§9. Проверка модельных предположений. Критерии согласия

Лекция 14

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

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

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

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

После разбиения $\mathcal{X}$ проводится сортировка выборочных данных по областям разбиений и подсчитываются количества $\nu_1, \ldots, \nu_2$, $\sum_1^n \nu_i = n$, данных, попавших в соответствующие области $A_1, \ldots A_r$. Вычисляются “теоретические” вероятности $p_i = P(A_i)$, $i = 1, \ldots, r$ попадания выборочных данных в эти области и вычисляется значение x^2 тестовой статистики

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

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

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

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

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

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

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

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

Если мы найдем характеристическую функцию $\varphi_Y(\mathbf{t})$, $\mathbf{t} = (t_1, \ldots, t_r)$, наблюдаемого вектора Y, то характеристическая функция $\varphi_\nu(\mathbf{t})$ вектора частот $\nu = (\nu_1, \ldots, \nu_r)$ будет вычисляться по формуле $\varphi_\nu(\mathbf{t}) = \varphi_Y^n(\mathbf{t})$, ибо характеристическая функция суммы независимых случайных величин равна произведению характеристических функций слагаемых (пункт 3^0 теоремы 12.1 курса ТВ). Но характеристическая функция вектора Y (напомним, $\sum_1^r X_j = 1$)

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

и поэтому

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

Теперь приступим к асимптотическому анализу характеристической функции вектора X нормированных частот

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

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

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

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

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

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

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

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

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

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

Рассмотрим квадратичную форму

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

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

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

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

Обратимся теперь к предельному распределению статистики $X^2 = \sum_1^r X_j$. Поскольку предельное распределение вектора X совпадает с распределением вектора Z, то предельное распределение статистики X^2 определяется распределением квадратичной формы $\sum_1^r Z_j^2$. Как известно, ортогональные преобразования не меняют суммы квадратов, поэтому $\sum_1^r Z_j^2 = \sum_1^r Y_j^2 = \sum_1^{r-1} Y_j$. Следовательно, предельное распределение статистики X^2 есть распределение суммы квадратов $r-1$ независимых случайных величин, имеющих общее стандартное нормальное распределение. По определению это – хи-квадрат распределение с $r-1$ степенями свободы. Теорема Пирсона доказана.

Лекция 15

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

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

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

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

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

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

при любом $x > 0$.

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

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

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

2^0. Критерий независимости хи-квадрат (таблицы сопряженности признаков). Следующая задача выявления зависимости между определенными признаками наблюдаемых объектов часто возникает в практических приложениях математической статистики. Предположим, что мы случайно выбрали n особей из некоторой этнической популяции, и хотим выяснить, существует ли зависимость между цветом волос и цветом глаз. Мы различаем $s \geq 2$ уровней первого признака (например, блондин, брюнет, шатен и рыжий) и $r \geq 2$ уровней второго (например, карие, серые, голубые и зеленые). Все n особей разбиваются на sr групп в соответствии с наличием тех или иных уровней каждого признака, и составляется следующая таблица частот особей в каждой группе.

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

Такие таблицы, в которых суммы

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

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

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

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

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

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

в которой суммирование распространяется на все rs групп таблицы сопряженности признаков. Понятно, что X^2 является тестовой статистикой только в случае известных значений $r+s-2$ параметров $p_{i\cdot} p_{\cdot j}$, $i = 1, \ldots, r$, $j = 1, \ldots, s$ (напомним, $\sum_1^r p_{i\cdot} = \sum_1^s p_{\cdot j} = 1$, так что с помощью этих соотношений два из $r+s$ параметров, например, $p_{r\cdot}$ и $p_{\cdot s}$, можно выразить через остальные $r+s-2$ параметров). В этом случае X^2 имеет в пределе $(n \to \infty)$ хи-квадрат распределение с $rs-1$ степенями свободы.

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

$$\hat{p}_{i\cdot} = \frac{\nu_{i\cdot}}{n},\ \hat{p}_{\cdot j} = \frac{\nu_{\cdot j}}{n}, \quad i = 1, \ldots, r,\ j = 1, \ldots, s,$$

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

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

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

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

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

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

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

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

Внутренняя сумма

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

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

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

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

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

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