Глава 18

СГРУППИРОВАННЫЕ ДАННЫЕ

Нередко данные, находящиеся в распоряжении исследователя, представляют собой таблицу количеств попаданий наблюдений в некоторые множества. В этой главе будут рассмотрены методы, позволяющие анализировать такие данные. Все они имеют в качестве предельного закона для статистики критерия *распределение хи-квадрат*, определенное в примере 3 гл. 11.

Ба! Знакомые все лица!

Фамусов в «Горе от ума» А. С. Грибоедова

Эти методы весьма универсальны, но одновременно довольно грубы из-за потери информации при группировке. Их можно рекомендовать для применения на предварительной стадии статистического анализа.

§ 1. ПРОСТАЯ ГИПОТЕЗА

Пусть $\xi_1,\ldots,\xi_n$ — выборка (см. § 1 гл. 4) из закона с функцией распределения $F(x)$. Разобьем множество значений ξ_1 на N промежутков (возможно, бесконечных) $\Delta_j = (a_j, b_j]$, $j = 1,\ldots,N$ (рис. 1).*) Положим $p_j = \mathbf{P}(\xi_1 \in \Delta_j)$, а случайные величины ν_j — равными количеству элементов выборки в Δ_j $(\nu_1 + \ldots + \nu_N = n)$.

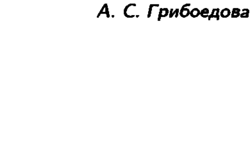

Рис. 1

Функция F неизвестна. Проверяется гипотеза

$H_0\colon F(x) = F_0(x)$,

где F_0 — заданная функция распределения. Если гипотеза верна, то согласно закону больших чисел (П6) *частоты попадания в промежутки* $\widehat{p}_j = \nu_j/n$ при достаточно больших n должны быть близки к соответствующим вероятностям $p_j^0 = F_0(b_j) - F_0(a_j)$.

В качестве *меры отклонения* от гипотезы H_0 Карл Пирсон в 1900 г. предложил статистику

$$X_n^2 = n \sum_{j=1}^{N} \frac{1}{p_j^0} \left(\widehat{p}_j - p_j^0\right)^2 = \sum_{j=1}^{N} \frac{\left(\nu_j - np_j^0\right)^2}{np_j^0}. \qquad (1)$$

Замечание 1. Первое представление в формуле (1) показывает, что X_n^2 есть взвешенная сумма квадратов отклонений частот от

*) Если множество значений ξ_1 является интервалом, то $a_j = b_{j-1}$.

гипотетических вероятностей. Для фиксированного промежутка в силу центральной предельной теоремы (П6) каждое отклонение асимптотически нормально (см. § 4 гл. 7) и имеет порядок малости $1/\sqrt{n}$. Множитель n перед суммой необходим для того, чтобы предельное распределение статистики не вырождалось в 0. Поскольку складываются квадраты отклонений с весами, обратно пропорциональными гипотетическим вероятностям (чтобы «уравнять» слагаемые между собой), представляется правдоподобным, что предельным законом будет распределение хи-квадрат — сумма квадратов независимых и одинаково распределенных по закону $\mathcal{N}(0,1)$ случайных величин.

Теорема 1. Если $0 < p_j^0 < 1$, $j = 1,\ldots,N$, то при $n \to \infty$

$X_n^2 \xrightarrow{d} \zeta \sim \chi^2_{N-1}$.

Вопрос 1.
Почему число степеней свободы предельного закона не совпадает с числом слагаемых в суммах из (1)?

Доказательство. Раскладывая независимые «шарики» ξ_i $(i = 1,\ldots,n)$ по «ящикам» Δ_j $(j = 1,\ldots,N)$ с вероятностями p_j^0 попадания в j-й «ящик» (см. § 5 гл. 10), получим

$$\mathsf{P}(\nu_1 = l_1,\ldots,\nu_N = l_N) = \frac{n!}{l_1!\ldots l_N!}\,(p_1^0)^{l_1}\ldots(p_N^0)^{l_N},$$

если все $l_j \geqslant 0$ и $l_1 + \ldots + l_N = n$, иначе вероятность равна 0.

Используя известную формулу возведения суммы в n-ю степень

$$(a_1 + \ldots + a_N)^n = \sum_{\substack{l_1\geqslant 0,\ldots,l_N\geqslant 0,\\ l_1+\ldots+l_N=n}} \frac{n!}{l_1!\ldots l_N!}\,a_1^{l_1}\ldots a_N^{l_N},$$

находим, что характеристическая функция (см. П9) случайного вектора $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_N)$ имеет вид

$$\psi_{\boldsymbol{\nu}}(\boldsymbol{t}) = \mathsf{M}e^{i\boldsymbol{t}^T\boldsymbol{\nu}} = \left(p_1^0e^{it_1} + \ldots + p_N^0e^{it_N}\right)^n, \quad \boldsymbol{t} = (t_1,\ldots,t_N). \tag{2}$$

Нетрудно убедиться, что для преобразованного случайного вектора $\boldsymbol{\nu}^* = (\nu_1^*,\ldots,\nu_N^*)$ с компонентами $\nu_j^* = (\nu_j - np_j^0)/\sqrt{n}$ характеристическая функция выглядит так:

$$\psi_{\boldsymbol{\nu}^*}(\boldsymbol{t}) = e^{-i\sqrt{n}\,\boldsymbol{t}^T\boldsymbol{p}^0}\left[1 + \sum_{j=1}^N p_j^0\left(e^{it_j/\sqrt{n}} - 1\right)\right]^n, \quad \boldsymbol{p}^0 = \left(p_1^0,\ldots,p_N^0\right).$$

Логарифмируя и раскладывая при $\varepsilon \to 0$ в ряды Тейлора функции $\ln(1+\varepsilon) = \varepsilon - \varepsilon^2/2 + O(\varepsilon^3)$ и $e^{i\varepsilon} = 1 + i\varepsilon - \varepsilon^2/2 + O(\varepsilon^3)$ (см. [82, с. 573]), получаем:

$$\ln\psi_{\boldsymbol{\nu}^*}(\boldsymbol{t}) = -i\sqrt{n}\,\boldsymbol{t}^T\boldsymbol{p}^0 + n\sum_{j=1}^N p_j^0\left(e^{it_j/\sqrt{n}} - 1\right) -$$

$$-\frac{n}{2}\left[\sum_{j=1}^N p_j^0\left(e^{it_j/\sqrt{n}} - 1\right)\right]^2 + O\left(1/\sqrt{n}\right) =$$

$$= -\frac{1}{2}\sum_{j=1}^N p_j^0t_j^2 + \frac{1}{2}\left(\sum_{j=1}^N p_j^0t_j\right)^2 + O\left(1/\sqrt{n}\right) = -\frac{1}{2}\,\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t} + O\left(1/\sqrt{n}\right),$$

где (см. П10)

$$\boldsymbol{\Sigma} = \|\sigma_{jk}\|_{N\times N}, \quad \sigma_{jk} = \begin{cases} p_j^0\left(1-p_j^0\right) & \text{при } k=j, \\ -p_j^0 p_k^0 & \text{при } k\neq j. \end{cases} \tag{3}$$

Отсюда следует, что предел $\psi_{\boldsymbol{\nu}^*}(\boldsymbol{t})$ при $n\to\infty$ есть характеристическая функция $\exp\left\{-\frac{1}{2}\,\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right\}$ многомерного нормального закона $\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$ (см. П9). (Неотрицательная определенность матрицы $\boldsymbol{\Sigma}$ устанавливается в задаче 5.) По теореме непрерывности из П9 распределение случайной величины $\boldsymbol{\nu}^*$ сходится к указанному закону.

Заметим, что ковариационная матрица $\boldsymbol{\Sigma}$ *вырождена* (П10). Причиной этого является линейная зависимость компонент вектора $\boldsymbol{\nu}^*$:

$$\nu_1^* + \ldots + \nu_N^* = 0. \tag{4}$$

Однако, ее подматрица $\boldsymbol{A}$ размера $(N-1)\times(N-1)$ уже не вырождена. Действительно, нетрудно убедиться, что обратной к ней служит матрица

$$\boldsymbol{A}^{-1} \equiv \boldsymbol{B} = \|b_{jk}\|_{(N-1)\times(N-1)}, \; b_{jk} = \begin{cases} 1/p_j^0 + 1/p_N^0 & \text{при } k=j, \\ 1/p_N^0 & \text{при } k\neq j. \end{cases}$$

Таким образом, для подвектора $\boldsymbol{c} = \left(\nu_1^*,\ldots,\nu_{N-1}^*\right)$ предельным будет невырожденный нормальный закон $\mathcal{N}(\boldsymbol{0},\boldsymbol{A})$. Согласно последнему утверждению из П9 и свойству 3 сходимости из П5

$$\boldsymbol{c}\boldsymbol{B}\boldsymbol{c}^T \xrightarrow{d} \zeta \sim \chi^2_{N-1} \quad \text{при } n\to\infty. \tag{5}$$

С другой стороны, из формул (1) и (4) имеем

$$X_n^2 = \sum_{j=1}^{N} \frac{1}{p_j^0}\left(\nu_j^*\right)^2 = \sum_{j=1}^{N-1} \frac{1}{p_j^0}\left(\nu_j^*\right)^2 + \frac{1}{p_N^0}\left(\nu_1^* + \ldots + \nu_{N-1}^*\right)^2.$$

Но правая часть совпадает с $\boldsymbol{c}\boldsymbol{B}\boldsymbol{c}^T$, что с учетом сходимости (5) завершает доказательство теоремы 1. ∎

Как отмечено в [32, с. 111], приближение распределения статистики X_n^2 с помощью закона χ^2_{N-1} является достаточно точным при $n\geqslant 50$ и $np_j^0\geqslant 5$ для всех $j=1,\ldots,N$.

Замечание 2. Последнее условие предназначено для того, чтобы обеспечивать возможность попадания хотя бы нескольких наблюдений ξ_i в каждый из промежутков Δ_j. Это необходимо для пригодности лежащего в основе теоремы 1 нормального приближения для распределения величин $\sqrt{n}\,(\widehat{p}_j - p_j^0)$: чем больше для заданного j *ожидаемое количество попаданий* np_j^0, тем приближение точнее. Поэтому число промежутков N не должно быть слишком большим. Однако, его не следует брать и очень малым, так как в этом случае набор вероятностей $p_1^0,\ldots,p_N^0$ недостаточно хорошо представляет гипотетическую функцию распределения $F_0(x)$. Обычно на практике берут $N \approx \log_2 n$.

Когда N выбрано, возникает вопрос, каким образом задавать промежутки $\Delta_j = (a_j, b_j]$. Если областью возможных значений случайной величины ξ_1 служит ограниченный интервал, то можно разбить его на *равные по длине части*. Альтернативным выбором (годящимся для неограниченных областей значений ξ_1) является разбиение действительной прямой на *равновероятные промежутки*, у которых $a_j = b_{j-1}$, а правые границы b_j находятся из уравнений $F_0(b_j) = j/N$, $j = 1,\ldots,N$.

Иногда N и p_j^0 не выбираются исследователем, а определяются самой изучаемой проблемой.

Г. И. Мендель (1822–1884), австрийский естествоиспытатель.

Пример 1. ***Генетические законы Менделя*** (см. [35, с. 563]). В экспериментах с селекцией гороха (1856–1863) Мендель наблюдал частоты различных видов семян, получаемых при скрещивании растений с круглыми желтыми семенами и растений с морщинистыми зелеными семенами. Эти данные и значения теоретических вероятностей, определяемые в соответствии с законом Менделя независимого расщепления признаков, приведены в следующей таблице:

Тип семян	Частота $\widehat{p}_j$	Вероятность p_j^0
Круглые и желтые	315/556	9/16
Морщинистые и желтые	101/556	3/16
Круглые и зеленые	108/556	3/16
Морщинистые и зеленые	32/556	1/16

Проверим гипотезу H_0 о согласованности частот с теоретическими вероятностями при помощи критерия хи-квадрат. Статистика критерия (см. формулу (1)) $X_n^2 \approx 0{,}47$. Из табл. Т3 получаем, что это значение находится между квантилями уровня 0,05 и 0,1 закона χ_3^2. Таким образом, согласие наблюдений с гипотезой H_0 очень хорошее.

Вопрос 2.
Чем подозрителен датчик псевдослучайных чисел, у которого в промежутки $\left(0,\frac{1}{4}\right]$, $\left(\frac{1}{4},\frac{1}{2}\right]$, $\left(\frac{1}{2},\frac{3}{4}\right]$ и $\left(\frac{3}{4},1\right]$ попали соответственно 504, 505, 492 и 499 точек?

§ 2. СЛОЖНАЯ ГИПОТЕЗА

Метод группировки наблюдений с последующим применением критерия хи-квадрат применим и для проверки сложной гипотезы H_0' о принадлежности неизвестной функции распределения элементов выборки некоторому заданному классу функций распределения $\mathcal{F} = \{F(x,\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k\}$.

В этом случае общая (при всевозможных $\boldsymbol{\theta} \in \Theta$) область значений ξ_1 также разбивается на N промежутков $\Delta_j = (a_j, b_j]$,

$j = 1,\ldots,N$. Как и ранее ν_j обозначает число элементов выборки в Δ_j. Однако теперь вероятности $\mathbf{P}(\xi_1 \in \Delta_j)$ при H_0' уже не будут заданы однозначно, а представляют собой функции от $\boldsymbol{\theta}$: $p_j(\boldsymbol{\theta}) = F(b_j,\boldsymbol{\theta}) - F(a_j,\boldsymbol{\theta})$ (рис. 2). Из-за этой зависимости от неизвестного параметра нельзя просто подставить $p_j(\boldsymbol{\theta})$ вместо p_j^0 в (1). Р. Фишер (1924 г.) доказал, что если подставить $p_j(\tilde{\boldsymbol{\theta}})$, где $\tilde{\boldsymbol{\theta}}$ — *оценка максимального правдоподобия, основанная на частотах* (определяемая ниже), то при некоторых условиях на класс $\mathcal{F}$ функций распределения (см. [32, с. 115]) статистика

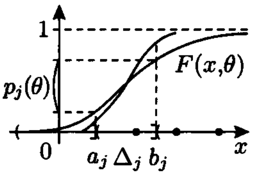

Рис. 2

$$\tilde{X}_n^2 = \sum_{j=1}^{N} \left(\nu_j - np_j(\tilde{\boldsymbol{\theta}})\right)^2 \Big/ \left[np_j(\tilde{\boldsymbol{\theta}})\right] \tag{6}$$

будет иметь в качестве предельного закона снова распределение хи-квадрат, только уже с $(N-1-k)$ степенями свободы, где k — размерность вектора $\boldsymbol{\theta}$.

Доказательство этой теоремы можно найти в [44, с. 462–470].

Определение. *Значением оценки максимального правдоподобия, основанной на частотах* $\tilde{\boldsymbol{\theta}}$, *служит вектор* $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)$, *на котором достигается максимум вероятности*

$$\mathbf{P}(\nu_1 = l_1,\ldots,\nu_N = l_N) = \frac{n!}{l_1!\ldots l_N!}\,[p_1(\boldsymbol{\theta})]^{l_1}\ldots[p_N(\boldsymbol{\theta})]^{l_N}.$$

Это равносильно максимизации по $\boldsymbol{\theta}$ *функции*

$$\sum_{j=1}^{N} l_j \ln p_j(\boldsymbol{\theta}) \tag{7}$$

или (для гладких моделей) решению системы, вообще говоря, нелинейных уравнений

$$\sum_{j=1}^{N} l_j\,\frac{\partial \ln p_j(\boldsymbol{\theta})}{\partial \theta_m} = 0, \quad m = 1,\ldots,k. \tag{8}$$

Пример 2. ***Критерий χ^2 для пуассоновской модели*** (см. [32, с. 116]). Положим $\pi_m(\theta) = e^{-\theta}\theta^m/m!$, $m \geqslant 0$. Возьмем промежутки $\Delta_j = [j-1,j)$, $j = 1,\ldots,N-1$; $\Delta_N = [N-1,\infty)$. Тогда вероятности $p_j(\theta) = \pi_{m-1}(\theta)$, $j = 1,\ldots,N-1$; $p_N(\theta) = \sum\limits_{m=N-1}^{\infty}\pi_m(\theta)$.

Так как θ — скалярный параметр, причем $(d/d\theta)\ln\pi_m(\theta) = m/\theta - 1$, то система (8) сводится к одному уравнению:

$$\sum_{j=0}^{N-2} l_{j+1}\,(j/\theta - 1) + l_N \sum_{m=N-1}^{\infty}(m/\theta - 1)\,\pi_m(\theta) \Big/ \sum_{m=N-1}^{\infty}\pi_m(\theta) = 0.$$

Поскольку $l_1 + \ldots + l_N = n$, отсюда получаем соотношение

$$\theta = \frac{1}{n}\left[\sum_{j=0}^{N-2} j\,l_{j+1} + l_N \sum_{m=N-1}^{\infty} m\,\pi_m(\theta) \Big/ \sum_{m=N-1}^{\infty}\pi_m(\theta)\right]. \tag{9}$$

Первый член в скобках равен сумме всех значений ξ_i, меньших или равных $N-2$. Второй член представляет собой $l_N\,\mathsf{M}(\xi_1 \mid \xi_1 \geqslant N-1)$ (см. П7). Он *приближенно* равен сумме всех значений ξ_i, которые больше или равны $N-1$. Поэтому решение $\tilde{\theta}$ уравнения (9) близко к среднему арифметическому $\overline{\xi}$ — оценке максимального правдоподобия параметра θ, построенной по всей выборке.

Вопрос 3. Почему $\overline{\xi}$ — ОМП для пуассоновской модели?

Применим критерий хи-квадрат к данным о падениях самолетов-снарядов в южной части Лондона во время второй мировой войны (см. [81, с. 177]). Опасность попадания в жилые дома вместо военных объектов велика при низкой точности стрельбы (при так называемой *стрельбе по площадной цели*).

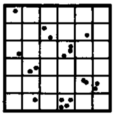

Рис. 3

Карта южной части Лондона была разбита на $n = 24\times 24 = 576$ небольших участков, каждый площадью 1/4 кв. км. На карте были отмечены места падения самолетов-снарядов (подобно рис. 3). В таблице ниже приведены количества участков l_{j+1} ровно с j падениями, $j = 0,1,\ldots,7$. Так как участков много, а вероятность попадания самолета-снаряда на отдельный участок мала, то при справедливости гипотезы о низкой точности стрельбы можно воспользоваться законом редких событий (см. § 1 гл. 5), согласно которому число попаданий на любой из участков есть (приближенно) пуассоновская случайная величина с некоторым общим для всех участков параметром θ. Мы также предположим, что попадания на разные участки независимы.

Общее число падений $M = \sum j\, l_{j+1} = 537$. Возьмем в качестве начальной оценки неизвестного параметра закона Пуассона *среднее число падений на один участок* $\widehat{\theta} = M/n \approx 0{,}932$. Тогда ожидаемые количества участков ровно с j падениями примерно равны $n\pi_j(\widehat{\theta})$.

j	0	1	2	3	4	5	6	7
l_{j+1}	229	211	93	35	7	0	0	1
$n\pi_j(\widehat{\theta})$	226,7	211,4	98,5	30,6	7,14	1,33	0,21	0,03
$n\pi_j(\tilde{\theta})$	228,6	211,3	97,6	30,1	8,46			

Прежде чем вычислять статистику критерия хи-квадрат, надо объединить последние 4 столбца таблицы для того, чтобы ожидаемое количество оказалось не меньше 5: $l_4 + \ldots + l_7 = 8$ и $n(\pi_4(\widehat{\theta}) + \ldots + \pi_7(\widehat{\theta})) = 8{,}71$.

Теперь заменим начальную оценку $\widehat{\theta}$ на $\tilde{\theta}$, максимизируя по θ функцию (7) на компьютере (удобно вычислять $p_5(\theta)$ по формуле $p_5(\theta) = 1 - p_1(\theta) - \ldots - p_4(\theta)$). Вероятно, проще всего уменьшать θ с шагом $h = 0{,}001$ до тех пор, пока функция возрастает. Ответ таков: $\tilde{\theta} = 0{,}924$ (отличие от $\widehat{\theta}$ составляет всего-навсего 0,008). Соответствующие ожидаемые количества приведены в третьей строке таблицы.

Значение статистики $\tilde{X}_n^2$ (см. формулу (6)) для таких данных равно 1,05. Поскольку $N = 5$ и $k = 1$, предельный закон должен иметь $N - k - 1 = 3$ степени свободы. Из табл. Т3 находим, что значение статистики попадает в интервал (0,58; 2,37), образованный 10% и 50% квантилями χ^2_3 (с помощью таблицы из [10, с. 140] уточняем, что фактический уровень значимости равен 0,79). Поэтому гипотеза о низкой точности стрельбы принимается. В [81, с. 177] отмечено:

> «Большинство населения верило в тенденцию точек падения скапливаться в нескольких местах. Если бы это было верно, то следовало бы ожидать бóльшую долю участков без попаданий либо с большим числом попаданий и меньшую долю участков промежуточного класса. Приведенная таблица показывает, что точки падения были совершенно случайными, все участки — равноправными; здесь мы имеем поучительную иллюстрацию того установленного факта, что неискушенному человеку случайность представляется регулярностью или стремлением к скоплению.»

Обратим внимание на *необходимость объединения маловероятных промежутков*: если оставить $N = 8$, то $\tilde{\theta} \approx \widehat{\theta} = 0{,}932$ и $\tilde{X}^2_n = 32{,}6$. Это значимо велико для χ^2_6 даже на уровне 10^{-5} (см. [10, с. 144]). Причиной резкого роста значения статистики является малая величина $np_8(\tilde{\theta}) \approx 0{,}03$, придающая слишком большой вес квадрату отклонения наблюдаемого количества $l_8 = 1$ от ожидаемого количества $np_8(\tilde{\theta})$.

Если данные предварительно группируются, то оценить $\boldsymbol{\theta}$ можно и до группировки наблюдений, например, методом максимального правдоподобия (см. § 4 гл. 9). Однако, как показывает следующий пример, в этом случае статистика $\tilde{X}^2_n$ будет сходиться, вообще говоря, к *другому предельному закону*.

Пример 3. ***Проверка нормальности по сгруппированным данным.*** Пусть $\xi_i \sim \mathcal{N}(\mu,\sigma^2)$, $i = 1,\ldots,n$, причем оба параметра μ и σ неизвестны. Для разбиения прямой на промежутки $\Delta_j = (a_j,b_j]$, $j = 1,\ldots,N$, оценим неизвестную функцию распределения $\Phi((x-\mu)/\sigma)$ при помощи $\Phi((x-\overline{\xi})/S)$, где $\overline{\xi} = \frac{1}{n}\sum \xi_i$ и $S^2 = \frac{1}{n}\sum(\xi_i-\overline{\xi})^2$. Чтобы вероятности попадания ξ_i в промежутки Δ_j были примерно одинаковы, возьмем в качестве b_j решения уравнений

$$\Phi((x-\overline{\xi})/S) = j/N, \quad j = 1,\ldots,N-1,$$

(см. табл. Т2 или приближение Хамакера для Φ^{-1} из § 5 гл. 4).

Далее подсчитаем ν_j — количества попаданий в построенные промежутки. Затем вычислим основанную на частотах оценку максимального правдоподобия $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1,\tilde{\theta}_2)$ при помощи численного поиска точки максимума функции (7), исходя из точки с координатами $(\overline{\xi},S)$. При этом для нахождения $p_j(\boldsymbol{\theta})$ понадобится запрограммировать приближенное вычисление $y = \Phi(x)$, например, с помощью **алгоритма Морана** (см. [58, с. 282]):

```
s = 0
t = x * Sqr(2) / 3
For i = 0 To 12
    z = i + 0.5
    s = s + Sin(z * t) * Exp(-z * z / 9) / z
Next i
y = 0.5 + s / 3.1415926536
```

(Он обеспечивает 9 точных десятичных цифр у $\Phi(x)$ при $|x| \leqslant 7$.)

Важно отметить, что сами оценки $\bar{\xi}$ и S использовать в формуле (6) *нельзя*. В [80, с. 322] указано, что в случае нарушения этого запрета статистика $\tilde{X}_n^2$ не будет (асимптотически) следовать распределению хи-квадрат с $N-3$ степенями свободы: график ее функции распределения пройдет несколько ниже графика функции распределения закона χ^2_{N-3}. Не будет она следовать и распределению хи-квадрат с $N-1$ степенями свободы (как было бы при точно известных параметрах). График ее функции распределения пройдет несколько выше.*)

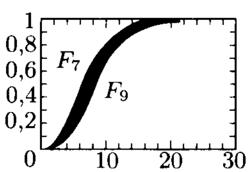

Рис. 4

В качестве иллюстрации на рис. 4 приведены графики функций F_7 и F_9 распределения законов χ^2_7 и χ^2_9 соответственно. Они ограничивают полосу, в которой будет проходить график функции распределения предельного закона для $\tilde{X}_n^2$ при $N = 10$, если для вычисления $p_j(\boldsymbol{\theta})$ использовать оценки $\bar{\xi}$ и S. Согласно табл. Т3 на уровне 0,95 ширина полосы равна $16{,}9 - 14{,}1 = 2{,}8$.

§ 3. ПРОВЕРКА ОДНОРОДНОСТИ

Допустим, что имеется k *независимых* между собой выборок размеров n_i из распределений F_i, $i = 1, \ldots, k$. Общее число наблюдений $n = n_1 + \ldots + n_k$. Проверим *гипотезу однородности*

$H_0''\colon F_1 = \ldots = F_k$

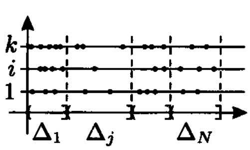

Рис. 5

с помощью критерия хи-квадрат. Для этого сгруппируем данные: разобьем общую для всех выборок область значений наблюдений на промежутки Δ_j, $j = 1, \ldots, N$, и для каждой пары индексов (i, j) подсчитаем величину ν_{ij} — *количество попаданий элементов i-й выборки в j-й промежуток* (рис. 5). В результате получим $k \times N$ таблицу (рис. 6), которую и будем анализировать в дальнейшем.

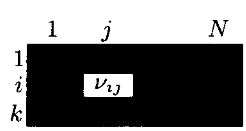

Рис. 6

Иногда данные с самого начала имеют дискретную структуру: в опытах наблюдается некоторый переменный признак, принимающий конечное число N значений (см. пример 4 ниже).

*) Как показали Чернов и Леман в 1954 г. (см. [13, с. 284]), статистика $\tilde{X}_n^2$ асимптотически распределена как сумма $\xi_1^2 + \ldots + \xi_{N-3}^2 + \gamma_1 \xi_{N-2}^2 + \gamma_2 \xi_{N-1}^2$, где ξ_i — независимые $\mathcal{N}(0,1)$-случайные величины; числа γ_1 и γ_2 лежат между 0 и 1 и зависят от проверяемого закона и способа разбиения на промежутки области возможных значений наблюдений.

Если гипотеза H_0'' верна, то ожидаемое количество наблюдений в ячейке с индексами i и j равно $n_i p_j$, где $\boldsymbol{p} = (p_1, \ldots, p_N)$ обозначает (неизвестный) вектор вероятностей попадания в промежутки Δ_j при справедливости гипотезы H_0''. Естественной оценкой для p_j служит $\widehat{p}_j = (\nu_{1j} + \ldots + \nu_{kj})/n$ — общая по всем выборкам частота попаданий в Δ_j (см. задачу 6). Тогда статистика

$$\widehat{X}_n^2 = \sum_{i=1}^{k} \sum_{j=1}^{N} (\nu_{ij} - n_i\widehat{p}_j)^2 / (n_i\widehat{p}_j) \tag{10}$$

измеряет отклонение наблюдаемых количеств от ожидаемых. Если справедлива гипотеза H_0'', то, как доказано в [44, с. 483], статистика $\widehat{X}_n^2$ сходится по распределению к хи-квадрат случайной величине с $(k-1)(N-1)$ степенями свободы при $\min\{n_1, \ldots, n_k\} \to \infty$.

Следующий любопытный пример из [72, с. 132] показывает, что к выводам, основанным на применении этого предельного результата, следует относиться с известной осторожностью.

Факты — упрямая вещь, но статистика гораздо сговорчивее.

Лоренс Питер

Пример 4. ***Парадокс критерия хи-квадрат*** **[72, с. 132].** Ниже приведены три таблицы, в которых отражено действие некоторого лекарства (способа лечения) только на мужчин, только на женщин и, наконец, на больных обоего пола (объединенные результаты).

Мужчины	B	$\overline{B}$
A	700	800
$\overline{A}$	80	130

Женщины	B	$\overline{B}$
A	150	70
$\overline{A}$	400	280

Вместе	B	$\overline{B}$
A	850	870
$\overline{A}$	480	410

Здесь A — принимавшие лекарство, $\overline{A}$ — не принимавшие лекарство, B — выздоровевшие, $\overline{B}$ — не выздоровевшие.

Заметим, что среди принимавших лекарство мужчин доля выздоровевших $700/(700+800) \approx 0{,}467$ больше, чем $80/(80+130) \approx$ $\approx 0{,}381$ — доля выздоровевших среди мужчин, не принимавших лекарство. Такая же картина и у женщин: $150/220 \approx 0{,}682 >$ $> 400/680 \approx 0{,}588$.

Рассчитано, что петербуржец, проживающий на солнцепеке, выигрывает двадцать процентов здоровья.

Козьма Прутков

Статистики $\widehat{X}_n^2$ (см. формулу (10)) для таблиц данных мужчин и женщин принимает значения 5,456 и 6,125. Из [10, с. 141] (см. также табл. Т3) для закона хи-квадрат с 1 степенью свободы находим, что фактические уровни значимости равны соответственно 0,020 и 0,013. Это говорит о существенности различия вероятностей выздоровления между теми, кто принимал лекарство и теми, кто его не принимал.

С другой стороны, как это ни странно, из таблицы с объединенными результатами следует, что доля выздоровевших больше среди тех людей, которые лекарство *не принимали* (!): $480/870 \approx 0{,}539 > 850/1720 \approx 0{,}494$, причем статистика $\widehat{X}_n^2$ для третьей таблицы равна 4,782, что значимо велико на уровне 0,029.

В [72, с. 133] Г. Секей пишет:

> «Аналогично, новое лекарство может оказаться эффективным в каждом из десяти различных госпиталей, но объединение результатов укажет на то, что это лекарство либо бесполезно, либо вредно».

Статистика — самая точная из всех лженаук.
Джин Ко

Причина парадокса заключается в непропорциональном представительстве в разных категориях: мужчины выздоравливают хуже, но лекарство испытывалось в основном на них.

Кроме того, число мужчин (210), не принимавших лекарство, недостаточно велико: согласно таблице, приведенной в книге Дж. Флейс «*Статистические методы для изучения таблиц долей и пропорций*», вероятность β ошибки II рода, для таких данных равна 50%. Чтобы обеспечить $\beta = 10\%$, необходимо иметь не менее 475 пациентов в этой категории.

ЗАДАЧИ

Опыт — лучший учитель.

1. Проверьте первый столбец табл. Т1 на равномерность с помощью критерия хи-квадрат.
2. В [10, с. 21] проводится анализ 2000 четырехзначных псевдослучайных чисел из книги М. Кадырова «Таблицы случайных чисел» (Ташкент, 1936). Первая цифра оказалась нулем у 160, тройкой — у 247, шестеркой — у 191, девяткой — у 185 чисел (остальные 1217 чисел начинались с других цифр). Стоит ли пользоваться такой таблицей?
3. Ниже приведены данные о количестве студентов двух групп, решивших в течение месяца занятий 0, 1–7, 8–15 и более 15 задач. Проверьте гипотезу о том, что студенты обеих групп одинаково активно решают задачи.

Число задач	0	1–7	8–15	> 15
Группа 1	9	8	5	4
Группа 2	3	5	9	11

4. Выведите теорему 1 при $N = 2$ непосредственно из центральной предельной теоремы (П6).
5*. Докажите неотрицательную определенность матрицы $\boldsymbol{\Sigma}$, задаваемой формулой (3), а) вычислив главные миноры (см. П10), б) установив, что она является ковариационной матрицей случайного вектора $\boldsymbol{\nu}^*$ из доказательства теоремы 1.
6*. Покажите при помощи метода неопределенных множителей Лагранжа (см. [46, с. 271]), что оценка $\widehat{p}_j$ из § 3 максимизирует функцию правдоподобия сгруппированной выборки при условии $p_1 + \ldots + p_N = 1$.

РЕШЕНИЯ ЗАДАЧ

Мало хотеть — надо уметь.

1. Поскольку длина столбца $n = 20$ возьмем $N = 4 \approx \log_2 n$ промежутков. При справедливости гипотезы равномерности равные

Повторные наблюдения

Критерий Мак-Немара (McNemar) для парных наблюдений

Таблицы вида 2 x 2 могут содержать сгруппированные данные бинарных повторных наблюдений. Например, v_{11} — студенты, выполнившие успешно тестовое задание и в начале, и в конце обучения, v_{12} — те, кто в начале выполнили, а в конце не выполнили и т. п. Пусть проверяемая гипотеза H_0 заключается в том, что вероятности $p_{12} \approx v_{12} / N$ и $p_{21} \approx v_{21} / N$ одинаковы. (Здесь $N = v_{11} + v_{12} + v_{21} + v_{22}$.) Предположим, что гипотеза H_0 не верна, а верна альтернатива H_1: $p_{21} > p_{12}$, равносильная неравенству

$$p_{11} + p_{21} > p_{11} + p_{12}.$$

Последнее неравенство означает, что вероятность успешного выполнения теста в конце больше вероятности успешного выполнения теста в начале. В таком случае величина v_{21} должна быть «заметно» больше, чем величина v_{12}. Значимость различия можно проверить с помощью *критерия Мак-Немара*, основанного на сходимости

$$(|\nu_{21} - \nu_{12}| - 1)^2 / (\nu_{12} + \nu_{21}) \to \chi_1^2,$$

в которой использована поправка –1, предложенная Эдвардсом (Edwards).

Cochran Q test

Критерий Кохрэна

Критерий Кохрэна обобщает критерий Мак-Немара на $k > 2$ зависимых дихотомических выборок. Типичный пример его применения — сравнение частот утвердительных ответов на k вопросов типа "да – нет" некоторой анкеты.

Статистика критерия Кохрэна имеет следующий вид:

$$Q = (k-1)\left[k\sum u_j^2 - N^2\right] \Big/ \left(kN - \sum v_i^2\right),$$

где v_i — число единиц в i-й строке, u_j — число единиц в j-м столбце, N — общее число единиц во всей таблице. В предположении однородности столбцов статистика Q имеет в качестве предельного закона распределение χ^2_{k-1}.

Поскольку $\sum_{j=1}^k u_j = N$, то числитель формулы, определяющей Q, пропорционален величине

$$\frac{1}{k}\sum u_j^2 - \left(\frac{N}{k}\right)^2,$$

которая представляет собой дисперсию случайной величины ξ, принимающей значения u_j с одинаковыми вероятностями $1/k$. Большие отличия между u_j приводят к большим значениям $\mathbf{D}\xi$.

Задачи

1. Вычислить математическое ожидание и дисперсию распределения хи-квадрат с k степенями свободы.

2. Доказать теорему Пирсона для случая $N = 2$. (*Используйте центральную предельную теорему*.)

3. Вычислить математическое ожидание статистики X_n^2, определённой формулой (1).

4. Объясните, почему предельное распределение статистики $\hat{X}_n^2$, определённой формулой (10), имеет $(k-1)(N-1) = kN - k - N + 1$ степеней свободы.

5*. Покажите при помощи неопределённых множителей Лагранжа, что оценка $\hat{p}_j$ максимизирует функцию правдоподобия сгруппированной выборки при условии $p_1 + \ldots + p_N = 1$.