theta_K)\, dx.$$

В первом интеграле все значения переменной интегрирования лежат в множестве D^*, следовательно,

$$L(x, \theta_K) > C \cdot L(x, \theta_H) \quad \text{для всех} \quad x \in \mathsf{D}^* \backslash \mathsf{D}.$$

Аналогично во втором интеграле $x \notin \mathsf{D}^*$, следовательно,

$$L(x, \theta_K) \leqslant C \cdot L(x, \theta_H) \quad \text{для всех} \quad x \in \mathsf{D} \backslash \mathsf{D}^*.$$

Подставляя эти оценки в $\Delta\beta$ и восстанавливая в обоих слагаемых интегралы по общей части $\mathsf{D}^* \cap \mathsf{D}$, имеем

$$\begin{aligned} \Delta\beta &\geqslant C\biggl(\int_{\mathsf{D}^* \backslash \mathsf{D}} L(x, \theta_H)\, dx - \int_{\mathsf{D} \backslash \mathsf{D}^*} L(x, \theta_H)\, dx \biggr) = \\ &= C\biggl(\int_{\mathsf{D}^*} L(x, \theta_H)\, dx - \int_{\mathsf{D}} L(x, \theta_H)\, dx \biggr). \end{aligned}$$

Интегралы в правой части равенства задают ошибки первого рода для критерия (1.10) и критерия с критическим множеством D. Для критерия (1.10) справедливо равенство (1.11), для второго ошибка не превосходит α_0,

$$\int_{\mathsf{D}^*} L(x, \theta_H)\, dx = \alpha_0, \qquad \int_{\mathsf{D}} L(x, \theta_H)\, dx \leqslant \alpha_0.$$

Отсюда

$$\Delta\beta = \beta^* - \beta \geqslant C \cdot (\alpha_0 - \alpha_0) = 0,$$

тем самым $\beta^* \geqslant \beta$ для любого критерия с ошибкой первого рода, не превосходящей α_0. Теорема доказана.

Замечание 1. Решение задачи (1.8) в классе (1.7) можно также получить, решая задачу на условный экстремум: мы максимизируем по критическим функциям φ величину β при условии, что $\alpha \leqslant \alpha_0$. Функция Лагранжа такой задачи с точностью до слагаемого, не зависящего от φ, имеет вид

$$\beta - c\alpha = \int_{\mathsf{R}} \varphi(x)\bigl(L(x, \theta_K) - cL(x, \theta_H)\bigr)\, dx,$$

где c – параметр Лагранжа (см. определения (1.6)). Максимум по φ функции Лагранжа с учётом того, что $\varphi(x) = 0$ или $\varphi(x) = 1$, очевиден: нужно положить

$$\varphi(x) = \begin{cases} 1, & \text{если } L(x,\theta_K) - cL(x,\theta_H) > 0, \\ 0, & \text{если } L(x,\theta_K) - cL(x,\theta_H) \leqslant 0, \end{cases}$$

что даёт критическое множество $\mathsf{D}^* = \{x\colon L(x,\theta_K) > cL(x,\theta_H)\}$ из леммы Неймана–Пирсона.

Замечание 2. Зададим на множестве реализаций R *отношение правдоподобия*

$$r(x) = \frac{L(x,\theta_K)}{L(x,\theta_H)}, \tag{1.12}$$

считая, что если $L(x,\theta_H)(x) = 0$, то $r(x) = +\infty$ в том смысле, что $r(x) > C$ для любого $C \in \mathbb{R}$. Если $L(x,\theta_K) = 0$, то $r(x) = 0$ и тем самым неравенство $r(x) > C$ не выполняется для любого $C \geqslant 0$. Выполнение двух равенств $L(x,\theta_H) = 0$ и $L(x,\theta_K) = 0$ невозможно для любого $x \in \mathsf{R}$ в силу определения (1.2) множества R. Тогда мы можем записать критическое множество из леммы Неймана–Пирсона как

$$\mathsf{D}^* = \{x \in \mathsf{R}\colon r(x) > C\}, \qquad 0 \leqslant r(x) \leqslant +\infty, \quad C \geqslant 0.$$

Если $L(x,\theta_H) = 0$, а $L(x,\theta_K) \neq 0$, то реализацию x можно получить, только когда верна альтернатива. Тем самым мы, разумеется, должны включить такие x в критическое множество любого критерия. Это согласуется с тем, что $r(x) = +\infty > C$. Наоборот, если $L(x,\theta_H) \neq 0$, а $L(x,\theta_K) = 0$, то такая реализация x однозначно свидетельствует в пользу гипотезы, и мы без сомнения не включаем её в критическое множество. Это согласуется с $r(x) = 0 \leqslant C$. Проверка статистической гипотезы является нетривиальной задачей только для тех реализаций, которые можно получить и в случае гипотезы, и в случае альтернативы.

Рассмотрим пример, когда доказанная нами теорема работает.

Пример 1. Пусть $\xi \sim \mathbf{N}(\mu, 1)$ и $H\colon \mu = 0$, $K\colon \mu = \mu_1$, причём $\mu_1 > 0$. Найдём НМК для этой задачи. Имеем

$$L(x;\mu) = \frac{1}{\sqrt{2\pi}}\, e^{-(x-\mu)^2/2}, \qquad x \in \mathbb{R}.$$

Пусть $C \geqslant 0$. Рассмотрим неравенство $L(x;\mu_1) > C \cdot L(x;0)$. Перепишем его через отношение правдоподобия:

$$r(x) = \frac{L(x;\mu_1)}{L(x;0)} = \frac{e^{-(x-\mu_1)^2/2}}{e^{-x^2/2}} = e^{2x\mu_1} \cdot e^{-\mu_1^2} > C.$$

Решение этого неравенства относительно x можно записать как $2x\mu_1 > C_1$ или как $x > C_0$ в силу $\mu_1 > 0$. Постоянные C_1 и C_0 определяются параметром μ_1 и постоянной C. Явные выражения для них через эти величины нетрудно найти, но они нам не потребуются, потому что если известно значение C_0, то неравенство $x > C_0$ полностью определяет критическое множество НМК:

$$D_* = \{x \in \mathbb{R} \colon x > C_0\}. \tag{1.13}$$

Чтобы найти C_0, воспользуемся уравнением (1.11):

$$\alpha_0 = \int_{\mathsf{D}^*} L(x;0)\,dx = \int_{C_0}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\,dx = 1 - \Phi(C_0). \tag{1.14}$$

Это уравнение в силу строгой монотонности интеграла вероятности имеет единственное решение C_0 для любого $\alpha_0 \in (0,1)$, и тем самым соотношения (1.13), (1.14) задают единственно возможный НМК для нашей задачи.

Условие $\mu_1 > 0$ является важным для построения НМК, при противоположном неравенстве мы получили бы качественно другое множество $\mathsf{D}^* = \{x < C_0\}$.

Замечание 3. Структура множества $\mathsf{D}^* = \{x > C_0\}$ достаточно очевидна: если $\mu_1 > 0$, то чем больше значение реализации x, тем скорее оно свидетельствует против гипотезы, потому что реализации нормального распределения концентрируются в окрестности среднего значения $\mu = \mu_1$ для альтернативы и $\mu = 0$ для гипотезы. Это же можно сказать и об общей формулировке леммы Неймана–Пирсона. Если интерпретировать значение $L(x,\theta)$ как вероятность того, что $\xi = x$, когда в распределении с.в. ξ стоит параметр θ, и «измерять» значения $L(x,\theta_K)$ в «единицах» $L(x,\theta_H)$, то опять же чем больше $L(x,\theta_K)$, тем скорее x свидетельствует в пользу альтернативы, а не гипотезы.

Замечание 4. В рассмотренном примере, как мы уже отмечали, важную роль играет неравенство $\mu_1 > 0$. С другой стороны, критическое множество НМК $\mathsf{D}^* = \{x > C_0\}$ полностью определяется уравнением (1.14), которое не зависит от μ_1. Таким образом, для любых $\mu_1 > 0$ НМК будет один и тот же.

Пример 2. Пусть $\xi \sim \mathbb{U}[a, b]$, т. е.

$$L(x; [a, b]) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b], \\ 0, & x \notin [a, b], \end{cases} \qquad -\infty < a < b < \infty,$$

и $H\colon [a, b] = [0, 2]$, $K\colon [a, b] = [-1, 1]$. Для краткости обозначим функции правдоподобия для гипотезы и альтернативы как $L_0(x)$ и $L_1(x)$. Множество реализаций есть $\mathsf{R} = [-1, 2]$, а отношение правдоподобия равно

$$r(x) = \frac{L_1(x)}{L_0(x)} = \begin{cases} +\infty, & x \in [-1, 0), \\ 1, & x \in [0, 1], \\ 0, & x \in (1, 2]. \end{cases}$$

Решаем неравенство $r(x) > C$ для $x \in \mathsf{R}$ и $C \geqslant 0$, получаем множество решений

$$\mathsf{D} = \{x\colon r(x) > C\} = \begin{cases} [-1, 1], & 0 \leqslant C < 1, \\ [-1, 0), & 1 \leqslant C < \infty. \end{cases}$$

Тогда при $0 \leqslant C < 1$

$$\alpha(C) = \int\limits_{\mathsf{D}} L_0(x)\, dx = \int\limits_{[-1,1]\cap[0,2]} \frac{1}{2}\, dx = \int_0^1 \frac{1}{2}\, dx = \frac{1}{2},$$

а при $C \geqslant 1$ мы получаем $\alpha(C) = 0$.

Мы видим, что уравнение $\alpha(C) = \alpha_0$ не имеет решения относительно C ни при каких α_0, кроме $\alpha_0 = 1/2$ и $\alpha_0 = 0$. Конечно, мы можем сказать, что для прочих $\alpha_0 < 1/2$ можно выбрать любое $C \geqslant 1$ и, следовательно, $\mathsf{D}^* = [-1, 0)$, что даст нам нулевую ошибку. Но в этом случае мощность критерия равна

$$\beta = \int\limits_{\mathsf{D}^*} L_1(x)\, dx = \int\limits_{[-1,0)\cap[-1,1]} \frac{1}{2}\, dx = \frac{1}{2},$$

и ошибка второго рода $\alpha_2 = 1 - \beta = 1/2$, т. е. примерно в половине случаев, пользуясь таким критерием, мы будем совершать ошибку, когда примем гипотезу. Это неприемлемо высокая доля ошибочных решений даже с учётом «бонуса» в виде нулевой ошибки первого рода.

Далее мы покажем, как можно решить проблему неразрешимости уравнения (1.11).

Рандомизированные критерии

Будем считать, что критическая функция $\varphi(x)$, $x \in \mathsf{R}$, не принимает два значения 1 и 0, а может быть равна любому числу из отрезка $[0, 1]$. Тогда для каждого фиксированного $x \in \mathsf{R}$ число $\varphi(x)$ можно интерпретировать как вероятность того, что, имея результат наблюдения x, мы отклоним гипотезу. Критерии, построенные таким способом, называются *рандомизированными*.

Важно понимать, что вероятностная модель для $\varphi(x)$ не имеет никакого отношения к вероятностной природе с.в. ξ. Можно сказать, что алгоритм принятия решения заключается в следующем: для заданного $\xi = x$ вычисляется значение $\varphi(x)$, и далее разыгрывается случайный эксперимент (бросается несимметричная монетка с вероятностью $\varphi(x)$ выпадения «орла»), в котором мы отклоняем гипотезу с вероятностью $\varphi(x)$ (если монетка выпала вверх «орлом»).

Покажем, что для рандомизированных критериев лемма Неймана–Пирсона может быть строго сформулирована и доказана. Переопределим множество (1.7) рассматриваемых критических функций как

$$\mathrm{Cr} = \mathrm{Cr}(\alpha_0) = \left\{ \varphi(\,\cdot\,)\colon \mathsf{R} \to [0,1],\ \int_{\mathsf{R}} \varphi(x) L(x, \theta_{\scriptscriptstyle H})\, dx \leqslant \alpha_0 \right\}, \quad (1.15)$$

где мы учли, что теперь для каждого $x \in \mathsf{R}$ значение $\varphi(x)$ может быть любым числом из отрезка $[0, 1]$. Задачу поиска НМК сохраним в виде (1.8).

Теорема 2. 1. *Решение задачи* (1.8) *в классе* (1.15) *имеет вид*

$$\varphi^*(x) = \begin{cases} 1, & \text{если } L(x, \theta_{\scriptscriptstyle K}) > C \cdot L(x, \theta_{\scriptscriptstyle H}), \\ a, & \text{если } L(x, \theta_{\scriptscriptstyle K}) = C \cdot L(x, \theta_{\scriptscriptstyle H}), \\ 0, & \text{если } L(x, \theta_{\scriptscriptstyle K}) < C \cdot L(x, \theta_{\scriptscriptstyle H}), \end{cases} \qquad x \in \mathsf{R}, \quad (1.16)$$

где постоянные $a \in [0,1]$ *и* $C \geqslant 0$ *находятся из уравнения*

$$\int_{\mathsf{R}} \varphi^*(x) L(x,\theta_H)\,dx = \alpha_0. \tag{1.17}$$

2. *Уравнение* (1.17) *имеет решение для любого* $\alpha_0 \in (0,1)$.

Мы видим, что по сравнению с теоремой 1 изменения коснулись значений критической функции на множестве

$$\{x \in \mathsf{R}\colon L(x,\theta_K) = C \cdot L(x,\theta_H)\},$$

а именно, раньше мы относили такие реализации к критическому множеству, т. е. при их наблюдении отклоняли гипотезу с вероятностью (уверенностью) единица, а теперь отклоняем с вероятностью (уверенностью) a.

Доказательство . Предположим, что уравнение (1.17) имеет решение. Пусть $\varphi(\,\cdot\,)\colon \mathsf{R} \to [0,1]$ – некоторая критическая функция из множества (1.15). Разобъём R на следующие подмножества:

$$\begin{aligned} X^+ &= \{x \in \mathsf{R}\colon \varphi^*(x) - \varphi(x) > 0\},\\ X^0 &= \{x \in \mathsf{R}\colon \varphi^*(x) - \varphi(x) = 0\},\\ X^- &= \{x \in \mathsf{R}\colon \varphi^*(x) - \varphi(x) < 0\}. \end{aligned}$$

Тогда для разности мощностей критериев с критическими функциями $\varphi^*(\,\cdot\,)$ и $\varphi(\,\cdot\,)$ имеем

$$\begin{aligned} \Delta\beta &= \int_{\mathsf{R}} \varphi^*(x) L(x,\theta_K)\,dx - \int_{\mathsf{R}} \varphi(x) L(x,\theta_K)\,dx =\\ &= \left(\int_{X^+} + \int_{X^0} + \int_{X^-}\right) [\varphi^*(x) - \varphi(x)] L(x,\theta_K)\,dx, \end{aligned}$$

причём интеграл по X^0 очевидно равен нулю. Отсюда

$$\Delta\beta = \int_{X^+} [\varphi^*(x) - \varphi(x)] L(x,\theta_K)\,dx + \int_{X^-} [\varphi^*(x) - \varphi(x)] L(x,\theta_K)\,dx.$$

Если $x \in X^+$, то $\varphi^*(x) > \varphi(x) \geqslant 0$, следовательно, $\varphi^*(x) \neq 0$. Это означает, что такой x не может соответствовать третьей строке в определении (1.16). В результате для $x \in X^+$ имеем неравенства

$$\begin{aligned} L(x,\theta_K) &\geqslant C \cdot L(x,\theta_H),\\ [\varphi^*(x) - \varphi(x)] L(x,\theta_K) &\geqslant C \cdot [\varphi^*(x) - \varphi(x)] L(x,\theta_H), \end{aligned} \tag{1.18}$$

поскольку $\varphi^*(x) - \varphi(x) > 0$ для $x \in X^+$.

Если $x \in X^-$, то $\varphi^*(x) < \varphi(x) \leqslant 1$, следовательно, $\varphi^*(x) \neq 1$. Это означает, что такой x не может соответствовать первой строке в определении (1.16). В результате для $x \in X^-$ имеем неравенства

$$\begin{aligned} L(x,\theta_K) &\leqslant C \cdot L(x,\theta_H), \\ [\varphi^*(x) - \varphi(x)]L(x,\theta_K) &\geqslant C \cdot [\varphi^*(x) - \varphi(x)]L(x,\theta_H), \end{aligned} \tag{1.19}$$

поскольку $\varphi^*(x) - \varphi(x) < 0$ для $x \in X^-$.

При этом по условию

$$\int_{\mathsf{R}} \varphi^*(x)L(x,\theta_H)\,dx = \alpha_0, \qquad \int_{\mathsf{R}} \varphi(x)L(x,\theta_H)\,dx \leqslant \alpha. \tag{1.20}$$

Объединяя вторые неравенства в (1.18), (1.19) и условия (1.20), а затем добавляя интеграл по X^0, равный нулю, получаем

$$\begin{aligned} \Delta\beta &= \left(\int_{X^+} + \int_{X^-}\right)[\varphi^*(x) - \varphi(x)]L(x,\theta_K)\,dx \geqslant \\ &\geqslant C\left(\int_{X^+} + \int_{X^-}\right)[\varphi^*(x) - \varphi(x)]L(x,\theta_H)\,dx = \\ &= C\int_{\mathsf{R}}[\varphi^*(x) - \varphi(x)]L(x,\theta_H)\,dx \geqslant C(\alpha_0 - \alpha_0) = 0. \end{aligned}$$

Тем самым $\Delta\beta \geqslant 0$ для любого рандомизированного критерия с ошибкой первого рода, не превосходящей α_0. Первый пункт теоремы доказан.

Докажем разрешимость уравнения (1.17). Пусть критическая функция $\varphi^*(x)$, $x \in \mathsf{R}$, имеет вид (1.16). В явном виде для $x \in \mathsf{R}$ уравнение (1.17) может быть записано как

$$\int\limits_{x:\ L(x,\theta_K) > C\cdot L(x,\theta_H)} L(x,\theta_H)\,dx + a \cdot \int\limits_{x:\ L(x,\theta_K) = C\cdot L(x,\theta_H)} L(x,\theta_H)\,dx = \alpha_0. \tag{1.21}$$

Рассмотрим отношение правдоподобия

$$r(x) = \frac{L(x,\theta_K)}{L(x,\theta_H)}, \quad 0 \leqslant r(x) \leqslant \infty, \qquad x \in \mathsf{R},$$

и введём с.в. $r(\xi)$ обычным образом: $r(\xi) = r(x)$, когда $\xi = x$.

Далее мы считаем, что $\xi \sim L(\,\cdot\,, \theta_H)$, тогда $r(\xi) < \infty$ с вероятностью единица. В последующих формулах мы не пишем нижний индекс θ_H у значка вероятности, поскольку все вероятности ниже в этом разделе вычисляются только по распределению с параметром θ_H.

Теперь вспомним, что интеграл от функции правдоподобия равен вероятности попадания с.в. ξ в область, по которой ведётся интегрирование,

$$\int\limits_X L(x, \theta)\, dx = P_\theta(\xi \in X),$$

Поэтому уравнение (1.21) можно записать как

$$P(r(\xi) > C) + a \cdot P(r(\xi) = C) = \alpha_0. \tag{1.22}$$

Для с.в. $r(\xi)$ введем функцию

$$\overline{F}(z) \stackrel{\text{def}}{=} P(r(\xi) > z), \qquad z \in \mathbb{R},$$

связанную с функцией распределения $F_r(z) = P(r(\xi) < z)$, $z \in \mathbb{R}$, известным равенством

$$\overline{F}(z) = 1 - P(r(\xi) \leqslant z) = 1 - F_r(z + 0), \qquad z \in \mathbb{R}.$$

Сравним свойства этих двух функций:

$F_r(\,\cdot\,)$ не убывает,	$\overline{F}(\,\cdot\,)$ не возрастает,
$F_r(\,\cdot\,)$ непрерывна слева,	$\overline{F}(\,\cdot\,)$ непрерывна справа,
$F_r(+\infty) = 1$,	$\overline{F}(+\infty) = 0$,
$F_r(C) = 0$ при $C \leqslant 0$,	$\overline{F}(0) = 1$ при $C < 0$

(напомним, что с.в. $r(\xi)$ положительна) и, кроме того,

$$P(r(\xi) = z) = F_r(z + 0) - F_r(z) = \overline{F}(z - 0) - \overline{F}(z).$$

Тогда уравнение (1.22) можно переписать как

$$\overline{F}(C) + a[\overline{F}(C - 0) - \overline{F}(C)] = \alpha_0. \tag{1.23}$$

Покажем, что уравнение (1.23) всегда имеет решение (C, a). Если существует C, такое что $\overline{F}(C) = \alpha_0$, то решением очевидно является пара C, $a = 0$. В результате получаем нерандомизированный критерий из теоремы 1.

Может ли для данного $\alpha_0 \in (0, 1)$ не существовать C, такое что $\overline{F}(C) = \alpha_0$? Да, может, если α_0 попало в разрыв значений функции $\overline{F}(\,\cdot\,)$, другими словами, при некотором $C > 0$ мы имеем

$$\overline{F}(C - 0) \geqslant \alpha_0, \qquad \overline{F}(C) < \alpha_0 \tag{1.24}$$

(первое неравенство нестрогое, потому что если $\overline{F}(C - 0) = \alpha_0$, то число $\overline{F}(C - 0)$ существует, но не является значением функции $\overline{F}(\,\cdot\,)$ в какой-либо точке). С другой стороны, если в точке C имеется разрыв, то

$$\overline{F}(C - 0) - \overline{F}(C) = P(r(\xi) = C) \neq 0.$$

Тогда мы выбираем

$$a = \frac{\alpha_0 - \overline{F}(C)}{\overline{F}(C - 0) - \overline{F}(C)}. \tag{1.25}$$

Видно, что $0 < a \leqslant 1$ в силу (1.24), и уравнение (1.22) удовлетворяется. Теорема доказана.

Решение (C, a) уравнения (1.17) даёт строго рандомизированный критерий, т. е. $0 < a < 1$ в формуле (1.25), если левое предельное значение $\overline{F}(C - 0) > \alpha_0$. Если $\overline{F}(C - 0) = \alpha_0$, то критерий нерандомизированный ($a = 1$), но критическое множество задаётся неравенством $L(x, \theta_K) \geqslant C \cdot L(x, \theta_H)$ в отличие от строгого неравенства в теореме 1.

Из теоремы 2 следует, что рандомизация может потребоваться, только если отношение правдоподобия $r(\xi)$ при верной гипотезе принимает какое-либо постоянное значение с ненулевой вероятностью. В противном случае, т. е. когда при любом C для $\xi \sim L(\,\cdot\,, \theta_H)$ мы имеем $P(r(\xi) = C) = 0$, НМК будет нерандомизированным.

1.2. Сложная альтернатива

В реальных экспериментах часто возникают задачи, в которой гипотеза простая, но альтернатива сложная, т. е.

$$H\colon\ \theta = \theta_0, \qquad K\colon\ \theta \in \Theta_K, \tag{1.26}$$

где множество Θ_K содержит более одного значения параметра. Исследователь, как правило, может точно сформулировать гипотезу, но допускает более или менее широкое множество альтернативных вариантов в случае, когда гипотеза неверна.

Как и ранее, будем задавать критерий с помощью критической функции $\varphi(x)$, $x \in \mathsf{R}$, равной единице на критическом множестве D и нулю на множестве $\mathsf{R}\backslash\mathsf{D}$ принятия гипотезы (или $\varphi(x) \in [0,1]$, если мы рассматриваем рандомизированные критерии). Ошибка критерия по-прежнему равна

$$\alpha = P_{\theta_0}(\xi \in \mathsf{D}) = \int_{\mathsf{D}} L(x;\theta_0)\,dx = \int_{\mathsf{R}} \varphi(x) L(x;\theta_0)\,dx.$$

Однако теперь мощность критерия – это не число, а функция, зависящая от $\theta \in \Theta_K$:

$$\beta(\theta) \stackrel{\text{def}}{=} P_\theta(\xi \in \mathsf{D}) = \int_{\mathsf{D}} L(x;\theta)\,dx = \int_{\mathsf{R}} \varphi(x) L(x;\theta)\,dx, \quad \theta \in \Theta_K.$$

Для каждого фиксированного $\theta_1 \in \Theta_K$ величина $\beta(\theta_1)$ равна мощности критерия проверки простой гипотезы $\theta = \theta_0$ против простой альтернативы $\theta = \theta_1$.

Будем придерживаться того же подхода, что и выше: попытаемся найти критерий с максимальной мощностью (вероятностью принять правильное решение, отклоняя гипотезу), одновременно контролируя вероятность ошибочного отклонения гипотезы. Вновь рассмотрим класс критериев $\mathrm{Cr} = \mathrm{Cr}(\alpha_0)$ с ошибкой, не превосходящей фиксированной заранее величины α_0.

Наша цель – максимизировать мощность критерия. Для каждого фиксированного $\theta_1 \in \Theta_K$ существует НМК, который строится по лемме Неймана–Пирсона. Однако вид этого критерия может оказаться различным для разных значений θ_1. Тогда мы не сможем найти критерий, который был бы наиболее мощным сразу для всех альтернатив: НМК для какого-либо альтернативного значения параметра уже не будет наиболее мощным для другого значения. Только если НМК имеет один и тот же вид для всех $\theta \in \Theta_K$, мы можем утверждать, что нашли равномерно наилучшее по всем $\theta \in \Theta_K$ решение задачи проверки гипотез. Говорят, что в этом случае существует *равномерно наиболее мощный критерий* (РНМК).

Хотя условие существования РНМК, кажется очень жёстким, не так мало задач проверки гипотез допускают такое решение. Обратимся к примеру 1, где для одномерной выборки из распределения $\mathbf{N}(\mu, 1)$ мы проверяли (простую) гипотезу $\mu = 0$ против (простой) альтернативы $\mu = \mu_1$ при условии $\mu_1 > 0$. НМК в этом случае задаётся следующим образом:

$$\mathsf{D} = \{x > C\}, \qquad \int_C^\infty \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx = \alpha_0.$$

Мы видим, что условия, определяющие НМК, не зависят от μ_1, следовательно, мы получили РНМК проверки простой гипотезы $\mu = 0$ против сложной альтернативы $\mu > 0$.

Однако если мы сформулируем альтернативу как $\mu_1 \neq 0$, то для такой задачи РНМК не существует. Это связано с тем, что для $\mu_1 < 0$ критическое множество НМК имеет вид $\{x < C\}$, а для $\mu_1 > 0$ – вид $\{x > C\}$. Таким образом, РНМК существует для «односторонних» сложных альтернатив $\mu_1 > 0$, $\mu_1 < 0$, но не для «двусторонней» сложной альтернативы $\mu_1 \neq 0$.

Очевидно, что достаточным условием существования РНМК является монотонность отношения правдоподобия. Пусть для всех значений $\theta_1 \in \Theta_K$

$$r(x; \theta_0, \theta_1) \stackrel{\text{def}}{=} \frac{L(x; \theta_1)}{L(x; \theta_0)} > C \quad \Longleftrightarrow \quad x > \widetilde{C}$$

или

$$r(x; \theta_0, \theta_1) > C \quad \Longleftrightarrow \quad x < \widetilde{C},$$

где $\widetilde{C}$ не зависит от x, но может зависеть от θ_0 и θ_1. Тогда, выбирая соответственно $\mathsf{D} = \{x < \widetilde{C}\}$ или $\mathsf{D} = \{x > \widetilde{C}\}$ и находя $\widetilde{C}$ из уравнения

$$P_{\theta_0}(\xi \in \mathsf{D}) = \alpha_0,$$

получаем РНМК.

Как решать задачу проверки гипотезы при сложной альтернативе, если понятно, что в классе $\mathrm{Cr}(\alpha_0)$ РНМК не существует? Понятно, что в этом случае нам потребуется как-то иначе поставить задачу, например сузить класс критериев или по-другому задать характеристики качества критерия. Некоторые наиболее важные подходы мы рассмотрим в следующих разделах.

1.3. Локально наиболее мощные критерии

Будем решать задачу проверки простой параметрической гипотезы $H\colon \theta = \theta_0$ при сложной альтернативе $K\colon \theta \neq \theta_0$, считая, что $|\theta - \theta_0|$ исчезающе мал (напомним, что мы рассматриваем $\theta \in \mathbb{R}$). Другими словами, наша цель – отличить гипотезу от близкой альтернативы. Заметим, что такая задача, как правило, является более сложной, чем задача для альтернативных значений θ, сильно отличающихся от θ_0. Естественно предположить, что чем больше отличаются θ и θ_0, тем сильнее отличаются распределения выборки для гипотезы и для альтернативы. Например, если предположить, что $\theta = M\xi_k$ – математическое ожидание элемента выборки и $D\xi_k < \infty$, то в силу неравенства Чебышёва реализации с.в. $\xi_1, \xi_2, \ldots, \xi_n$ с большой вероятностью лежат в окрестности значения θ. Когда $|\theta - \theta_0|$ много больше размера этой окрестности, мы практически безошибочно сможем различить гипотезу и альтернативу.

Пусть, как и выше $L(x, \theta)$, $x \in \mathbb{R}^n$, $\theta \in \Theta$, есть функция правдоподобия (плотность вероятности) n-мерной выборки. Рассмотрим произвольный нерандомизированный критерий

$$\varphi(x) = \begin{cases} 1, & \text{если } \ x \in \mathsf{D}, \\ 0, & \text{если } \ x \notin \mathsf{D}, \end{cases} \qquad x \in \mathsf{R},$$

и введём функцию мощности в окрестности θ_0:

$$\beta(\theta) = \int_{\mathsf{D}} L(x;\theta)\,dx = \int_{\mathsf{R}} \varphi(x) L(x;\theta)\,dx, \qquad |\theta - \theta_0| \to 0.$$

Заметим, что в точке $\theta = \theta_0$ значение $\beta(\theta_0)$ – это ошибка критерия,

$$\beta(\theta_0) = \int_{\mathsf{D}} L(x;\theta_0)\,dx = \int_{\mathsf{R}} \varphi(x) L(x;\theta_0)\,dx.$$

Предположим, что функцию мощности в окрестности θ_0 можно разложить по формуле Тейлора до второго порядка,

$$\beta(\theta) = \beta(\theta_0) + \beta'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\beta''(\theta_0)(\theta - \theta_0)^2 + o(\theta - \theta_0)^2,$$

Поставим следующую задачу. Требуется найти критерий

$$\varphi^*_{\mathrm{loc}}(x) = \begin{cases} 1, & \text{если } \ x \in \mathsf{D}^*_{\mathrm{loc}}, \\ 0, & \text{если } \ x \notin \mathsf{D}^*_{\mathrm{loc}}, \end{cases}$$

на котором для всех θ в бесконечно малой окрестности значения θ_0 достигается максимум приближённой мощности

$$\tilde{\beta}(\theta) = \beta(\theta_0) + \beta'(\theta_0)(\theta - \theta_0) + \frac{1}{2}\beta''(\theta_0)(\theta - \theta_0)^2 \tag{1.27}$$

в классе всех критериев с ошибкой первого рода $\beta(\theta_0) = \alpha_0$. Критерий φ^*_{loc} называется *локально наиболее мощным* (ЛНМК).

Сформулируем три альтернативы: две односторонние

$$K_1\colon \theta > \theta_0, \quad \theta \to \theta_0 + 0 \quad \text{и} \quad K_2\colon \theta < \theta_0, \quad \theta \to \theta_0 - 0$$

и одну двустороннюю $K_3\colon \theta \neq \theta_0$, $\theta \to \theta_0$. Заметим, что при малых значениях $|\theta - \theta_0|$ третье слагаемое в правой части (1.27) много меньше, чем второе. Это дает нам основание в случае односторонних альтернатив пренебречь третьим слагаемым и искать

$$\max\bigl(\alpha_0 + \beta'(\theta_0)(\theta - \theta_0)\bigr).$$

С учётом знака разности $\theta - \theta_0$ получаем, что локально наиболее мощный критерий для односторонних альтернатив K_1 и K_2 определяется следующими экстремумами:

$$K_1\colon \ \max \beta'(\theta_0), \qquad K_2\colon \ \min \beta'(\theta_0). \tag{1.28}$$

Для двусторонней альтернативы рассмотрим следующий условный максимум:

$$\max\left\{\alpha_0 + \frac{1}{2}\beta''(\theta_0)(\theta - \theta_0)^2 \,\middle|\, \beta'(\theta_0) = 0\right\}$$

или, эквивалентно,

$$\max\bigl\{\beta''(\theta_0) \,\big|\, \beta'(\theta_0) = 0\bigr\}. \tag{1.29}$$

Условие $\beta'(\theta_0) = 0$ обеспечивает независимость максимальной мощности от знака разности $\theta - \theta_0$.

Заметим, что, поскольку мы ищем критерий, равномерно наиболее мощный по параметрам θ, лежащим в малой окрестности значения θ_0, естественно рассматривать критерии, у которых критическое множество D не зависит от $\theta \neq \theta_0$. Поэтому мы можем

считать, что

$$\beta'(\theta_0) = \left(\frac{d}{d\theta}\int_{\mathsf{D}} L(x;\theta)\,dx\right)\Big|_{\theta=\theta_0} = \int_{\mathsf{D}} \frac{\partial L}{\partial\theta}(x;\theta_0)\,dx,$$
$$\beta''(\theta_0) = \left(\frac{d^2}{d\theta^2}\int_{\mathsf{D}} L(x;\theta_0)\,dx\right)\Big|_{\theta=\theta_0} = \int_{\mathsf{D}} \frac{\partial^2 L}{\partial\theta^2}(x;\theta_0)\,dx$$

(дополнительно полагая, что внесение производной под знак интеграла корректно). Тем самым задачи (1.28) и (1.29) сводятся к следующим задачам на условный экстремум:

$$\begin{aligned}
&K_1\colon \max_D\left\{\int_{\mathsf{D}} \frac{\partial L}{\partial\theta}(x;\theta_0)\,dx \;\middle|\; \int_{\mathsf{D}} L(x;\theta_0)\,dx = \alpha_0\right\},\\
&K_2\colon \min_D\left\{\int_{\mathsf{D}} \frac{\partial L}{\partial\theta}(x;\theta_0)\,dx \;\middle|\; \int_{\mathsf{D}} L(x;\theta_0)\,dx = \alpha_0\right\},\\
&K_3\colon \max_D\left\{\int_{\mathsf{D}} \frac{\partial^2 L}{\partial\theta^2}(x;\theta_0)\,dx \;\middle|\; \int_{\mathsf{D}} L(x;\theta_0)\,dx = \alpha_0,\ \int_{\mathsf{D}} \frac{\partial L}{\partial\theta}(x;\theta_0)\,dx = 0\right\}.
\end{aligned} \tag{1.30}$$

Для решения этих задач нам потребуется *обобщённая лемма Неймана–Пирсона.*

Теорема 3. *Пусть заданы некоторые функции*

$$f_0(x), f_1(x), \ldots, f_m(x), \qquad x \in \mathbb{R}^n,$$

и числа $a_1, \ldots, a_m$. *Рассмотрим все подмножества* D *в* $\mathbb{R}^n$, *удовлетворяющие условиям*

$$\int_{\mathsf{D}} f_j(x)\,dx = a_j, \qquad j = 1, \ldots, m. \tag{1.31}$$

Предположим, что существуют постоянные $c_1, \ldots, c_m$, *при которых множество*

$$\mathsf{D}^* = \bigl\{x \in \mathbb{R}^n\colon f_0(x) > c_1 f_1(x) + \cdots + c_m f_m(x)\bigr\}$$

удовлетворяет условиям (1.31). *Тогда*

$$\int_{\mathsf{D}^*} f_0(x)\,dx \geqslant \int_{\mathsf{D}} f_0(x)\,dx$$

для любого множества D, *удовлетворяющего условиям* (1.31).

Доказательство. По аналогии с доказательством фундаментальной леммы Неймана–Пирсона рассмотрим разность

$$\Delta = \int_{\mathsf{D}^*} f_0(x)\,dx - \int_{\mathsf{D}} f_0(x)\,dx = \left(\int_{\mathsf{D}^*\setminus\mathsf{D}} - \int_{\mathsf{D}\setminus\mathsf{D}^*}\right) f_0(x)\,dx,$$

где мы учли, что интегралы по общей части $\mathsf{D}^* \cap \mathsf{D}$ сокращаются. По определению множества D^* имеем

$$\begin{aligned} f_0(x) &> c_1 f_1(x) + \cdots + c_m f_m(x) \quad \text{для} \quad x \in \mathsf{D}^*\setminus\mathsf{D}, \\ f_0(x) &\leqslant c_1 f_1(x) + \cdots + c_m f_m(x) \quad \text{для} \quad x \in \mathsf{D}\setminus\mathsf{D}^*. \end{aligned}$$

Отсюда

$$\Delta \geqslant \left(\int_{\mathsf{D}^*\setminus\mathsf{D}} - \int_{\mathsf{D}\setminus\mathsf{D}^*}\right) \sum_{j=1}^m c_j f_j(x)\,dx = \left(\int_{\mathsf{D}^*} - \int_{\mathsf{D}}\right) \sum_{j=1}^m c_j f_j(x)\,dx =$$
$$= \sum_{j=1}^m c_j \int_{\mathsf{D}^*} f_j(x)\,dx - \sum_{j=1}^m c_j \int_{\mathsf{D}} f_j(x)\,dx = \sum_{j=1}^m c_j a_j - \sum_{j=1}^m c_j a_j = 0.$$

Таким образом, $\Delta \geqslant 0$, и теорема доказана.

Заменяя все знаки неравенств на противоположные, получаем ещё один вариант леммы Неймана–Пирсона.

Теорема 4. *В условиях теоремы* 3 *множество*

$$D_* = \bigl\{x \in \mathbb{R}^n\colon f_0(x) < c_1 f_1(x) + \cdots + c_m f_m(x)\bigr\}$$

доставляет

$$\min \int_D f_0(x)\,dx$$

по всем множествам D, *удовлетворяющим условиям* (1.31).

Заметим, что теорема 1 (фундаментальная лемма Неймана–Пирсона) есть частный случай теоремы 3 при $m = 1$, если положить

$$f_0(x) = L(x, \theta_K), \quad f_1(x) = L(x, \theta_H), \quad x \in \mathbb{R}, \qquad a_1 = \alpha_0.$$

Применим теоремы 3 и 4 к задачам (1.30). Положив в теореме 3

$$f_0(x) = \frac{\partial L}{\partial \theta}(x; \theta_0), \quad f_1(x) = L(x; \theta_0), \quad x \in \mathbb{R}, \qquad a_1 = \alpha_0,$$

получаем критическое множество ЛНМК в случае альтернативы $\theta > \theta_0$:

$$\mathsf{D}^*_{\text{loc}} = \left\{ x \in \mathbb{R}^n \colon \frac{\partial L}{\partial \theta}(x;\theta_0) > C \cdot L(x;\theta_0) \right\}, \tag{1.32}$$

где постоянная C определяется из уравнения

$$\int_{\mathsf{D}^*_{\text{loc}}} L(x;\theta_0)\,dx = \alpha_0. \tag{1.33}$$

Аналогичным образом из теоремы 4 получаем критическое множество ЛНМК в случае альтернативы $\theta < \theta_0$:

$$\mathsf{D}^*_{\text{loc}} = \left\{ x \in \mathbb{R}^n \colon \frac{\partial L}{\partial \theta}(x;\theta_0) < C \cdot L(x;\theta_0) \right\}, \tag{1.34}$$

где постоянная C опять же определяется из уравнения (1.33).

Для двусторонней альтернативы положим в теореме 3

$$f_0(x) = \frac{\partial^2 L}{\partial \theta^2}(x;\theta_0), \quad f_1(x) = L(x;\theta_0), \quad f_2(x) = \frac{\partial L}{\partial \theta}(x;\theta_0)$$

и соответственно $a_1 = \alpha_0$, $a_2 = 0$. Получим

$$\mathsf{D}^*_{\text{loc}} = \left\{ x \in \mathbb{R}^n \colon \frac{\partial^2 L}{\partial \theta^2}(x;\theta_0) > C_1 \cdot L(x;\theta_0) + C_2 \cdot \frac{\partial L}{\partial \theta}(x;\theta_0) \right\}, \tag{1.35}$$

где постоянные C_1 и C_2 определяются из системы двух уравнений: уравнения (1.33) и уравнения

$$\int_{\mathsf{D}^*_{\text{loc}}} \frac{\partial L}{\partial \theta}(x;\theta_0)\,dx = 0. \tag{1.36}$$

1.4. Инвариантные критерии

Рассмотрим критические функции, аргументом которых является не реализация с.в. $\xi = (\xi_1, \ldots, \xi_n)$, а некоторая разумным образом заданная функция от неё (статистика).

Пример 3. Пусть $\xi \sim \mathbf{N}(\mu, 1)$ и $H \colon \mu = 0$, $K \colon \mu \neq 0$. В этой задаче присутствует очевидная симметрия относительно замены $\xi \to -\xi$. При этой замене распределение с.в. остаётся нормальным и, что особенно важно, соответствующая замена параметра $\mu \to -\mu$ не

изменяет гипотезу и альтернативу. Это даёт нам основание рассматривать в качестве статистики какую-либо чётную функцию от ξ. В многомерном случае $\xi = (\xi_1, \ldots, \xi_n)$ в качестве такой статистики выбирают $\sum_{k=1}^n \xi_k^2$, что в одномерном случае даёт ξ^2. Для простоты формул мы возьмём в сущности эквивалентную статистику $t(\xi) = |\xi|$. Понятно, что для $x > 0$

$$F_{|\xi|}(x) = P(|\xi| < x) = P(-x < \xi < x) = F_\xi(x) - F_\xi(-x),$$
$$p_{|\xi|}(x) = \frac{dF_{|\xi|}(x)}{dx} = p_\xi(x) + p_\xi(-x),$$

где $F_\xi(\,\cdot\,)$ и $p_\xi(\,\cdot\,)$ – функция распределения и плотность вероятности с.в. ξ, и $p_{|\xi|}(x) = 0$ при $x < 0$. Подставляя явный вид нормальной плотности, для гипотезы получаем в качестве функции правдоподобия статистики $|\xi|$ функцию

$$L(x;0) = \frac{2}{\sqrt{2\pi}}\, e^{-x^2/2}, \qquad x > 0,$$

а для альтернативы

$$L(x;\mu) = \frac{1}{\sqrt{2\pi}}(e^{-(x-\mu)^2/2} + e^{-(-x-\mu)^2/2}) =$$
$$= \frac{1}{\sqrt{2\pi}}\, e^{-(x^2+\mu^2)/2}(e^{\mu x} + e^{-\mu x}) = \frac{2}{\sqrt{2\pi}}\, e^{-(x^2+\mu^2)/2}\,\mathrm{ch}(\mu x),$$

где $x > 0$. Отношение правдоподобия имеет вид

$$r(x,\mu) = \frac{L(x;\mu)}{L(x;0)} = e^{-\mu^2/2}\,\mathrm{ch}(\mu x), \qquad x > 0.$$

Решая неравенство $r(x,\mu) > C$ при $x > 0$, получаем $x|\mu| > C_1$ или $x > C_0$. При решении неравенств мы, как обычно, не обращаем внимание на зависимость постоянных от μ. Если теперь выбрать постоянную C_0 из условия

$$\alpha_0 = \int_{C_0}^\infty L(x;0)\,dx = \int_{C_0}^\infty \frac{2}{\sqrt{2\pi}}\, e^{-x^2/2}\,dx = 2(1 - \Phi(C_0)),$$

то в классе всех критериев, зависящих от статистики $|\xi|$, мы получим РНМК проверки гипотезы $\mu = 0$ против сложной альтернативы $\mu \neq 0$ с ошибкой первого рода не больше α_0. Этот критерий отклоняет гипотезу, если $|\xi| > C_0$.

Сформулируем такой подход математически более строго. Будем считать выборку $\xi = (\xi_1, \ldots, \xi_n)$ случайным вектором со значениями в евклидовом пространстве $\mathcal{R}^n$. Пусть G – некоторая группа[2)] взаимно однозначных преобразований пространства $\mathcal{R}^n$. Для $x \in \mathcal{R}^n$ и преобразования $g \in G$ мы записываем результат преобразования вектора x как gx. Любая функция $t(\,\cdot\,)$ на $\mathcal{R}^n$, удовлетворяющая условию

$$t(gx) = t(x) \quad \text{для любого} \quad x \in \mathcal{R}^n \quad \text{и любого} \quad g \in G, \qquad (1.37)$$

называется инвариантом группы G. Множество значений функции $t(\,\cdot\,)$ может быть произвольным, но, имея в виду использование инвариантов для построения критериев, мы считаем, что $t(x)$ – это вещественное число.

Среди инвариантов выделяют максимальный инвариант $T(x)$, $x \in \mathcal{R}^n$, такой что из $T(x_1) = T(x_2)$ следует, что $x_2 = gx_1$ для некоторого $g \in G$. Справедливо следующее утверждение.

Предложение 1. *Функция $t(x)$, $x \in \mathcal{R}^n$, является инвариантом тогда и только тогда, когда она зависит от x через максимальный инвариант, т. е. на множестве значений максимального инварианта $T(x)$, $x \in \mathcal{R}^n$, существует функция $S(\,\cdot\,)$, такая что $t(x) = S(T(x))$ для всех $x \in \mathcal{R}^n$.*

Доказательство . Если $t(x) = S(T(x))$, то для любого $g \in G$

$$t(gx) = S(T(gx)) = S(T(x)) = t(x), \qquad x \in \mathcal{R}^n,$$

и $t(\,\cdot\,)$ – инвариант.

Докажем обратное утверждение. Пусть $t(\,\cdot\,)$ – инвариант. Тогда мы можем задать функцию $S(\,\cdot\,)$ следующим образом: для каждого значения $T(x)$ положим

$$S(T(x)) = t(x), \qquad x \in \mathcal{R}^n.$$

Это определение корректно, т. е. каждому $x \in \mathcal{R}^n$ отвечает единственное значение $S(T(x))$, поскольку

$$T(x_2) = T(x_1) \quad \Longrightarrow \quad x_2 = gx_1 \quad \Longrightarrow \quad t(x_2) = t(gx_1) = t(x_1).$$

[2)] В наших рассуждениях определение группы и её свойства не играют важной роли, и мы их опускаем. Подробности можно найти в многочисленных учебниках по теории групп.

Вернёмся к задачам проверки гипотез. Будем называть распределение выборки *инвариантным относительно группы* G, если для любого преобразования $g \in G$ и любого $\theta \in \Theta$ найдётся единственное значение параметра $\bar{g}\theta \in \Theta$, такое что $\xi \sim L(\,\cdot\,;\theta)$ влечёт $g\xi \sim L(\,\cdot\,;\bar{g}\theta)$. Другими словами, преобразование выборки порождает такое преобразование её распределения, которое не меняет вид функции правдоподобия, но, вообще говоря, меняет значение параметра, однако оставляет его в множестве Θ. Мы, конечно, дополнительно предполагаем, что $g\xi$ является с.в., т. е. имеет распределение, для любого преобразования $g \in G$.

Задача проверки гипотез называется инвариантной относительно группы G, если:

1) распределение выборки инвариантно относительно G;

2) для $\theta \in \Theta_H$ имеем $\bar{g}\theta \in \Theta_H$, для $\theta \in \Theta_K$ имеем $\bar{g}\theta \in \Theta_K$, т. е. индуцированное преобразованием $g \in G$ преобразование $\bar{g}$ множества параметров не изменяет гипотезу и альтернативу.

Для инвариантной задачи проверки гипотез естественно рассматривать *инвариантные критерии*, критическая функция которых есть инвариант группы G. Согласно доказанному утверждению все инвариантные критерии по сути зависят от максимального инварианта (опять же мы полагаем, что максимальный инвариант $T(\xi)$ есть с.в.).

Обобщим пример 3, взяв n-мерную выборку из нормального распределения $\mathbf{N}(\mu, I)$, где $\mu \in \mathcal{R}^n$, и поставим задачу проверки гипотезы $\mu = 0$ (нулевой вектор) против альтернативы $\mu \neq 0$. Эта задача инвариантна относительно группы ортогональных преобразований евклидова пространства (поворотов и отражений). В самом деле, если U – ортогональный оператор и $\xi \sim \mathbf{N}(\mu, I)$, то $U\xi \sim \mathbf{N}(U\mu, I)$ и $U\mu = 0$ тогда и только тогда, когда $\mu = 0$. Максимальный инвариант этой группы – квадрат нормы вектора,

$$T(x) = \|x\|^2 = \sum_{k=1}^{n} x_k^2, \qquad x = \langle x_1, \ldots, x_n \rangle.$$

Если $\mu = 0$, то $T(\xi) \sim \mathbf{X}_n^2$. Если $\mu \neq 0$, то статистика $T(\xi)$ имеет так называемое нецентральное распределение хи-квадрат, которое при заданном n полностью определяется параметром нецентральности $\|\mu\|^2$. Известно, что отношение правдоподобия для этих двух

распределений монотонно возрастает, следовательно, в классе инвариантных критериев существует РНМК, задающийся как

$$\mathsf{D} = \{\|x\|^2 \geqslant C\}, \qquad \int_C^\infty p_n(x)\,dx = \alpha_0,$$

где $p_n(\,\cdot\,)$ – плотность вероятности для распределения $\mathbf{X}_n^2$.

В заключение этого раздела приведём максимальные инварианты некоторых групп преобразований евклидова пространства $\mathcal{R}^n$.

Группа сдвигов на константу:

$$gx = g\langle x_1, \ldots, x_n\rangle = \langle x_1 + a, \ldots, x_n + a\rangle, \qquad a \in \mathbb{R}.$$

Максимальный инвариант $T(x) = \langle x_1 - x_n, \ldots, x_{n-1} - x_n\rangle$. В данном случае максимальный инвариант – это $(n-1)$-мерный вектор. Конечно, существуют и другие максимальные инварианты, например, можно брать разности $x_k - x_1$ для $k = 2, \ldots, n$.

Группа преобразований масштаба:

$$gx = g\langle x_1, \ldots, x_n\rangle = \langle ax_1, \ldots, ax_n\rangle, \qquad a \in \mathbb{R}.$$

Максимальный инвариант $T(x) = \langle x_1/x_n, \ldots, x_{n-1}/x_n\rangle$. В данном случае нам нужно потребовать, чтобы для соответствующего случайного вектора $\langle \xi_1, \ldots, \xi_n\rangle$ было выполнено условие $P(\xi_n = 0) = 0$, тем самым мы можем исключить из рассмотрения случай деления на ноль в координатах максимального инварианта.

Группа перестановок элементов множества $\{x_1, \ldots, x_n\}$ вещественных чисел:

$$g(x) = g(x_1, \ldots, x_n) = (x_{i_1}, \ldots, x_{i_n}), \qquad x = (x_1, \ldots, x_n),$$

где $i_1, \ldots, i_n$ – перестановка натуральных чисел $1, \ldots, n$. Максимальный инвариант $T(x) = x_1 + \cdots + x_n$.

1.5. Несмещённые, байесовские и максиминные критерии

Вернёмся к общей задаче проверки параметрической сложной гипотезы $H\colon\ \theta \in \Theta_H$ против сложной альтернативы $K\colon\ \theta \in \Theta_K$

и рассмотрим класс критериев с максимальной ошибкой первого рода не выше α_0:

$$\mathrm{Cr}^* = \mathrm{Cr}^*(\alpha_0) = \Big\{\varphi(\,\cdot\,)\colon \mathsf{R} \to [0,1],\ \sup_{\theta\in\Theta_H} \int_{\mathsf{R}} \varphi(x)L(x;\theta)\,dx \leqslant \alpha_0\Big\},$$

Для любого такого критерия

$$\alpha(\theta) = \int_{\mathsf{R}} \varphi(x)L(x;\theta)\,dx \leqslant \alpha_0 \quad \text{для всех} \quad \theta \in \Theta_H.$$

При этом мощность критерия

$$\beta(\theta) = \int_{\mathsf{R}} \varphi(x)L(x;\theta)\,dx, \qquad \theta \in \Theta_K,$$

может оказаться и больше α_0, и меньше α_0, но в последнем случае, отклоняя гипотезу, мы совершаем ошибку чаще, чем принимаем правильное решение. Такой критерий вряд ли можно считать удовлетворительным.

Несмещённые критерии

Рассмотрим критерии, в которых мощность всегда не меньше ошибки первого рода.

Определение 1. Критерий φ называется *несмещённым*, если

$$\begin{cases} \alpha(\theta) \leqslant \alpha_0 & \text{для всех } \theta \in \Theta_H, \\ \beta(\theta) \geqslant \alpha_0 & \text{для всех } \theta \in \Theta_K. \end{cases} \tag{1.38}$$

Видно, что несмещённый критерий принадлежит $\mathrm{Cr}^*(\alpha_0)$.

Существует тривиальный (рандомизированный) несмещённый критерий: положим $\varphi(x) = \alpha_0$ для всех $x \in \mathsf{R}$. Для такого критерия при любых θ (как для гипотезы, так и для альтернативы)

$$\int_{\mathsf{R}} \varphi(x)L(x;\theta)\,dx = \alpha_0 \cdot \int_{\mathsf{R}} L(x;\theta)\,dx = \alpha_0.$$

Условие (1.38) можно заменить более простым. Пусть функции $\alpha(\,\cdot\,)$ и $\beta(\,\cdot\,)$ непрерывно зависят от своего аргумента θ при

$\theta \in \Theta_H$ и $\theta \in \Theta_K$ соответственно. Определим Θ_{HK} – множество общих граничных точек гипотезы и альтернативы, т. е. множество точек, которые являются предельными и для Θ_H, и для Θ_K. Если, например, $\Theta_H = [\theta_1, \theta_2]$ и $\Theta_K = (-\infty, \theta_1) + (\theta_2, \infty)$, то $\Theta_{HK} = \{\theta_1, \theta_2\}$. Тогда условие несмещённости (1.38) влечёт

$$\alpha(\theta) = \beta(\theta) = \alpha_0 \quad \text{для всех} \quad \theta \in \Theta_{HK}. \tag{1.39}$$

Тем самым множество несмещённых критериев содержится в множестве критериев, удовлетворяющих условию (1.39). Следовательно, если среди всех критериев, удовлетворяющих (1.39), существует РНМК φ^*, то его мощность $\beta^*(\theta)$ не меньше мощности любого несмещённого критерия,

$$\beta^*(\theta) \geqslant \beta(\theta) \geqslant \alpha_0 \quad \text{для каждого} \quad \theta \in \Theta_K;$$

здесь $\beta(\theta)$ – значение мощности какого либо несмещённого критерия.

Мы доказали следующее утверждение.

Предложение 2. *Если $\alpha(\,\cdot\,)$ и $\beta(\,\cdot\,)$ непрерывно зависят от своего аргумента θ при $\theta \in \Theta_H$ и $\theta \in \Theta_K$ соответственно, а критерий φ^* является РНМК в классе всех критериев из $\mathrm{Cr}^*(\alpha_0)$, удовлетворяющих условию* (1.39), *то он является несмещённым (удовлетворяет условию* (1.38)*) и равномерно наиболее мощным в классе всех несмещённых критериев с ошибкой первого рода не выше α_0.*

Используя это предложение, мы можем искать РНМК в классе несмещённых критериев из $\mathrm{Cr}^*(\alpha_0)$, проверяя условие (1.39), которое проще, чем (1.38).

Пример 4. Пусть $\xi \sim \mathbf{E}(\theta)$, $\Theta = \{\theta > 0\}$ и

$$H\colon \theta \in [\theta_1, \theta_2], \qquad K\colon \theta \notin [\theta_1, \theta_2],\ \theta > 0,$$

где $0 < \theta_1 < \theta_2$. Отношение правдоподобия для экспоненциального распределения равно

$$r(x; \theta, \theta') = \frac{L(x; \theta')}{L(x; \theta)} = \frac{\theta' e^{-\theta' x}}{\theta e^{-\theta x}} = \frac{\theta'}{\theta}\, e^{-(\theta' - \theta)x}, \quad x > 0,$$

Понятно, что

$$r(x;\theta,\theta') > C \quad \Longleftrightarrow \quad (\theta'-\theta)x < c_1,$$

что для $\theta' > \theta$ даёт критическое множество вида $\{x < c\}$, а для $\theta' < \theta$ мы имеем критическое множество вида $\{x > c\}$. Таким образом, для двусторонней альтернативы не существует РНМК.

Условие (1.39) – это система уравнений в точках θ_1, θ_2 на границе гипотезы и альтернативы:

$$\beta(\theta_1) = \int_0^\infty \varphi(x)\theta_1 e^{-\theta_1 x}\,dx = \int_{\mathsf{D}} \theta_1 e^{-\theta_1 x}\,dx = \alpha_0,$$
$$\beta(\theta_2) = \int_0^\infty \varphi(x)\theta_2 e^{-\theta_2 x}\,dx = \int_{\mathsf{D}} \theta_2 e^{-\theta_2 x}\,dx = \alpha_0.$$

Доказательство того, что среди критериев, удовлетворяющих этому условию, существует РНМК, весьма непростое, и мы его опустим. Однако попробуем найти вид критического множества на основе свойств экспоненциального распределения. Мы знаем, что если $\xi \sim \mathbf{E}(\theta)$, то $M\xi = 1/\theta$. Таким образом, при $\theta \in [\theta_1, \theta_2]$ мы скорее будем наблюдать промежуточные реализации $x \in [c_1, c_2]$, чем близкие к нулю или очень большие. Это даёт нам основание выбрать $[c_1, c_2]$ как множество принятия гипотезы. Оказывается, что именно такой критерий является РНМК среди критериев, удовлетворяющих условию (1.39).

С учётом предложения 2 несмещённый РНМК с ошибкой первого рода не выше α_0 имеет вид

$$\varphi^*(x) = \begin{cases} 1, & \text{если } x \in [c_1, c_2], \\ 0, & \text{если } x \notin [c_1, c_2], \end{cases}$$

где постоянные $0 < c_1 < c_2$ определяются из системы уравнений

$$\int_{c_1}^{c_2} \theta_1 e^{-\theta_1 x}\,dx = 1 - \alpha_0,$$
$$\int_{c_1}^{c_2} \theta_2 e^{-\theta_2 x}\,dx = 1 - \alpha_0.$$

Решение системы существует и единственно, поэтому в $\mathrm{Cr}^*(\alpha_0)$ существует несмещённый РНМК.

Байесовские критерии

Байесовским называется подход, в котором искомому параметру сначала приписывается некоторое априорное вероятностное распределение, а после получения результата наблюдений это распределение пересчитывается в апостериорное (условное при заданном результате наблюдений) по формуле Байеса.

Рассмотрим для простоты случай, когда θ – одномерная с.в., принимающая значения $t \in \Theta$. Мы зададим распределения этой с.в. отдельно на множествах Θ_H и Θ_K и будем считать, что на этих множествах с.в. θ имеет плотности вероятности $g_H(\,\cdot\,)$ и $g_K(\,\cdot\,)$ соответственно,

$$\int_{\Theta_H} g_H(t)\,dt = 1, \qquad \int_{\Theta_K} g_K(t)\,dt = 1.$$

Они задают априорное распределение параметра[3)]. В отличие от рандомизированного критерия, априорное распределение с.в. θ не связано с конкретным наблюдением $x \in \mathsf{R}$, и по сути мы имеем совместное распределение двух с.в.: выборки ξ и параметра θ. В этих терминах то, что мы раньше обозначали как $L(x;\theta)$, есть значение условной плотности вероятности выборки в точке x при условии, что параметр фиксирован. Мы теперь можем написать

$$L(x;t) = p_{\xi|\theta}(x|t), \qquad x \in \mathsf{R}, \quad t \in \Theta_H \text{ или } t \in \Theta_K.$$

Если задан критерий φ проверки гипотезы $\theta \in \Theta_H$ против альтернативы $\theta \in \Theta_K$, то ошибку и мощность такого критерия теперь естественно усреднить по распределению с.в. θ. При фиксированном $\theta = t$ ошибка и мощность задаются как интеграл

$$P_t(\xi \in \mathsf{D}) = \int_{\mathsf{D}} L(x;t)\,dx = \int_{\mathsf{R}} \varphi(x) L(x;t)\,dx,$$

где $t \in \Theta_H$ для ошибки критерия и $t \in \Theta_K$ для его мощности. При

[3)] Распределение с.в. θ также можно задать на всём множестве Θ; после сужения этого распределения на множества Θ_H и Θ_K и умножения на соответствующее нормировочные множители мы приходим к распределениям на множествах гипотезы и альтернативы.

этом средние ошибка и мощность равны

$$\bar\alpha = \int_{\Theta_H} \alpha(t) g_H(t)\,dt = \int_{\Theta_H} \biggl[\int_{\mathsf{R}} \varphi(x) L(x;t)\,dx\biggr] g_H(t)\,dt$$
$$\bar\beta = \int_{\Theta_K} \beta(t) g_K(t)\,dt = \int_{\Theta_K} \biggl[\int_{\mathsf{R}} \varphi(x) L(x;t)\,dx\biggr] g_K(t)\,dt.$$

Поменяем в этих выражениях порядок интегрирования, получим

$$\begin{aligned} \bar\alpha &= \int_{\mathsf{R}} \varphi(x) \biggl[\int_{\Theta_H} L(x;t) g_H(t)\,dt\biggr]\,dx = \int_{\mathsf{R}} \varphi(x) \bar p_H(x)\,dx, \\ \bar\beta &= \int_{\mathsf{R}} \varphi(x) \biggl[\int_{\Theta_K} L(x;t) g_K(t)\,dt\biggr]\,dx = \int_{\mathsf{R}} \varphi(x) \bar p_K(x)\,dx, \end{aligned} \tag{1.40}$$

где мы ввели обозначения $\bar p_H(x)$ и $\bar p_K(x)$ для интегралов в квадратных скобках,

$$\begin{aligned} \bar p_H(x) &= \int_{\Theta_H} L(x;t) g_H(t)\,dt, \\ \bar p_K(x) &= \int_{\Theta_K} L(x;t) g_K(t)\,dt. \end{aligned} \tag{1.41}$$

Заметим, что $\bar p_H(x) \geqslant 0$, при этом

$$\int_{\mathsf{R}} \bar p_H(x)\,dx = \int_{\Theta_H} g_H(t)\,dt \int_{\mathsf{R}} L(x;t)\,dx = \int_{\Theta_H} g_H(t)\,dt = 1$$

и аналогично для $\bar p_K(\,\cdot\,)$, т. е. $\bar p_H(\,\cdot\,)$ и $\bar p_K(\,\cdot\,)$ имеют смысл плотностей вероятности. Это также можно понять из следующих рассуждений: мы имеем $L(x;t)g_H(t) = p_{\xi|\theta}(x|t)p_\theta(t) = p_{\xi,\theta}(x,t)$ для $t \in \Theta_H$; аналогично $L(x;t)g_K(t) = p_{\xi,\theta}(x,t)$ для $t \in \Theta_K$. Интегрируя эти выражения соответственно по $t \in \Theta_H$ и по $t \in \Theta_K$, мы получаем (безусловные) плотности вероятности с.в. ξ в точке x для гипотезы и альтернативы.

Заметим также, что плотности (1.41) можно интерпретировать как усреднённые по распределениям параметров гипотезы и альтернативы функции правдоподобия $L(x;\,\cdot\,)$ (плотности вероятности выборки) при фиксированном x,

$$\bar p_H(x) = M_\theta L(x;\theta), \quad \theta \in \Theta_H, \qquad \bar p_K(x) = M_\theta L(x;\theta), \quad \theta \in \Theta_K$$

(здесь нижний индекс у математического ожидания означает, что оно вычисляется по распределению с.в. θ).

Мы видим, что формулы (1.40) в точности соответствуют ошибке и мощности критерия φ в задаче проверки простой гипотезы $\xi \sim \bar{p}_H(\,\cdot\,)$ против простой альтернативы $\xi \sim \bar{p}_K(\,\cdot\,)$. Для этой задачи в классе всех критериев с $\bar{\alpha} \leqslant \alpha_0$ по лемме Неймана–Пирсона существует НМК φ^*. Этот критерий характеризуется допустимым размером ошибки α_0 и максимально возможной при такой ошибке мощностью:

$$\begin{aligned} \bar{\alpha}_0 &= \int_{\mathsf{R}} \varphi^*(x) \left[\int_{\Theta_H} L(x;t) g_H(t)\, dt \right] dx = \int_{\mathsf{R}} \varphi^*(x) \bar{p}_H(x)\, dx, \\ \bar{\beta}^* &= \int_{\mathsf{R}} \varphi^*(x) \left[\int_{\Theta_K} L(x;t) g_K(t)\, dt \right] dx = \int_{\mathsf{R}} \varphi^*(x) \bar{p}_K(x)\, dx. \end{aligned} \tag{1.42}$$

Недостатком предложенного решения является его зависимость от априорных распределений $g_H(\,\cdot\,)$ и $g_K(\,\cdot\,)$, задание которых неочевидно. Далее мы отмечаем эту зависимость от распределений, положив $\varphi^* = \varphi^*_g$. Мощность данного НМК обозначим как $\bar{\beta}^*_g$.

Максиминные критерии.

Подойдём к нашей задаче с другой стороны. На время забудем о том, что $g_H(\,\cdot\,)$ и $g_K(\,\cdot\,)$ – распределения на множестве параметров, но продолжим использовать критерий φ^*_g, найденный по лемме Неймана–Пирсона в байесовском подходе. Никто не может запретить нам применить этот критерий φ^*_g для проверки простой гипотезы $\theta = t \in \Theta_H$ против простой альтернативы $\theta = t' \in \Theta_K$ (назовём такую задачу проверки гипотез локальной). При этом качество критерия задаётся локальными ошибкой и мощностью

$$\alpha^*_g(t) = \int_{\mathsf{R}} \varphi^*_g(x) L(x;t)\, dx, \qquad \beta^*_g(t') = \int_{\mathsf{R}} \varphi^*_g(x) L(x;t')\, dx.$$

Теперь вспомним, что на самом деле гипотеза и альтернатива сложные, и рассмотрим наихудшую (наибольшую) по $t \in \Theta_H$ локальную ошибку и наихудшую (наименьшую) по $t' \in \Theta_K$ локальную мощность, т. е. величины

$$\sup_{t \in \Theta_H} \alpha^*_g(t), \qquad \inf_{t' \in \Theta_K} \beta^*_g(t'), \tag{1.43}$$

и примем их за характеристики качества нашего (байесовского) критерия φ^*_g как критерия проверки сложной гипотезы $\theta \in \Theta_H$ против сложной альтернативы $\theta \in \Theta_K$ в небайесовской постановке.

Ещё раз обратим внимание на то, что мы используем один и тот же критерий для решения разных задач: проверки простой гипотезы $\xi \sim \bar{p}_H(\,\cdot\,)$ против простой альтернативы $\xi \sim \bar{p}_K(\,\cdot\,)$ и для проверки сложной гипотезы $\xi \sim L(\,\cdot\,;\theta)$, $\theta \in \Theta_H$ против сложной альтернативы $\theta \in \Theta_K$. При этом характеристики качества критерия в этих двух случаях задаются по-разному, формулами (1.42) и (1.43), и их значения, вообще говоря, могут отличаться. Связь между этими характеристиками даётся следующей теоремой.

Теорема 5. *Пусть $\alpha_0 \in (0,1)$ фиксировано. Предположим, что найдутся такие распределения $g_H^*(\,\cdot\,)$ и $g_K^*(\,\cdot\,)$, что в* (1.43)

$$\sup_{t\in\Theta_H} \alpha_{g^*}^*(t) \leqslant \alpha_0, \qquad \inf_{t'\in\Theta_K} \beta_{g^*}^*(t') = \bar{\beta}_{g^*}^* \tag{1.44}$$

(*наихудшая мощность в небайесовской постановке и мощность байесовского НМК для g^*-распределений совпадают*). *Тогда*:

1) *для любого критерия φ, такого что*

$$\sup_{t\in\Theta_H} \int_R \varphi(x) L(x;t)\,dx \leqslant \alpha_0, \tag{1.45}$$

его минимальная по $t \in \Theta_K$ мощность не больше, чем $\bar{\beta}_{g^*}^*$;

2) *распределения $g_H^*(\,\cdot\,)$ и $g_K^*(\,\cdot\,)$ являются наихудшими в том смысле, что для любых других распределений $g_H(\,\cdot\,)$ и $g_K(\,\cdot\,)$ мощность $\bar{\beta}_g^*$ соответствующего НМК критерия φ_g^* проверки байесовской гипотезы с ошибкой не выше α_0 не ниже мощности критерия $\varphi_{g^*}^*$:*

$$\bar{\alpha}_g^* \leqslant \alpha_0, \qquad \bar{\beta}_g^* \geqslant \bar{\beta}_{g^*}^*. \tag{1.46}$$

Доказательство. 1. Пусть заданы распределения $g_H^*(\,\cdot\,)$ и $g_K^*(\,\cdot\,)$, удовлетворяющие (1.44). Рассмотрим любой критерий φ проверки гипотезы $\theta \in \Theta_H$ против альтернативы $\theta \in \Theta_K$ с ошибкой, удовлетворяющей (1.45). Тогда для любого $t \in \Theta_H$

$$\alpha(t) = \int_{\mathsf{R}} \varphi(x) L(x;t)\,dx \leqslant \alpha_0. \tag{1.47}$$

В силу нормировки плотности $g_H^*(\,\cdot\,)$ параметра θ отсюда имеем

$$\int_{\Theta_H} \alpha(t) g_H^*(t)\,dt \leqslant \int_{\Theta_H} \alpha_0 \cdot g_H^*(t)\,dt = \alpha_0.$$

С другой стороны, меняя порядок интегрирования, получаем

$$\int_{\Theta_H} \alpha(t) g_H^*(t)\,dt = \int_{\Theta_H} \biggl[\int_{\mathsf{R}} \varphi(x) L(x;t)\,dx \biggr] g_H^*(t)\,dt = \\ = \int_{\mathsf{R}} \varphi(x) \biggl[\int_{\Theta_H} L(x;t) g_H^*(t)\,dt \biggr]\,dx = \int_{\mathsf{R}} \varphi(x) \bar{p}_H^*(x)\,dx.$$

Следовательно,

$$\int_{\mathsf{R}} \varphi(x) \bar{p}_H^*(x)\,dx \leqslant \alpha_0.$$

Таким образом, φ является критерием проверки простой гипотезы $\xi \sim \bar{p}_H^*(\,\cdot\,)$ с ошибкой не выше α_0. Если мы теперь возьмём в качестве альтернативы $\xi \sim \bar{p}_K^*(\,\cdot\,)$, где

$$\bar{p}_K^*(x) = \int_{\Theta_K} L(x;t') g_K^*(t')\,dt',$$

то мощность критерия φ будет не выше мощности НМК $\varphi_{g^*}^*$, которая равна $\bar{\beta}_{g^*}^*$. В результате получаем

$$\bar{\beta}_{g^*}^* \geqslant \int_{\mathsf{R}} \varphi(x) \bar{p}_K^*(x)\,dx = \int_{\mathsf{R}} \varphi(x) \biggl[\int_{\Theta_K} L(x;t') g_K^*(t')\,dt' \biggr]\,dx = \\ = \int_{\Theta_K} \biggl[\int_{\mathsf{R}} \varphi(x) L(x;t')\,dx \biggr] g_K^*(t')\,dt' = \int_{\Theta_K} \beta(t') g_K^*(t')\,dt' \geqslant \\ \geqslant \inf_{t' \in \Theta_K} \beta(t') \cdot \int_{\Theta_K} g_K^*(t')\,dt' = \inf_{t' \in \Theta_K} \beta(t'). \tag{1.48}$$

Первый пункт теоремы доказан.

2. Для произвольных распределений $g_H(\,\cdot\,)$ и $g_K(\,\cdot\,)$ напомним обозначения

$$\int_{\Theta_H} L(x;t) g_H(t)\,dt = \bar{p}_H(x), \qquad \int_{\Theta_K} L(x;t) g_K(t)\,dt = \bar{p}_K(x).$$

Напомним также, что для этих распределений существует критерий φ_g^* проверки гипотезы $\xi \sim \bar{p}_H(\,\cdot\,)$ против $\xi \sim \bar{p}_K(\,\cdot\,)$, который задается как НМК по лемме Неймана–Пирсона .

С другой стороны, если мы возьмём в этой же задаче критерий $\varphi^*_{g^*}$, то

$$\int_{\mathsf{R}} \varphi^*_{g^*}(x)\bar{p}_{H}(x)\,dx = \int_{\Theta_H}\left[\int_{\mathsf{R}} \varphi^*_{g^*}(x)L(x;t)\,dx\right] g_H(t)\,dt \leqslant$$
$$\leqslant \sup_{t\in\Theta_H}\left[\int_{\mathsf{R}} \varphi^*_{g^*}(x)L(x;t)\,dx\right]\cdot\int_{\Theta_H} g_H(t)\,dt \leqslant \alpha_0\cdot 1.$$

Последнее неравенство выполнено в силу первого из условий (1.44) и условия нормировки плотности вероятности $g_{H}(\,\cdot\,)$. Таким образом, ошибка критерия $\varphi^*_{g^*}$ проверки простой гипотезы $\xi\sim\bar{p}_{H}(\,\cdot\,)$ против простой альтернативы $\xi\sim\bar{p}_{K}(\,\cdot\,)$ удовлетворяет неравенству

$$\int_{\mathsf{R}} \varphi^*_{g^*}(x)\bar{p}_{H}(x)\,dx \leqslant \alpha_0.$$

При альтернативе $\xi\sim\bar{p}_{K}(\,\cdot\,)$ мощность критерия $\varphi^*_{g^*}$

$$\int_{\mathsf{R}} \varphi^*_{g^*}(x)\bar{p}_{K}(x)\,dx \leqslant \bar{\beta}^*_g,$$

потому что мощность $\bar{\beta}^*_g$ максимальна в силу леммы Неймана–Пирсона. Повторяя преобразования (1.48) для $\varphi^*_{g^*}$ и $\bar{p}_{K}(\,\cdot\,)$, получаем

$$\bar{\beta}^*_g \geqslant \int_{\mathsf{R}} \varphi^*_{g^*}(x)\bar{p}_{K}(x)\,dx = \int_{\Theta_K}\left[\int_{\mathsf{R}} \varphi^*_{g^*}(x)L(x;t')\,dx\right] g_K(t')\,dt' \geqslant$$
$$\geqslant \inf_{t'\in\Theta_K}\left[\int_{\mathsf{R}} \varphi^*_{g^*}(x)L(x;t')\,dx\right]\cdot\int_{\Theta_K} g_K(t')\,dt' = \inf_{t'\in\Theta_K}\beta^*_{g^*}(t') = \bar{\beta}^*_{g^*};$$

при получении последнего равенства мы учли второе соотношение в (1.44). Теорема доказана.

Из этой теоремы напрямую вытекает следующий вывод. Рассмотрим задачу проверки сложной гипотезы $\theta\in\Theta_H$ против сложной альтернативы $\theta\in\Theta_K$. Будем характеризовать качество критерия φ наихудшей по $t\in\Theta_H$ локальной ошибкой первого рода и наихудшей по $t\in\Theta_K$ локальной мощностью:

$$\alpha_{\max}(\varphi) \stackrel{\text{def}}{=} \sup_{t\in\Theta_H}\alpha(t) = \sup_{t\in\Theta_H}\int_{\mathsf{R}}\varphi(x)L(x;t)\,dx,$$
$$\beta_{\min}(\varphi) \stackrel{\text{def}}{=} \inf_{t'\in\Theta_K}\beta(t') = \inf_{t'\in\Theta_K}\int_{\mathsf{R}}\varphi(x)L(x;t')\,dx.$$

Можно назвать эти величины соответственно ошибкой и мощностью критерия φ проверки сложной гипотезы $\theta \in \Theta_H$ против сложной альтернативы $\theta \in \Theta_K$. Зная их, мы можем контролировать качество решения для наименее благоприятных и, следовательно, для любых значений параметра θ. Критерий $\varphi_{\max\min}$, такой что

$$\beta_{\min}(\varphi_{\max\min}) = \max_{\varphi} \beta_{\min}(\varphi), \qquad \alpha_{\max}(\varphi_{\max\min}) \leqslant \alpha_0,$$

называется *максиминным* критерием проверки сложной гипотезы $\theta \in \Theta_H$ против сложной альтернативы $\theta \in \Theta_K$ с ошибкой не выше α_0.

Согласно теореме 5 в классе всех критериев с ошибкой не выше α_0 максимальная мощность достигается на критерии (для простоты запишем нерандомизированный вариант леммы Неймана–Пирсона)

$$\varphi(x) = \begin{cases} 1, & \text{если } \bar{p}^*_K(x) > C\bar{p}^*_H(x), \\ 0, & \text{если } \bar{p}^*_K(x) \leqslant C\bar{p}^*_H(x), \end{cases}$$

где C находится из уравнения

$$\int\limits_{x\colon \bar{p}^*_K(x) > C\bar{p}^*_H(x)} \varphi(x)\bar{p}^*_H(x)\,dx = \alpha_0;$$

эта мощность равна

$$\int\limits_{x\colon \bar{p}^*_K(x) > C\bar{p}^*_H(x)} \varphi(x)\bar{p}^*_K(x)\,dx = \beta_0.$$

Здесь

$$\bar{p}^*_H(x) = \int_{\Theta_H} L(x;t)g^*_K(t)\,dt, \qquad \bar{p}^*_K(x) = \int_{\Theta_K} L(x;t')g^*_K(t')\,dt',$$

а распределения $g^*_H(\,\cdot\,)$ и $g^*_K(\,\cdot\,)$ на множествах значений параметра являются наихудшими в том смысле, что для любых других распределений $g_H(\,\cdot\,)$ и $g_K(\,\cdot\,)$ мощность аналогичного критерия больше или равна β_0, когда ошибка первого рода для этого критерия не выше α_0.

Таким образом, задача поиска максиминного критерия сводится к нахождению наихудших априорных распределений значений параметра θ на множествах Θ_H и Θ_K. Однако нельзя сказать, что вторая задача решается проще, чем первая.

2. Непараметрические гипотезы

В этом разделе статистическая гипотеза формулируется как предположение о виде распределения выборки без введения какой-либо идентификации этого распределения с помощью параметра. Заметим, что если при этих условиях речь идёт о проверке простой гипотезы против простой альтернативы, т. е. о выборе между двумя конкретными и заранее заданными распределениями, то можно найти НМК по лемме Неймана–Пирсона в точности так же, как при проверке гипотезы о значении параметра. В самом деле, пусть гипотеза заключается в том, что выборка ξ имеет плотность вероятности $p_0(\,\cdot\,)$, альтернативную плотность вероятности обозначим как $p_1(\,\cdot\,)$. Тогда (нерандомизированный) НМК с ошибкой первого рода, равной α_0, задаётся как

$$\varphi^*(x) = \begin{cases} 1, & \text{если } \ p_1(x) > C_0 \cdot p_0(x), \\ 0, & \text{если } \ p_1(x) \leqslant C_0 \cdot p_0(x), \end{cases} \quad x \in \mathsf{R},$$

где постоянная $C_0 \geqslant 0$ находится из уравнения

$$\int_{\mathsf{R}} \varphi^*(x) p_0(x)\, dx = \alpha_0.$$

По сути дела номер $k = 0, 1$ в $p_k(\,\cdot\,)$ играет роль параметра распределения. Эта идея без труда переносится на случай сложных гипотез и/или альтернатив, если можно как-то параметризовать (например, пронумеровать) распределения, чтобы отличать одно от другого. В этом случае работают все рассмотренные нами выше методы.

Особенность задач проверки непараметрических гипотез состоит в том, что в этих задачах не формулируется класс альтернативных распределений. Как мы уже отмечали, такая ситуация очень часто складывается на практике. Исследователь обычно может более или менее конкретно сформулировать предположение о том, как должно, по его мнению, выглядеть распределение выборки (гипотеза), но не может дать ответ на вопрос о том, как будет вы-

глядеть это распределение, когда предположение не выполняется (сформулировать альтернативу).

Далее мы рассматриваем выборку $\xi = (\xi_1, \dots, \xi_n)$ объёма n, где, как обычно, с.в. $\xi_1, \dots, \xi_n$ независимы и одинаково распределены. Распределение каждой из этих с.в. в общем случае задаётся функцией распределения $F(\,\cdot\,)$ (в случае абсолютно непрерывного распределения – плотностью вероятности $p(\,\cdot\,)$, в случае дискретного распределения – распределением вероятностей $P(\,\cdot\,)$). Тогда мы пишем $\xi \sim F(\,\cdot\,)$ ($\xi \sim p(\,\cdot\,)$ или $\xi \sim P(\,\cdot\,)$ соответственно). Заметим, что эти обозначения несколько несогласованные: в формуле $\xi \sim F(\,\cdot\,)$ с.в. имеет размерность n, а $F(\,\cdot\,)$ – функция одномерного распределения. Но, поскольку все элементы выборки имеют одно и то же распределение, это не приведёт к недоразумениям. Можно сказать, что в данном случае значок $\sim$ заменяет слова «есть выборка из распределения, заданного функцией».

2.1. Критерии согласия

Всюду далее мы рассматриваем простую гипотезу (её часто называют нулевой)

$$H\colon\ \xi \sim F_0(\,\cdot\,). \tag{2.1}$$

Решение задачи проверки гипотезы $\xi \sim F_0(\,\cdot\,)$ опирается на проверку «удалённости» эмпирического распределения от теоретического. Введём функционал $d(\,\cdot\,)$ со значениями в $\mathbb{R}_+ = \{x \geqslant 0\}$, в каком-либо смысле характеризующий расстояние между двумя произвольными функциями, действующими из $\mathbb{R}$ в $\mathbb{R}$. Если этот функционал удовлетворяет определённым аксиомам (которые в нашем курсе несущественны), то он называется метрикой. Наиболее часто используются следующие метрики: равномерная

$$d_\infty(F_1, F_2) = \sup_{x \in \mathbb{R}} \big|F_1(x) - F_2(x)\big|, \tag{2.2}$$

квадратичная или евклидова

$$d_2(F_1, F_2) = \sqrt{\int_{\mathbb{R}} \big(F_1(x) - F_2(x)\big)^2\, dx}, \tag{2.3}$$

а также метрика

$$d_1(F_1, F_2) = \int_{\mathbb{R}} \big|F_1(x) - F_2(x)\big|\, dx. \tag{2.4}$$

В математической статистике эти метрики (при некоторой их модификации) используются для определения расстояния между функциями распределения или плотностями вероятности.

Для любой выборки $\xi = (\xi_1, \ldots, \xi_n)$ мы можем найти выборочную функцию распределения,

$$F^*(x) = \frac{1}{n}\sum_{j=1}^{n} I_{(-\infty,x)}(\xi_j) = \frac{\nu_\xi(x)}{n}, \qquad x \in \mathbb{R}, \tag{2.5}$$

где $I_{(-\infty,x)}$ – индикатор множества $(-\infty, x)$ и $\nu_\xi(x)$ – с.в., равная количеству элементов выборки ξ, попавших в полубесконечный интервал $(-\infty, x)$. Можно сказать, что $F_0(\,\cdot\,)$ – теоретическая функция распределения выборки, а $F^*(\,\cdot\,)$ – эмпирическая функция распределения.

Замечание 5. Следует помнить, что в $F^*(x)$ аргумент x неслучаен, $x \in \mathbb{R}$, но само значение $F^*(x)$ есть с.в., которая зависит от n-мерной выборки, т. е. по сути дела

$$F^*(x) = \Psi(\xi_1, \ldots, \xi_n; n; x), \tag{2.6}$$

где случайность величины Ψ определяется её зависимостью от ξ, а $x \in \mathbb{R}$ и $n \in \mathbb{N}$ выступают как параметры этой с.в. Нетрудно найти распределение с.в. Ψ при фиксированном x: очевидно, что

$$P\left(\Psi = \frac{m}{n}\right) = P(\nu_\xi(x) = m) = C_n^m p^m (1-p)^{n-m}, \quad m = 0, 1, \ldots, n, \tag{2.7}$$

где $p = P(\xi_k < x) = F(x)$ есть значение функции распределения выборки в точке x. Таким образом, $F^*(x) \sim \frac{1}{n}\mathbf{B}(n, p)$ с $p = F(x)$.

Если мы каким-то образом введём расстояние $d(F_0, F^*)$ между $F_0(\,\cdot\,)$ и $F^*(\,\cdot\,)$, то в силу зависимости функции $F^*(\,\cdot\,)$ от выборки ξ получим, что значение расстояния случайно, и при фиксированной гипотезе мы можем записать $d(F_0, F^*) = \delta(\xi)$. Разумно выбирать такие функционалы расстояния, чтобы $\delta(\xi)$ полностью определялась выборкой (при фиксированной гипотезе) и была с.в., т. е. являлась статистикой. По смыслу расстояния с вероятностью единица должно выполняться неравенство $\delta(\xi) \geqslant 0$.

Желательно, чтобы расстояние $\delta(\xi)$ удовлетворяло следующему естественному требованию: когда гипотеза неверна, вероятность

$P(\delta(\xi) > C)$ должна быть достаточно велика по сравнению с вероятностью $P_0(\delta(\xi) > C)$ при верной гипотезе. Это означает, что если гипотеза верна, то теоретическая и эмпирическая функция с большой вероятностью расположены рядом в смысле введённого расстояния, а если гипотеза неверна, то мы можем скорее ожидать большие значения этого расстояния, чем малые.

Мы будем строить нерандомизированные критерии, зависящие от выборки через статистику $\delta(\xi)$ следующим образом:

$$\varphi(x) = \begin{cases} 1, & \text{если } \delta(x) > C_0, \\ 0, & \text{если } \delta(x) \leqslant C_0. \end{cases} \tag{2.8}$$

Постоянная C_0 должна определяться из условия малости ошибки первого рода – вероятности отклонить гипотезу $\xi \sim F_0(\,\cdot\,)$, когда гипотеза верна. Если α_0 – приемлемый размер ошибки, то, решая относительно $C > 0$ уравнение

$$\alpha_0 = P_0(\delta(\xi) > C) = \int\limits_{x\colon \delta(x) > C} dF_0(x), \tag{2.9}$$

получаем искомый критерий. Он называется *критерием согласия* (с гипотезой $\xi \sim F_0(\,\cdot\,)$).

Критерии согласия почти всегда носят асимптотический характер: точные утверждения об их свойствах справедливы при $n \to \infty$, где n – размерность выборки. Для предельного при $n \to \infty$ распределения с.в. $\delta(\xi)$ при верной гипотезе находим постоянную C из уравнения

$$\alpha_0 = P_{\lim}(\delta(\xi) > C), \tag{2.10}$$

где вероятность $P_{\lim}$ отвечает этому предельному распределению, и получаем критерий согласия. Он отклоняет гипотезу всякий раз, когда $\delta(\xi) > C$, и имеет в пределе $n \to \infty$ ошибку α_0. Для многих известных критериев согласия часто можно получить важный асимптотический результат о поведении мощности критерия: при $n \to \infty$ мощность критерия, заданного формулами (2.8), (2.10), стремится к единице для любой альтернативы $F \neq F_0$.

Критерий Колмогорова. Пусть $H\colon \xi \sim F_0(\,\cdot\,)$. Будем опираться на расстояние (2.2) между теоретической и эмпирической

функциями распределения и рассмотрим статистику, отличающуюся от (2.2) только числовым множителем:

$$\delta_n(\xi) = \sqrt{n} \sup_{x \in \mathbb{R}} \big| F_0(x) - F^*(x) \big|. \tag{2.11}$$

Она называется *статистикой Коломогорова.*

Замечание 6. Множитель $\sqrt{n}$ вводится для того, чтобы при верной гипотезе в пределе $n \to \infty$ статистика $\delta_n(\xi)$ имела нетривиальное распределение. Можно показать, что без этого множителя

$$\sup_{x \in \mathbb{R}} \big| F_0(x) - F^*(x) \big| \xrightarrow{\mathrm{P}} 0, \qquad n \to \infty.$$

Это утверждение следует из того, что, как мы отмечали в замечании 5, $F^*(x) \sim \frac{1}{n}\mathbf{B}(n,p)$, следовательно, при верной гипотезе

$$MF^*(x) = p = F_0(x), \qquad DF^*(x) = \frac{pq}{n} \to 0, \quad n \to \infty,$$

поэтому $\big| F_0(x) - F^*(x) \big| \xrightarrow{\mathrm{P}} 0$ при $n \to \infty$ для любого фиксированного x. Умножение на $\sqrt{n}$ обеспечивает ненулевую и конечную дисперсию предельного распределения.

А. Н. Колмогоров нашёл предельное распределение статистики $\delta_n(\xi)$. Оказалось, что оно является одним и тем же для любой функции распределения и не зависит ни от каких параметров. Теперь это распределение называют *распределением Колмогорова.*

Теорема 6 (Колмогорова)**.** *Пусть функция $F_0(\,\cdot\,)$ не имеет разрывов и $D_n(y) = P_0(\delta_n(\xi) < y)$, $y > 0$, – функция распределения статистики* (2.11) *при верной гипотезе для фиксированного n. Тогда для любого фиксированного $y > 0$*

$$D_n(y) \underset{n\to\infty}{\longrightarrow} \sum_{k=-\infty}^{\infty} (-1)^k e^{-k^2 y^2} \stackrel{\mathrm{def}}{=} K(y).$$

Значения функции $K(y)$, $y > 0$, представлены во многих руководствах по математической статистике и вычислительных программах.

Критерий φ, зависящий от выборки через статистику $\delta_n(\xi)$ и заданный как

$$\varphi(x) = \begin{cases} 1, & \text{если } \delta_n(x) > C_0, \\ 0, & \text{если } \delta_n(x) \leqslant C_0, \end{cases} \qquad x \in \mathbb{R}^n, \tag{2.12}$$

где C_0 находится из уравнения

$$\int\limits_{y>C_0} K(y)\,dy = \alpha_0, \tag{2.13}$$

называется *критерием Колмогорова* проверки непараметрической гипотезы $\xi \sim F_0(\,\cdot\,)$ с (асимптотической при $n \to \infty$) ошибкой α_0.

Докажем, что мощность этого критерия для любой альтернативы стремится к единице при $n \to \infty$.

Теорема 7. *Пусть $\xi \sim F_1(\,\cdot\,)$, где $F_1(\,\cdot\,)$ – какая-либо функция распределения, не совпадающая с $F_0(\,\cdot\,)$ в том смысле, что*

$$\sup_{x\in\mathbb{R}} \bigl|F_0(x) - F_1(x)\bigr| \neq 0. \tag{2.14}$$

Тогда для любого фиксированного $C > 0$

$$\beta_n \stackrel{\text{def}}{=} P_1(\delta_n(\xi) > C) \underset{n\to\infty}{\longrightarrow} 1. \tag{2.15}$$

В (2.15), как обычно, нижний индекс P_1 означает, что вероятность рассчитывается для $\xi \sim F_1(\,\cdot\,)$, и отмечена зависимость мощности от n, т. е. от размера выборки.

Доказательство. При условии (2.14) найдётся $x_0 \in \mathbb{R}$, такой что

$$\bigl|F_0(x_0) - F_1(x_0)\bigr| = \Delta_0 > 0.$$

С другой стороны, если $\xi \sim F_1(\,\cdot\,)$, то имеется сходимость по вероятности $F^*(x_0) \underset{n\to\infty}{\overset{P_1}{\longrightarrow}} F_1(x_0)$, следовательно,

$$\bigl|F_0(x_0) - F^*(x_0)\bigr| \underset{n\to\infty}{\overset{P_1}{\longrightarrow}} \bigl|F_0(x_0) - F_1(x_0)\bigr| = \Delta_0$$

(напомним, что в этих предельных переходах $x_0 \in \mathbb{R}$ фиксирован, а $F^*(x_0)$ – с.в., зависящая от n-мерной выборки ξ, см. формулу (2.5)

и замечание 5). Перепишем в развёрнутом виде последний предел: для любого $\varepsilon > 0$

$$P_1\Big(\Big|\,\big|F_0(x_0) - F^*(x_0)\big| - \Delta_0\Big| < \varepsilon\Big) = \\ = P_1\Big(\Delta_0 - \varepsilon < \big|F_0(x_0) - F^*(x_0)\big| < \Delta_0 + \varepsilon\Big) \underset{n\to\infty}{\longrightarrow} 1. \quad (2.16)$$

Выберем и зафиксируем положительное $\varepsilon < \Delta_0$. Введём три последовательности случайных событий

$$A_n = \Big\{\xi\colon\ \Delta_0 - \varepsilon < \big|F_0(x_0) - F^*(x_0)\big| < \Delta_0 + \varepsilon\Big\},$$
$$B_n = \Big\{\xi\colon\ \big|F_0(x_0) - F^*(x_0)\big| > \Delta_0 - \varepsilon\Big\},$$
$$C_n = \Big\{\xi\colon\ \sup_{x\in\mathbb{R}}\big|F_0(x) - F^*(x)\big| > \Delta_0 - \varepsilon\Big\}.$$

Здесь $n = 1, 2\ldots$, а зависимость от n-мерной выборки ξ содержится в с.в. $F^*(x_0)$ и $F^*(x)$. Очевидно, что A_n влечёт B_n, а B_n влечёт C_n. Поэтому

$$P_1(A_n) \leqslant P_1(B_n) \leqslant P_1(C_n) \leqslant 1.$$

Поскольку $P_1(A_n) \to 1$ при $n \to \infty$ в силу (2.16), мы получаем $P_1(C_n) \to 1$ при $n \to \infty$.

Итак, существует число $\Delta_1 = \Delta_0 - \varepsilon > 0$, определяющееся только параметром Δ_0, т. е. функциями распределения $F_1(\,\cdot\,)$ и $F_0(\,\cdot\,)$, такое что

$$P_1\Big(\sup_{x\in\mathbb{R}}\big|F_0(x) - F^*(x)\big| > \Delta_1\Big) \underset{n\to\infty}{\longrightarrow} 1.$$

Статистика Колмогорова $\delta_n(\xi)$ отличается от точной верхней грани в последнем соотношении только множителем $\sqrt{n}$, поэтому

$$P_1\Big(\delta_n(\xi) > \sqrt{n}\Delta_1\Big) = P_1\Big(\sqrt{n}\sup_{x\in\mathbb{R}}\big|F_0(x) - F^*(x)\big| > \sqrt{n}\Delta_1\Big) \underset{n\to\infty}{\longrightarrow} 1.$$

Нам осталось переписать полученный предел в терминах мощности (2.15) критерия Колмогорова. Пусть $C > 0$ фиксировано. Найдём и зафиксируем натуральное N_1, такое что $\sqrt{N_1}\Delta_1 > C$. Тогда для любого $n \geqslant N_1$ мы имеем цепочку следствий

$$\delta_n(\xi) > \sqrt{n}\Delta_1 \quad\Longrightarrow\quad \delta_n(\xi) > \sqrt{N_1}\Delta_1 \quad\Longrightarrow\quad \delta_n(\xi) > C,$$

откуда $P_1(\delta_n(\xi) > C) \geqslant P_1(\delta_n(\xi) > \sqrt{n}\Delta_1)$ и

$$\lim_{n\to\infty} \beta_n = \lim_{n\to\infty} P_1(\delta_n(\xi) > C) = \lim_{\substack{n\to\infty,\\ n\geqslant N_1}} P_1(\delta_n(\xi) > C) \geqslant$$
$$\geqslant \lim_{\substack{n\to\infty,\\ n\geqslant N_1}} P_1(\delta_n(\xi) > \sqrt{n}\Delta_1) = 1. \tag{2.17}$$

Поскольку $\beta_n \leqslant 1$ для любого n, получаем утверждение теоремы.

Критерий хи-квадрат (Пирсона). Этот критерий, пожалуй, используется наиболее часто. Для его использования выборочные данные группируют, разбивая множество $\mathbb{R}$ на r непересекающихся подмножеств ΔX_k, $k = 1, \ldots, r$ (как правило, в качестве этих подмножеств выбирают интервалы), и подсчитывают количество элементов выборки, попавших в каждое из ΔX_k:

$$\nu_k = \nu_k(\xi) = \sum_{j=1}^{n} I_{\Delta X_k}(\xi_j), \quad k = 1, \ldots, r, \qquad \sum_{k=1}^{r} \nu_k = n. \tag{2.18}$$

Замечание 7. Статистика $\nu = (\nu_1, \ldots, \nu_r)$ имеет полиномиальное распределение: для $m = (m_1, \ldots, m_r)$, где $m_1 \geqslant 0, \ldots, m_r \geqslant 0$ и $m_1 + \cdots + m_r = n$,

$$P(\nu = m) = P(\nu_1 = m_1, \ldots, \nu_r = m_r) = \frac{n!}{m_1! \ldots, m_r!} p_1^{m_1} \ldots p_r^{m_r}.$$

Здесь $p_k = P(\xi_j \in \Delta X_k)$, $k = 1, \ldots, r$, для любой с.в. ξ_j из выборки $\xi = (\xi, \ldots, \xi_n)$ и, конечно, $p_1 + \cdots + p_r = 1$.

Вернёмся к проверке гипотезы $\xi \sim F_0(\,\cdot\,)$. Введём обозначение $p_k = P_0(\xi_j \in \Delta X_k)$ для $k = 1, \ldots, r$ и любого фиксированного $j = 1, \ldots, n$. Поскольку $\xi_1, \ldots, \xi_n$ одинаково распределены, вероятность p_k не зависит от j. Из с.в. $\nu_k(\xi)$, заданных в (2.18), и вероятностей p_k составим статистику

$$\chi^2(\xi) = \sum_{k=1}^{r} \frac{(\nu_k(\xi) - np_k)^2}{np_k}. \tag{2.19}$$

Для простоты последующих формул опустим зависимость от ξ и будем писать

$$\chi^2 = \sum_{k=1}^{r} \frac{(\nu_k - np_k)^2}{np_k}. \tag{2.20}$$

Часто можно встретить другие формы записи статистики хи-квадрат:

$$\chi^2 = \sum_{k=1}^{r} \frac{(O_k - E_k)^2}{E_k} = n \sum_{k=1}^{r} \frac{(\nu_k/n - p_k)^2}{p_k}. \tag{2.21}$$

В первой сумме просто введены обозначения $O_k = \nu_k$ и $E_k = np_k$ для наблюдаемого (observed) количества попаданий в интервал ΔX_k и для его математического ожидания (expectation). Вторая сумма получается после тривиальных алгебраических преобразований статистики (2.20) и показывает, что χ^2 контролирует определённым образом заданное расстояние между эмпирическими частотами ν_k/n и теоретическими вероятностями p_k, $k = 1, \ldots, r$.

Справедлива следующая теорема.

Теорема 8. *Если гипотеза $\xi \sim F_0(\,\cdot\,)$ верна, то при $n \to \infty$ распределение статистики* (2.19) *стремится к распределению хи-квадрат с $r-1$ степенями свободы.*

Проведём доказательство в простейшем случае $r = 2$.

Доказательство. Поскольку $\nu_1 + \cdots + \nu_r = n$ и $p_1 + \cdots + p_r = 1$, в данном случае имеем $\nu_2 = n - \nu_1$ и $p_2 = 1 - p_1$. Тогда

$$\begin{aligned}
\chi^2 &= \frac{(\nu_1 - np_1)^2}{np_1} + \frac{(\nu_2 - np_2)^2}{np_2} = \\
&= \frac{(\nu_1 - np_1)^2}{np_1} + \frac{(n - \nu_1 - n(1-p_1))^2}{n(1-p_1)} = \\
&= \frac{(1-p_1)(\nu_1 - np_1)^2 + p_1(\nu_1 - np_1)^2}{np_1(1-p_1)} = \frac{(\nu_1 - np_1)^2}{np_1(1-p_1)}.
\end{aligned}$$

При этом $\nu_1 = \nu_1(\xi) = \sum_{j=1}^{n} I_{\Delta X_1}(\xi_j)$, где с.в. $I_{\Delta X_1}(\xi_1), \ldots, I_{\Delta X_1}(\xi_n)$ независимы, одинаково распределены,

$$P_0(I_{\Delta X_1}(\xi_j) = 1) = P_0(\xi_j \in \Delta X_1) = p_1, \quad P_0(I_{\Delta X_1}(\xi_j) = 0) = 1 - p_1,$$

таким образом, $MI_{\Delta X_1}(\xi_j) = p_1$ и $DI_{\Delta X_1}(\xi_j) = p_1(1-p_1)$. Для краткости обозначив $I_{\Delta X_1}(\xi_j)$ просто как I_j, имеем

$$\frac{(\nu_1 - np_1)^2}{np_1(1-p_1)} = \left(\sum_{j=1}^{n} \frac{I_j - MI_j}{\sqrt{n \cdot DI_j}} \right)^2.$$

По центральной предельной теореме имеется сходимость по распределению

$$\sum_{j=1}^{n} \frac{I_j - MI_j}{\sqrt{n \cdot DI_j}} \xrightarrow[n\to\infty]{\mathrm{d}} \nu_* \sim \mathbf{N}(0,1).$$

Отсюда тривиальным образом получаем, что для любого $y > 0$

$$P_0\left(\frac{(\nu_1 - np_1)^2}{np_1(1-p_1)} < y\right) = P_0\left(\left|\sum_{j=1}^{n} \frac{I_j - MI_j}{\sqrt{n \cdot DI_j}}\right| < \sqrt{y}\right) \underset{n\to\infty}{\longrightarrow}$$
$$\underset{n\to\infty}{\longrightarrow} P(\,|\nu_*| < \sqrt{y}\,) = P(\nu_*^2 < y) = F_{\chi_1^2}(y),$$

где $F_{\chi_1^2}(\,\cdot\,)$ – функция распределения хи-квадрат с одной степенью свободы. Теорема доказана.

Доказательство для общего r основано на тех же идеях, но технически более сложное.

Критерий хи-квадрат или, иначе, критерий Пирсона задаётся как следующий критерий φ, зависящий от выборки через статистику $\chi^2 = \chi^2(\xi)$ и обладающий (асимптотической при $n \to \infty$) ошибкой α_0:

$$\varphi(x) = \begin{cases} 1, & \text{если } \chi^2(x) > C_0, \\ 0, & \text{если } \chi^2(x) \leqslant C_0, \end{cases} \qquad x \in \mathbb{R}^n, \tag{2.22}$$

где C_0 находится из уравнения

$$\int_{y > C_0} p_{\chi^2_{r-1}}(y)\, dy = \alpha_0. \tag{2.23}$$

В критерии (2.22) статистика $\chi^2(\xi)$ определяется формулами (2.19), (2.18), а $p_{\chi^2_{r-1}}(\,\cdot\,)$ – плотность вероятности для распределения хи-квадрат с $r-1$ степенями свободы.

Заметим, что критерий хи-квадрат в сущности проверяет не гипотезу $\xi \sim F_0(\,\cdot\,)$, а тот факт, что

$$p_k = P_0(\xi \in \Delta X_k) = \int\limits_{x \in \Delta X_k} dF_0(x), \qquad k = 1, \ldots, r. \tag{2.24}$$

В частности, если мы имеем другое распределение, $\xi \sim F(\,\cdot\,)$, но $P(\xi \in \Delta X_k) = p_k$ для всех $k = 1, \ldots, r$, то с помощью критерия

хи-квадрат такое распределение неотличимо от исходного. Поэтому говорить о мощности этого критерия можно только для тех альтернатив $\xi \sim F(\,\cdot\,)$, у которых вектор вероятностей

$$p' = \langle p'_1, \ldots, p'_r \rangle, \qquad p'_k = \int\limits_{x \in \Delta X_k} dF(x), \quad k = 1, \ldots, r,$$

не совпадает с вектором $p = \langle p_1, \ldots, p_r \rangle$, заданным в (2.24).

Справедлива теорема, аналогичная теореме (7).

Теорема 9. *Пусть для гипотезы вектор $p = \langle p_1, \ldots, p_r \rangle$ не совпадает с вектором $p' = \langle p'_1, \ldots, p'_r \rangle$ для альтернативы в том смысле, что $p_{k_0} \neq p'_{k_0}$ хотя бы при одном k_0. Тогда для любого фиксированного $C > 0$*

$$\beta_n \stackrel{\text{def}}{=} P'\left(\sum_{k=1}^{r} \frac{(\nu_k - np_k)^2}{np_k} > C \right) \underset{n\to\infty}{\longrightarrow} 1. \tag{2.25}$$

В (2.25) вероятность $P'(\,\cdot\,)$ рассчитывается при условии, что вектор $\nu = \langle \nu_1, \ldots, \nu_r \rangle$ эмпирических частот имеет полиномиальное распределение (см. замечание 7), отвечающее альтернативе, т. е. $P'(\xi_j \in \Delta X_k) = p'_k$ для $k = 1, \ldots, r$ и любой с.в. ξ_j из выборки ξ. Зависимость от n скрыта, в том числе, в зависимости $\nu_k = \nu_k(\xi)$ от выборки $\xi = (\xi, \ldots, \xi_n)$.

Доказательство. Предположим, что верна альтернатива. Частота сходится к вероятности:

$$\frac{\nu_{k_0}}{n} \underset{n\to\infty}{\xrightarrow{P'}} p'_{k_0}, \qquad \frac{(\nu_{k_0}/n - p_{k_0})^2}{p_{k_0}} \underset{n\to\infty}{\xrightarrow{P'}} \frac{(p'_{k_0} - p_{k_0})^2}{p_{k_0}} \stackrel{\text{def}}{=} \Delta X_0 > 0.$$

Далее, учитывая цепочку следствий

$$\left| \frac{(\nu_{k_0}/n - p_{k_0})^2}{p_{k_0}} - \Delta X_0 \right| < \varepsilon \quad \Longleftrightarrow$$

$$\Longleftrightarrow \quad \Delta X_0 - \varepsilon < \frac{(\nu_{k_0}/n - p_{k_0})^2}{p_{k_0}} < \Delta X_0 + \varepsilon \quad \Longrightarrow$$

$$\Longrightarrow \quad \frac{(\nu_{k_0}/n - p_{k_0})^2}{p_{k_0}} > \Delta X_0 - \varepsilon = \Delta X_1 \quad \Longrightarrow$$

$$\Longrightarrow \quad \sum_{k=1}^{r} \frac{(\nu_k/n - p_k)^2}{p_k} > \Delta X_1$$

и представляя статистику хи-квадрат как $\chi^2 = n \sum\limits_{k=1}^{r} \frac{(\nu_k/n - p_k)^2}{p_k}$, получаем

$$P'(\chi^2 > n\Delta X_1) = P'\left(\sum_{k=1}^{r} \frac{(\nu_k/n - p_k)^2}{p_k} > \Delta X_1 \right) \geqslant$$
$$\geqslant P'\left(\left| \frac{(\nu_{k_0}/n - p_{k_0})^2}{p_{k_0}} - \Delta X_0 \right| < \varepsilon \right) \underset{n\to\infty}{\longrightarrow} 1.$$

Отсюда аналогично (2.17) имеем $\beta_n \to 1$.

При достаточно мелком разбиении на интервалы $\Delta X_1, \ldots, \Delta X_r$ мы можем считать, что $p_1, \ldots, p_r$ при больших r достаточно хорошо описывают распределение гипотезы $\xi \sim F_0(\,\cdot\,)$. Но при фиксированном размере n нашей выборки, может случиться так, что в интервал ΔX_k попадает мало элементов выборки, и тогда частота ν_k/n может сильно отличаться от вероятности (2.24) даже в случае верной гипотезы. Это приведёт к большому значению статистики хи-квадрата и к отклонению гипотезы.

Критерий хи-квадрат при наличии мешающих параметров. Мы знаем, что распределения часто зависят от некоторых параметров, которые исследователю неизвестны. Если нас по-прежнему интересует общий вид распределения, а не значения параметров, то мы получаем сложную гипотезу

$$H\colon\ \xi \sim F_0(\,\cdot\,;\theta), \qquad \theta = (\theta_1, \ldots, \theta_m) \in \Theta \subset \mathbb{R}^m, \tag{2.26}$$

в которой параметр θ исследователя не интересует, но увеличивает сложность решаемой задачи. Поэтому этот параметр часто называют мешающим. Например, требуется проверить предположение, что выборка распределена нормально, $\xi \sim \mathbf{N}(\mu, \sigma^2)$ с какими-либо значениями μ и σ^2.

При проверке сложной гипотезы (2.26) нам придётся учесть, что статистика хи-квадрат приобретает зависимость от θ:

$$\chi^2 = \chi^2(\xi;\theta) = \sum_{k=1}^{r} \frac{[\nu_k(\xi) - np_k(\theta)]^2}{np_k(\theta)}, \tag{2.27}$$

$$p_k(\theta) = \int\limits_{x\in\Delta_k} dF_0(x;\theta), \qquad k = 1, \ldots, r. \tag{2.28}$$

Естественная идея решения этой проблемы – сначала получить оценку $\hat{\theta} = \hat{\theta}(\xi)$ значения параметра имеющейся выборке, а потом применить критерий хи-квадрат, в котором гипотетические вероятности рассчитываются при $\theta = \hat{\theta}$. Однако в этом случае найти распределение статистики (2.27) и, следовательно, вычислить ошибки критерия гораздо труднее, чем при отсутствии мешающего параметра: χ^2 зависит от выборки ξ не только через с.в. $\nu_1, \ldots, \nu_r$, но и через оценку параметра $\hat{\theta}$.

При определённых условиях на гладкость функций $p_k(\theta)$, $\theta \in \Theta$, справедлива следующая теорема (Фишера).

Теорема 10. *Пусть в задаче проверки гипотезы* (2.26) *с m-мерным параметром $\theta \in \Theta$ оценка параметра $\theta_*(\xi)$ определяется как точка минимума статистики хи-квадрат для каждой реализации $\xi = (x_1, \ldots, x_n)$,*

$$\chi^2(x_1, \ldots, x_n; \theta_*) = \min_{\theta \in \Theta} \chi^2(x_1, \ldots, x_n; \theta). \tag{2.29}$$

Если гипотеза (2.26) *верна, то при $n \to \infty$ распределение статистики*

$$\chi^2(\xi; \theta_*) = \sum_{k=1}^{r} \frac{[\nu_k(\xi) - np_k(\theta_*)]^2}{np_k(\theta_*)}, \qquad \theta_* = \theta_*(\xi),$$

стремится к распределению хи-квадрат с $r - 1 - m$ степенями свободы.

Доказательство теоремы математически очень глубокое и весьма непростое технически, поэтому мы его не приводим, но можем сделать некоторые выводы.

Мы видим, что асимптотическое распределение статистики χ^2 по-прежнему есть распределение хи-квадрат, но поменялось число степеней свободы. В простой гипотезе без параметра θ число степеней свободы было равно $r - 1$. Одна степень свободы «пропала», потому что с.в. $\nu_1, \ldots, \nu_r$ не меняются независимо, а подчинены условию $\nu_1 + \cdots + \nu_r = n$. Когда m параметров $\theta_1, \ldots, \theta_m$ оценивается по выборке, мы «теряем» ещё m степеней свободы и получаем число степеней свободы $r - 1 - m$.

Критерий хи-квадрат проверки гипотезы (2.26) теперь можно задать как

$$\varphi(x) = \begin{cases} 1, & \text{если } \chi^2(x;\theta_*) > C_0, \\ 0, & \text{если } \chi^2(x;\theta_*) \leqslant C_0, \end{cases} \qquad x \in \mathbb{R}^n,$$

$$\int\limits_{y>C_0} p_{\chi^2_{r-1-m}}(y)\,dy = \alpha_0.$$

Решение задачи на минимум (2.29) может оказаться сложным и, как правило, получается численными методами. При достаточно гладких функциях $p_k(\theta)$, $\theta \in \Theta$, нам требуется решить систему уравнений

$$-\frac{n}{2}\frac{\partial \chi^2}{\partial \theta_j} = \sum_{k=1}^{r}\left(\frac{\nu_k - np_k(\theta)}{p_k(\theta)} + \frac{[\nu_k - np_k(\theta)]^2}{p_k^2(\theta)}\right)\frac{\partial p_k(\theta)}{\partial \theta_j} = 0, \quad (2.30)$$

где $j = 1, \ldots, m$.

Строго говоря, в уравнениях (2.30) следует писать n_k вместо с.в. $\nu_k = \nu_k(\xi)$, где неотрицательные целые числа n_k, $k = 1, \ldots, r$, удовлетворяют условию $n_1 + \cdots + n_r = n$ и тогда определяют реализацию с.в. $\nu = (\nu_1, \ldots, \nu_r)$. Однако, чтобы не вводить новые обозначения, далее мы пишем ν_k, имея в виду, что уравнения должны быть выполнены для любой фикисированниой реализации с.в. ν, т. е. с вероятностью единица при любом распределении выборки ξ.

Если $\xi \sim F_0(\,\cdot\,;\theta)$, то $\nu_k \approx np_k(\theta)$ при больших n, точнее, мы можем утверждать, что $P_\theta(|\nu_k/n - p_k(\theta)| < \varepsilon)$ стремится к единице при $n \to \infty$ для любого $\varepsilon > 0$). Тогда вторым, квадратичным по $\nu_k - np_k(\theta)$, слагаемым в высоких круглых скобках в (2.30) можно пренебречь по сравнению с первым, линейным $\nu_k - np_k(\theta)$. Получаем систему

$$\sum_{k=1}^{r}\frac{\nu_k - np_k(\theta)}{p_k(\theta)}\,\frac{\partial p_k(\theta)}{\partial \theta_j} = 0, \qquad j = 1, \ldots, m. \quad (2.31)$$

Заметим, что $p_1(\theta) + \cdots + p_k(\theta) = 1$ для любого $\theta \in \Theta$, поэтому

$$\sum_{k=1}^{r}\frac{np_k(\theta)}{p_k(\theta)}\,\frac{\partial p_k(\theta)}{\partial \theta_j} = n\sum_{k=1}^{r}\frac{\partial p_k(\theta)}{\partial \theta_j} = n\frac{\partial}{\partial \theta_j}\sum_{k=1}^{r}p_k = n\frac{\partial}{\partial \theta_j}1 = 0,$$

и последняя система превращается в

$$\sum_{k=1}^{r} \frac{\nu_k}{p_k(\theta)} \frac{\partial p_k(\theta)}{\partial \theta_j} = 0, \qquad j = 1, \dots, m.$$

Её можно переписать как

$$\sum_{k=1}^{r} \frac{\partial}{\partial \theta_j} \ln[p_k(\theta)]^{\nu_k} = \frac{\partial}{\partial \theta_j} \ln\Bigg(\prod_{j=1}^{r} [p_k(\theta)]^{\nu_k} \Bigg) = 0,$$

где $j = 1, \dots, m$. Эта система эквивалента системе уравнений максимального правдоподобия

$$\frac{\partial}{\partial \theta_j} \ln L(\nu; \theta) = 0, \qquad j = 1, \dots, m, \tag{2.32}$$

когда $L(\,\cdot\,; \theta)$ – функция правдоподобия полиномиального распределения с.в. ν:

$$\begin{aligned} L(n_1, \dots, n_r; \theta) &= P_\theta(\nu_1 = n_1, \dots, \nu_r = n_r) = \\ &= \frac{n_1! \,\dots\, n_r!}{n!} [p_1(\theta)]^{n_1} \dots [p_r(\theta)]^{n_r} \end{aligned}$$

для $n_k \geqslant 0$, $k = 1, \dots, r$, с условием $n_1 + \dots + n_r$.

Таким образом, если обосновать замену системы (2.30) на систему (2.32), мы можем искать оценку мешающего параметра θ как оценку $\hat{\theta} = \hat{\theta}(\nu)$ максимально правдоподобия статистики $\nu = \nu(\xi)$. Теорема Фишера с заменой θ_* на $\hat{\theta}$ остаётся справедливой.

Критерий омега-квадрат. Существует много других критериев согласия. Некоторые из них представляют собой модификации и уточнения критериев Коломогорова и хи-квадрат, некоторые основаны на статистиках, которые также контролируют расстояние между теоретическими и эмпирическими вероятностями, но отличаются от статистик Колмогорова и хи-квадрат. Один из таких критериев – это критерий омега-квадрат, который также называют критерием Крамера–Мизеса–Смирнова. Рассмотрим его подробнее.

Пусть $\xi = (\xi_1, \dots, \xi_n)$ – выборка, как обычно, составленная из независимых одинаково распределённых с.в. Пусть $F(\,\cdot\,)$ – одномерная функция распределения любого элемента выборки, а $F^*(\,\cdot\,)$ –

выборочная функция распределения, полученная по выборке объёма n. Введём статистику, основанную на интегральной квадратичной метрике:

$$\omega^2 = n \int_{-\infty}^{\infty} [F(x) - F^*(x)]^2 \, dF(x). \tag{2.33}$$

Её можно привести к более удобному для расчётов виду. Предположим, что значения выборки, полученные в эксперименте, суть $\xi_1 = x_1, \ldots, \xi_n = x_n$ и упорядочим их по возрастанию (построим вариационный ряд), члены которого обозначим стандартным образом как $x_{(k)}$, $k = 1, \ldots, n$. Если $F(\,\cdot\,)$ непрерывна в каждой точке, то $P(\xi_i = \xi_j) = 0$ для любых $i \neq j$. Тогда мы можем считать, что члены вариационного ряда строго возрастают,

$$x_{(1)} < \cdots < x_{(n)}.$$

После этого легко записать выборочную функцию распределения:

$$F^*(x) = \begin{cases} 0, & x \leqslant x_{(1)}, \\ k/n, & x_{(k)} < x \leqslant x_{(k+1)}, \quad k = 1, \ldots, n-1, \\ 1, & x > x_{(n)}. \end{cases} \tag{2.34}$$

Теперь найдём интеграл в (2.33):

$$\begin{aligned} I = \int_{-\infty}^{\infty} [F(x) - F^*(x)]^2 \, dF(x) &= \int_{-\infty}^{\infty} F^2(x) \, dF(x) - \\ &- 2 \int_{-\infty}^{\infty} F(x) F^*(x) \, dF(x) + \int_{-\infty}^{\infty} [F^*(x)]^2 \, dF(x) = \\ = I_0 - 2I_1 + I_2. \end{aligned}$$

Для I_0 с учётом $F(-\infty) = 0$, $F(\infty) = 1$ имеем

$$I_0 = \int_{-\infty}^{\infty} F^2(x) \, dF(x) = \frac{1}{3} F^3(x) \Big|_{-\infty}^{\infty} = \frac{1}{3}. \tag{2.35}$$

Чтобы найти I_1, разобьём область интегрирования на интервалы точками $x_{(k)}$, $k = 1, \ldots, n$, и подставим (2.34). Получим

$$\begin{aligned} I_1 &= \sum_{k=1}^{n-1} \frac{k}{n} \int_{x_{(k)}}^{x_{(k+1)}} F(x) \, dF(x) + \int_{x_{(n)}}^{\infty} F(x) \, dF(x) = \\ &= \frac{1}{2n} \sum_{k=1}^{n-1} k(F_{k+1}^2 - F_k^2) + 1 - \frac{1}{2} F_n^2, \end{aligned}$$

где введено краткое обозначение $F(x_{(k)}) = F_k$, $k = 1, \ldots, n$. Найдём сумму в правой части:

$$\sum_{k=1}^{n-1} k(F_{k+1}^2 - F_k^2) = \sum_{k=1}^{n-1}\big((k+1)F_{k+1}^2 - kF_k^2\big) - \sum_{k=1}^{n-1} F_{k+1}^2 = \\ = \sum_{k=1}^{n-1}(k+1)F_{k+1}^2 - \sum_{k=1}^{n-1} kF_k^2 - \sum_{k=2}^{n} F_k^2.$$

Мы видим, что в разности первой и второй суммы все слагаемые, кроме крайних, сокращаются,

$$\sum_{k=1}^{n-1}(k+1)F_{k+1}^2 - \sum_{k=1}^{n-1} kF_k^2 = nF_n^2 - 1 \cdot F_1^2,$$

в результате для I_1 получаем

$$I_1 = \frac{1}{2n} \cdot nF_n^2 - \frac{1}{2n} \cdot F_1^2 - \frac{1}{2n}\sum_{k=2}^{n} F_k^2 + 1 - \frac{1}{2}F_n^2.$$

Первое и последнее слагаемые в правой части сокращаются, а второе можно включить в сумму. Итак, получаем

$$I_1 = 1 - \frac{1}{2n}\sum_{k=1}^{n} F_k^2. \tag{2.36}$$

Интеграл I_2 вычисляется аналогично:

$$I_2 = \sum_{k=1}^{n-1} \frac{k^2}{n^2} \int_{x_{(k)}}^{x_{(k+1)}} dF(x) + \int_{x_{(n)}}^{\infty} dF(x) = \\ = \frac{1}{n^2}\sum_{k=1}^{n-1} k^2(F_{k+1} - F_k) + 1 - F_n = \\ = \frac{1}{n^2}\left(\sum_{k=1}^{n-1}\big((k+1)^2(F_{k+1} - F_k) - \sum_{k=1}^{n-1}(2k+1)F_{k+1}\right) + 1 - F_n = \\ = \frac{1}{n^2}(n^2F_n - 1^2 \cdot F_1) - \frac{1}{n^2}\sum_{k=1}^{n-1}(2k+1)F_{k+1} + 1 - F_n.$$

После сокращения членов $\pm F_n$ и замены индекса суммирования k на $k-1$ окончательно получаем

$$I_2 = 1 - \frac{1}{2n} - \frac{1}{n^2}\sum_{k=1}^{n-1}(2k-1)F_k. \tag{2.37}$$

Собирая результаты (2.35)–(2.37), имеем

$$\begin{aligned} I &= \frac{1}{3} - 1 + \frac{1}{n}\sum_{k=1}^{n} F_k^2 + 1 - \frac{1}{2n} - \frac{1}{n^2}\sum_{k=1}^{n-1}(2k-1)F_k = \\ &= \frac{1}{3} + \frac{1}{n}\sum_{k=1}^{n}\left(F_k^2 - \frac{2k-1}{n}F_k\right). \end{aligned}$$

В принципе это уже достаточно удобное для расчёта выражение, но мы можем привести его ещё более компактному и наглядному виду, если дополним выражение в высоких круглых скобках до полного квадрата,

$$I = \frac{1}{3} + \frac{1}{n}\sum_{k=1}^{n}\left(F_k - \frac{2k-1}{2n}\right)^2 - \frac{1}{n}\sum_{k=1}^{n}\left(\frac{2k-1}{2n}\right)^2, \tag{2.38}$$

и вычислим в явном виде последнюю сумму:

$$\begin{aligned} \sum_{k=1}^{n}\left(\frac{2k-1}{2n}\right)^2 &= \frac{1}{n^2}\sum_{k=1}^{n}\left(k - \frac{1}{2}\right)^2 = \\ &= \frac{1}{n^2}\left(\sum_{k=1}^{n} k^2 - \sum_{k=1}^{n} k + \frac{1}{4}\sum_{k=1}^{n} 1\right) = \\ &= \frac{1}{n^2}\left(\frac{n(n+1)(2n+1)}{6} - \frac{n(n-1)}{2} + \frac{n}{4}\right) = \\ &= \frac{1}{n^2}\frac{4n^3+n}{12} = \frac{n}{3} + \frac{1}{12n}. \end{aligned}$$

Подставим этот результат в (2.38) и получим

$$I = \frac{1}{12n^2} + \frac{1}{n}\sum_{k=1}^{n}\left(F_k - \frac{2k-1}{2n}\right)^2. \tag{2.39}$$

Если посмотреть внимательно на слагаемые суммы, то мы увидим, что выражение в высоких круглых скобках равно

$$F(x_{(k)}) - \frac{1}{2}\left(\frac{k-1}{n} + \frac{k}{n}\right) = F(x_{(k)}) - \frac{F^*(x_{(k-1)}) + F^*(x_{(k)})}{2}$$

(считаем, что $F(x_{(0)}) = 0$), таким образом, интеграл I контролирует расстояние между значением теоретической функции распределения и среднего арифметического двух соседних значений выборочной функции. Если распределение выборки действительно описывается непрерывной функцией $F(\,\cdot\,)$, то эта разность должна быть мала при больших n.

Статистика омега-квадрат (2.33) записывается следующим образом:

$$\omega^2 = \omega^2(\xi) = \frac{1}{12n} + \sum_{k=1}^{n}\left(F(\xi_{(k)}) - \frac{2k-1}{2n}\right)^2, \qquad (2.40)$$

где $\xi_{(1)} < \cdots < \xi_{(n)}$ есть вариационный ряд выборки. Поиск распределения этой статистики в случае $\xi \sim F(\,\cdot\,)$ является сложной задачей, но его асимптотика при $n \to \infty$ известна, она не зависит от распределения выборки. Это даёт основания сформулировать критерий стандартным образом:

$$\varphi(x) = \begin{cases} 1, & \text{если } \omega^2(x) > C_0, \\ 0, & \text{если } \omega^2(x) \leqslant C_0, \end{cases} \qquad x \in \mathbb{R}^n,$$

$$\int\limits_{y > C_0} p_{\omega^2}(y)\,dy = \alpha_0.$$

Здесь $p_{\omega^2}(\,\cdot\,)$ – асимптотическая плотность вероятности статистики омега-квадрат (2.40), когда $n \to \infty$ и $\xi \sim F(\,\cdot\,)$.

По сравнению с критерием хи-квадрат критерий омега-квадрат обладает тем преимуществом, что не требует группировки данных по r интервалам разбиения, но он значительно более вычислительно затратный, поскольку необходимо упорядочить элементы выборки, что при больших её объёмах может потребовать значительных ресурсов.

Подведём итог. Помимо рассмотренных нами подробно классических критериев Колмогорова, Пирсона (хи-квадрат) и Крамера–

Мизеса–Смирнова (омега-квадрат) имеется много других критериев согласия. Некоторые из них представляют собой просто уточнение или модификацию классических критериев, некоторые основаны на новых статистиках. Все эти критерии определяются статистиками, так или иначе контролирующими расхождение между теоретическими и эмпирическим вероятностями. Статистики должны вести себя «правильно», т. е. иметь известное распределение в случае верной гипотезы и принимать преимущественно большие значения, если гипотеза неверна. Эти свойства, как правило, носят асимптотический характер: они хорошо работают при больших объёмах выборки, на практике их обычно используют, если объём выборки $n \gtrsim 50$.

2.2. Критерии однородности

Критерии однородности служат для проверки гипотез о совпадении распределений или параметров распределений нескольких (в простейшем случае двух) независимых выборок. Такие задачи возникают при проверке предположения о том, что на одну выборку оказано некоторое воздействие, в другой, контрольной, это воздействие отсутствует. Предполагается, что воздействие изменяет распределение первой выборки по сравнению с распределением контрольной выборки.

Сначала рассмотрим две задачи проверки гипотез о совпадении параметров распределений.

Критерий Стьюдента. Предположим, что даны две независимые выборки $\xi = (\xi_1, \ldots, \xi_n)$ и $\xi' = (\xi'_1, \ldots, \xi'_m)$, где $n, m > 1$. Независимость выборок означает, что ξ и ξ' – независимые многомерные с.в. и, как обычно, мы предполагаем, что с.в., составляющие каждую выборку в отдельности, независимы и одинаково распределены.

Пусть с.в. ξ_i, $i = 1, \ldots, n$, имеют нормальное распределение со средним μ, а с.в. ξ'_j, $j = 1, \ldots, m$, имеют нормальное распределение со средним μ'. Дисперсию σ^2 мы считаем одинаковой для обеих выборок, но неизвестной. Поставим задачу проверки гипотезы о равенстве средних:

$$H\colon\ \mu = \mu', \qquad K\colon\ \mu \neq \mu'.$$

Введём обозначения для выборочных средних и дисперсии:

$$\begin{aligned} \bar\xi &= \frac{1}{n}\sum_{i=1}^{n}\xi_i, & \bar\xi' &= \frac{1}{m}\sum_{j=1}^{m}\xi'_i, \\ \hat\sigma_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\bar\xi)^2, & \hat\sigma_m^2 &= \frac{1}{m-1}\sum_{j=1}^{m}(\xi'_j-\bar\xi')^2. \end{aligned} \tag{2.41}$$

Напомним также, что выборочные дисперсии можно преобразовать как

$$\begin{aligned} \hat\sigma_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\bar\xi)^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n}\xi_i^2 - 2\bar\xi\sum_{i=1}^{n}\xi_i + n(\bar\xi)^2\Big) = \\ &= \frac{1}{n-1}\Big(\sum_{i=1}^{n}\xi_i^2 - 2n(\bar\xi)^2 + n(\bar\xi)^2\Big) = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i^2-(\bar\xi)^2) \end{aligned}$$

и аналогично

$$\hat\sigma_m^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\xi'_j-\bar\xi')^2 = \frac{1}{m-1}\sum_{j=1}^{m}((\xi'_j)^2-(\bar\xi')^2).$$

Также введём обозначения

$$\mathring{\xi}_i = \frac{\xi_i-\mu}{\sqrt{\sigma^2}}, \quad i=1,\dots,n, \qquad \mathring{\xi}'_i = \frac{\xi'_j-\mu'}{\sqrt{\sigma^2}}, \quad j=1,\dots,m,$$

тогда все эти с.в. независимы и каждая из них имеет стандартное нормальное распределение $\mathbf{N}(0,1)$.

Далее воспользуемся известными распределениями.

Теорема 11. *Статистика*

$$t(\xi,\xi') = \frac{\sqrt{\frac{nm}{n+m}}\big[(\bar\xi-\mu)-(\bar\xi'-\mu')\big]}{\sqrt{\dfrac{(n-1)\hat\sigma_n^2+(m-1)\hat\sigma_m^2}{n+m-2}}} \tag{2.42}$$

имеет распределение Стьюдента $\mathbf{T}_{n+m-2}$ *с* $n+m-2$ *степенями свободы.*

Доказательство. Рассмотрим наши выборки как единый $(n+m)$-мерный случайный вектор и положим

$$\vec{\xi} = \langle \xi_1, \ldots, \xi_n, \xi'_1, \ldots, \xi'_m \rangle, \qquad \vec{\xi}^{\,\circ} = \langle \mathring{\xi}_1, \ldots, \mathring{\xi}_n, \mathring{\xi}'_1, \ldots, \mathring{\xi}'_m \rangle.$$

Введём в евклидовом пространстве $\mathcal{R}_{n+m}$ два $(n+m)$-мерных вектора

$$\vec{e}_1 = \Big\langle \underbrace{\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}}_{n}, \underbrace{0, \ldots, 0}_{m} \Big\rangle, \quad \vec{e}_2 = \Big\langle \underbrace{0, \ldots, 0}_{n} \ \underbrace{\frac{1}{\sqrt{m}}, \ldots, \frac{1}{\sqrt{m}}}_{m}, \Big\rangle.$$

Очевидно, что $\|\vec{e}_1\|^2 = \|\vec{e}_2\|^2 = 1$ и $(\vec{e}_1, \vec{e}_2) = 0$. Представим разность $(\bar{\xi} - \mu) - (\bar{\xi}' - \mu')$ в виде

$$(\bar{\xi} - \mu) - (\bar{\xi}' - \mu') = c_1(\vec{\xi}^{\,\circ}, \vec{e}_1) + c_2(\vec{\xi}^{\,\circ}, \vec{e}_2) \tag{2.43}$$

и найдём коэффициенты c_1 и c_2. Для этого запишем равенства

$$\bar{\xi} - \mu = \frac{1}{n}\sum_{i=1}^{n} \xi_i - \mu = \frac{1}{n}\sum_{i=1}^{n} (\xi_i - \mu),$$
$$\bar{\xi}' - \mu' = \frac{1}{m}\sum_{j=1}^{m} \xi_i - \mu' = \frac{1}{m}\sum_{i=1}^{n} (\xi'_i - \mu'),$$

следовательно,

$$(\bar{\xi} - \mu) - (\bar{\xi}' - \mu') = \frac{1}{n}\sum_{i=1}^{n} (\xi_i - \mu) - \frac{1}{m}\sum_{i=1}^{n} (\xi'_i - \mu'). \tag{2.44}$$

с другой стороны,

$$(\vec{\xi}^{\,\circ}, \vec{e}_1) = \sum_{i=1}^{n} \frac{\xi_i - \mu}{\sqrt{\sigma^2}} \cdot \frac{1}{\sqrt{n}}, \qquad (\vec{\xi}^{\,\circ}, \vec{e}_2) = \sum_{j=1}^{m} \frac{\xi'_j - \mu'}{\sqrt{\sigma^2}} \cdot \frac{1}{\sqrt{m}}, \tag{2.45}$$

откуда

$$c_1(\vec{\xi}^{\,\circ}, \vec{e}_1) + c_2(\vec{\xi}^{\,\circ}, \vec{e}_2) = \frac{c_1}{\sqrt{n\sigma^2}} \sum_{i=1}^{n} (\xi_i - \mu) + \frac{c_2}{\sqrt{m\sigma^2}} \sum_{j=1}^{m} (\xi'_i - \mu'). \tag{2.46}$$

Приравнивая правые части выражений (2.44) и (2.46), получаем

$$c_1 = \sqrt{\frac{\sigma^2}{n}}, \qquad c_2 = -\sqrt{\frac{\sigma^2}{m}}. \tag{2.47}$$

Векторы $\vec{e}_1$ и $\vec{e}_2$ ортонормированные, следовательно, случайные векторы $(\vec{\xi}^{\,\circ}, \vec{e}_1)$ и $(\vec{\xi}^{\,\circ}, \vec{e}_2)$ независимы и имеют независимые стандартно нормально распределённые координаты. Поэтому

$$c_1(\vec{\xi}^{\,\circ}, \vec{e}_1) + c_2(\vec{\xi}^{\,\circ}, \vec{e}_2) \sim \mathbf{N}(0, c_1^2 + c_2^2).$$

и

$$\frac{1}{\sqrt{c_1^2 + c_2^2}}\big(c_1(\vec{\xi}^{\,\circ}, \vec{e}_1) + c_2(\vec{\xi}^{\,\circ}, \vec{e}_2)\big) \sim \mathbf{N}(0, 1).$$

Теперь подставим явные выражения (2.47) для постоянных c_1, c_2, тогда

$$\frac{1}{\sqrt{c_1^2 + c_2^2}} = \frac{\sigma^2}{n} + \frac{\sigma^2}{n} = \sigma^2 \frac{n+m}{nm},$$

и из (2.43) мы получаем

$$\begin{aligned} &\frac{1}{\sqrt{c_1^2 + c_2^2}}\big(c_1(\vec{\xi}^{\,\circ}, \vec{e}_1) + c_2(\vec{\xi}^{\,\circ}, \vec{e}_2)\big) = \\ &\qquad = \sqrt{\frac{1}{\sigma^2}}\sqrt{\frac{nm}{n+m}}\big((\bar{\xi} - \mu) - (\bar{\xi}' - \mu')\big) \sim \mathbf{N}(0, 1). \end{aligned} \tag{2.48}$$

Мы видим, что с точностью до множителя $1/\sqrt{\sigma^2}$ эта с.в., распределённая стандартно нормально, совпадает с числителем с.в. (2.42).

Обратимся к знаменателю с.в. (2.42). Введём в евклидовом пространстве $\mathcal{R}_{n+m}$ ортогональный проектор Π_2 на двумерное подпространство, натянутое на векторы $\vec{e}_1$, $\vec{e}_2$. На любой $\vec{x} \in \mathcal{R}_{n+m}$ он действует как

$$\Pi_2 \vec{x} = (\vec{x}, \vec{e}_1)\vec{e}_1 + (\vec{x}, \vec{e}_2)\vec{e}_2,$$

при этом

$$\|\Pi_2 \vec{x}\,\|^2 = (\vec{x}, \vec{e}_1)^2 + (\vec{x}, \vec{e}_2)^2, \qquad \|\vec{x} - \Pi_2 \vec{x}\,\|^2 = \|\vec{x}\|^2 - \|\Pi_2 \vec{x}\,\|^2.$$

Подставим в эти равенства $\vec{\xi}^{\circ}$ вместо $\vec{x}$ и получим с учётом (2.45)

$$\|\Pi_2\vec{\xi}^{\circ}\|^2 = \frac{1}{n\sigma^2}\Big(\sum_{i=1}^{n}(\xi_i-\mu)\Big)^2 - \frac{1}{m\sigma^2}\Big(\sum_{j=1}^{m}(\xi_i-\mu)\Big)^2 = \\ = \frac{n}{\sigma^2}(\bar{\xi}-\mu)^2 + \frac{m}{\sigma^2}(\bar{\xi}'-\mu')^2.$$

Отсюда

$$\|\vec{\xi}^{\circ} - \Pi_2\vec{\xi}^{\circ}\|^2 = \|\vec{\xi}^{\circ}\|^2 - \|\Pi_2\vec{\xi}^{\circ}\|^2 = \\ = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\xi_i-\mu)^2 + \frac{1}{\sigma^2}\sum_{j=1}^{m}(\xi_i'-\mu')^2 - \frac{n}{\sigma^2}(\bar{\xi}-\mu)^2 + \frac{m}{\sigma^2}(\bar{\xi}'-\mu')^2 = \\ = \frac{1}{\sigma^2}\Big\{\sum_{i=1}^{n}[(\xi_i-\mu)^2 - (\bar{\xi}-\mu)^2] + \sum_{j=1}^{m}[(\xi_i'-\mu')^2 - (\bar{\xi}'-\mu')^2]\Big\}.$$

Простые алгебраические преобразования, аналогичные преобразованиям выборочных дисперсий, дают

$$\sum_{i=1}^{n}[(\xi_i-\mu)^2 - (\bar{\xi}-\mu)^2] = \sum_{i=1}^{n}(\xi_i-\bar{\xi})^2 = (n-1)\hat{\sigma}_n^2$$

и точно так же

$$\sum_{j=1}^{m}[(\xi_i'-\mu')^2 - (\bar{\xi}'-\mu')^2] = \sum_{i=1}^{n}(\xi_i'-\bar{\xi}')^2 = (m-1)\hat{\sigma}_m^2.$$

Таким образом, получаем

$$\|\vec{\xi}^{\circ} - \Pi_2\vec{\xi}^{\circ}\|^2 = \frac{1}{\sigma^2}\big((n-1)\hat{\sigma}_n^2 + (m-1)\hat{\sigma}_m^2\big). \qquad (2.49)$$

Видно, что $\|\vec{\xi}^{\circ} - \Pi_2\vec{\xi}^{\circ}\|$ с точностью до множителя $1/\sigma^2$ совпадает со знаменателем с.в. (2.42).

Итак, имеем, что с.в. $\|\vec{\xi}^{\circ} - \Pi_2\vec{\xi}^{\circ}\|^2$ имеет распределение хи-квадрат с $n+m-2$ степенями свободы,

$$\|\vec{\xi}^{\circ} - \Pi_2\vec{\xi}^{\circ}\|^2 \sim \mathbf{X}_{n+m-2},$$

и не зависит от $(\bar{\xi}-\mu)-(\bar{\xi}'-\mu') = c_1(\vec{\xi}^{\,\circ},\vec{e}_1)+c_2(\vec{\xi}^{\,\circ},\vec{e}_2)$, поскольку оператор $I-\Pi_2$ проецирует любой $(n+m)$-мерный вектор на подпространство, ортогональное векторам $\vec{e}_1$, $\vec{e}_2$. Тогда с учётом (2.48) и (2.49) статистика (2.42) может быть представлена в виде

$$t(\xi,\xi') = \frac{\sqrt{\sigma^2}\big((\bar{\xi}-\mu)-(\bar{\xi}'-\mu')\big)}{\sqrt{\sigma^2\dfrac{\|\vec{\xi}^{\,\circ}-\Pi_2\vec{\xi}^{\,\circ}\|^2}{n+m-2}}} = \frac{\nu_0}{\sqrt{\dfrac{\chi^2_{n+m-2}}{n+m-2}}},$$

где с.в. $\nu_0 \sim \mathbf{N}(0,1)$ и с.в. $\chi^2_{n+m-2} \sim \mathbf{X}_{n+m-2}$ независимы. По определению $t(\xi,\xi')$ имеет распределение Стьюдента с $n+m-2$ степенями свободы. Теорема доказана.

На основании этой теоремы и с учетом того, что мы сформулировали двустороннюю гипотезу $\mu \neq \mu'$, задаём следующий критерий с ошибкой первого рода, равной α_0:

$$\varphi(x,x') = \begin{cases} 1, & \text{если } |t(x,x')| > C_0, \\ 0, & \text{если } |t(x,x')| \leqslant C_0, \end{cases} \qquad x \in \mathbb{R}^n, \quad x' \in \mathbb{R}^m,$$

$$\int\limits_{|y|>C_0} p_{n+m-2}(y)\,dy = \alpha_0,$$

где $p_{n+m-2}(\,\cdot\,)$ – плотность вероятности распределения Стьюдента с $n+m-2$ степенями свободы.

Если гипотеза односторонняя, $\mu < \mu'$ или $\mu > \mu'$, то критические множества имеют соответственно вид $t(x,x') > C_0$ или $t(x,x') < C_0$, где постоянная C_0 получается из уравнения на ошибку критерия с надлежащей заменой области интегрирования.

Критерий Стьюдента также применяют для проверки гипотезы о сравнении средних для выборок, не имеющих нормального распределения, как асимптотический в пределе больших объёмов выборок.

2.3. Проверка независимости

Пусть дана двумерная выборка объёма n, т. е. набор числовых пар $(x_1,y_1),\ldots,(x_n,y_n)$, которые являются независимыми реализациями двумерной с.в. (ξ,η). Гипотеза, которую мы собираемся

проверять, формулируется так: с.в. ξ и η независимы. Никаких дополнительных предположений о распределении этих с.в. мы не делаем.

Критерий хи-квадрат проверки независимости. Идея такого критерия очень близка к идее критерия согласия хи-квадрат из раздела 2.1. Мы группируем данные отдельно по x и по y, подсчитываем количество реализаций, попавших в каждый интервал группировки и сравниваем эмпирические частоты с теми, которые должны быть, если верна гипотеза.

Пусть X – диапазон значений с.в. ξ, а Y – диапазон значений с.в. η. Разобьём X и Y на соответственно $r > 1$ и $s > 1$ непересекающихся интервалов, $X = \sum_{i=1}^{r} \Delta X_r$, $Y = \sum_{j=1}^{s} \Delta Y_j$.

Составим следующую таблицу:

	ΔX_1	ΔX_2	$\dots$	ΔX_r	$\sum_{i=1}^{r}$
ΔY_1	ν_{11}	ν_{12}	$\dots$	ν_{1r}	$\tilde\nu_1$
ΔY_2	ν_{21}	ν_{22}	$\dots$	ν_{2r}	$\tilde\nu_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
ΔY_s	ν_{s1}	ν_{s2}	$\dots$	ν_{sr}	$\tilde\nu_s$
$\sum_{j=1}^{s}$	ν_1	ν_2	$\dots$	ν_r	n

В этой таблице:

• ν_{ij} есть количество реализаций (x_k, y_k), в которых $x_k \in \Delta X_i$ и $y_k \in \Delta Y_j$, $i = 1, \dots, r$ и $j = 1, \dots, s$; очевидно, $\sum_{i=1}^{r} \sum_{j=1}^{s} \nu_{ij} = n$;

• ν_i есть количество реализаций (x_k, y_k), в которых $x_k \in \Delta X_i$, $i = 1, \dots, r$; очевидно, $\sum_{j=1}^{s} \nu_{ij} = \nu_i$;

• $\tilde\nu_j$ есть количество реализаций (x_k, y_k), в которых $y_k \in \Delta Y_j$, $j = 1, \dots, s$; очевидно, $\sum_{i=r}^{s} \nu_i = \tilde\nu_j$.

Частота попадания в прямоугольник разбиения на плоскости или в интервал на прямой при $n \to \infty$ сходится к соответствующей вероятности,

$$\frac{\nu_{ij}}{n} \to P(\xi \in \Delta X_i, \eta \in \Delta Y_j) \stackrel{\text{def}}{=} p_{ij},$$

$$\frac{\nu_i}{n} \to P(\xi \in \Delta X_i) \stackrel{\text{def}}{=} p_i, \qquad \frac{\tilde\nu_j}{n} \to P(\eta \in \Delta Y_j) \stackrel{\text{def}}{=} \tilde p_j,$$

и, если верна гипотеза о независимости,

$$P(\xi \in \Delta X_i, \eta \in \Delta Y_j) = P(\xi \in \Delta X_i) P(\eta \in \Delta Y_j) \tag{2.50}$$

для всех $i = 1, \ldots, r$ и $j = 1, \ldots, s$.

Заменим нашу исходную задачу о проверке гипотезы о независимости задачей проверки равенств (2.50). С математической точки зрения это проверка гипотезы

$$H: \quad p_{ij} = p_i \tilde{p}_j \quad \text{для всех} \quad i = 1, \ldots, r, \;\; j = 1, \ldots, s \tag{2.51}$$

по следующим данным:
• реализации $(\nu_{11}, \nu_{12}, \ldots, \nu_{rs})$ с.в., имеющей полиномиальное распределение с вероятностями $(p_{11}, p_{12}, \ldots, p_{rs})$;
• реализации $(\nu_1, \ldots, \nu_r)$ с.в., имеющей полиномиальное распределение с вероятностями $(p_1, \ldots, p_r)$;
• реализации $(\tilde{\nu}_1, \ldots, \tilde{\nu}_s)$ с.в., имеющей полиномиальное распределение с вероятностями $(\tilde{p}_1, \ldots, \tilde{p}_s)$.
Вероятности в полиномиальных распределениях произвольны, за исключением условия их неотрицательности и условий нормировки

$$\sum_{i=1}^{r} \sum_{j=1}^{s} p_{ij} = 1, \quad \sum_{i=1}^{r} p_i = 1, \quad \sum_{j=1}^{s} \tilde{p}_j = 1. \tag{2.52}$$

Аналогичные равенства справедливы и для реализаций (т. е. реализации не являются независимыми):

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \nu_{ij} = n, \quad \sum_{i=1}^{r} \nu_i = n, \quad \sum_{j=1}^{s} \tilde{\nu}_j = n.$$

При проверке гипотезы (2.51) будем принимать решение в зависимости от близости произведений частот $(\nu_i/n) \cdot (\tilde{\nu}_j/n)$ к частоте ν_{ij}/n или, что то же самое, от близости $\nu_i \tilde{\nu}_j / n$ к ν_{ij}. Введем следующее «расстояние» между наборами $\{\nu_i \tilde{\nu}_j / n\}$ и $\{\nu_{ij}\}$:

$$\chi^2 = n \sum_{k=1}^{r} \frac{(\nu_i \tilde{\nu}_j / n - \nu_{ij})^2}{\nu_i \tilde{\nu}_j}. \tag{2.53}$$

(сравните с формулой (2.21)).

Справедлива следующая теорема.

Теорема 12. *Если гипотеза* (2.51) *верна, то при* $n \to \infty$ *распределение статистики* (2.53) *стремится к распределению хи-квадрат с* $(r-1)(s-1)$ *степенями свободы.*

Заметим, что если гипотеза верна, то вероятности $(p_1, \dots, p_r)$ и $(\tilde{p}_1, \dots, \tilde{p}_s)$ полностью определяют вероятности $(p_{11}, p_{12}, \dots, p_{rs})$, а первое равенство в (2.52) является следствием двух других. При этом второе и третье равенства в (2.52) приводят к двум линейным условиям на вероятности, в них участвующие, и тем самым «съедают» одну степень свободы для каждого из наборов одномерных вероятностей. Поэтому с.в. в предельном распределении имеет $(r-1)(s-1)$ степеней свободы.

Статистический критерий строится стандартным образом:

$$\varphi(\chi^2) = \begin{cases} 1, & \text{если } \chi^2 > C_0, \\ 0, & \text{если } \chi^2 \leqslant C_0, \end{cases}$$

где C_0 находится из уравнения

$$\int_{y>C_0} p_{\chi^2_{(r-1)(s-1)}}(y)\, dy = \alpha_0.$$

Статистика χ^2 определяется формулами (2.53), а $p_{\chi^2_{(r-1)(s-1)}}(\,\cdot\,)$ – плотность вероятности для распределения хи-квадрат с $(r-1)(s-1)$ степенями свободы.

Анализ корреляций. Напомним, что если с.в. ξ и η независимы, то $\operatorname{cor}(\xi, \eta) = 0$. Здесь коэффициент корреляции

$$\operatorname{cor}(\xi, \eta) = \frac{M(\xi - M\xi)(\eta - M\eta)}{\sqrt{D\xi\, D\eta}} = \frac{M\xi\eta - M\xi M\eta}{\sqrt{D\xi\, D\eta}}. \tag{2.54}$$

Обратное следствие, вообще говоря неверно, равенство нулю коэффициента корреляции (некоррелированность с.в.) не обязательно влечёт независимость. Тем не менее довольно часто решение задачи проверки гипотезы о независимости выборок строят на основании выборочного коэффициента корреляции. Понятно, что достаточно далёкое от нуля значение этого коэффициента является серьёзным основанием для отклонения гипотезы.

Коэффициент (2.54) удовлетворяет условию $-1 \leqslant \operatorname{cor}(\xi, \eta) \leqslant 1$, причём равенство $|\operatorname{cor}(\xi, \eta)| = 1$ возможно тогда и только тогда, когда одна из с.в. ξ и η с вероятностью единица есть линейная функция от другой. Поэтому часто говорят, что коэффициент корреляции измеряет величину линейной связи между с.в.

Выборочный коэффциент корреляции получается из (2.54) заменой математических ожиданий на выборочные средние. Пусть, как и выше, $(x_1, y_1), \ldots, (x_n, y_n)$ – двумерная выборка объёма n. Выборочный коэффициент корреляции задаётся как

$$r(x,y) = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{S_x^2\, S_y^2}} = \frac{\frac{1}{n}\sum_{k=1}^{n} x_k y_k - \bar{x}\bar{y}}{\sqrt{S_x^2\, S_y^2}}, \qquad (2.55)$$

где

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

суть выборочные среднее и дисперсия для $x = (x_1, \ldots, x_n)$ и аналогично для $y = (y_1, \ldots, y_n)$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k, \qquad S_y^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^2 = \frac{1}{n}\sum_{k=1}^{n} y_k^2 - \bar{y}^2$$

Заметим, что для выборочного коэффициента корреляции (2.55), как и для коэффициента корреляции (2.54), имеются два представления, и тот факт, что они дают один тот же результат, легко проверяется прямым вычислением. Кроме того, подставив S_x^2 и S_y^2 и умножив числитель и знаменатель в (2.55) на n, получим ещё одну формулу для расчёта выборочного коэффициента корреляции:

$$r(x,y) = \frac{\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{k=1}^{n} x_k^2 - n\bar{x}^2\right)\left(\sum_{k=1}^{n} y_k^2 - n\bar{y}^2\right)}}. \qquad (2.56)$$

Пусть $x_1, \ldots, x_n$ – реализации н.с.в. $\xi_1, \ldots, \xi_n$, а $y_1, \ldots, y_n$ – реализации н.с.в. $\eta_1, \ldots, \eta_n$, имеющих следующие нормальные распределения:

$$\xi_k \sim \mathbf{N}(\mu, \sigma^2), \quad \eta_k \sim \mathbf{N}(\tilde{\mu}, \tilde{\sigma}^2), \quad k = 1, \ldots, n. \qquad (2.57)$$

Заметим, что тривиальные алгебраические преобразования дают

$$\frac{\xi_k - \bar{\xi}}{\sqrt{\sum_{k=1}^{n}(\xi_k - \bar{\xi})^2}} = \frac{\frac{\xi_k - \mu}{\sqrt{\sigma^2}} - \overline{\frac{\xi - \mu}{\sqrt{\sigma^2}}}}{\sqrt{\sum_{k=1}^{n}\left(\frac{\xi_k - \mu}{\sqrt{\sigma^2}} - \overline{\frac{\xi - \mu}{\sqrt{\sigma^2}}}\right)^2}},$$

где

$$\overline{\xi-\mu}=\frac{1}{n}\sum_{k=1}^{n}(\xi_k-\mu)=\bar{\xi}-\mu,$$

и аналогично для $\eta_1,\dots\eta_n$. Поэтому распределение статистики выборочного коэффициента корреляции не зависит от параметров нормальных распределений:

$$r(\xi,\eta)=\frac{\frac{1}{n}\sum_{k=1}^{n}(\xi_k-\bar{\xi})(\eta_k-\bar{\eta})}{\sqrt{\frac{1}{n}\sum_{k=1}^{n}(\xi_k-\bar{\xi})^2\frac{1}{n}\sum_{k=1}^{n}(\eta_k-\bar{\eta})^2}}=$$

$$=\frac{\frac{1}{n}\sum_{k=1}^{n}(\overset{\circ}{\xi}_k-\overset{\circ}{\bar{\xi}})(\overset{\circ}{\eta}_k-\overset{\circ}{\bar{\eta}})}{\sqrt{\frac{1}{n}\sum_{k=1}^{n}(\overset{\circ}{\xi}_k-\overset{\circ}{\bar{\xi}})^2\frac{1}{n}\sum_{k=1}^{n}(\overset{\circ}{\eta}_k-\overset{\circ}{\bar{\eta}})^2}},\qquad(2.58)$$

где для каждого $k=1,\dots,n$

$$\overset{\circ}{\xi}_k=\frac{\xi_k-\mu}{\sqrt{\sigma^2}}\sim\mathbf{N}(0,1),\quad\overset{\circ}{\eta}_k=\frac{\eta_k-\tilde{\mu}}{\sqrt{\tilde{\sigma}^2}}\sim\mathbf{N}(0,1).\qquad(2.59)$$

С другой стороны, известно, что нормально распределённые с.в. $\xi_1,\dots,\xi_n,\eta_1,\dots,\eta_n$ независимы в совокупности тогда и только тогда, когда все коэффициенты корреляции равны нулю, т. е. $\operatorname{cor}(\xi_k,\eta_j)=0$ для любых $k,j=1,\dots,n$ и $\operatorname{cor}(\xi_k,\xi_j)=\operatorname{cor}(\eta_k,\eta_j)=0$ для любых $k\neq j$. Таким образом, в случае нормального распределения выборки гипотеза о независимости сводится к гипотезе о некоррелированности данных. Большие значения выборочного коэффициента корреляции $r=r(\xi,\eta)$ очевидно свидетельствуют против гипотезы, поэтому критерий с ошибкой α_0 можно было бы задать как

$$\varphi(r)=\begin{cases}1, & \text{если } |r|>C_0,\\ 0, & \text{если } |r|\leqslant C_0,\end{cases},\qquad\int_{|u|>C_0}^{\infty}p_r(u)\,dt=\alpha_0.$$

Здесь $p_r(\,\cdot\,)$ –плотность вероятности выборочного коэффициента корреляции $r(\xi,\eta)$, где с.в. $\xi_1,\dots,\xi_n,\eta_1,\dots,\eta_n$ независимы в совокупности и распределены нормально в соответствии с (2.57). Заметим, что в силу соотношений (2.58), (2.59) можно заменить ξ_k и η_k на $\overset{\circ}{\xi}_k$ и $\overset{\circ}{\eta}_k$ и считать, что все с.в. распределены стандартно нормально. Однако даже с такой заменой применение этого критерия затруднительно из-за сложного вида функции $p_r(\,\cdot\,)$.

Тем не менее существует удобный критерий для проверки независимости нормальных выборок с помощью выборочного коэффициента корреляции. Справедлива следующая теорема.

Теорема 13. *Пусть с.в.* $\xi_1, \dots, \xi_n, \eta_1, \dots, \eta_n$ *независимы в совокупности и имеют нормальные распределения вида* (2.57). *Пусть* $r(\xi, \eta)$ – *выборочный коэффициент корреляции* (2.58). *Тогда статистика*

$$\tau(\xi, \eta) = \frac{r(\xi, \eta)}{\sqrt{\frac{1}{n-2}(1 - r^2(\xi, \eta))}}$$

имеет распределение Стьюдента $\mathbf{T}_{n-2}$ *с* $n-2$ *степенями свободы.*

Соответственно, критерий задаётся как

$$\varphi(\tau) = \begin{cases} 1, & \text{если } |\tau| > C_0, \\ 0, & \text{если } |\tau| \leqslant C_0, \end{cases}, \qquad \int_{|t|>C_0}^{\infty} p_{n-2}(t)\, dt = \alpha_0.$$

Рассмотрим подробнее статистику $\tau(\xi, \eta)$ из теоремы 13. Будем понимать выборки как случайные векторы со значениями в n-мерном пространстве: $\xi = \langle \xi_1, \dots, \xi_n \rangle$, $\eta = \langle \eta_1, \dots, \eta_n \rangle$. Если в (2.58) сократить множитель $1/n$, то выборочный коэффициент корреляции можно переписать как

$$r(\xi, \eta) = \frac{(\xi - \bar{\xi}, \eta - \bar{\eta})}{\|\xi - \bar{\xi}\| \cdot \|\eta - \bar{\eta}\|} = \frac{(\overset{\circ}{\xi} - \overset{\circ}{\bar{\xi}}, \overset{\circ}{\eta} - \overset{\circ}{\bar{\eta}})}{\|\overset{\circ}{\xi} - \overset{\circ}{\bar{\xi}}\| \cdot \|\overset{\circ}{\eta} - \overset{\circ}{\bar{\eta}}\|}. \tag{2.60}$$

Более того, вводя ортогональный проектор Π_1, действующий по правилу $\Pi_1 \xi = (\xi, e_1) e_1$, где $e_1 = \langle 1/\sqrt{n}, \dots, 1/\sqrt{n} \rangle$, приведем равенство (2.60) к виду

$$r(\xi, \eta) = \frac{\big((I - \Pi_1)\overset{\circ}{\xi}, (I - \Pi_1)\overset{\circ}{\eta}\big)}{\|(I - \Pi_1)\overset{\circ}{\xi}\| \cdot \|(I - \Pi_1)\overset{\circ}{\eta}\|}.$$

Другими словами, коэффициент корреляции $r(\xi, \eta)$ есть косинус угла между векторами $(I - \Pi_1)\overset{\circ}{\xi}$ и $(I - \Pi_1)\overset{\circ}{\eta}$,

$$r(\xi, \eta) = \cos \varphi(\xi, \eta), \qquad 0 \leqslant \varphi(\xi, \eta) \leqslant \pi,$$

что согласуется с условием $|r(\xi,\eta)| \leqslant 1$ с вероятностью единица. Тогда $\sqrt{1-r^2(\xi,\eta)} = \sin\varphi(\xi,\eta)$, и статистика в теореме 13 принимает весьма элегантный вид:

$$\tau(\xi,\eta) = \sqrt{n-2}\operatorname{ctg}\varphi(\xi,\eta), \qquad 0 < \varphi(\xi,\eta) < \pi.$$

Здесь мы исключили из рассмотрения случай, когда с вероятностью единица $(I-\Pi_1)\mathring{\xi} = \mathrm{const}\cdot(I-\Pi_1)\mathring{\eta}$ (или, эквивалентно, $|r(\xi,\eta)|=1$), приводящий к $\sin\varphi(\xi,\eta)=0$ и бесконечным значениям статистики $\tau(\xi,\eta)$. В остальных случаях $\tau(\xi,\eta)$ конечна с вероятностью единица.

Существуют и другие методы проверки независимости, например точный тест Фишера, который позволяет принимать решение в случае малых выборок, но вычислительно достаточно сложен. Также существуют различные статистики, помимо выборочного коэффициента корреляции, позволяющие сделать вывод о наличии или отсутствии связи между двумя выборками. В этом пособии мы ограничимся двумя представленными выше критериями, основанными на теоремах 12 и 13.

2.4. P-value или надёжность гипотезы

Величина, которая называется p-value (p-значение, p-уровень значимости, p-критерий), в последнее время стала очень широко применяться в статистических исследованиях. Однако определения этой величины весьма туманны и зачастую просто искажают её природу. В некоторых монографиях и учебниках это понятие также вводится и называется по-разному: «критический уровень» (Э. Леман), «фактически достигаемый уровень семейства критериев» (А. А. Боровков), «реально достигнутый уровень значимости критерия» (Н. И. Чернова). В этом разделе мы попытаемся объяснить, что такое p-value, и исследовать её свойства.

Начнём с очень простого примера. Хорошо известно, что для проверки простой гипотезы $H\colon \xi \sim \mathbf{N}(0,1)$ против сложной альтернативы $K\colon \xi \sim \mathbf{N}(\mu,1)$, $\mu > 0$ (мы рассматриваем одномерную выборку ξ), существует РНМК, критическое множество которого имеет вид $D_\alpha = \{z > C_\alpha\}$, где C_α связана с ошибкой α первого

рода (вероятностью ошибочно отклонить гипотезу) равенством

$$\alpha = P_0(\xi \in D_\alpha) = P_0(\xi > C_\alpha) = \int_{C_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz. \qquad (2.61)$$

Задавая приемлемый размер ошибки $\alpha = \alpha_0$ и находя из уравнения (2.61) значение C_{α_0}, мы получаем РНМК, который отклоняет гипотезу всякий раз, когда реализация $\xi = x$ больше C_{α_0}.

Если $x \leqslant C_{\alpha_0}$, то мы гипотезу принимаем. Однако, очевидно, имеется некоторая разница между реализацией $x \ll C_{\alpha_0}$ и реализацией $x \approx C_{\alpha_0}$, $x \leqslant C_{\alpha_0}$. По всей видимости, в первом случае мы будем принимать гипотезу с большей уверенностью, чем во втором, потому что значение $x \ll C_{\alpha_0}$ находится «глубоко внутри» множества принятия гипотезы, а значение $x \approx C_{\alpha_0}$ близко к границе множества принятия гипотезы и критического множества. Достаточно немного увеличить величину α_0, и согласно нашему критерию реализация $\xi = x$ уже будет противоречить гипотезе.

Возникает идея для выборки x количественно охарактеризовать эту степень уверенности в принятии гипотезы величиной

$$\int_{x}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz.$$

В самом деле, для $x \ll C_{\alpha_0}$ эта величина достаточно большая (и тем больше, чем дальше x от C_{α_0}), а для $x \approx C_{\alpha_0}$ мы имеем $\alpha(x) \approx \alpha_0$, т. е. величина $\alpha(x)$ мала (больше, но почти равна ошибке критерия, которую, конечно, выбирают малой).

Теперь будем считать реализацию x фиксированной и рассмотрим семейство критических множеств $D_\alpha = \{z > C_\alpha\}$, $\alpha \in (0, 1)$, нашего РНМК. Поставим задачу найти минимальное значение α при условии, что наша реализация $x \in D_\alpha$, т. е. будем искать

$$\alpha(x) = \inf_{D_\alpha:\ x \in D_\alpha} \alpha. \qquad (2.62)$$

Нетрудно заметить, что

$$\alpha(x) = \sup_{C_\alpha:\ C_\alpha < x} \int_{C_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz. \qquad (2.63)$$

Действительно, интеграл

$$\alpha = \int_{C_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz$$

тем меньше, чем больше C_α. При этом мы требуем, чтобы реализация x лежала в множестве D_α, т. е. $x > C_\alpha$. Таким образом, чтобы найти точную нижнюю грань в (2.63), нужно найти точную верхнюю грань значений C_α, которые удовлетворяют неравенству $C_\alpha < x$; очевидно, что эта точная верхняя грань равна x и

$$\alpha(x) = \sup_{C_\alpha\colon C_\alpha < x} \int_{C_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz. \qquad (2.64)$$

Величина $\alpha(x)$, заданная уравнением (2.62), и равна p-value для данного критерия и данной реализации x. Часто говорят, что она равна минимальной вероятности ошибочно отклонить гипотезу по наблюдению $\xi = x$.

Можно также провести рассуждения «в обратном направлении», начав с величины p-value. Предположим, что по реализации x мы нашли $\alpha(x)$ по формуле (2.63). Теперь поставим вопрос: какова ошибка критерия, который обязательно отклоняет гипотезу по реализации x? Имеем $x \in D_\alpha$, следовательно, $x > C_\alpha$, и тогда

$$\alpha = \int_{C_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz \geqslant \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz = \alpha(x).$$

Таким образом, если $\alpha(x)$ велика, то ошибаться, отклоняя гипотезу по реализации x, мы будем с вероятностью $\alpha \geqslant \alpha(x)$, т. е. с очень большой вероятностью. Если, наоборот, $\alpha(x)$ мала, то, отклоняя гипотезу по реализации x, мы ошибаемся не с такой большой вероятностью.

Однако все эти рассуждения верны для фиксированной реализации x и тем самым для фиксированной величины $\alpha(x)$. Поэтому рассуждения о какой-либо вероятности ошибочного решения теряют смысл. В самом деле, для понимания, чему равна вероятность, нам нужно много раз применить свой критерий (и тогда доля случаев ошибочных решений будет примерно равна вероятности), но всякий раз мы будем иметь разные реализации x и, следовательно, разные значения $\alpha(x)$. Другими словами, p-value – это случайная величина, заданная как функция $\alpha(\xi)$ от случайной выборки, и любая аккуратная интерпретация p-value требует исследования её распределения.

P-value как случайная величина систематически и достаточно полно была исследована в работах Ю. П. Пытьева и его учеников.

Он назвал эту случайную величину *надёжностью статистической гипотезы* (точнее, но длиннее было бы назвать эту величину надёжностью решения о принятии статистической гипотезы).

Вернёмся к рассмотренному примеру и рассмотрим с.в $\alpha(\xi)$, которая рассчитывается по формуле (2.64) для $\alpha(x)$, когда $\xi = x$. Найдём распределение надёжности в случаях, когда верна гипотеза и когда верна альтернатива. Для нашего примера

$$\alpha(\xi) = \int_{\xi}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz.$$

Сначала заметим, что для любого распределения с.в. ξ и для любого $b \in (0,1)$ в силу строго монотонной зависимости интеграла от нижнего предела

$$P(\alpha(\xi) < b) = P(\xi > x_b), \quad \text{где} \quad \alpha(x_b) = b. \tag{2.65}$$

Перепишем последнее уравнение как

$$b = \int_{x_b}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz = P_0(\xi > x_b). \tag{2.66}$$

Пусть верна гипотеза. Сравнивая уравнение (2.66) с первым равенством в (2.65), видим, что $P_0(\alpha(\xi) < b) = b$ для любого $b \in (0,1)$. Таким образом, если верна гипотеза, то функция распределения и плотность вероятности надёжности задаются как

$$F_\alpha(b) = b, \quad p_\alpha(b) = F'_\alpha(b) = 1, \qquad 0 < b < 1.$$

Мы получили следующий результат: *при верной простой гипотезе надёжность распределена равномерно*, $\alpha(\xi) \sim \mathbf{U}(0,1)$.

Пусть теперь верна альтернатива, $\xi \sim \mathbf{N}(\mu, 1)$, $\mu > 0$. Из (2.65) получаем

$$F_\alpha(b) = P(\alpha(\xi) < b) = P(\xi > x_b) = \int_{x_b}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-(z-\mu)^2/2}\, dz,$$

$$p_\alpha(b) = \frac{dF_\alpha(b)}{db} = -\frac{1}{\sqrt{2\pi}}\, e^{-(x_b-\mu)^2/2} \cdot \frac{dx_b}{db}.$$

При этом уравнение (2.66) для x_b остаётся прежним, дифференцируя его по b, имеем

$$1 = \frac{d}{db} \int_{x_b}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, dz = -\frac{1}{\sqrt{2\pi}}\, e^{-x_b^2/2} \cdot \frac{dx_b}{db}.$$

Отсюда

$$\frac{dx_b}{db} = \left(-\frac{1}{\sqrt{2\pi}}\, e^{-x_b^2/2}\right)^{-1},$$

что после подстановки в формулу для $p_\alpha(b)$ даёт

$$p_\alpha(b) = \frac{e^{-(x_b-\mu)^2/2}}{e^{-x_b^2/2}} = e^{2x_b\mu} \cdot e^{-\mu^2/2}.$$

Из (2.66) следует, что если $b \to 1$, то $x_b \to -\infty$, если $b \to 0$, то $x_b \to +\infty$. Отсюда получаем предельное поведение плотности вероятности надёжности при верной альтернативе $\mu > 0$:

$$\lim_{b\to 1} p_\alpha(b) = \lim_{x_b\to -\infty} e^{2x_b\mu} \cdot e^{-\mu^2/2} = 0,$$
$$\lim_{b\to 0} p_\alpha(b) = \lim_{x_b\to +\infty} e^{2x_b\mu} \cdot e^{-\mu^2/2} = +\infty.$$

Таким образом, при *верной альтернативе надёжность с большой вероятностью близка к* 0 *и с малой вероятностью близка к* 1.

Полученные результаты показывают, что для рассмотренного примера интуитивные представления о p-value в целом верны. Если $\alpha(\xi)$ велика, то это свидетельствует в пользу верной гипотезы, но не потому, что при верной гипотезе распределение надёжности сосредоточено в области значений, близких к 1, а потому что при верной альтернативе такие значения маловероятны. Напротив, низкие значения надёжности свидетельствуют в пользу альтернативы, но опять же потому, что при верной альтернативе вероятность наблюдать низкие значения $\alpha(\xi)$ достаточно высока по сравнению с той же вероятностью при верной гипотезе.

Однако всё это верно лишь в рассмотренном случае. Для других задач и других критериев требуется отдельное исследование распределения надёжности.

Тем не менее можно доказать некоторое общее утверждение. Рассмотрим простую гипотезу $H\colon \xi \sim L_0(\,\cdot\,)$ и предположим, что задано семейство нерандомизированных критериев для её проверки, параметризованное величиной ошибки $\alpha \in (0,1)$. Обозначим критические множества для таких критериев как D_α, $\alpha \in (0,1)$:

$$P_0(\xi \in D_\alpha) = \int_{D_\alpha} L_0(x)\, dx = \alpha \quad \text{для каждого} \quad \alpha \in (0,1). \quad (2.67)$$

.

Наложим на семейство критических множеств $\{D_\alpha\}$ следующие естественные условия:

1) для любого $\alpha \in (0,1)$ существует множество D_α;

2) семейство монотонно по α: если $\alpha_1 \leqslant \alpha_2$, то $D_{\alpha_1} \subset D_{\alpha_2}$.

Зададим надёжность гипотезы как

$$\alpha(\xi) = \inf_{\alpha\colon\, \xi \in D_\alpha} \alpha = \inf_{D_\alpha\colon\, \xi \in D_\alpha} P_0(\xi \in D_\alpha). \tag{2.68}$$

Теорема 14. *Если при верной гипотезе функция распределения надёжности* $F_\alpha(b) = P_0(\alpha(\xi) < b)$, $0 \leqslant b \leqslant 1$, *непрерывна всюду на интервале* $(0,1)$, *то* $\alpha(\xi) \sim \mathbf{U}(0,1)$.

Доказательство. Рассмотрим множество $\{\xi\colon \alpha(\xi) < b\}$ при фиксированном $b \in (0,1)$. Если $\alpha(\xi) < b$, то по определению точной нижней грани среди α, по которым вычисляется точная нижняя грань (2.68) (тех, для которых $\xi \in D_\alpha$), найдётся значение $\alpha = a$, такое что $\alpha(\xi) = a < b$. Тогда $\xi \in D_a \subset D_b$, следовательно, $\xi \in D_b$.

Теперь, наоборот, пусть $\{\xi \in D_b\}$, тогда значение $b \in (0,1)$ попадает в множество $\{\alpha \in (0,1)\colon \xi \in D_\alpha\}$, следовательно,

$$\alpha(\xi) = \inf_{\alpha\colon\, \xi \in D_\alpha} \alpha \leqslant b.$$

Таким образом, мы имеем цепочку следствий

$$\alpha(\xi) < b \quad \Longrightarrow \quad \xi \in D_b \quad \Longrightarrow \quad \alpha(\xi) \leqslant b,$$

и в терминах вероятностей для любого распределения с.в. ξ

$$P(\alpha(\xi) < b) \leqslant P(D_b) \leqslant P(\alpha(\xi) \leqslant b\}.$$

Отсюда при верной гипотезе в силу $P_0(D_b) = b$

$$F_\alpha(b) = P_0(\alpha(\xi) < b) \leqslant b \leqslant P_0(\alpha(\xi) \leqslant b) = F_\alpha(b+0) = F_\alpha(b),$$

где самое правое равенство есть следствие непрерывности функции распределения надёжности. Как результат, имеем $F_\alpha(b) = b$ для всех $b \in (0,1)$. Теорема доказана.

Список литературы

1. Пытьев Ю. П., Шишмарев И. А. Теория вероятностей, математическая статистика и элементы теории возможностей для физиков. — Издательство МГУ, Москва, 2023. —410 с.

2. Боровков А. А. Математическая статистика — Издательство «Лань», СанктПетербург, 2010. — 704 с.

3. Чернова Н. И. Математическая статистика: Учебное пособие — Издательство НГУ, Новосибирск, 2007. — 148 с.

4. Коршунов Д. А., Чернова Н. И. Сборник задач и упражнений по математической статистике. Учебное пособие. 2-е изд., испр. — Издательство во Института математики СО РАН, Новосибирск, 2004. — 128 с.

5. Горяинов В. Б., Павлов И. В., Цветкова Г. М. Математическая статистика: Учебник для вузов. — Издательство МГТУ им. Н. Э. Баумана, Москва, 2001. — 424 с..