Оглавление

Лекция 1

Математическая статистика

1.1 Основные выборочные характеристики

Основные понятия

Математическая статистика — наука о математических методах, позволяющих по статистическим данным, например по реализациям случайной величины, построить теоретико-вероятностную модель исследуемого явления. Задачи математической статистики являются, в некотором смысле, обратными к задачам теории вероятностей. Центральным понятием математической статистики является выборка.

Определение 1.1. ***Однородной выборкой (выборкой) объема*** n при $n \geqslant 1$ называется случайный вектор $Z_n = (X_1, \ldots, X_n)$, компоненты которого X_i, $i = \overline{1,n}$, называемые ***элементами выборки***, являются независимыми случайными величинами с одной и той же функцией распределения $F(x)$. Будем говорить, что выборка Z_n ***соответствует*** функции распределения $F(x)$.

Определение 1.2. ***Реализацией выборки*** называется неслучайный вектор $z_n = (x_1, \ldots, x_n)$, компонентами которого являются реализации соответствующих элементов выборки X_i, $i = \overline{1,n}$.

Из определений 1.1 и 1.2 вытекает, что реализацию выборки z_n можно также рассматривать как последовательность $x_1, \ldots, x_n$ из n реализаций одной и той же случайной величины X, полученных в серии из n независимых одинаковых опытов, проводимых в одинаковых условиях. Поэтому можно говорить, что выборка Z_n *порождена наблюдаемой* случайной величиной X, имеющей распределение $F_X(x) = F(x)$.

Определение 1.3. Если компоненты вектора Z_n независимы, но их распределения $F_1(x_1), \ldots, F_n(x_n)$ различны, то такую выборку называют ***неоднородной***.

Определение 1.4. Множество S всех реализаций выборки Z_n называется ***выборочным пространством***.

Выборочное пространство может быть всем n-мерным евклидовым пространством $\mathbb{R}^n$ или его частью, если случайная величина X непрерывна, а также может состоять из конечного или счетного числа точек из $\mathbb{R}^n$, если случайная величина X дискретна.

На практике при исследовании конкретного эксперимента распределения $F_1(x_1), \ldots, F_n(x_n)$ случайных величин $X_1, \ldots, X_n$ редко бывают известны полностью. Часто априори (до опыта) можно лишь утверждать, что распределение $F_{Z_n}(z_n) = F_1(x_1) \cdot \ldots \cdot F_n(x_n)$ случайного вектора Z_n принадлежит некоторому классу (семейству) $\mathscr{F}$.

Определение 1.5. Пара $(S, \mathscr{F})$ называется ***статистической моделью*** описания серии опытов, порождающих выборку Z_n.

Определение 1.6. Если распределения $F_{Z_n}(z_n, \theta)$ из класса $\mathscr{F}$ определены с точностью до некоторого векторного параметра $\theta \in \Theta \subset \mathbb{R}^s$, то такая статистическая модель называется ***параметрической*** и обозначается $(S_\theta, F_{Z_n}(z_n, \theta))$, $\theta \in \Theta \subset \mathbb{R}^s$.

В некоторых случаях выборочное пространство может не зависеть от неизвестного параметра θ распределения $F_{Z_n}(z_n,\theta)$.

В зависимости от вида статистической модели в математической статистике формулируются соответствующие задачи по обработке информации, содержащейся в выборке.

Определение 1.7. Случайная величина $Z=\varphi(Z_n)$, где $\varphi(z_n)$ — произвольная функция, определенная на выборочном пространстве S и не зависящая от распределения $F_{Z_n}(z_n,\theta)$, называется ***статистикой***.

Вариационный ряд

Определение 1.8. Упорядочим элементы реализации выборки $x_1,\ldots,x_n$ по возрастанию: $x_{(1)}\leqslant x_{(2)}\leqslant\ldots\leqslant x_{(n)}$, где нижний индекс соответствует номеру элемента в упорядоченной последовательности. Обозначим через $X_{(k)}$ случайную величину, которая при каждой реализации z_n выборки Z_n принимает k-е (по нижнему номеру) значение $x_{(k)}$, $k=\overline{1,n}$. Упорядоченную последовательность случайных величин $X_{(1)}\leqslant\ldots\leqslant X_{(n)}$ называют ***вариационным рядом выборки***.

Определение 1.9. Элементы $X_{(k)}$ вариационного ряда называются ***порядковыми статистиками***, а крайние члены вариационного ряда $X_{(1)}$, $X_{(n)}$ — ***экстремальными порядковыми статистиками***.

Например, для $k=1$ функция $\varphi(z_n)$ для статистики $X_{(1)}=\varphi(Z_n)$ определяется следующим образом:

$$\varphi(z_n)=\min\left\{x_k:\ k=\overline{1,n}\right\}.$$

Определение 1.10. Порядковая статистика $X_{([np]+1)}$ с номером $[np]+1$, где $[\cdot]$ — целая часть числа, называется *выборочной квантилью уровня p*.

Выборочная (эмпирическая) функция распределения

Пусть $X_1,\ldots,X_n$ независимые наблюдения случайной величины X с функцией распределения $F(x)$. Обозначим через $\nu_n(x)$ случайную величину, реализация которой для каждого $x\in\mathbb{R}$ и соответствующей реализации $z_n=(x_1,\ldots,x_n)$ выборки $Z_n=(X_1,\ldots,X_n)$ равна количеству наблюдений (т.е. количеству чисел $x_1,\ldots,x_n$), оказавшихся не больше x.

Отметим, что $\nu_n(x)$ при фиксированном x является биномиальной случайной величиной с вероятностью "успеха" $p=\mathsf{P}\{X\leqslant x\}=F(x)$.

Определение 1.11. Функцию

$$\hat{F}_n(x)=\frac{\nu_n(x)}{n},\quad x\in\mathbb{R},\tag{1.1}$$

будем называть ***выборочной (эмпирической) функцией распределения*** случайной величины X.

Для каждого фиксированного $x\in\mathbb{R}^1$ случайная величина $\widehat{F}_n(x)$ является статистикой, реализациями которой являются числа $0,\ 1/n,\ 2/n,\ldots,n/n$, и при этом

$$\mathsf{P}\left\{\widehat{F}_n(x)=\frac{k}{n}\right\}=\mathsf{P}\{\nu_n(x)=k\}=C_n^k p^k(1-p)^{n-k},\quad k=\overline{0,n},\quad p=\mathsf{P}\{X\leqslant x\}=F(x).$$

Любая реализация выборочной функции $\widehat{F}_n(x)$ является ступенчатой функцией, характерный вид которой показан на рис. 1.1. В точках $x_{(1)}<\ldots<x_{(n)}$, где, напомним, $x_{(k)}$ — реализация порядковой статистики $X_{(k)}$, реализация выборочной функции $\widehat{F}_n(x)$ имеет скачки величиной $1/n$ и является непрерывной справа.

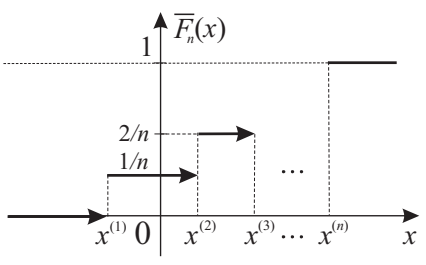


Рис 1.1.

Свойства $\widehat{F}_n(x)$

1) $\mathsf{E}\left[\widehat{F}_n(x)\right] = F(x)$, для любого $x \in \mathbb{R}^1$ и любого $n \geqslant 1$.

Действительно, при фиксированном x выборочная функция распределения $\widehat{F}_n(x)$ является частотой $\frac{\nu_n}{n}$, математическое ожидание которой равно $p = F(x)$.

2) Согласно УЗБЧ (в форме Колмогорова)

$$\widehat{F}_n(x) - F(x) \xrightarrow{\text{п.н.}} 0$$

при $n \to \infty$ для любого $x \in \mathbb{R}^1$.

3) Усиленный вариант свойства 2) даёт ***теорема Гливенко–Кантелли***

$$\sup_{x \in \mathbb{R}^1} \left|\widehat{F}_n(x) - F(x)\right| \xrightarrow{\text{п.н.}} 0$$

при $n \to \infty$. Доказательство данного свойства можно найти в учебнике А.Н. Ширяева "Вероятность".

4) $\left(\widehat{F}_n(x) - F(x)\right) \Big/ \sqrt{\frac{F(x)(1-F(x))}{n}} \xrightarrow{d} U$ при $n \to \infty$, где случайная величина U имеет распределение $\mathcal{N}(0;1)$.

Так как $\frac{F(x)(1-F(x))}{n}$ является дисперсией случайной величины $\widehat{F}_n(x)$, то данное свойство вытекает из ***теоремы Муавра–Лапласа***.

Свойства 2) и 3) свидетельствуют о том, что при увеличении числа испытаний n происходит сближение выборочной функции распределения $\widehat{F}_n(x)$ с функцией распределения $F(x)$ случайной величины X. Свойство 4) позволяют оценить скорость этого сближения в зависимости от объема n выборки.

Гистограмма

Рассмотрим процедуру ***группировки*** выборки. Для этого действительную ось $\mathbb{R}^1 = (-\infty,\infty)$ разделим точками $\alpha_0,\ldots,\alpha_{l+1}$ на $l+1$ непересекающихся полуинтервалов (***разрядов***) $\Delta_k = [\alpha_k,\alpha_{k+1})$, $k = \overline{0,l}$, таким образом, что $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_l < \alpha_{l+1} = +\infty$, $\alpha_1 \leqslant x_{(1)}$, $\alpha_l \geqslant x_{(n)}$. Обычно длина разрядов Δ_k, $k = \overline{1,l-1}$, выбирается одинаковой, т. е. равной $h_k = (\alpha_l - \alpha_1)/(l-1)$. Используя реализацию вариационного ряда $x_{(1)} < \ldots < x_{(n)}$, для каждого k-го разряда Δ_k, $k = \overline{1,l-1}$, вычислим частоту попадания элементов реализации выборки в этот разряд. Получаем $\overline{p}_k = n_k/n$, где n_k — число элементов реализации выборки z_n, попавших в k-й разряд. Если рассмотреть выборку Z_n и случайное число N_k элементов этой выборки, попавших в k-й разряд, то получим набор случайных величин $\widehat{p}_k = N_k/n$.

$[\alpha_1,\alpha_2)$	$\cdots$	$[\alpha_{l-1},\alpha_l]$
$\overline{p}_1$	$\cdots$	$\overline{p}_{l-1}$

Таблица 1.1.

Определение 1.12. Последовательность пар $(\Delta_k,\widehat{p}_k)$, $k = \overline{1,l-1}$, называется ***статистическим рядом***, а его реализация $(\Delta_k,\overline{p}_k)$, $k = \overline{1,l-1}$, представляется в виде табл. 1.1.

Изобразим графически статистический ряд.

Определение 1.13. На оси Ox отложим разряды и на них, как на основании, построим прямоугольники с высотой, равной $\overline{p}_k/h_k$, $k = \overline{1,l-1}$. Тогда площадь каждого прямоугольника будет равна $\overline{p}_k$. Полученная фигура называется ***столбцовой диаграммой***, а кусочно-постоянная функция $\overline{f}_n(x)$, образованная верхними гранями полученных прямоугольников, — ***гистограммой*** (рис. 1.2).

При этом полагают $\overline{f}_n(x) = 0$ для всех $x < \alpha_1$ и $x \geqslant \alpha_l$, так как $n_0 = 0$ и $n_l = 0$.

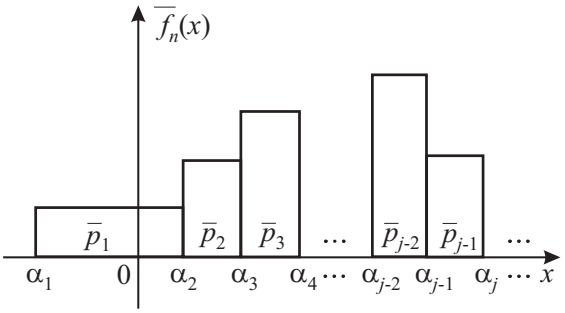


Рис 1.2.

Пусть плотность распределения $f(x)$ непрерывна и ограничена. При выборе числа интервалов группировки можно воспользоваться формулой Стерджесса $l = 1 + [3.321\lg n]$, где [] — целая часть числа. Тогда выборочная плотность распределения $\widehat{f}_n(x)$, реализациями которой служат гистограммы $\overline{f}_n(x)$, сходится по вероятности к плотности $f(x)$ наблюдаемой случайной величины, т. е. $\widehat{f}_n(x) \xrightarrow{\text{P}} f(x)$ при $n \to \infty$ для любого $x \in \mathbb{R}^1$. Таким образом, при достаточно «мелком» разбиении отрезка $[\alpha_1, \alpha_l]$ и при большом объеме выборки n высоты построенных прямоугольников можно принимать в качестве приближенных значений плотности $f(x)$ в средних точках соответствующих интервалов. Из этого следует, что гистограмму можно рассматривать как статистический аналог плотности распределения наблюдаемой случайной величины X. Используя гистограмму, неизвестную плотность можно аппроксимировать кусочно-постоянной функцией.

Определение 1.14. Сглаженную гистограмму в виде ломанной, у которой прямые линии последовательно соединяют середины верхних граней прямоугольников, образующих столбцовую диаграмму, называют *полигоном частот*.

Пример 1.1. В 1889–1890 гг. был измерен рост 1000 взрослых мужчин (рабочих московских фабрик). Результаты измерений представлены в табл. 1.2. По имеющимся наблюдениям требуется построить гисторамму.

Рост [см]	143–146	146–149	149–152	152–155	155–158
Число мужчин	1	2	8	26	65
Рост [см]	158–161	161–164	164–167	167–170	170–173
Число мужчин	120	181	201	170	120
Рост [см]	173–176	176–179	179–182	182–185	185–188
Число мужчин	64	28	10	3	1

Таблица 1.2

Р е ш е н и е: Пусть случайная величина X — рост взрослого мужчины.

В данной задаче группировка выборки уже проведена: действительная ось разделена на 17 полуинтервалов Δ_k, $k = \overline{0,16}$, где $\Delta_0 = (-\infty, 143)$, $\Delta_{16} = [188, +\infty)$, а остальные 15 полуинтервалов имеют одинаковую длину $h = 3$. Во второй строке таблицы приведены числа n_k, $k = \overline{1,15}$, равные количеству элементов выборки, попавших в k-й разряд. Вычислим частоту попадания в k-й полуинтервал, $k = \overline{1,15}$: $\overline{p}_k = n_k/1000$, и построим реализацию статистического ряда (см. табл. 1.3).

Δ_k	143–146	146–149	149–152	152–155	155–158
$\overline{p}_k$	0,001	0,002	0,008	0,026	0,065
Δ_k	158–161	161–164	164–167	167–170	170–173
$\overline{p}_k$	0,12	0,181	0,201	0,17	0,12
Δ_k	173–176	176–179	179–182	182–185	185–188
$\overline{p}_k$	0,064	0,028	0,01	0,003	0,001

Таблица 1.3

Теперь на оси OX отложим разряды Δ_k, $k = \overline{1,15}$, и на них, как на основании, построим прямоугольники высотой $\overline{p}_k/h$ (рис. 1.3).

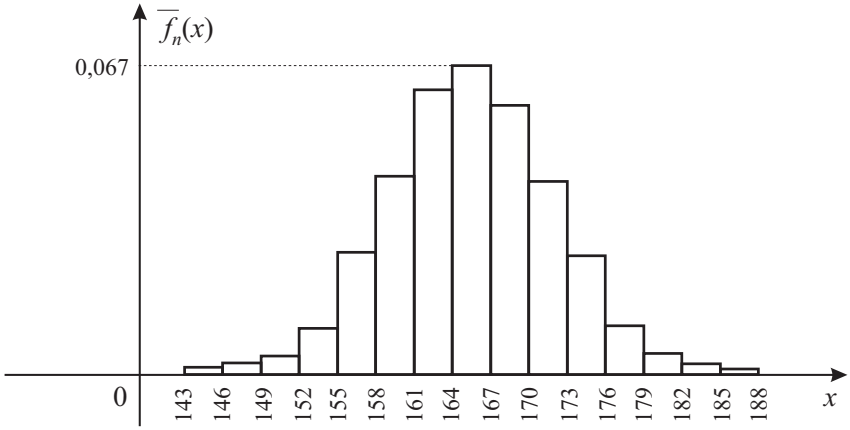


Рис 1.3.

Лекция 2

Точечные оценки

2.1 Выборочные моменты

Пусть имеется выборка $Z_n = (X_1, \ldots, X_n)$, которая порождена случайной величиной X с функцией распределения $F_X(x)$.

Определение 2.1. Для выборки Z_n объема n ***выборочными начальными*** и ***центральными моментами порядка*** r случайной величины X называются следующие случайные величины:

$$\widehat{\mu}_r = \frac{1}{n}\sum_{k=1}^{n}(X_k)^r, \quad r = 1,2,\ldots;$$
$$\widehat{\nu}_r = \frac{1}{n}\sum_{k=1}^{n}(X_k - \widehat{\mu}_1)^r, \quad r = 2,3,\ldots$$

Определение 2.2. ***Выборочным средним*** и ***выборочной дисперсией*** случайной величины X называются соответственно

$$\overline{X} = \widehat{\mu}_1 = \frac{1}{n}\sum_{k=1}^{n}X_k,$$
$$s_X^2 = \widehat{\nu}_2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2.$$

В дальнейшем мы будем также использовать обозначения $\widehat{m}_X = \overline{X}$, $\widehat{d}_X = s_X^2$. Пусть имеется также выборка $V_n = (Y_1, \ldots, Y_n)$, порожденная случайной величиной Y с функцией распределения $F_Y(y)$.

Определение 2.3. ***Выборочной ковариацией*** случайных величин X и Y называют

$$\widehat{k}_{XY} = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y}).$$

Определение 2.4. ***Выборочным коэффициентом корреляции*** случайные величины X и Y называют

$$\widehat{\rho}_{XY} = \frac{\widehat{k}_{XY}}{s_X s_Y}.$$

Пусть существуют исследуемые моменты μ_r, ν_r. Тогда справедливы следующие свойства.

Свойства выборочных моментов

1) $\mathsf{E}[\widehat{\mu}_r] = \mu_r$ для любого $n \geqslant 1$ и для всех $r = 1,2,\ldots$ Действительно,

$$\mathsf{E}[\widehat{\mu}_r] = \frac{1}{n}\sum_{k=1}^{n}\mathsf{E}[X_k^r] = \frac{1}{n}(n\mu_r) = \mu_r.$$

2) $\widehat{\mu}_r \xrightarrow{\text{п.н.}} \mu_r$ при $n \to \infty$ для всех $r = 1,2,\ldots$ Это свойство вытекает из теоремы Колмогорова.

3) $\widehat{\nu}_r \xrightarrow{\text{п.н.}} \nu_r$ при $n \to \infty$ для всех $r = 2,3,\ldots$ Используя разложение бинома Ньютона, получим

$$\widehat{\nu}_r = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^r = \frac{1}{n}\sum_{k=1}^{n}\sum_{i=0}^{r} C_r^i X_k^i (-\overline{X})^{r-i} = \\ = \sum_{i=0}^{r} C_r^i (-\overline{X})^{r-i}\left(\frac{1}{n}\sum_{k=1}^{n} X_k^i\right) = \sum_{i=0}^{r} C_r^i (-\overline{X})^{r-i}\widehat{\mu}_i.$$

Используя свойство 2), устанавливаем, что $\widehat{\mu}_i \xrightarrow{\text{п.н.}} \mu_i$, $i = 1,\ldots,r$. Проводя обратное преобразование по биному Ньютона, получаем требуемое утверждение.

4) $\mathsf{D}\left[\overline{X}\right] = d_X/n$, где $d_X = \mathsf{D}[X]$. В самом деле, воспользовавшись независимостью случайных величин X_k, $k = 1,\ldots,n$, находим

$$\mathsf{D}\left[\overline{X}\right] = \mathsf{D}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}\sum_{k=1}^{n}\mathsf{D}[X_k] = \frac{1}{n^2}(nd_X) = \frac{d_X}{n}.$$

5) $\mathsf{E}\left[s^2\right] = \frac{n-1}{n}\,d_X$.

Докажем это свойство на семинаре.

6) $\left(\overline{X} - m_X\right) \Big/ \sqrt{d_X/n} \xrightarrow{d} U_1$ при $n \to \infty$, где случайная величина U_1 имеет распределение $\mathscr{N}(0;1)$. Поскольку последовательность X_i, $i = 1,2,\ldots$, образована независимыми одинаково распределенными случайными величинами, то $\mathsf{E}\left[\overline{X}\right] = m_X$ по свойству 1), и $\mathsf{D}\left[\overline{X}\right] = d_X/n$ по свойству 4). Таким образом, к последовательности

$$S_n = \frac{1}{\sqrt{d_X n}}\sum_{k=1}^{n}(X_k - m_X) = \frac{1}{\sqrt{d_X/n}}\left(\overline{X} - m_X\right)$$

применима ЦПТ.

Второе и третье свойства указывают на то, что с увеличением объема выборки выборочные моменты будут сколь угодно близки к соответствующим теоретическим моментам.

Из первого свойства вытекает, что математические ожидания выборочных начальных моментов совпадают с соответствующими значениями начальных моментов случайной величины X, т. е. в этом смысле обладают свойством «несмещенности». А математическое ожидание выборочной дисперсии s^2 не совпадает с дисперсией d_X случайной величины X, т. е. в этом смысле случайная величина s^2 является «смещенной» выборочной характеристикой d_X. Поэтому часто вместо s^2 используют «***исправленную***» выборочную дисперсию $\widetilde{s}_X^2 = \frac{n}{n-1}\,s^2$, для которой $\mathsf{E}\left[\widetilde{s}_X^2\right] = d_X$.

Пример 2.1. В метеорологии принято характеризовать температуру месяца ее средним значением (среднее значение температуры месяца равно сумме температур всех дней данного месяца, деленной на число дней в этом месяце). В табл. 2.1 приведены значения средней температуры января в Саратове и Алатыре.

Год	1891	1892	1893	1894	1895	1896	1897
Саратов	−19,2	−14,8	−19,6	−11,1	−9,4	−16,9	−13,7
Алатырь	−21,8	−15,4	−20,8	−11,3	−11,6	−19,2	−13,0
Год	1899	1911	1912	1913	1914	1915	
Саратов	−4,9	−13,9	−9,4	−8,3	−7,9	−5,3	
Алатырь	−7,4	−15,1	−14,4	−11,1	−10,5	−7,2	

Таблица 2.1

Требуется по данным реализациям найти:

а) выборочное среднее и выборочную дисперсию средней температуры января в Саратове и Алатыре;

б) выборочный коэффициент корреляции средней температуры января в Саратове и средней температуры января в Алатыре.

Р е ш е н и е. Пусть случайная величина X — средняя температура января в Саратове, а случайная величина Y — средняя температура января в Алатыре. В табл. 2.1 приведена реализация $x_1,\ldots,x_{13}$ выборки $X_1,\ldots,X_{13}$, порожденной случайной величиной X, и реализация $y_1,\ldots,y_{13}$ выборки $Y_1,\ldots,Y_{13}$, порожденной случайнойвеличиной Y. Выборочное среднее $\overline{X}$ случайной величины X для данной реализации $x_1,\ldots,x_{13}$ равно

$$\overline{X} = \frac{1}{13}\sum_{i=1}^{13} x_i \approx -11{,}87,$$

а выборочное среднее $\overline{Y}$ случайной величины Y равно

$$\overline{Y} = \frac{1}{13}\sum_{i=1}^{13} y_i \approx -13{,}75.$$

Выборочная дисперсия s_X^2 случайной величины X для данной реализации $x_1,\ldots,x_{13}$ равна

$$s_X^2 = \frac{1}{13}\sum_{i=1}^{13} (x_i - (-11{,}87))^2 \approx 22{,}14,$$

а выборочная дисперсия s_Y^2 случайной величины Y равна

$$s_Y^2 = \frac{1}{13}\sum_{i=1}^{13} (y_i - (-13{,}75))^2 \approx 20{,}09.$$

Выборочный коэффициент корреляции случайной величины X и Y:

$$\widehat{\rho}_{XY} = \frac{\sum_{i=1}^{13} (x_i - (-11{,}87))\,(y_i - (-13{,}75))}{13\sqrt{22{,}14}\sqrt{20{,}09}} \approx 0{,}95.$$

О т в е т. $\overline{X} \approx -11{,}87$, $\overline{Y} \approx -13{,}75$, $s_X^2 \approx 22{,}14$, $s_Y^2 \approx 20{,}09$, $\widehat{\rho}_{XY} \approx 0{,}95$.

2.2 Основные распределения в статистике

Распределение хи-квадрат

Определение 2.5. Пусть U_k, $k = \overline{1,n}$, — набор из n независимых нормально распределенных случайных величин, $U_k \sim \mathcal{N}(0;1)$. Тогда случайная величина

$$X_n = \sum_{k=1}^{n} U_k^2$$

имеет ***распределение хи-квадрат*** (***χ^2-распределение***) с n ***степенями свободы***, что обозначается как $X_n \sim \chi^2(n)$.

С в о й с т в а р а с п р е д е л е н и я х и-к в а д р а т $\chi^2(n)$

1) Случайная величина X_n имеет следующую плотность распределения:

$$f(x,n) = \begin{cases} \frac{1}{2^{(n/2)}\Gamma(n/2)}\, x^{(n/2)-1} e^{-x/2}, & x > 0, \\ 0, & x \leqslant 0, \end{cases}$$

где $\Gamma(m) = \int\limits_0^{+\infty} y^{m-1} e^{-y}\, dy$ — гамма-функция. Графики функций $f(x,n)$ (рис. 2.1), называемые ***кривыми Пирсона***, асимметричны и начиная с $n > 2$ имеют один максимум в точке $x = n - 2$.

2) Случайная величина $X_n \sim \chi^2(n)$ имеет следующие моменты:

$$\mathsf{E}[X_n] = n, \qquad \mathsf{D}[X_n] = 2n.$$

3) Сумма любого числа m независимых случайных величин X_k, $k = \overline{1,m}$, имеющих распределение хи-квадрат с n_k степенями свободы, имеет распределение хи-квадрат с $n = \sum\limits_{k=1}^{m} n_k$ степенями свободы.

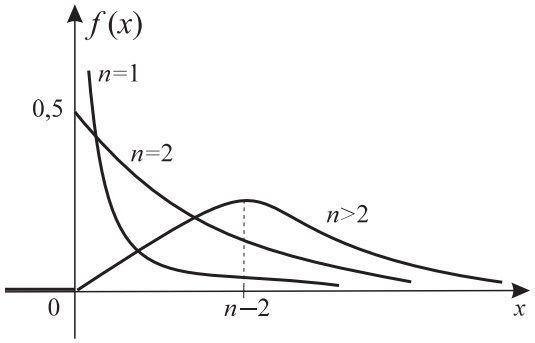


Рис 2.1

4) Распределение хи-квадрат обладает свойством ***асимптотической нормальности***:

$$\frac{X_n - n}{\sqrt{2n}} \xrightarrow{d} U \qquad \text{при } n \to \infty,$$

где случайная величина U имеет распределение $\mathcal{N}(0;1)$. Это означает, что при достаточно большом объеме n выборки можно приближенно считать $X_n \sim \mathcal{N}(n;2n)$. Фактически эта аппроксимация имеет место уже при $n \geqslant 30$.

Пример 2.2. Приведем пример, в котором возникает распределение хи-квадрат. Пусть выборка Z_n соответствует нормальному распределению $\mathcal{N}(m;\sigma^2)$. Рассмотрим выборочную дисперсию

$$s_X^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2,$$

где $\overline{X}$ — выборочное среднее. Тогда случайная величина $Y_n = ns_X^2 \big/ \sigma^2$ имеет распределение $\chi^2(n-1)$ и не зависит от $\overline{X}$.

Распределение Стьюдента

Определение 2.6. Пусть U и X_n — независимые случайные величины, $U \sim \mathcal{N}(0;1)$, $X_n \sim \chi^2(n)$. Тогда случайная величина $T_n = U \Big/ \sqrt{X_n/n}$ имеет ***распределение Стьюдента с n степенями свободы***, что обозначают как $T_n \sim \mathbf{t}(n)$.

Свойства распределения Стьюдента $\mathbf{t}(n)$

1) Случайная величина T_n имеет плотность распределения

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

Графики плотностей $f(t,n)$ (рис. 2.2), называемые ***кривыми Стьюдента***, симметричны при всех $n = 1,2,\ldots$ относительно оси ординат.

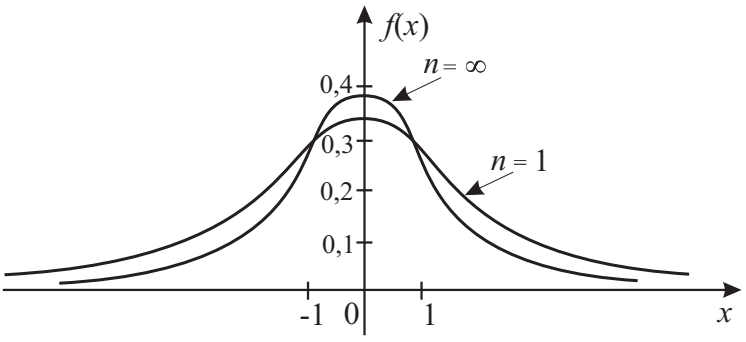


Рис 2.2

2) Случайная величина T_n имеет математическое ожидание, равное $\mathsf{E}[T_n] = 0$ для всех $n \geqslant 2$, и дисперсию $\mathsf{D}[T_n] = n/(n-2)$ при $n > 2$. При $n = 2$ дисперсия $\mathsf{D}[T_n] = +\infty$.

3) При $n = 1$ распределение Стьюдента называется ***распределением Коши***, плотность которого равна

$$f(t,1) = \frac{1}{\pi}\frac{1}{1+t^2}.$$

Математическое ожидание и дисперсия случайной величины T_1, имеющей распределение Коши, не существуют, так как бесконечен предел

$$\lim_{a\to\infty} I(a) = +\infty,$$

где $I(a) = \frac{1}{\pi}\int\limits_0^a \frac{t}{t^2+1}\,dt$.

4) Можно показать, что при $n \to \infty$ распределение $\mathbf{t}(n)$ асимптотически нормально, т. е. $T_n \xrightarrow{d} U$, где случайная величина U имеет распределение $\mathcal{N}(0;1)$. При $n \geqslant 30$ распределение Стьюдента $\mathbf{t}(n)$ практически не отличается от $\mathcal{N}(0;1)$.

Пример 2.3. Приведем пример, в котором встречается распределение Стьюдента. Пусть выборка Z_n соответствует нормальному распределению $\mathcal{N}(m;\sigma^2)$. Пусть $\overline{X}$ — выборочное среднее, а s_X^2 — выборочная дисперсия. Тогда случайная величина

$$T_n = \sqrt{n-1}\,\frac{\overline{X}-m}{\sqrt{s_X^2}}$$

имеет распределение Стьюдента $\mathbf{t}(n-1)$.

Распределение Фишера

Определение 2.7. Пусть независимые случайные величины X_n и X_m имеют распределения хи-квадрат соответственно с n и m степенями свободы. Тогда случайная величина $V_{n,m} = \frac{X_n/n}{X_m/m}$ имеет ***распределение Фишера с n и m степенями свободы***, что записывают как $V_{n,m} \sim \mathbf{F}(n;m)$.

Свойства распределения Фишера $\mathbf{F}(n;m)$

1) Случайная величина $V_{n,m}$ имеет плотность $f(v,n,m) = 0$ при $v \leqslant 0$ и

$$f(v,n,m) = \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}\, n^{\frac{n}{2}} m^{\frac{m}{2}}\, \frac{v^{\frac{n}{2}-1}}{(m+nv)^{\frac{n+m}{2}}} \quad \text{при } v > 0.$$

Графики функции $f(v,n,m)$, называемые ***кривыми Фишера***, асимметричны и при $n > 2$ достигают максимальных значений в точках $v = \frac{(n-2)m}{(m+2)n}$, близких к единице при больших значениях m и n. Типовой вид кривой Фишера приведен на рис. 2.3.

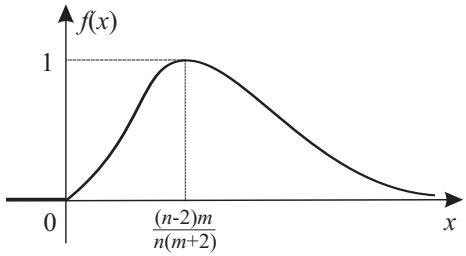


Рис 2.3

2) Случайная величина $V_{n,m}$ имеет следующие моменты:

$$\mathsf{E}[V_{n,m}] = \frac{m}{m-2} \text{ при } m > 2, \qquad \mathsf{D}[V_{n,m}] = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)} \text{ при } m > 4.$$

Пример 2.4. Пусть $Z_n = (X_1,\ldots,X_n)$ — выборка объема n, порожденная случайнойвеличиной X с нормальным распределением $\mathcal{N}(m_X;\sigma^2)$, а $W_n = (Y_1,\ldots,Y_m)$ — выборка объема m, порожденная случайнойвеличиной Y с нормальным распределением $\mathcal{N}(m_Y;\sigma^2)$, и случайные величины Z_n и W_n независимы. Тогда случайная величина $V_{n,m}$, образованная отношением «исправленных» выборочных дисперсий случайных величин X и Y, т. е.

$$V_{n,m} = \frac{\frac{1}{n-1}\sum\limits_{k=1}^{n}\left(X_k-\overline{X}\right)^2}{\frac{1}{m-1}\sum\limits_{j=1}^{m}\left(Y_j-\overline{Y}\right)^2},$$

имеет распределение $\mathbf{F}(n-1;m-1)$.

2.3 Определения и свойства оценок

Определение 2.8. ***Параметром распределения*** $\theta \in \Theta \subset \mathbb{R}^1$ случайной величины X называется любая числовая характеристика этой случайной величины (математическое ожидание, дисперсия и т. п.) или любая константа, явно входящая в выражение для функции распределения.

В общем случае будем предполагать, что параметр распределения θ может быть векторным, т. е. $\theta \in \Theta \subset \mathbb{R}^s$.

В случае параметрической статистической модели $(S_\theta, F_{Z_n}(z_n,\theta))$ таким параметром распределения может служить неизвестный вектор $\theta \in \Theta \subset \mathbb{R}^s$, характеризующий распределение $F_{Z_n}(z_n,\theta)$.

Пусть имеется выборка $Z_n = (X_1,\ldots,X_n)$ с реализацией $z_n = (x_1,\ldots,x_n)$.

Определение 2.9. ***Точечной*** **(*выборочной*)** ***оценкой*** неизвестного параметра распределения $\theta \in \Theta \subset \mathbb{R}^s$ называется произвольная статистика $\widehat{\theta}(Z_n)$, построенная по выборке Z_n и принимающая значения в множестве Θ.

Ясно, что существует много разных способов построения точечной оценки, которые учитывают тип статистической модели. Для параметрической и непараметрической моделей эти способы могут быть различны. Рассмотрим некоторые свойства, которые характеризуют качество введенной оценки.

Определение 2.10. Оценка $\widehat{\theta}(Z_n)$ параметра θ называется ***несмещенной***, если ее математическое ожидание равно θ, т. е. $\mathsf{E}\left[\widehat{\theta}(Z_n)\right] = \theta$ для любого $\theta \in \Theta$.

Определение 2.11. Оценка $\widehat{\theta}(Z_n)$ параметра θ называется ***асимптотически несмещенной***, если $\lim\limits_{n\to\infty} \mathsf{E}\left[\widehat{\theta}(Z_n)\right] = \theta$ для любого $\theta \in \Theta$.

Определение 2.12. Оценка $\widehat{\theta}(Z_n)$ параметра θ называется ***состоятельной***, если она сходится по вероятности к θ, т. е. $\widehat{\theta}(Z_n) \xrightarrow{\text{P}} \theta$ при $n \to \infty$ для любого $\theta \in \Theta$.

Определение 2.13. Оценка $\widehat{\theta}(Z_n)$ параметра θ называется ***сильно состоятельной***, если она сходится почти наверное к θ, т. е. $\widehat{\theta}(Z_n) \xrightarrow{\text{п.н.}} \theta$ при $n \to \infty$ для любого $\theta \in \Theta$.

Очевидно, что если оценка сильно состоятельная, то она является также состоятельной.

Пример 2.5. Оценка $\widehat{\theta}_1(Z_n) = \overline{X}$ неизвестного математического ожидания $\theta_1 = m_X$ случайной величины X является несмещенной (по свойству 1 выборочных моментов).
А оценка $\widehat{\theta}_2(Z_n) = s_X^2$ неизвестной дисперсии $\theta_2 = d_X$ — смещенной, так как $\mathsf{E}\left[s_X^2\right] = \frac{n-1}{n} d_X$ (по свойству 5 выборочных моментов). Оценка $\widehat{\theta}_2(Z_n) = \widetilde{s}_X^2$ является несмещенной оценкой d_X по определению $\widetilde{s}_X^2$.
Если существуют моменты m_X, d_X, то сильная состоятельность оценок $\overline{X}$, $\widehat{d}_X$ гарантируется свойствами 2), 3) выборочных моментов.

Свойствами состоятельности и несмещенности могут обладать сразу несколько оценок неизвестного параметра θ.

Определение 2.14. Несмещенная оценка $\widehat{\theta}^*(Z_n)$ скалярного параметра θ называется ***эффективной***, если $\mathsf{D}\left[\widehat{\theta}^*(Z_n)\right] \leqslant \mathsf{D}\left[\widehat{\theta}(Z_n)\right]$ для всех несмещенных оценок $\widehat{\theta}(Z_n)$ параметра θ, т. е. ее дисперсия минимальна по сравнению с дисперсиями других несмещенных оценок при одном и том же объеме n выборки Z_n.

Вообще говоря, дисперсии несмещенных оценок могут зависеть от параметра θ. В этом случае под эффективной оценкой понимается такая, для которой вышеприведенное неравенство является строгим хотя бы для одного значения параметра θ.

Определение 2.15. Параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n,\theta))$, $\theta \in \Theta \subset \mathbb{R}^1$, называется *регулярной*, если выполняются следующие условия:

1) функция $L(z_n,\theta) = \prod\limits_{k=1}^{n} f_X(x_k,\theta) > 0$ для всех $\theta \in \Theta$ и $z_n \in S_\theta$ дифференцируема по параметру $\theta \in \Theta$;

2) для любого измеримого множества $A \subset S$ выполняется условие

$$\frac{\partial}{\partial\theta}\int\limits_A L(z_n,\theta)\,dz_n = \int\limits_A \frac{\partial}{\partial\theta}L(z_n,\theta)\,dz_n.$$

Из условия 2) вытекает, в частности, что в случае регулярной модели выборочное пространство $S_\theta = S$, т. е. не зависит от неизвестного параметра θ.

Пример 2.6. Пусть выборка Z_n соответствует равномерному распределению $R(0;b)$ с неизвестным параметром $\theta = b$. В этом случае параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n,\theta))$ не является регулярной, так как выборочное пространство S_θ определяется на отрезках $[0,b]$, а следовательно, зависит от параметра b.

Определение 2.16. В случае регулярной статистической модели $(S, F_{Z_n}(z_n,\theta))$, $\theta \in \Theta \subset \mathbb{R}^1$, величина

$$I_n(\theta) = \mathsf{E}\left[\left(\frac{\partial \ln L(Z_n,\theta)}{\partial\theta}\right)^2\right]$$

называется *информацией Фишера* о параметре $\theta \in \Theta$, содержащейся в выборке Z_n.

Замечание 2.1. Можно показать, что

$$I_n(\theta) = nI_1(\theta),$$

где $I_1(\theta)$ -информация Фишера о параметре, содержащаяся в одном наблюдении.

В случае регулярной модели $(S, F_{Z_n}(z_n,\theta))$ для любой несмещенной оценки $\widehat{\theta}(Z_n)$ параметра $\theta \in \Theta \subset \mathbb{R}$ справедливо *неравенство Рао–Крамера*

$$\mathsf{D}\left[\widehat{\theta}(Z_n)\right] \geqslant \frac{1}{I_n(\theta)}.$$

Это неравенство дает нижнюю границу для дисперсии несмещенной оценки.

Определение 2.17. Несмещенная оценка $\widehat{\theta}(Z_n)$ параметра $\theta \in \Theta \subset \mathbb{R}$ называется *R-эффективной оценкой*, если для этой оценки в неравенстве Рао–Крамера достигается равенство, т. е.

$$\mathsf{D}\left[\widehat{\theta}(Z_n)\right] = \frac{1}{I_n(\theta)}.$$

Определение 2.18. Эффективностью несмещённой оценки $\widehat{\theta}(Z_n)$ называется величина

$$e(\widehat{\theta}(Z_n)) = \frac{1}{I_n(\theta)\mathsf{D}(\widehat{\theta}(Z_n))}.$$

Если R-эффективная оценка существует, то она является также эффективной в смысле минимума дисперсии (см. определение 2.14).

Способ построения R-эффективных оценок вытекает из *критерия эффективности*, состоящего в следующем. Оценка $\widehat{\theta}(Z_n)$ параметра θ является R-эффективной тогда и только тогда, когда существует некоторая функция $a(\theta)$, такая что выполняется равенство

$$\widehat{\theta}(Z_n) - \theta = a(\theta)\sum_{k=1}^{n}\frac{\partial \ln f(X_k,\theta)}{\partial\theta}.$$

Пример 2.7. Пусть случайная величина X имеет нормальное распределение $\mathbf{N}(\theta;\sigma_X^2)$ с неизвестным параметром $\theta = m_X$. Найдем R-эффективную оценку параметра θ. С этой целью вычислим

$$\sum_{k=1}^{n}\frac{\partial \ln f(X_k,\theta)}{\partial\theta} = \frac{1}{\sigma_X^2}\sum_{k=1}^{n}(X_k-\theta) = \frac{n}{\sigma_X^2}\left(\frac{1}{n}\sum_{k=1}^{n}X_k - \theta\right) = \frac{n}{\sigma_X^2}(\widehat{m}_X - \theta).$$

В этом случае $a(\theta) = n/\sigma_X^2$ и R-эффективной оценкой m_X является выборочное среднее $\widehat{m}_X$. Как отмечалось выше, R-эффективная оценка является также эффективной. Поэтому $\widehat{m}_X$ является эффективной оценкой m_X.

Модель	$I_n(\theta)$	$\widehat{\theta}(Z_n)$	$\mathsf{D}[\widehat{\theta}(Z_n)]$
$\mathbf{N}(\theta;\sigma_X^2)$	$\frac{n}{\sigma_X^2}$	$\widehat{m}_X = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$	$\frac{\sigma_X^2}{n}$
$\mathbf{N}(m_X;\theta)$	$\frac{n}{2\theta^2}$	$\widehat{d}_X = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - m_X)^2$	$\frac{2\theta^2}{n}$
$\mathbf{Bi}(k;\theta)$	$\frac{kn}{\theta(1-\theta)}$	$\frac{\widehat{m}_X}{k} = \frac{1}{nk}\sum\limits_{i=1}^{n} X_i$	$\frac{\theta(1-\theta)}{kn}$
$\Pi(\theta)$	$\frac{n}{\theta}$	$\widehat{m}_X = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$	$\frac{\theta}{n}$

Таблица 2.2

Пример 2.8. Приведем примеры R-эффективных (а следовательно, и эффективных) оценок $\widehat{\theta}(Z_n)$ неизвестных параметров θ некоторых распространенных распределений, дисперсии $\mathsf{D}\left[\widehat{\theta}(Z_n)\right]$ этих оценок, а также значения информациии Фишера $I_n(\theta)$ (см. табл. 2.2). Из таблицы видно, что $\mathsf{D}\left[\widehat{\theta}(Z_n)\right] = \frac{1}{I_n(\theta)}$. Следовательно, согласно определению 2.17, эти оценки являются R-эффективными.

Пример 2.9. Приведем пример, когда эффективная оценка неизвестного параметра существует, а R-эффективная оценка не существует. Пусть выборка Z_n соответствует распределению $\mathbf{N}(m;\sigma^2)$ с неизвестными параметрами $\theta_1 = m$, $\theta_2 = \sigma^2$. В данном случае параметрическая модель является регулярной. Из п. 2.1 следует, что статистика

$$\widehat{s}_X(Z_n) = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \widehat{m}_X)^2$$

является несмещенной оценкой неизвестной дисперсии $d_X = \sigma^2$. Как указано в примере 2.2 из п. 2.2, случайная величина

$$Y_n = \frac{(n-1)\widehat{s}_X(Z_n)}{\sigma^2}$$

имеет распределение хи-квадрат $\chi^2(n-1)$, а следовательно, её дисперсия равна $\mathsf{D}[Y_n] = 2(n-1)$. Поэтому

$$\mathsf{D}\left[\widehat{s}_X(Z_n)\right] = \frac{2\sigma^4}{n-1}.$$

Согласно неравенству Рао-Крамера нижняя граница дисперсии любой несмещенной оценки в этой статистической модели равна $2\sigma^4/n$. Таким образом, $\widehat{s}_X(Z_n)$ не является R-эффективной оценкой. Но можно показать, что эта оценка является эффективной, т. е. $\widehat{s}_X(Z_n)$ имеет минимальную дисперсию. Таким образом, нижняя граница в неравенстве Рао-Крамера не достигается, а следовательно, R-эффективной оценки не существует.

Робастные оценки

Определение 2.19. Пусть $\widehat{\theta}_n$ — оценка параметра θ, построенная по выборке $Z_n = (X_1,\ldots,X_n)$. Затем в выборку добавлено ещё одно наблюдение x, и построена оценка $\widehat{\theta}_{n+1}$. Тогда кривой чувствительности (sensitivity curve), измеряющей дифференциальное влияние наблюдения x на оценку $\widehat{\theta}_{n+1}$, называется функция

$$SC_{n+1}(x) = \frac{\widehat{\theta}_{n+1} - \widehat{\theta}_n}{1/(n+1)} = (n+1)(\widehat{\theta}_{n+1} - \widehat{\theta}_n).$$

Определение 2.20. Назовём оценку $\widehat{\theta}_n$ B-робастной, если кривая чувствительности этой оценки является ограниченной функцией.

Задача 2.1. Покажите, что: 1) выборочное среднее не является B-робастной оценкой, 2) выборочная медиана является B-робастной оценкой.

Определение 2.21. Пороговой точкой (breakdown point) ε_n^* оценки $\widehat{\theta}_n$, построенной по выборке $X_1,\ldots,X_n$, называется величина

$$\varepsilon_n^*(\widehat{\theta}_n, X_1,\ldots,X_n) = \frac{1}{n}\max\left\{m : \max_{i_1,\ldots,i_m}\sup_{y_1,\ldots,y_m}\left|\widehat{\theta}_n(Z_1,\ldots,Z_n)\right| < \infty\right\}$$

где выборка $Z_1,\ldots,Z_n$ получена заменой наблюдений $X_{i_1},\ldots,X_{i_m}$ произвольными значениями $y_1,\ldots,y_m$.

Оценки, у которых пороговая точка близка к 0.5, называют оценками с высокой пороговой точкой (high breakdown point).

Лекция 3

Методы оценивания параметров

На практике часто удается предсказать вид распределения наблюдаемой случайной величины с точностью до неизвестных параметров $\theta = (\theta_1, \ldots, \theta_s)$, т. е. для непрерывной случайной величины X оказывается известной плотность $f(x,\theta)$, а для дискретной случайной величины X — вероятности $p(x_i,\theta) = \mathsf{P}_\theta\{X = x_i\}$, $i = \overline{1,m}$, где x_i, $i = \overline{1,m}$, — возможные значения случайной величины X. Например, может быть $\theta_1 = m_X$, $\theta_2 = d_X$ при $s = 2$. Эти неизвестные параметры требуется оценить по имеющейся выборке Z_n. Расскажем о двух распространённых методах оценивания параметров - методе максимального правдоподобия (ММП) и методе моментов.

Метод максимального правдоподобия

Определение 3.1. ***Функцией правдоподобия*** для неизвестного параметра $\theta \in \Theta \subset \mathbb{R}^s$ называется: в случае непрерывной наблюдаемой случайной величины X — плотность распределения

$$L(z_n,\theta_1,\ldots,\theta_s) = f_{Z_n}(z_n,\theta_1,\ldots,\theta_s) = \prod_{k=1}^{n} f_X(x_k,\theta_1,\ldots,\theta_s),$$

где $f_X(x,\theta_1,\ldots,\theta_s)$ — плотность распределения случайной величины X, а в случае дискретной наблюдаемой случайной величины X — произведение вероятностей

$$L(z_n,\theta_1,\ldots,\theta_s) = \prod_{k=1}^{n} P_X(x_k,\theta_1,\ldots,\theta_s),$$

где $P_X(x_k,\theta_1,\ldots,\theta_s)$ — вероятность события $\{X = x_k\}$.

Рассмотрим параметрическую статистическую модель $(S_\theta, F_{Z_n}(z_n,\theta))$, $\theta \in \Theta \subset \mathbb{R}^s$, для которой известна функция правдоподобия $L(z_n,\theta_1,\ldots,\theta_s)$.

Определение 3.2. ***Оценкой максимального правдоподобия*** (МП-***оценкой***) параметра $\theta \in \Theta$ называется статистика $\widehat{\theta}(Z_n)$, максимизирующая для каждой реализации z_n функцию правдоподобия, т. е.

$$\widehat{\theta}(z_n) = \arg\max_{\theta\in\Theta} L(z_n,\theta).$$

Способ построения МП-оценки называется ***методом максимального правдоподобия***.

Поскольку функция правдоподобия $L(z_n,\theta)$ и её логарифм $\ln L(z_n,\theta)$ достигают максимума при одних и тех же значениях θ, то часто вместо $L(z_n,\theta)$ рассматривают логарифмическую функцию правдоподобия $\ln L(z_n,\theta)$.

Определение 3.3. ***Логарифмической функцией правдоподобия*** для неизвестного параметра $\theta \in \Theta \subset \mathbb{R}^s$ называется функция $\ln L(z_n,\theta_1,\ldots,\theta_s)$.

В случае дифференцируемости функции $\ln L(z_n,\theta)$ по θ МП-оценку можно найти, решая относительно $\theta_1,\ldots,\theta_s$ систему ***уравнений правдоподобия***

$$\begin{cases} \frac{\partial \ln L(z_n,\theta_1,\ldots,\theta_s)}{\partial \theta_1} = 0, \\ \ldots \\ \frac{\partial \ln L(z_n,\theta_1,\ldots,\theta_s)}{\partial \theta_s} = 0. \end{cases}$$

Пример 3.1. Если, например, случайная величина X имеет нормальное распределение $\mathcal{N}(m_X;\sigma_X^2)$ с неизвестным математическим ожиданием $\theta = m_X$, то легко установить, что оценкой максимального правдоподобия параметра $\theta = m_X$ при любых σ_X является выборочное среднее $\overline{X}$. Действительно, в этом случае

$$L(z_n,m_X) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left\{-\frac{(x_k-m_X)^2}{2\sigma_X^2}\right\} = \frac{1}{(2\pi)^{n/2}\sigma_X^n} \exp\left\{-\sum_{k=1}^{n}\frac{(x_k-m_X)^2}{2\sigma_X^2}\right\}.$$

Дифференцируя функцию $\ln L(z_n,m_X)$ по m_X и приравнивая нулю получаемое выражение, находим уравнение, решение которого

$$\widehat{\theta}(Z_n) = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}.$$

Пример 3.2. Пусть случайная величина X имеет равномерное распределение $\mathscr{R}(a;b)$ с неизвестными параметрами $\theta_1 = a$, $\theta_2 = b$. В данном случае плотность распределения $f(x,a,b) = 1/(b-a)$, если $x \in [a,b]$, и $f(x,a,b) = 0$, если $x \notin [a,b]$. Оценим параметры a и b методом максимального правдоподобия. Функция правдоподобия в данном случае имеет вид

$$L(x_1,\ldots,x_n,a,b) = \prod_{k=1}^{n} f(x_k,a,b) = \left(\frac{1}{b-a}\right)^n,$$

если $a \leqslant x_k \leqslant b$ для всех $k = \overline{1,n}$, и $L(x_1,\ldots,x_n,a,b) = 0$ в остальных случаях. Таким образом, функция правдоподобия отлична от нуля, если неизвестные параметры a и b удовлетворяют неравенствам

$$b \geqslant x_{(n)} = \max_{k=\overline{1,n}} x_k, \qquad a \leqslant x_{(1)} = \min_{k=\overline{1,n}} x_k.$$

При этом функция $L(x_1,\ldots,x_n,a,b)$ достигает максимума по a и b, когда разность $b-a$ оказывается минимально возможной, не нарушающей полученные неравенства, т. е. в случае достижения в них равенств. Таким образом получаем МП-оценки неизвестных параметров a и b:

$$\widehat{a}(Z_n) = X_{(1)}, \qquad \widehat{b}(Z_n) = X_{(n)},$$

где $X_{(n)}$ и $X_{(1)}$ — экстремальные порядковые статистики.

Пример 3.3. Пусть случайная величина X имеет распределение Пуассона $\Pi(a)$ с неизвестным параметром $\theta = a$. Построим МП-оценку параметра a. Функция правдоподобия в этом случае равна

$$L(z_n,a) = \prod_{k=1}^{n} \frac{a^{x_k}e^{-a}}{x_k!},$$

поэтому логарифмическая функция правдоподобия равна

$$\ln L(z_n,a) = \sum_{k=1}^{n} x_k \ln a - na - \ln(x_1!\cdot\ldots\cdot x_n!).$$

Решая соответствующее уравнение правдоподобия

$$\frac{\partial}{\partial a}\ln L(z_n,a) = 0,$$

находим МП-оценку неизвестного параметра a:

$$\widehat{a}(Z_n) = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}.$$

Пример 3.4. Найдем МП-оценку для вероятности p «успеха» в схеме испытаний Бернулли. В этом случае имеем распределение Бернулли $\mathbf{Bi}(1;p)$ с неизвестным параметром $\theta = p$. Поэтому

$$L(z_n,a) = \prod_{k=1}^{n} p^{x_k}(1-p)^{1-x_k},$$

где $x_k = 1$, если в k-м испытании был «успех», и $x_k = 0$ — в противном случае. Решая уравнение правдоподобия

$$\frac{\partial}{\partial p}\ln L(z_n,p) = \sum_{k=1}^{n}\left(\frac{x_k}{p} - \frac{1-x_k}{1-p}\right) = 0$$

относительно параметра p, находим МП-оценку:

$$\widehat{p}(Z_n) = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}.$$

Метод моментов

Исторически первым для оценивания неизвестных параметров был предложен следующий метод. Пусть имеется параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n,\theta))$, $\theta \in \Theta \subset \mathbb{R}^s$. Предположим, что у наблюдаемой случайной величины X, порождающей выборку Z_n, существуют начальные моменты $\mu_i = \mathsf{E}[X^i]$, $i = \overline{1,s}$. Тогда в общем случае от неизвестных параметров будут зависеть и начальные моменты, т. е. $\mu_i = \mu_i(\theta)$.

Пусть $\widehat{\mu}_i$, $i = \overline{1,s}$, — выборочные начальные моменты. Рассмотрим систему уравнений

$$\mu_i(\theta) = \widehat{\mu}_i, \quad i = \overline{1,s},$$

и предположим, что её можно разрешить относительно параметров $\theta_1,\ldots,\theta_s$, т. е. найти функции $\widehat{\theta}_i = \varphi_i(\widehat{\mu}_1,\ldots,\widehat{\mu}_s)$, $i = \overline{1,s}$.

Определение 3.4. Решение полученной системы уравнений $\widehat{\theta}_i = \varphi_i(\widehat{\mu}_1,\ldots,\widehat{\mu}_s)$, $i = \overline{1,s}$, называется ***оценкой*** параметра θ, найденной по ***методу моментов***, или ММ-***оценкой***.

Если функции $\varphi_1(\cdot),\ldots,\varphi_s(\cdot)$ непрерывны, то ММ-оценки являются состоятельными.

Замечание 3.1. Уравнения метода моментов часто оказываются более простыми по сравнению с уравнениями правдоподобия, и их решение не связано с большими вычислительными трудностями.

Пример 3.5. Пусть $Z_n = (X_1,\ldots,X_n)$ — выборка, соответствующая нормальному распределению $\mathcal{N}(m;\sigma^2)$ с неизвестными параметрами $\theta_1 = m$ и $\theta_2 = \sigma^2$. Оценим параметры m и σ^2 с помощью метода моментов. В данном случае $\mu_1 = m$, $\mu_2 = m^2 + \sigma^2$ и система уравнений для метода моментов принимает вид

$$\begin{cases} m = \widehat{\mu}_1 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k, \\ m^2 + \sigma^2 = \widehat{\mu}_2 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k^2. \end{cases}$$

Решая эту систему, находим ММ-оценки:

$$\widehat{\theta}_1(Z_n) = \widehat{\mu}_1, \qquad \widehat{\theta}_2(Z_n) = \widehat{\mu}_2 - \widehat{\mu}_1^2.$$

Пример 3.6. Пусть $Z_n = (X_1,\ldots,X_n)$ — выборка, соответствующая равномерному распределению $\mathscr{R}(a;b)$ с неизвестными параметрами $\theta_1 = a$ и $\theta_2 = b$. Оценим параметры a и b с помощью метода моментов. Поскольку для данного распределения

$$\mu_1 = \frac{a+b}{2}, \qquad \mu_2 = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2,$$

то система уравнений метода моментов принимает вид

$$\begin{cases} \dfrac{a+b}{2} = \widehat{\mu}_1 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k, \\ \dfrac{(b-a)^2}{12} + \left(\dfrac{a+b}{2}\right)^2 = \widehat{\mu}_2 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k^2. \end{cases}$$

Решая эту систему, получаем ММ-оценки неизвестных параметров:

$$\widehat{a}(Z_n) = \widehat{\mu}_1 - \sqrt{3s_X^2}, \qquad \widehat{b}(Z_n) = \widehat{\mu}_1 + \sqrt{3s_X^2},$$

где $s_X^2 = \frac{1}{n}\sum\limits_{k=1}^{n}(X_k - \overline{X})^2$.

Оценки, полученные в примере 3.5 с помощью метода моментов, совпадают с МП-оценками, найденными в примере 3.1 из п. 3, а ММ-оценки в примере 3.6 не совпадают с МП-оценками, построенными в примере 3.2 из того же пункта.

Лекция 4

Интервальные оценки

Основные понятия

Пусть имеется параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n,\theta))$, $\theta \in \Theta \subset \mathbb{R}^1$, и по выборке $Z_n = (X_1,\ldots,X_n)$, соответствующей распределению $F(x,\theta)$ наблюдаемой случайной величины X, требуется оценить неизвестный параметр θ. Вместо точечных оценок, рассмотренных ранее, рассмотрим другой тип оценок неизвестного параметра $\theta \in \Theta \subset \mathbb{R}^1$.

Определение 4.1. Интервал $[\theta_1(Z_n),\theta_2(Z_n)]$ со случайными концами, «накрывающий» с вероятностью $1-\alpha$, $0<\alpha<1$, неизвестный параметр θ, т. е.

$$\mathsf{P}\{\theta_1(Z_n) \leqslant \theta \leqslant \theta_2(Z_n)\} = 1-\alpha,$$

называется ***доверительным интервалом*** (или ***интервальной оценкой***) уровня надежности $1-\alpha$ параметра θ.

Аналогично определяется доверительный интервал для произвольной функции от параметра θ.

Определение 4.2. Число $\delta = 1-\alpha$ называется ***доверительной вероятностью*** или ***уровнем доверия (надежности)***.

Определение 4.3. Доверительный интервал $[\theta_1(Z_n),\theta_2(Z_n)]$ называется ***центральным***, если выполняются следующие условия:

$$\mathsf{P}\{\theta \geqslant \theta_2(Z_n)\} = \frac{\alpha}{2}, \qquad \mathsf{P}\{\theta_1(Z_n) \geqslant \theta\} = \frac{\alpha}{2}.$$

Часто вместо двусторонних доверительных интервалов рассматривают односторонние доверительные интервалы, полагая $\theta_1(Z_n) = -\infty$ или $\theta_2(Z_n) = +\infty$.

Определение 4.4. Интервал, границы которого удовлетворяют условию

$$\mathsf{P}\{\theta \geqslant \theta_2(Z_n)\} = \alpha \qquad (\text{или} \quad \mathsf{P}\{\theta_1(Z_n) \geqslant \theta\} = \alpha),$$

называется соответственно ***правосторонним*** (или ***левосторонним***) ***доверительным интервалом***.

Рассмотрим метод построения доверительных интервалов, основанный на центральной (опорной) статистике.

Использование центральной статистики

Определение 4.5. Функция $G(Z_n,\theta)$ случайной выборки Z_n, такая что её распределение $F_G(y)$ не зависит от параметра θ и при любом значении z_n функция $G(z_n,\theta)$ является непрерывной и строго монотонной по θ, называется ***центральной (опорной) статистикой*** для параметра θ.

Зная распределение $F_G(y)$ центральной статистики $G(Z_n,\theta)$, можно найти числа g_1 и g_2, удовлетворяющие условию

$$\mathsf{P}\{g_1 \leqslant G(Z_n,\theta) \leqslant g_2\} = 1-\alpha.$$

Тогда границы доверительного интервала $[\theta_1(Z_n),\theta_2(Z_n)]$ для параметра θ могут быть найдены, если разрешить, учитывая свойства функции $G(z_n,\theta)$, следующие неравенства:

$$g_1 \leqslant G(Z_n,\theta) \leqslant g_2.$$

В частности, если $G(z_n,\theta)$ — монотонно возрастающая по θ функция, то

$$\theta_1(Z_n) = G^{-1}(Z_n,g_1), \qquad \theta_2(Z_n) = G^{-1}(Z_n,g_2),$$

где $G^{-1}(z_n,g_1)$ — функция, обратная по отношению к $G(z_n,\theta)$. Если $G(z_n,\theta)$ — монотонно убывающая по θ функция, то

$$\theta_1(Z_n) = G^{-1}(Z_n,g_2), \qquad \theta_2(Z_n) = G^{-1}(Z_n,g_1).$$

Применим данный метод для построения доверительных интервалов неизвестных параметров нормального распределения $\mathcal{N}(m_X;\sigma_X^2)$. С этой целью сформулируем утверждение, с помощью которого можно определить центральные статистики для неизвестых параметров m_X, σ_X^2.

Теорема 4.1 (теорема Фишера)**.** *Пусть $Z_n = (X_1,\ldots,X_n)$ — выборка, порожденная случайной величиной $X \sim \mathcal{N}(m_X;\sigma_X^2)$, а $\overline{X}$ и s_X^2 — выборочные среднее и дисперсия.*

Тогда

1) случайная величина $\frac{(\overline{X}-m_X)\sqrt{n}}{\sigma_X}$ имеет распределение $\mathcal{N}(0;1)$;

2) случайная величина ns_X^2/σ_X^2 имеет распределение $\chi^2(n-1)$;

3) случайные величины $\overline{X}$ и s_X^2 независимы;

4) $\frac{(\overline{X}-m_X)\sqrt{n-1}}{s_X}$ имеет распределение Стьюдента $\mathbf{t}(n-1)$.

Пример 4.1. По выборке Z_n из нормального распределения требуется построить доверительный интервал для неизвестного математического ожидания m_X при известной дисперсии σ_X^2. Из приведенного выше утверждения следует, что случайная величина $\frac{(\overline{X}-m_X)\sqrt{n}}{\sigma_X}$ имеет нормальное распределение $\mathcal{N}(0;1)$, которое не зависит от m_X, и, кроме того, функция $G(Z_n,m_X) = (\overline{X}-m_X)\sqrt{n}/\sigma_X$ является непрерывной и убывающей по m_X. Это значит, что указанная случайная величина является центральной статистикой. Поэтому доверительный интервал для неизвестного m_X можно построить, если найти такие величины g_1 и g_2, что

$$\mathsf{P}\left(g_1 \leqslant \frac{(\overline{X}-m_X)\sqrt{n}}{\sigma_X} \leqslant g_2\right) = 1-\alpha.$$

Заметим, что данное условие неоднозначно определяет g_1, g_2. Выберем доверительный интервал минимальной длины. Учитывая симметрию относительно оси OY плотности стандартного нормального распределения, можно показать, что такой интервал будет иметь минимальную длину, если положить $g_1 = -g_2$, и при этом он оказывается центральным. Таким образом, получаем следующий доверительный интервал:

$$\left(\overline{X} - \frac{\sigma_X}{\sqrt{n}}u_\gamma; \overline{X} + \frac{\sigma_X}{\sqrt{n}}u_\gamma\right),$$

где u_γ — квантиль уровня $\gamma = 1-\alpha/2$ стандартного нормального распределения $\mathcal{N}(0;1)$. В данном случае длина доверительного интервала равна $\Delta = 2u_\gamma\sigma_X/\sqrt{n}$ и не случайна. Поэтому, задавшись значениями любых двух из трех величин Δ, α, n, можно определить значение третьей величины.

Пример 4.2. Теперь по выборке Z_n из нормального распределения построим доверительный интервал для неизвестного математического ожидания m_X при неизвестной дисперсии σ_X^2.Используя утверждение 4) теоремы 4.1, получаем, что распределение случайной величины $\frac{(\overline{X}-m_X)\sqrt{n-1}}{s_X}$ не зависит от параметра m_X, а функция $G(Z_n,m_X) = (\overline{X}-m_X)\sqrt{n-1}\big/ s_X$ является непрерывной и убывающей по m_X. Следовательно, функция $G(Z_n,m_X) = (\overline{X}-m_X)\sqrt{n-1}\big/ s_X$ является центральной статистикой. По аналогии с предыдущим примером получим центральный доверительный интервал для неизвестного m_X и в случае, когда величина σ_X неизвестна:

$$\left(\overline{X} - \frac{s_X}{\sqrt{n-1}}\, t_{\gamma,n-1}; \overline{X} + \frac{s_X}{\sqrt{n-1}}\, t_{\gamma,n-1}\right),$$

где $t_{\gamma,n-1}$ — квантиль уровня $\gamma = 1-\alpha/2$ распределения Стьюдента $\mathbf{t}(n-1)$.

В отличие от предыдущего примера, длина доверительного интервала случайна и зависит от случайной величины s_X.

Заметим также, что при $n \geqslant 30$ распределение Стьюдента близко к стандартному нормальному распределению, и поэтому при больших объёмах выборки ($n \geqslant 30$) в последней формуле можно использовать квантили стандартного нормального распределения.

Пример 4.3. Центральный доверительный интервал для неизвестного параметра σ_X^2 случайной величины $X \sim \mathcal{N}(m_X;\sigma_X^2)$ при известном m_X имеет следующий вид

$$\left(\frac{\sum\limits_{i=1}^{n}(x_i-m_X)^2}{\chi^2_{1-\alpha/2;n}}; \frac{\sum\limits_{i=1}^{n}(x_i-m_X)^2}{\chi^2_{\alpha/2;n}}\right),$$

где $\chi^2_{1-\alpha/2;,n}$ и $\chi^2_{\alpha/2;,n}$ — квантили уровней $1-\alpha/2$ и $\alpha/2$ распределения $\chi^2(n)$. Покажите это самостоятельно.

Пример 4.4. Центральный доверительный интервал для неизвестного параметра σ_X^2 случайной величины $X \sim \mathcal{N}(m_X;\sigma_X^2)$ при неизвестном m_X можно получить, используя утверждение 2) теоремы 4.1,

$$\left(\frac{n}{\chi^2_{1-\alpha/2;n-1}}\, s_X^2; \frac{n}{\chi^2_{\alpha/2;,n-1}}\, s_X^2\right),$$

где $\chi^2_{1-\alpha/2;,n-1}$ и $\chi^2_{\alpha/2;,n-1}$ — квантили уровней $1-\alpha/2$ и $\alpha/2$ распределения $\chi^2(n-1)$.

Пример 4.5. Пусть $Z_m = (X_1,\ldots,X_m)$ и $V_n = (Y_1,\ldots,Y_n)$ — две независимые выборки, порождённые случайными величинами с распределениями $\mathcal{N}(\mu_1,\sigma_1^2)$ и $\mathcal{N}(\mu_2,\sigma_2^2)$ соответственно. Дисперсии σ_1^2 и σ_1^2 известны, а математические ожидания μ_1 и μ_2 — неизвестны.Требуется построить доверительный интервал для параметра $\theta = \mu_1 - \mu_2$. Покажите, что случайная функция

$$G(Z_m,V_n,\theta) = \frac{(\overline{X}-\overline{Y}) - (\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}}$$

является центральной статистикой для параметра θ. Найдите распределение этой статистики и постройте центральный доверительный интервал уровня надёжности $1-\alpha$ для параметра θ.

Теорема 4.2. *Пусть $Z_m = (X_1,\ldots,X_m)$ и $V_n = (Y_1,\ldots,Y_n)$ — две независимые выборки, порождённые случайными величинами с распределениями $\mathcal{N}(\mu_1,\sigma^2)$ и $\mathcal{N}(\mu_2,\sigma^2)$ соответственно. Параметры μ_1, μ_2 и σ^2 неизвестны. Обозначим*

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \qquad S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i-\overline{X})^2,$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2,$$

$$S_{X,Y} = \sqrt{\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}}.$$

Тогда:

1) случайная величина $\dfrac{S_X^2}{S_Y^2}$ *имеет распределение Фишера с числом степеней свободы* $m-1$ *и* $n-1$*;*

2) случайная величина

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_{X,Y}\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

имеет распределение Стьюдента с $m+n-2$ *степенями свободы.*

Пример 4.6. Пользуясь утверждением 2 теоремы4.2, постройте центральный доверительный интервал для разности средних значений $\theta = \mu_1 - \mu_2$ двух независимых гауссовских величин с одинаковыми, но неизвестными дисперсиями.

Пример 4.7. Построим доверительный интервал для неизвестного параметра b равномерного распределения $\mathscr{R}(0;b)$. Можно показать, что случайная величина $G(Z_n,b) = \left(\frac{X^{(n)}}{b}\right)^n$ имеет распределение $\mathscr{R}(0;1)$ для любого $b>0$. Кроме того, функция $G(z_n,b)$ — убывающая по b. Следовательно, $G(Z_n,b)$ является центральной статистикой. Тогда получаем условие

$$\mathsf{P}\left\{g_1 \leqslant \left(\frac{X^{(n)}}{b}\right)^n \leqslant g_2\right\} = 1-\alpha$$

для некоторых чисел g_1, g_2. Разрешая двойное неравенство в этом вероятностном условии относительно b, получаем следующий доверительный интервал:

$$\left(\frac{X^{(n)}}{g_2^{1/n}}; \frac{X^{(n)}}{g_1^{1/n}}\right).$$

Отметим, что этот интервал будет иметь наименьшую длину, если $g_2 = 1$, а g_1 является квантилью уровня α распределения $\mathscr{R}(0;1)$, т. е. $g_1 = \alpha$.

Асимптотический доверительный интервал.

Пусть $\hat{\theta}_n$ состоятельная оценка параметра θ, построенная по выборке $X_1,...,X_n$ и при $n \to \infty$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} U, U \sim N(0, \sigma^2(\theta)),$$

где асимптотическая дисперсия $\sigma^2(\theta)$ непрерывна по θ.

Тогда асимптотический доверительный интервал для параметра θ уровня надёжности $1-\alpha$ имеет вид

$$\left(\hat{\theta}_n - z_{1-\alpha/2}\sigma(\hat{\theta}_n)/\sqrt{n}; \hat{\theta}_n + z_{1-\alpha/2}\sigma(\hat{\theta}_n))/\sqrt{n}\right),$$

где $z_{1-\alpha/2}$ —квантиль уровня $(1-\alpha/2)$ стандартного нормального распределения.

Пример 4.8. Выборка $X_1,...,X_n$ порождена биномиальной случайной величиной $Bi(1,p)$ с неизвестным параметром p. Построить асимптотический доверительный интервал для параметра p уровня надёжности $1-\alpha$.

Р е ш е н и е. Состоятельной, несмещённой и эффективной оценкой для вероятности «успеха» p в схеме испытаний Бернулли является частота $\hat{p}$. Нетрудно показать, что $E\hat{p} = p, D\hat{p} = \frac{p(1-p)}{n}$. Согласно ЦПТ при большом объёме выборки n имеем, что

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} U, U \sim N\big(0, p(1-p)\big).$$

Тогда асимптотический доверительный интервал для параметра p уровня надёжности $1-\alpha$ имеет следующий вид:

$$\mathsf{P}(\hat{p}-z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}<p<\hat{p}+z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}})=1-\alpha,$$

где $z_{1-\alpha/2}$ — квантиль уровня $(1-\alpha/2)$ стандартного нормального распределения.

Задача 4.1. В прошлом учебном году на втором курсе ОПИ ФКН обучалось 286 студентов. На экзамене по ТВиМС 42 человека получили неудовлетворительные оценки. Постройте асимптотический доверительный интервал уровня надёжности 0.95 для вероятности «успеха». («Успехом» назовём событие, состоящее в том, что студент не получит неудовлетворительной оценки.)

На семинарских занятиях решим следующие задачи.

Задача 4.2. Проводится две серии независимых испытаний по схеме Бернулли. В первой серии проведено n_1 испытаний, в второй n_2 испытаний. Предполагается, что n_1 и n_2 достаточно велики. Построить асимптотический доверительный интервал для разности вероятностей «успеха» в двух сериях.

Задача 4.3. Два года назад на втором курсе ОПИ ФКН обучалось 252 студента. На экзамене по ТВиМС 29 человек получили неудовлетворительные оценки. В прошлом учебном году на втором курсе ОПИ ФКН обучалось 286 студентов. На экзамене по ТВиМС 42 человека получили неудовлетворительные оценки. Постройте асимптотический доверительный интервал уровня надёжности 0.95 для разности вероятностей «успеха« по выборкам прошлого и позапрошлого года.

Типовые задачи

Пример 4.9. Используя данные примера 2.1, построить центральный доверительный интервал уровня доверия 0,95 для неизвестного математического ожидания средней температуры января в Саратове (случайная величина X). Будем предполагать, что случайная величина X имеет гауссовское распределение.

Р е ш е н и е. Согласно примеру 4.2 центральный доверительный интервал для неизвестного математического ожидания m_X уровня доверия 0,95 определяется из следующего соотношения

$$\mathsf{P}(\overline{X}-\sqrt{\frac{s_X^2}{n-1}}\,t_{0,975,n-1}\leqslant m_X\leqslant\overline{X}+\sqrt{\frac{s_X^2}{n-1}}\,t_{0,975,n-1})=0{,}95,$$

где объем выборки $n=13$, $\overline{X}=-11.87$, $s_X^2=22.14$ (результат решения примера 2.1), а квантиль распределения Стьюдента $t_{0.975,12}=2.18$ находится по таблице.

Сделав необходимые вычисления, получаем доверительный интервал для m_X : $[-14.43,-8.9]$.

О т в е т. $[-14.43,-8.9]$.

Пример 4.10. Используя данные типовой примера 2.1, построить центральный доверительный интервал уровня доверия 0.95 для неизвестной дисперсии средней температуры января в Саратове (случайная величина X). Будем предполагать, что случайная величина X имеет гауссовское распределение.

Р е ш е н и е. Согласно примеру 4.4 центральный доверительный интервал для неизвестной дисперсии σ_X^2 уровня доверия 0.95 найдём, используя следующее соотношение

$$\mathsf{P}(\frac{n}{\chi^2_{0.975,n-1}}s_X^2\leqslant\sigma_X^2\leqslant\frac{n}{\chi^2_{0.025,n-1}}s_X^2)=0.95,$$

где объем выборки $n=13$, $s_X^2=22.14$ (результат решения примера 2.1), а $\chi^2_{0.975,12}=23.3$, $\chi^2_{0.025,12}=4.4$ — квантили распределения $\chi^2(12)$, которые находятся по таблице.

Таким образом, получаем доверительный интервал для σ_X^2 : $[11.4,60.38]$.

О т в е т. $[11.4,60.38]$.

Лекция 5

Проверка статистических гипотез

Основные понятия

Определение 5.1. ***Статистической гипотезой*** H (или просто ***гипотезой***) называется любое предположение относительно вида распределения,параметров распределения или свойств закона распределения наблюдаемой в эксперименте случайной величины X.

Определение 5.2. Проверяемая гипотеза называется ***основной*** (или ***нулевой***) и обозначается H_0. Гипотеза, конкурирующая с H_0, называется ***альтернативной*** и обозначается H_1.

Определение 5.3. Статистическая гипотеза H_0 называется ***простой***, если она однозначно определяет параметр или распределение случайной величины X. В противном случае гипотеза H_0 называется ***сложной***.

Пример 5.1. По выборке Z_n требуется проверить гипотезу H_0 о том, что $m_X = m_0$, где m_0 — некоторое фиксированное число, против гипотезы H_1 о том, что $m_X \neq m_0$. Или проверить гипотезу H_0 против гипотезы H_2 о том, что $m_X > m_0$.

Одна из основных задач математической статистики состоит в проверке соответствия результатов эксперимента предполагаемой гипотезе H_0.

Определение 5.4. ***Статистическим критерием*** (***критерием согласия, критерием значимости*** или ***решающим правилом***) проверки гипотезы H_0 называется правило, в соответствии с которым по реализации z_n выборки Z_n гипотеза H_0 принимается или отвергается.

Для этого выборочное пространство S должно быть разбито на две непересекающиеся области - доверительную область S_0 и критическую область S_1, так что $S_0 + S_1 = S$. При попадании реализации выборки z_n в область S_0 принимается основная гипотеза H_0, если же z_n попадает в область S_1, то принимается альтернативная гипотеза H_1. На практике построение n- мерных областей S_0 и S_1, как правило, является слишком трудоёмкой задачей. Поэтому доверительную и критические области задают с помощью статистики критерия.

Определение 5.5. ***Статистикой критерия*** называют некоторую числовую функцию $T = \varphi(Z_n) = \varphi(X_1,...,X_n)$, выборки Z_n, обладающую тем свойством, что её закон распределения $F_T(z)$ полностью известен в том случае, когда проверяемая гипотеза H_0 верна.

Обозначим распределение статистики $T = \varphi(Z_n)$ в случае справедливости основной гипотезы через $F_T(z|H_0)$. Поскольку распределение статистики критерия при справедливости основной гипотезы H_0 известно, то можно явно указать одномерную область в $\mathbb{R}^1$, в которую данная статистика будет попадать с указанной вероятностью. С помощью этой статистики строится процедура (правило) проверки гипотезы.

Определение 5.6. ***Критической областью*** $\overline{G}$ статистического критерия называется область реализаций t статистики $T = \varphi(Z_n)$, при которых гипотеза H_0 отвергается.

Определение 5.7. ***Доверительной областью*** G статистического критерия называется область значений t статистики T, при которых гипотеза H_0 принимается.

Например, в качестве статистического критерия можно использовать правило:

1) если значение $t = \varphi(z_n)$ статистики $T = \varphi(Z_n)$ лежит в критической области $\overline{G}$, то гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1;

2) если реализация $t = \varphi(z_n)$ статистики $T = \varphi(Z_n)$ лежит в доверительной области G, то гипотеза H_0 принимается.

При реализации этого правила возникают ошибки двух видов.

Определение 5.8. ***Ошибкой 1-го рода*** называется событие, состоящее в том, что гипотеза H_0 отвергается, когда она верна.

Определение 5.9. ***Ошибкой 2-го рода*** называется событие, состоящее в том, что принимается гипотеза H_0, когда верна гипотеза H_1.

Определение 5.10. ***Уровнем значимости*** статистического критерия называется вероятность ошибки 1-го рода $\alpha = \mathsf{P}\{T \in \overline{G} | H_0 \text{ верна}\}$. Вероятность ошибки 1-го рода α может быть вычислена, если известно распределение $F_T(z|H_0)$ статистики T.

Вероятность ошибки 2-го рода равна $1-\beta = \mathsf{P}\{Z \in G | H_1 \text{ верна}\}$ и может быть вычислена, если известно распределение $F(z|H_1)$ статистики T при справедливости гипотезы H_1.

Ясно, что с уменьшением вероятности α ошибки 1-го рода возрастает вероятность β ошибки 2-го рода, и наоборот, т. е. при выборе критической и доверительной областей должен достигаться определенный компромисс. Поэтому часто при фиксированной вероятности ошибки 1-го рода критическая область выбирается таким образом, чтобы вероятность ошибки 2-го рода была минимальна.

Определение 5.11. Пусть критерий предназначен для проверки параметрической гипотезы $H_0 : \theta = \theta_0$ против альтернативной гипотезы $H_1 : \theta \neq \theta_0$. Тогда функция

$$\beta(\theta, \overline{G}) = \mathsf{P}\{T \in \overline{G}\}$$

называется ***функцией мощности критерия***.

Определение 5.12. Критерий, предназначенный для проверки гипотезы $H_0 : \theta = \theta_0$ против альтернативной гипотезы $H_1 : \theta \in \Theta_1, \theta_0 \cap \Theta_1 = \varnothing$, называется ***несмещённым***, если

$$\beta(\theta_0) \leqslant \alpha,$$

$$\beta(\theta) > \alpha \quad \text{для любого } \theta \in \Theta_1.$$

Определение 5.13. Критерий, предназначенный для проверки гипотезы $H_0 : \theta = \theta_0$, называется ***состоятельным*** для альтернативной гипотезы $H_1 : \theta \in \Theta_1, \theta_0 \cap \Theta_1 = \varnothing$, если $\beta(\theta, \overline{G}) \longrightarrow 1$ при $n \longrightarrow \infty$ для любого $\theta \in \Theta_1$ и $\alpha > 0$.

Определение 5.14. Пусть имеются два критерия уровня значимости α с критическими областями $\overline{G}_\alpha$ и $\overline{G}^*_\alpha$, предназначенные для проверки гипотезы $H_0 : \theta = \theta_0$, против альтернативной гипотезы $H_1 : \theta \in \Theta_1, \theta_0 \cap \Theta_1 = \varnothing$. Если

$$\beta(\theta_0, \overline{G}_\alpha) \geqslant \beta(\theta_0, \overline{G}^*_\alpha)$$

и

$$\beta(\theta, \overline{G}_\alpha) \leqslant \beta(\theta, \overline{G}^*_\alpha)$$

для $\theta \in \Theta_1$, причём последнее неравенство является строгим хотя бы для одного значения $\theta \in \Theta_1$. Тогда критерий с критической областью $\overline{G}^*_\alpha$ равномерно мощнее критерия с критической областью $\overline{G}_\alpha$.

Лемма Неймана — Пирсона. Пусть выборка $X_1,...,X_n$ соответствует распределению $f(x,\theta)$ с неизвестным параметром θ. Наиболее мощный критерий уровня значимости α, $\alpha > 0$, для проверки простой гипотезы $H_0 : \theta = \theta_0$ против простой альтернативной гипотезы $H_1 : \theta = \theta_1 > \theta_0$ задаётся критической областью вида:

$$S^*_{1\alpha} = \{(x_1,...,x_n) : T(x_1,...,x_n) \geqslant c_\alpha\},$$

где статистика $T(X) = \frac{L(X_1,...,X_n,\theta_1)}{L(X_1,...,X_n,\theta_0)}$, а константа c_α находится из условия $P_{\theta_0}(T(X) \geqslant c_\alpha) = \alpha$.

Замечание 5.1. Критическая область $\overline{G}^*_\alpha$ реализаций статистики $T(X)$ будет иметь вид $\overline{G}^*_\alpha = [c_\alpha; +\infty)$.

Задача 5.1. Выборка $X_1,...,X_n$ соответствует нормальному распределению $N(m,\sigma^2)$ с известной дисперсией σ^2 и неизвестным средним m. Постройте наиболее мощный критерий для проверки гипотезы $H_0 : m = m_0$ против простой альтернативной гипотезы $H_1 : m = m_1 > m_0$.

Ответ:

$$S^*_{1\alpha} = \left\{ (x_1,...,x_n) : T(x_1,...,x_n) \geqslant m_0 + \frac{\sigma z_{1-\alpha}}{\sqrt{n}} \right\},$$

$T(X) = \overline{X}$.

Проверка статистической гипотезы может быть подразделена на следующие этапы:

1) сформулировать проверяемую гипотезу H_0 и альтернативную к ней гипотезу H_1;

2) выбрать уровень значимости α;

3) выбрать статистику T для проверки гипотезы H_0;

4) найти распределение $F(z|H_0)$ статистики T, при условии что гипотеза H_0 верна;

5) построить, в зависимости от формулировки гипотезы H_1 и уровня значимости α, критическую область $\overline{G}$;

6) получить реализацию выборки наблюдений $x_1,\ldots,x_n$ и вычислить реализацию $t = \varphi(x_1,\ldots,x_n)$ статистики T критерия;

7) принять статистическое решение на уровне доверия $1-\alpha$: если $t \in \overline{G}$, то отклонить гипотезу H_0 как не согласующуюся с результатами наблюдений, а если $t \in G$, то принять гипотезу H_0 как не противоречащую результатам наблюдений.

Проверка гипотезы о значении параметра

Пусть имеется параметрическая статистическая модель $(S_\theta, F_{Z_n}(z_n,\theta))$, $\theta \in \Theta \subset \mathbb{R}^1$, т. е. выборка Z_n соответствует распределению $F(x,\theta)$ с неизвестным параметром θ. Проверим простую гипотезу H_0, состоящую в том, что $\theta = \theta_0$, где θ_0 — некоторое фиксированное число из множества Θ.

Формулировка альтернативной гипотезы H_1 и уровень значимости α определяют размер и положение критической области $\overline{G}$ на множестве значений статистики T. Например, если альтернативная гипотеза H_1 формулируется как $\theta > \theta_0$ (или $\theta < \theta_0$), то критическая область размещается на правом (или левом) «хвосте» распределения статистики T, т. е.

$$\overline{G} = \{t > z_{1-\alpha}\} \quad (\text{или } \overline{G} = \{t < z_\alpha\}),$$

где $z_{1-\alpha}$ и z_α — квантили уровней $1-\alpha$ и α соответственно распределения $F_T(x|H_0)$ статистики T, при условии что верна гипотеза H_0. В этом случае статистический критерий называется ***односторонним***. Если альтернативная гипотеза H_1 формулируется как $\theta \neq \theta_0$, то критическая область $\overline{G}$ размещается на обоих «хвостах» распределения статистики T, т. е. определяется совокупностью неравенств

$$\overline{G} = \left\{t < z_{\alpha/2}\right\} \cup \left\{t > z_{1-\alpha/2}\right\},$$

где $z_{\alpha/2}$ и $z_{1-\alpha/2}$ — квантили уровней $\alpha/2$ и $1-\alpha/2$ соответственно распределения $F(x|H_0)$. В этом случае критерий называется ***двусторонним***.

Лекция 6

Проверка гипотез о параметрах в одновыборочной гауссовской модели и биномиальных моделях

Рассмотрим процедуру проверки параметрической гипотезы на примере одной из "старинных"статистических задач - задаче с дихотомическими (двоичными) данными.

Пример 6.1. ***Биномиальный критерий.***

Рассмотрим следующую статистическую модель. Проводится серия из n испытаний, удовлетворяющая условиям:

1. всякое испытание имеет два исхода – либо «успех», либо «неудача»;
2. все испытания независимы;
3. вероятность p «успеха» в одном испытании не изменяется от опыта к опыту.

Пусть случайная величина X - число «успехов» в n испытаниях, тогда X имеет биномиальное распределение $Bi(n,p)$. Как мы уже знаем, неизвестную вероятность p можно оценить частотой «успехов» $\hat{p} = \frac{X}{n}$, и оценка $\hat{p}$ обладает следующими свойствами:

- несмещенная, т.е. $\mathsf{E}[\hat{p}] = p$;
- сильно состоятельная;
- асимптотически нормальная: $\sqrt{n}(\hat{p} - p) \sim N(0, p(1-p))$ при $n \to \infty$.

В практических задачах бывает важно не только оценить вероятность «успеха» p, но и проверить гипотезу о том, что p равна некоторой заданной величине p_0.

Построим критерий, называемый биномиальным критерием, для проверки гипотезы $H_0 : p = p_0$ против альтернатив следующего вида:

$H_1 : p < p_0$;

$H_2 : p > p_0$;

$H_3 : p \neq p_0$.

Рассмотрим статистику $T(X) = X$, где X – количество «успехов» в n испытаниях. Согласно рассматриваемой модели, при справедливости гипотезы $H_0 : p = p_0$ статистика $T(X)$ будет иметь биномиальное распределение $\mathbf{Bi}(n, p_0)$.

Квантили биномиального распределения с известными параметрами n и p_0 табулированы. Выбрав уровень значимости α, можно построить доверительные и критические области биномиального критерия для альтернатив вида H_1, H_2, H_3. Понятно, что в пользу альтернативы H_1 будут говорить малые значения статистики $T(X)$, в пользу альтернативы H_2 – большие значения статистики $T(X)$, в пользу альтернативы H_3 большие и малые значения статистики $T(X)$. Таким образом, доверительные и критические области определяются следующим образом:

Альтернативная гипотеза	Доверительная область	Критическая область
$H_1 : p < p_0$	$(z_\alpha; n]$	$[0; z_\alpha]$
$H_2 : p > p_0$	$[0; z_\alpha]$	$[z_\alpha; n]$
$H_3 : p \neq p_0$	$(z_{\alpha/2}; z_{1-\alpha/2})$	$[0; z_{\alpha/2}] \cup [z_{1-\alpha/2}; n]$

Таблица 6.1

где z_γ – γ квантиль распределения $Bi(n, p_0)$.

Если наблюдений достаточно много, то согласно теореме Муавра-Лапласа, статистика $T(X)$ будет иметь асимптотически гауссовское распределение. То есть при $n \geqslant 30$ можно в качестве статистики биномиального критерия выбрать статистику

$$\tilde{T}(X) = \frac{X - np_0}{\sqrt{np_0q_0}},$$

распределение которой при справедливости H_0 будет стандартным гауссовским. Доверительные и критические области для критерия со статистикой $\tilde{T}(X)$ будет иметь вид:

Альтернативная гипотеза	Доверительная область	Критическая область
$H_1 : p < p_0$	$(z_\alpha; +\infty)$	$(-\infty; z_\alpha]$
$H_2 : p > p_0$	$(-\infty; z_\alpha)$	$[z_\alpha; +\infty)$
$H_3 : p \neq p_0$	$(z_{\alpha/2}; z_{1-\alpha/2})$	$(-\infty; z_{\alpha/2}] \cup [z_{1-\alpha/2}; +\infty)$

Таблица 6.2

Замечание 6.1. Доказано, что построенные биномиальные критерии является состоятельными для альтернатив вида H_1, H_2, H_3.

Задача 6.1. (о леди, дегустирующей чай) Согласно принятой в Англии традиции чаепития, в чашку следует сначала наливать молоко, а потом – чай. Считается, что настоящая английская леди умеет отличить «правильный» чай от «неправильного». Чтобы выяснить дегустаторские качества дамы, был проведен следующий эксперимент: в течение 30 дней даме каждый день подавали пару чашек чая, в одну из которых сначала был налит чай, а в другую – молоко. Дама 21 раз верно указала «правильный» чай. Можно ли (на уровне доверия 0.95) считать ее хорошим дегустатором?

Р е ш е н и е.

Пусть p – вероятность выбора «правильной» чашки чая.

Тогда утверждение о том, что $p = 0.5$ соответствует ситуации, при которой выбор «правильной» чашки осуществляется случайным образом. Если же $p > 0.5$ то это означает, что выбор чашки основан на каких-то предпочтениях, т.е. неслучаен. Теперь можем провести процедуру проверки гипотезы.

1. В качестве основной гипотезы следует выбрать простую гипотезу $H_0 : p = 0.5$, а в качестве альтернативной – сложную гипотезу $H_A : p > 0.5$.

2. Пусть уровень значимости $\alpha = 0.05$.

3. Выберем статистику $\tilde{T}(X) = \frac{X - np_0}{\sqrt{np_0q_0}}$, где $p_0 = 0.5$

4. Известно, что при $n \geqslant 30$ $\tilde{T}(X)\big|_{H_0} \sim N(0,1)$.

5. Альтернативной гипотезе должны соответствовать большие значения статистики $T(X)$, т.е. критическая область расположена справа. Критическая точка $z_{0.95} = 1.65$.

6. Вычислим реализацию статистики $\tilde{T}(x) = \frac{21 - 30 \cdot 0.5}{\sqrt{30 \cdot 0.5 \cdot 0.5}} = 2.19$.

7. Реализация статистики попала в критическую область, следовательно, гипотеза H_0 отвергается. Следовательно, на уровне доверия 0.95 можно считать, что леди – хороший дегустатор.

Пример 6.2. ***Проверка гипотезы о равенстве параметров двух биномиальных совокупностей.*** Для решения этой задачи необходимо дать определение гипергеометрического распределения.

Пусть имеется N деталей, среди которых N_1 бракованных деталей. Из N деталей случайным образом выбирают n деталей. Пусть случайная величина ξ – количество бракованных деталей среди выбранных. Найдём распределение случайной величины ξ.

$$\begin{array}{ccccc} N & = & N_1 & + & (N-N_1) \\ \downarrow & & \downarrow & & \downarrow \\ n & = & \xi & + & (n-\xi) \end{array}$$

Построим ряд распределения этой случайной величины:

$$P(\xi = k) = \frac{C_{N_1}^k \cdot C_{N-N_1}^{n-k}}{C_N^n}, \quad k = 0, 1, \ldots, \min(n, N_1).$$

Распределение случайной величины ξ называется гипергеометрическим распределением с параметрами (N,n,N_1).

Рассмотрим статистическую модель описывающую двоичные исходы в двух сериях испытаний. Для данной модели справедливы следующие предположения:

1. все испытания независимы;
2. исход каждого испытания двоичен: A («успех») или $\overline{A}$ («неудача»);
3. Вероятность «успеха» p_1 в первой серии и вероятность «успеха» p_2 во второй серии не изменяются от опыта к опыту.

Пусть X_1 – число «успехов» в первой серии испытаний, тогда $X_1 \sim Bi(n_{1\bullet},p_1)$, где $n_{1\bullet}$ – количество испытаний в первой серии. Пусть X_2 – число «успехов» во второй серии испытаний, тогда $X_2 \sim Bi(n_{2\bullet},p_2)$, где $n_{2\bullet}$ – количество испытаний во второй серии. Пусть X – число «успехов» в обеих сериях, а $n = n_{1\bullet} + n_{2\bullet}$ – количество проведённых испытаний в обеих сериях. Проверим гипотезу $H_0 : p_1 = p_2 = p$ против альтернатив $H_1 : p_1 < p_2$, $H_2 : p_1 > p_2$, $H_3 : p_1 \neq p_2$.

Справедливость гипотезы H_0 означает, что случайная величина $X \sim Bi(n,p)$.

Пусть в первой серии произошло n_{11} «успехов» и n_{12} «неудач», во второй серии n_{21} «успехов» и n_{22} «неудач». По результатам наблюдений строим таблицу:

№ серии	$\boldsymbol{A}$	$\bar{\boldsymbol{A}}$	
1	n_{11}	n_{12}	$n_{1\bullet}$
2	n_{21}	n_{22}	$n_{2\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Таблица 6.3

Вычислим условную вероятность того, что в первой серии появилось n_{11} «успехов» при условии, что в обеих сериях было $n_{\bullet 1}$ «успехов».

$$P(X_1 = n_{11} | X = n_{\bullet 1}) = \frac{P(X_1 = n_{11}, X_2 = n_{21})}{P(X = n_{\bullet 1})} = \frac{P(X_1 = n_{11})\,P(X_2 = n_{21})}{P(X = n_{\bullet 1})} =$$

$$= \frac{C_{n_{1\bullet}}^{n_{11}} \cdot p_1^{n_{11}} \cdot q_1^{n_{12}} \cdot C_{n_{2\bullet}}^{n_{21}} \cdot p_2^{n_{21}} \cdot q_2^{n_{22}}}{C_n^{n_{\bullet 1}} \cdot p^{n_{\bullet 1}} \cdot q^{n-n_{\bullet 1}}}.$$

Пусть $p_1 = p_2 = p$. Тогда

$$P(X_1 = n_{11} | X = n_{\bullet 1}) = \frac{C_{n_{1\bullet}}^{n_{11}} \cdot p^{n_{\bullet 1}} \cdot q^{n-n_{\bullet 1}} \cdot C_{n_{2\bullet}}^{n_{21}}}{C_n^{n_{\bullet 1}} \cdot p^{n_{\bullet 1}} \cdot q^{n-n_{\bullet 1}}} = \frac{C_{n_{1\bullet}}^{n_{11}} \cdot C_{n_{2\bullet}}^{n_{21}}}{C_n^{n_{\bullet 1}}}.$$

Итак, при справедливости H_0 условное распределение случайной величины X_1 является гипергеометрическим распределением с параметрами $(n,n_{\bullet 1},n_{1\bullet})$.

Если $\frac{n_{\bullet i}n_{j\bullet}}{n} < 5$, $\forall i,j = 1,2$, то в качестве статистики критерия выбирают $T(X) = X_1$.

Если количество наблюдений велико, то для проверки гипотезы H_0 используется другой критерий.

Оценим вероятности p_1 и p_2 как $\hat{p}_1 = \frac{n_{11}}{n_{1\bullet}}$ и $\hat{p}_2 = \frac{n_{21}}{n_{2\bullet}}$ соответственно. При большом количестве испытаний статистики $\hat{p}_1$, $\hat{p}_2$ имеют гауссовское распределение.

Заметим, что

$$\mathsf{E}[\hat{p}_1] = p_1, \mathsf{D}[\hat{p}_1] = \frac{n_{1\bullet}p_1q_1}{n_{1\bullet}^2} = \frac{p_1q_1}{n_{1\bullet}};$$

$$\mathsf{E}[\hat{p}_2] = p_2, \mathsf{D}[\hat{p}_2] = \frac{n_{2\bullet}p_2q_2}{n_{2\bullet}^2} = \frac{p_2q_2}{n_{2\bullet}}.$$

При справедливости $H_0 : p_1 = p_2$ и $min(n_{1\bullet}, n_{2\bullet}) \to \infty$.

$$T(x) = \left.\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\mathsf{D}(\hat{p}_1 - \hat{p}_2)}}\right|_{H_0: p_1 = p_2 = p} \sim \mathbf{N}(0,1),$$

где

$$\mathsf{D}(\hat{p}_1 - \hat{p}_2) = \mathsf{D}(\hat{p}_1 + (-\hat{p}_2)) = (\mathsf{D}\hat{p}_1 + \mathsf{D}\hat{p}_2) = \frac{pq}{n_{1\bullet}} + \frac{pq}{n_{2\bullet}} = pq\left(\frac{1}{n_{1\bullet}} + \frac{1}{n_{2\bullet}}\right).$$

Вероятности p и $q = 1 - p$ неизвестны. Их следует оценить соответсвующими частотами, используя обе серии наблюдений.

Получим

$$\widehat{\mathsf{D}}(\hat{p}_1 - \hat{p}_2) = \frac{n_{\bullet 1}}{n}\left(1 - \frac{n_{\bullet 1}}{n}\right) \cdot \left(\frac{1}{n_{1\bullet}} + \frac{1}{n_{2\bullet}}\right).$$

Теперь при $n \to \infty$ статистика

$$T(x) = \left.\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\widehat{\mathsf{D}}(\hat{p}_1 - \hat{p}_2)}}\right|_{H_0: p_1 = p_2} \sim \mathbf{N}(0,1).$$

Доверительные и критические области этого критерия при различных альтернативах будут иметь следующий вид:

Альтернативная гипотеза	**Доверительная область**	**Критическая область**
$H_1 : p_1 < p_2$	$(z_\alpha; +\infty)$	$(-\infty; z_\alpha]$
$H_2 : p_1 > p_2$	$(-\infty; z_{1-\alpha})$	$[z_{1-\alpha}; +\infty)$
$H_3 : p_1 \neq p_2$	$(z_{\alpha/2}; z_{1-\alpha/2})$	$(-\infty; z_{\alpha/2}] \cup [z_{1-\alpha/2}; +\infty)$

Таблица 6.4

где z_γ – $\gamma-$квантиль стандартного гауссовского распределения.

Задача 6.2. При изучении эффективности лекарства от аллергии обследовалось две группы людей. Были получены следующие результаты:

Серия	**заболели**	**не заболели**
принимали лекарство	3	172
не принимали лекарство	32	168

Таблица 6.5

Указывают ли эти результаты на эффективность лекарства?

Уровень значимости принять равным 0.05.

Р е ш е н и е. Назовем «успехом» событие, состоящее в том, что пациент не заболел, а «неудачей» – событие, состоящее в том, что пациент заболел.

Серия	«успех»	«неудача»	
принимали лекарство	172	3	175
не принимали лекарство	168	32	200
	340	35	375

Таблица 6.6

Пусть p_1 – вероятность не заболеть для принимавших лекарство, а p_2 – вероятность не заболеть для не принимавших лекарство.

Проверим гипотезу $H_0 : p_1 = p_2$ (вероятность не заболеть одинакова для принимавших и не принимавших лекарство);

$H_1 : p_1 > p_2$ (среди принимавших лекарство вероятность не заболеть больше, чем среди не принимавших лекарство).

Оценим вероятности p_1 и p_2 соответсвующими частотами:

$$\hat{p}_1 = \frac{172}{175} = 0.98,\ \hat{p}_2 = \frac{168}{200} = 0.84.$$

$$\widehat{\mathsf{D}}\,[\hat{p}_1 - \hat{p}_2] = \frac{340}{375}\left(1 - \frac{340}{375}\right)\left(\frac{1}{175} + \frac{1}{200}\right) = 0.03^2.$$

$$T\,(x) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\widehat{\mathsf{D}}\,(\hat{p}_1 - \hat{p}_2)}} = \frac{0.14}{0.03} = 4.6.$$

Построим доверительную и критическую область. Критическая область расположена справа, критическая точка $z_{1-\alpha} = z_{0.95} = 1.65$

Значение статистики попало в критическую область, следовательно, на уровне значимости $\alpha = 0.05$ гипотеза H_0 отвергается в пользу гипотезы H_1.

Пример 6.3. ***Проверка гипотезы о среднем значении гауссовской случайной величины.*** Пусть известно, что случайная величина X имеет нормальное распределение. Требуется, используя реализацию выборки z_n, проверить гипотезу H_0, состоящую в том, что $m_X = m_0$ (m_0 — некоторое фиксированное число), против альтернативной гипотезы H_A о том, что $m_X \neq m_0$. Возможны два случая: дисперсия σ_X^2 известна и неизвестна. Статистики T для обоих случаев можно выбрать, используя утверждение из п. 4.4. Представим эти случаи в виде табл. 6.7.

Предположение	**Статистика T критерия**	**Распределение $F(z\|H_0)$**	**Доверительная область G критерия**
σ_X^2 известна	$\dfrac{(\overline{X} - m_0)\sqrt{n}}{\sigma_X}$	$\mathcal{N}(0;1)$	$[-u_\gamma, u_\gamma]$
σ_X^2 неизвестна	$\dfrac{(\overline{X} - m_0)\sqrt{n-1}}{\sqrt{s^2}}$	$\mathbf{t}(n-1)$	$\left[-t_{\gamma,n-1}, t_{\gamma,n-1}\right]$

Таблица 6.7

Для каждого случая в соответствии с утверждениями 1) и 4) теоремы 4.1 получаем свою доверительную область, где u_γ, $t_{\gamma,n-1}$ — квантили уровня $\gamma = 1 - \alpha/2$ распределений $\mathcal{N}(0;1)$ и $\mathbf{t}(n-1)$ соответственно.

Задача 6.3. Постройте доверительные и критические области для проверки гипотезы $H_0 : m = m_0$ против альтернативных гипотез $H_1 : m < m_0$ и $H_2 : m > m_0$.

Пример 6.4. ***Проверка гипотезы о дисперсии гауссовской случайной величины.*** Пусть случайная величина X нормально распределена, а ее дисперсия неизвестна. Требуется на основе реализации z_n выборки Z_n, порожденной случайной величиной X, проверить гипотезу H_0 о том, что $\sigma_X^2 = \sigma_0^2$ (σ_0 — некоторое фиксированное число), против альтернативной гипотезы H_A, состоящей в том, что $\sigma_X^2 \neq \sigma_0^2$.

Возможны два случая: m_X — известно или m_X — неизвестно. Представим эти случаи в виде следующей таблицы:

Предположение	**Статистика T критерия**	**Распределение $F(z\|H_0)$**	**Доверительная область G критерия**
m_X известно	$\dfrac{\sum\limits_{k=1}^{n}(X_k - m_X)^2}{\sigma_0^2}$	$\chi^2(n)$	$[\chi^2_{1-\gamma,n}, \chi^2_{\gamma,n}]$
m_X неизвестно	$\dfrac{ns^2}{\sigma_0^2}$	$\chi^2(n-1)$	$[\chi^2_{1-\gamma,n-1}, \chi^2_{\gamma,n-1}]$

Таблица 6.8

Здесь $\chi^2_{\gamma,k}$, $\chi^2_{1-\gamma,k}$ — квантили уровня $\gamma = 1-\alpha/2$ и $1-\gamma$ распределения $\chi^2(k)$ с k степенями свободы, где $k = n, n-1$.

Задача 6.4. Постройте доверительные и критические области для проверки гипотезы H_0 : $\sigma_X^2 = \sigma_0^2$ против альтернативных гипотез $H_1 : \sigma_X^2 < \sigma_0^2$ и $H_2 : \sigma_X^2 > \sigma_0^2$.

Пример 6.5. Используя данные примера 2.1, проверить на уровне значимости $\alpha = 0.05$ гипотезу H_0, состоящую в том, что математическое ожидание m_X средней температуры января в Саратове (случайная величина X) равно -13.75, т. е. что $m_X = -13.75$, против альтернативной гипотезы H_1, состоящей в том, что $m_X \neq -13.75$. Будем предполагать, что случайная величина X имеет гауссовское распределение.

Р е ш е н и е: Дисперсия σ_X^2 случайной величины X неизвестна, поэтому выберем для проверки гипотезы H_0 статистику

$$T = \frac{(\widehat{m}_X - m_0)\sqrt{n-1}}{s_X}.$$

В данной задаче $n = 13$, $m_0 = -13.75$ а $\widehat{m}_X = -11.87$ и $s_X^2 = 22.14$ (результат решения примера 2.1). Согласно примеру 6.3 распределением $F(z|H_0)$ статистики T при справедливости H_0 является распределение Стьюдента $\mathbf{S}(n-1)$.

Таким образом, доверительная область G имеет вид

$$G = \left[-t_{0.975,12}, t_{0.975,12}\right] = [-2.18, 2.18].$$

Вычисляя значение t статистики T для данной реализации выборки, имеем

$$t = \frac{(-11.87 + 13.75)\sqrt{12}}{\sqrt{22.14}} \approx 1.38.$$

Таким образом, значение t статистики T попадает в доверительную область G, и, следовательно, на уровне доверия $1-\alpha = 0.95$ можно считать, что результаты наблюдений не противоречат гипотезе H_0, состоящей в том, что $m_X = -13.75$.

О т в е т. Гипотеза H_0 на уровне значимости $\alpha = 0.05$ принимается.

Пример 6.6. Используя данные примера 2.1, проверить на уровне доверия $1-\alpha = 0.95$ гипотезу H_0, состоящую в том, что дисперсия σ_X^2 средней температуры января в г. Саратове (случайная величина X) равна 20, т. е. что $\sigma_X^2 = 20$, против альтернативной гипотезы H_1, состоящей в том, что $\sigma_X^2 \neq 20$. Будем предполагать, что случайная величина X имеет гауссовское распределение.

Р е ш е н и е: Математическое ожидание m_X случайной величины X неизвестно, поэтому выберем для проверки гипотезы H_0 статистику

$$T = \frac{n\widehat{d}_X}{\sigma_0^2}.$$

В данной задаче $n = 13$, $\sigma_0^2 = 20$, а $s_X^2 = 22.14$ (результат решения примера 2.1). Согласно примеру 6.4 распределением $F(z|H_0)$ статистики Z при справедливости H_0 является распределение $\chi^2(12)$.

Таким образом, доверительная область G имеет вид

$$G = \left[\chi^2_{0.025,12}, \chi^2_{0.975,12}\right] = [4.4, 23.3].$$

Вычисляя значение t статистики T для данной реализации выборки, имеем

$$t = \frac{13 \cdot 22.14}{20} = 14.39.$$

Таким образом, значение t статистики T попадает в доверительную область G, и, следовательно, на уровне доверия $1 - \alpha = 0.95$ можно считать, что результаты наблюдений не противоречат гипотезе H_0, состоящей в том, что $\sigma_X^2 = 20$.

О т в е т. Гипотеза H_0 на уровне значимости $\alpha = 0.05$ принимается.

Пример 6.7. Пусть $Z_m = (X_1, \ldots, X_m)$ и $V_n = (Y_1, \ldots, Y_n)$ — две независимые выборки, порождённые случайными величинами с распределениями $\mathcal{N}(\mu_1, \sigma^2)$ и $\mathcal{N}(\mu_2, \sigma^2)$ соответственно. Параметры μ_1, μ_2 неизвестны, а параметр σ^2 известен.Опишите процедуру проверки гипотезы H_0 о равенстве средних значений $\mu_1 = \mu_2$ против альтернативной гипотезы $H_1 : \mu_1 \neq \mu_2$.

Пример 6.8. Пусть $Z_m = (X_1, \ldots, X_m)$ и $V_n = (Y_1, \ldots, Y_n)$ — две независимые выборки, порождённые случайными величинами с распределениями $\mathcal{N}(\mu_1, \sigma^2)$ и $\mathcal{N}(\mu_2, \sigma^2)$ соответственно. Параметры μ_1, μ_2 и σ^2 неизвестны.Используя утверждение теоремы4.2, опишите процедуру проверки гипотезы H_0 о равенстве средних значений $\mu_1 = \mu_2$ против альтернативной гипотезы $H_1 : \mu_1 \neq \mu_2$.

Лекция 7

Проверка гипотезы о виде закона распределения.

Пусть имеется реализация z_n выборки Z_n, порожденной случайной величиной X с неизвестной функцией распределения $F(x)$. Требуется проверить гипотезу H_0, состоящую в том, что случайная величина X имеет определенный закон распределения $\overline{F}(x,\theta)$ (например, нормальный, равномерный и т. д.). Истинный закон распределения $F(x)$ неизвестен. Закон распределения $\overline{F}(x,\theta)$ часто называют гипотетическим (т.е. соответствующим основной гипотезе H_0). Для проверки такой гипотезы можно использовать различные статистические критерии.

7.1 Критерий Колмогорова

Данный критерий применяется в тех ситуациях, когда гипотетическая функция распределения является непрерывной, и параметры распределения этой функции известны. Т.е. основная гипотеза является простой $H_0 : X \sim \overline{F}(x,\theta_0) = F_0(x)$. Статистика этого критерия имеет вид:

$$D_n = \sup_{x\in\mathbb{R}^1} \left|\widehat{F}_n(x) - F_0(x)\right|,$$

где $\widehat{F}_n(x)$ -эмпирическая функция распределения, построенная по выборке Z_n.

Как было показано выше, эмпирическая функция распределения $\widehat{F}_n(x)$ является несмещённой и сильно состоятельной оценкой для функции распределения наблюдаемой случайной величины. Поэтому при большом объёме выборки максимальное расхождение между эмпирической и гипотетической функциями распределения (в случае правильно выбранной гипотетической функции) не должно существенно отличаться от нуля.

А именно, при $n \to \infty$ и справедливости гипотезы $H_0 : X \sim F_0(x)$ статистика $\sqrt{n}D_n$ имеет распределение Колмогорова с функцией распределения следующего вида

$$K(t) = \sum_{j=-\infty}^{\infty} (-1)^j \exp(-2j^2t^2), \qquad t \geqslant 0.$$

В соответствии с критерием Колмогорова гипотеза H_0 принимается (т. е. реализация выборки z_n согласуется с гипотезой H_0) на уровне надежности $1-\alpha$, если $\sqrt{n}D_n \in G = [0,k_{1-\alpha}]$, где $k_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения Колмогорова. Если же $\sqrt{n}D_n \in \overline{G}$, то гипотеза H_0 отвергается.

Замечание 7.1. Распределение статистики $\sqrt{n}D_n$ достаточно хорошо аппроксимируется распределением Колмогорова уже при $n > 20$.

Замечание 7.2. При объёмах выборки $n \leqslant 20$ есть таблицы квантилей точного распределения статистики D_n в случае справедливости гипотезы H_0.

Замечание 7.3. Распределение статистики D_n (в случае справедливости гипотезы H_0) не зависит от вида функции $F_0(x)$.

Замечание 7.4. Квантили распределения Колмогорова: $k_{0,95} = 1{,}358; k_{0,99} = 1{,}628$.

7.2 Критерий хи-квадрат К. Пирсона для проверки простой гипотезы о виде распределения случайной величины

Вычисление статистики D_n может оказаться довольно трудоёмкой задачей. Поэтому для проверки простой гипотезы $H_0 : X \sim \overline{F}(x,\theta_0) = F_0(x)$ о том, что наблюдаемая случайная величина имеет распределение $F_0(x)$ с известными параметрами, можно применять и другой критерий, предложенный Карлом Пирсоном.

Для того, чтобы воспользоваться этим критерием, проводят группировку выборки по $l+1$ интервалу Δ_k, $k = 1,...,l+1$, (например, так, как это было сделано при построении гистограммы) и находят частоты $\frac{n_k}{n}$ попадания элементов выборки в каждый из интервалов Δ_k, $k = 1,...,l+1$. Теперь проверка гипотезы H_0 сводится к проверке гипотезы о том, что наблюдаемый вектор частот имеет полиномиальное распределение с известным вектором вероятностей $p_1^{(0)},...,p_{l+1}^{(0)}$, где $p_k^{(0)}$ - гипотетическая вероятность попадания в интервал Δ_k, $k = 1,...,l+1$. В качестве статистики, характеризующей отклонение частот от соответствующих гипотетических вероятностей, выбирают

$$X^2 = \sum_{k=1}^{l+1} \frac{n}{p_k^{(0)}} \left(\frac{n_k}{n} - p_k^{(0)} \right)^2 = \sum_{k=1}^{l+1} \frac{n_k^2}{np_k^{(0)}} - n.$$

Распределение этой статистики при справедливости гипотезы H_0 даётся следующей теоремой.

Теорема 7.1. *Если* $0 < p_k^{(0)} < 1$ *для* $k = 1,...,l+1$, *то при* $n \longrightarrow \infty$ *и справедливости гипотезы* H_0 *статистика* X^2 *имеет распределение хи-квадрат с* l *степенями свободы* $\chi^2(l)$.

Доказательство этой теоремы можно найти, например, в учебниках Ивченко Г.И., Медведева Ю.И. "Математическая статистика "Введение в математическую статистику".

Понятно, что в пользу альтернативной гипотезы будут свидетельствовать "большие"значения статистики X^2 (т.е., "большие"отклонения наблюдаемых наблюдаемых частот от соответствующих гипотетических вероятностей). Поэтому критическая область $\overline{G}$ критерия хи-квадрат принимает вид $\overline{G} = (x_{1-\alpha}(l), +\infty)$, где $x_{1-\alpha}(l)$ — квантиль уровня $1-\alpha$ распределения $\chi^2(l)$, α — заданный уровень значимости (обычно $\alpha = 0.05$).

Пример 7.1. Г. Мендель проводил опыты, в которых изучал наследование признаков у семян гороха. Одно из скрещиваемых растений имело гладкие (признак A) и жёлтые (признак B) семена, другое - морщинистые (признак a) и зелёные (признак b). В первом поколении все растения имели гладкие жёлтые семена (фенотип AB). Во втором поколении произошло расщепление: кроме фенотипа AB появились также гладкие зелёные (фенотип Ab), морщинистые жёлтые (фенотип aB) и морщинистые зеленые (фенотип ab). Согласно теории наследственности Менделя вероятности появления фенотипов AB, Ab,aB,ab равны соответственно 9/16, 3/16,3/16 и 1/16. Проведя скрещивание, Мендель получил во втором поколении: 315 гладких жёлтых семян, 108 гладких зелёных, 101 морщинистое жёлтое и 32 морщинистых зелёных. Требуется проверить гипотезу о соответствии опытных данных указанному распределению на уровне значимости 0.05.

Р е ш е н и е: Применим критерий хи-квадрат для проверки простой гипотезы H_0 о том, что наблюдаемые данные имеют полиномиальное распределение с 4 возможными исходами, вероятности которых равны 9/16, 3/16,3/16 и 1/16. Соответствующие частоты равны 0,556; 0,194; 0,182 и 0,058.

Вычислим реализацию статистики хи-квадрат

$$X^2 = \sum_{k=1}^{4} \frac{n}{p_k^{(0)}} \left(\frac{n_k}{n} - p_k^{(0)} \right)^2 = 0{,}042 + 0{,}145 + 0{,}083 + 0{,}221 = 0{,}49.$$

При справедливости гипотезы H_0 указанная статистика должна иметь распределение хи-квадрат с 3 степенями свободы. Найдём $x_{0.95}(3)$ — квантиль уровня $1-\alpha$ распределения

$\chi^2(3)$. По таблице $x_{0.95}(3) = 7{,}81$. Следовательно, критическая область имеет вид: $(7{,}81;\infty)$. Вычисленное значение статистики попадает в доверительную область,т.е. критерий хи-квадрат не отвергает гипотезу H_0. Это означает, что имеется достаточно хорошая согласованность между гипотезой и наблюдаемыми данными.

О т в е т. Гипотеза H_0 принимается на уровне значимости $\alpha = 0.05$.

7.3 Критерий хи-квадрат К. Пирсона для проверки сложной гипотезы о виде распределения случайной величины

Правило проверки заключается в следующем.

1) Формулируется гипотеза H_0, состоящая в том, что случайная величина X имеет распределение определенного вида $\overline{F}(x,\theta_1,\ldots,\theta_s)$ с s неизвестными параметрами $\theta_1,\ldots,\theta_s$ (например, m и σ^2 для нормального распределения, a и b — для равномерного и т. д.).

2) По реализации z_n выборки Z_n методом максимального правдоподобия находятся оценки $\widehat{\theta}_1,\ldots,\widehat{\theta}_s$ неизвестных параметров $\theta_1,\ldots,\theta_s$.

3) Действительная ось $\mathbb{R}^1$ разбивается на $l+1$ непересекающихся полуинтервалов (разрядов) $\Delta_0,\ldots,\Delta_l$ следующим образом. Действительная ось $\mathbb{R}^1 = (-\infty,\infty)$ разделяется точками $\alpha_0,\ldots,\alpha_{l+1}$, образуя таким образом $l+1$ непересекающихся полуинтервалов $\Delta_k = [\alpha_k,\alpha_{k+1})$, $k = \overline{0,l}$, при этом $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_l < \alpha_{l+1} = +\infty$. Обычно выбирают $\alpha_1 \leqslant x_{(1)}$, $\alpha_l \geqslant x_{(n)}$ так, как это сделано при построении гистограммы в п. 1.11. Подсчитывается число n_k элементов выборки, попавших в каждый k-й разряд Δ_k, $k = \overline{1,l-1}$, за исключением Δ_0 и Δ_l. Полагается $n_0 = n_l = 0$.

4) Вычисляются гипотетические вероятности p_k попадания случайной величины X в полуинтервалы Δ_k, $k = \overline{0,l}$.

$$p_k = \overline{F}\left(\alpha_{k+1},\widehat{\theta}_1,\ldots,\widehat{\theta}_s\right) - \overline{F}\left(\alpha_k,\widehat{\theta}_1,\ldots,\widehat{\theta}_s\right)$$

Если у распределения $\overline{F}(x,\theta_1,\ldots,\theta_s)$ имеется плотность $\overline{f}(x,\theta_1,\ldots,\theta_s)$, то вероятности p_k могут быть вычислены следующим образом:

$$p_k = \int\limits_{\alpha_k}^{\alpha_{k+1}} \overline{f}\left(x,\widehat{\theta}_1,\ldots,\widehat{\theta}_s\right) dx,$$

где $\alpha_0 = -\infty$, $\alpha_{l+1} = +\infty$.

5) Вычисляется реализация статистики критерия хи-квадрат по формуле

$$X^2 = \varphi(z_n) = np_0 + \sum_{k=1}^{l-1} \frac{(n_k - np_k)^2}{np_k} + np_l.$$

6) Известно, что при соблюдении некоторых естественных условий регулярности и достаточно большом объеме n выборки Z_n распределение $F(x|H_0)$ статистики $X^2 = \varphi(Z_n)$ хорошо аппроксимируется распределением $\chi^2(l-s)$ с $l-s$ степенями свободы, где s — количество неизвестных параметров предполагаемого закона распределения $\overline{F}(x,\theta_1,\ldots,\theta_s)$, а $l+1$ — количество разрядов, вероятность попадания в которые ненулевая. Тогда критическая область $\overline{G}$ принимает вид $\overline{G} = (x_{1-\alpha}(l-s),+\infty)$, где $x_{1-\alpha}(l-s)$ — квантиль уровня $1-\alpha$ распределения $\chi^2(l-s)$, α — заданный уровень значимости (обычно $\alpha = 0.05$).

7) В соответствии с критерием хи-квадрат гипотеза H_0 принимается (т. е. реализация выборки z_n согласуется с гипотезой H_0) на уровне надежности $1-\alpha$, если $\varphi(z_n) \in G = [0,x_{1-\alpha}(l-s)]$. Если же $\varphi(z_n) \in \overline{G}$, то гипотеза H_0 отвергается.

Замечание 7.5. Если при разбиении на полуинтервалы Δ_k оказалось, что $np_k < 5$ для $k = \overline{1,l-1}$, то рекомендуется объединить соседние полуинтервалы.

Если при обработке наблюдений имеется только реализация статистического ряда, то вычисляя выборочные моменты считают все выборочные значения, попавшие в k-й интервал, равными середине этого интервала. Это вносит известную ошибку, особенно заметную при малом числе интервалов. Для уменьшения ошибок, вносимых группировкой, применяют ***поправки Шеппарда***. Если все интервалы Δ_k имеют длину, равную h, то с учетом поправок Шеппарда первые четыре выборочных момента $\widehat{\nu}'_i$, $i=\overline{1,4}$, соответственно равны

$$\widehat{\nu}'_1 = \widehat{\nu}_1, \qquad \widehat{\nu}'_2 = \widehat{\nu}_2 - \frac{1}{12}h^2,$$
$$\widehat{\nu}'_3 = \widehat{\nu}_3 - \frac{1}{4}\widehat{\nu}_1 h^2, \qquad \widehat{\nu}'_4 = \widehat{\nu}_4 - \frac{1}{2}\widehat{\nu}_2 h^2 + \frac{7}{240}h^4.$$

Пример 7.2. В течение Второй мировой войны на южную часть Лондона упало 535 снарядов. Территория южного Лондона была разделена на 576 участков площадью 0,25 км^2. В следующей таблице приведено количество участков n_k, на каждый из которых упало по k снарядов:

k	0	1	2	3	4	5
n_k	299	211	93	35	7	1

Таблица 7.1

Требуется с помощью критерия хи-квадрат проверить на уровне доверия $1-\alpha = 0{,}95$ гипотезу H_0, состоящую в том, что случайная величина X (число снарядов, упавших на один участок) распределена по закону Пуассона.

Р е ш е н и е: Распределение Пуассона $\mathbf{\Pi}(\theta)$ имеет один параметр, МП-оценка этого параметра $\widehat{\theta} = \widehat{m}_X \approx 0{,}93$ (см. пример 3.3).

В данной задаче действительная ось естественным образом разбивается на 8 непересекающихся полуинтервалов Δ_k: $(-\infty,0)$, $[0,1)$, $[1,2)$,..., $[5,6)$, $[6,+\infty)$.

Гипотетические вероятности p_k попадания пуассоновской случайной величины X в k-й полуинтервал вычисляются по следующим формулам:

$$p_k = \frac{e^{-\widehat{\theta}}\widehat{\theta}^{k-1}}{(k-1)!} \qquad \text{для } k=\overline{1,6},$$
$$p_7 = 1 - \sum_{k=1}^{6} p_k, \qquad p_0 = \mathsf{P}\{X<0\} = 0.$$

Вычисленные значения вероятностей указаны в табл. 7.2.

k	1	2	3	4	5	6	7
Δ_k	$[0,1)$	$[1,2)$	$[2,3)$	$[3,4)$	$[4,5)$	$[5,6)$	$[6,\infty)$
p_k	0,394	0,366	0,171	0,053	0,012	0,002	0,002

Таблица 7.2

Вычисляя значения z статистики критерия хи-квадрат, получим

$$z = \sum_{k=1}^{6} \frac{(n_k - np_k)^2}{np_k} + np_7 \approx 2{,}2.$$

При справедливости гипотезы H_0 статистика Z имеет распределение $\chi^2(5)$. Тогда критическая область $\overline{G}$ имеет вид $\overline{G} = (x_{0,95}(5), +\infty) = (11{,}1, +\infty)$, а доверительная область — $G = [0,11{,}1]$.

Так как вычисленное по выборке значение статистики попадает в доверительную область G, то с вероятностью 0,95 можно утверждать, что опытные данные согласуются с гипотезой H_0.

О т в е т. Гипотеза H_0 принимается на уровне значимости $\alpha = 0.05$.

Пример 7.3. Используя данные примера 1.1, проверить на уровне значимости $\alpha = 0{,}95$ гипотезу H_0, состоящую в том, что рост взрослого мужчины (случайная величина X) имеет нормальное распределение.

Р е ш е н и е: Нормальное распределение $\mathcal{N}(\theta_1;\theta_2^2)$ имеет два неизвестных параметра: $\theta_1 = m_X$ и $\theta_2^2 = \sigma_X^2$. МП-оценкой параметра θ_1 является выборочное среднее $\widehat{m}_X$, а параметра θ_2^2 — выборочная дисперсия $\widehat{d}_X$. С учетом поправок Шеппарда

$$\widehat{m}_X = 165{,}77, \qquad \widehat{d}_X = 34{,}24.$$

Вычислим гипотетические вероятности p_k, $k=\overline{1,15}$, попадания гауссовской случайной величины X в полуинтервалы Δ_k (которые определены в задаче 1.1) по приближенной формуле

$$p_k = h \cdot \frac{1}{\sqrt{2\pi}\sqrt{\widehat{d}_X}} \exp\left\{-\frac{(\bar{x}_k - \widehat{m}_X)^2}{2\widehat{d}_X}\right\},$$

где $\bar{x}_k$ — середина интервала Δ_k, а h — длина интервала Δ_k, которая в данной задаче равна 3 для всех $k=\overline{1,15}$. Результаты вычислений для $k=\overline{1,15}$ приведены в табл. 7.3. Вероятность попадания в интервал $\Delta_0=(-\infty,143)$ равна $p_0=4\times10^{-5}$, а в интервал $\Delta_{16}=[188,+\infty)$ — $p_{16}=4\cdot10^{-5}$. Вычисляя реализацию z статистики Z, получим

Δ_k	143–146	146–149	149–152	152–155	155–158
p_k	0,0003	0,002	0,007	0,023	0,058
Δ_k	158–161	161–164	164–167	167–170	170–173
p_k	0,115	0,175	0,204	0,184	0,127
Δ_k	173–176	176–179	179–182	182–185	185–188
p_k	0,067	0,027	0,009	0,002	0,0004

Таблица 7.3

$$z = np_0 + \sum_{k=1}^{15} \frac{(n_k - np_k)^2}{np_k} + np_{16} \approx 7{,}177.$$

При справедливости гипотезы H_0 статистика Z имеет распределение $\chi^2(14)$. Тогда критическая область $\overline{G}$ имеет вид $\overline{G}=(x_{0,95}(14),+\infty)=(23{,}7,+\infty)$, а доверительная область — $G=[0,23{,}7]$.

Так как вычисленное по выборке значение статистики попадает в доверительную область G, то с вероятностью 0,95 можно утверждать, что опытные данные согласуются с гипотезой H_0.

О т в е т. Гипотеза H_0 принимается на уровне значимости $\alpha=0.05$.

Лекция 8

Проверка гипотезы однородности в двухвыборочной модели

На практике часто встречаются задачи, в которых требуется выяснить, привело ли некоторое воздействие, усовершенствование или обработка к изменению наблюдаемого показателя. Например, привело ли предложение усовершенствования производственного процесса к увеличению выпуска продукции; увеличивается ли урожайность пшеницы при внесении в почву удобрений; является ли разработанная биодобавка эффективным средством для снижения (увеличения) массы тела; повышается ли точность измерительного прибора после проведённой наладки. Для решения таких задач сначала требуется провести эксперимент (измерение нужного показателя) и получить две независимые выборки. Если при проведении эксперимента воздействие (усовершенствование производства, внесение удобрений, приём биодобавки, ремонт аппаратуры) к измеряемой величине не применяется, то полученные экспериментальные данные представляют выборку, которую принято называть контрольной. Наблюдения, полученные в ходе эксперимента с воздействием, составляют опытную выборку. Понятно, что полученные выборки будут различаться. Вопрос состоит в следующем: можно ли приписать эти различия случайной изменчивости наблюдаемого показателя или контрольная и опытная выборки имеют значимое различие? Если удаётся установить, что законы распределения контрольной и опытной выборок одинаковы, то это означает, что применяемое воздействие не изменяет наблюдаемый показатель. Если же законы распределения этих выборок различаются, то это различие можно приписать эффекту производственного воздействия.

Отметим, что задача выявления различий может возникать и в ситуациях, когда выборки являются измерениями однотипных показателей, полученных в результате двух различных способов «обработки». Например, нужно выяснить различается ли по прочности сталь, производимая двумя различными методами; различается ли урожайность пшеницы, при применении двух различных удобрений; различается ли некоторый экономический показатель деятельности предприятий в двух регионах страны. Под «обработкой» в этих примерах понимают метод изготовления стали, сорт удобрения, региональный фактор. Различие законов распределения можно приписать эффекту влияния «обработки» на измеряемый показатель.

Все описанные задачи сводятся к решению проблемы проверки одинаковой распределённости (однородности) двух случайных величин, порождающих две выборки. Для решения таких задач применяют статистические критерии проверки гипотезы об однородности в двухвыборочной модели. Рассмотрим пять наиболее важных и распространенных критериев. Почему же существует так много критериев для решения этой проблемы? Это обстоятельство, в основном, связано с двумя аспектами.

Первый — это характер неоднородности, который может определяться физическим смыслом задачи. Например, применение удобрений производится с целью повышения урожайности, употребление биодобавки — для снижения или наоборот увеличения веса. Понятно, что в этих случаях случайные величины, порождающие две выборки, различаются лишь средними значениями, а соответствующие им распределения — сдвигом. Если же требуется выяснить, одинакова ли точность двух однотипных не имеющих систематической

ошибки измерительных приборов, то распределения выборок, соответствующих показаниям двух приборов, порождающих эти выборки, могут различаться лишь дисперсиями. В случаях, когда известно, что неоднородность выборок связана со сдвигом или растяжением (сжатием), применяют специальные критерии, предназначенные только для таких типов неоднородности. Если характер неоднородности априори неизвестен, то существуют критерии, которые позволяют проверять однородность двух выборок против любых возможных альтернатив.

Второй аспект связан с ограничениями, накладываемыми на статистическую модель. Так, если априори известно или может быть проверено, что наблюдения имеют нормальное распределение, то оптимальными критериями проверки однородности являются критерии, называемые классическими. К сожалению, на практике закон распределения выборок редко бывает известен. Для этих ситуаций разработаны непараметрические критерии, которые не основаны на предположении о том, что выборки имеют некоторое определённое параметрическое распределение. К таким критериям относятся, например, ранговые критерии.

Здесь у вас может возникнуть естественный вопрос о том, можно ли сравнить разные критерии проверки однородности? Какой из многочисленных критериев следует выбрать для решения конкретной задачи? В разделе 7.7 учебного пособия «Прикладные методы анализа статистических данных» (Горяинова Е.Р., Панков А.Р., Платонов Е.А.) обсуждается проблема сравнения асимптотических эффективностей критериев при разных распределениях наблюдаемых величин.

8.1 Теоретические положения

Пусть выборка $\mathbb{X}_m = [X_1, \ldots, X_m]^T$ соответствует распределению $F_X(t)$, а выборка $\mathbb{Y}_n = [Y_1, \ldots, Y_n]^T$ распределению $F_Y(t)$.

Определение 8.1. Выборка $\mathbb{X}_m$ и $\mathbb{Y}_n$ называются *однородными, если* $F_X(t) = F_Y(t)$ *для любого* $t \in \mathbb{R}^1$.

Статистическую гипотезу вида

$$H_0 \colon F_X(t) = F_Y(t), \quad \forall t \in \mathbb{R}^1. \tag{8.1}$$

называют *гипотезой об однородности* выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$.

Альтернативной гипотезой общего вида для H_0 является гипотеза

$$H_1 \colon \exists t \in \mathbb{R}^1, \quad \text{такое что } F_X(t) \neq F_Y(t). \tag{8.2}$$

Неоднородность выборок может быть обусловлена разными причинами. Рассмотрим два важнейших типа неоднородности, которые можно описать, использую понятия сдвига и сжатия (растяжения) распределений $F_X(t)$ и $F_Y(t)$.

Пусть неоднородность выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$ состоит в том, что распределения $F_X(t)$ и $F_Y(t)$ различаются лишь сдвигом на некоторую величину θ , а именно:

$$F_Y(t) = F_X(t - \theta), \quad \forall t \in \mathbb{R}^1. \tag{8.3}$$

В этом случае будем говорить, что неоднородность выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$ вызвана наличием сдвига. Если неоднородность выборок обусловлена лишь сдвигом (8.3), то гипотеза об однородности формулируется следующим образом:

$$H_0 \colon \theta = 0, \tag{8.4}$$

а альтернативные гипотезы имеют вид:

$$H_1 \colon \theta < 0, \qquad H_2 \colon \theta > 0, \qquad H_3 \colon \theta \neq 0.$$

Замечание 8.1. Если $\mathsf{E}X < \infty$, то нетрудно показать, что $\theta = \mathsf{E}Y - \mathsf{E}X$. Тогда гипотеза H_1 означает, что $\mathsf{E}Y < \mathsf{E}X$, гипотеза H_2 означает, что $\mathsf{E}Y > \mathsf{E}X$, гипотеза $H_3 : \mathsf{E}Y \neq \mathsf{E}X$.

Пусть выборка $\mathbb{X}_m$ соответствует распределению $F_X(t-\mu)$, а выборка $\mathbb{Y}_n$ — распределению

$$F_Y(t-\mu) = F_X\left(\frac{t-\mu}{\Delta}\right), \qquad \Delta > 0,$$

где функция $F(t)$ удовлетворяет условию

$$\int_{-\infty}^{+\infty} t dF(t) = 0,$$

а μ — мешающий параметр сдвига. В этом случае математическое ожидание случайных величин X и Y, порождающих выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$, совпадают, а неоднородность выборок вызвана растяжением (сжатием).

Замечание 8.2. Рассмотренную модель можно описать и другим способом. Пусть μ — мешающий параметр сдвига, который предполагается одинаковым для случайных величин X и Y, порождающих выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$. Обозначим через $\widetilde{X}$ случайную величину $X-\mu$ и через $\widetilde{Y}$ — случайную величину $Y-\mu$. Пусть $F_{\widetilde{X}}(t)$ — функция распределения случайной величины $\widetilde{X}$. Тогда функция распределения случайной величины $\widetilde{Y}$ имеет вид $F_{\widetilde{X}}(\frac{t}{\Delta})$.

Гипотеза об однородности в этом случае имеет вид

$$H_0 \colon \Delta = 1, \tag{8.5}$$

а альтернативные к ней гипотезы вид

$$H_1 \colon \Delta < 1, \qquad H_2 \colon \Delta > 1, \qquad H_3 \colon \Delta \neq 1.$$

Замечание 8.3. Пусть $\mathsf{D}\{X\} < \infty$. Тогда можно показать, что $\Delta^2 = \frac{DY}{DX}$. Таким образом, справедливость гипотезы H_1 означает, что $\mathsf{D}\{X\} > \mathsf{D}\{Y\}$, из H_2 следует, что $\mathsf{D}\{X\} < \mathsf{D}\{Y\}$ и из H_3 следует, что $\mathsf{D}\{X\} \neq \mathsf{D}\{Y\}$.

Для проверки гипотезы об однородности используют различные критерии, применимость которых обусловлена различными требованиями, предъявляемыми к выборкам.

Определение 8.2. *Рангом* элемента выборки называют номер места, которое занимает элемент в вариационном ряду.

Процедуру определения рангов всех элементов выборки называют ранжированием.

Определение 8.3. Совокупность совпадающих наблюдений называется *связкой*. Количество наблюдений в связке называют *размером связки*. При ранжировании всем элементам связки присваивается *средний ранг*.

Средний ранг связки определяется следующим образом: если связке предшествует k элементов вариационного ряда, и связка имеет размер m, то средний ранг этой связки равен $\frac{1}{m}\sum\limits_{i=k+1}^{k+m} i$.

Заметим, что средний ранг может принимать как целые, так и дробные значения.

Критерии, базирующиеся на предположении о гауссовости выборок, принято называть *классическими*. Критерии, статистики которых являются функциями рангов наблюдений, называются *ранговыми критериями*.

Для проверки гипотезы об однородности вида (8.3) – (8.4) можно использовать, например, критерий Стьюдента и ранговый критерий Вилкоксона, а для проверки гипотезы вида (8.5) — критерий Фишера и ранговый критерий Ансари–Брэдли. Подробно рассмотрим эти критерии.

8.2 Проверка гипотезы об однородности против альтернатив о сдвиге

Критерий Стьюдента

Пусть справедливы следующие положения:

1) выборка $\mathbb{X}_m$ соответствует распределению $\mathcal{N}(m_X;\sigma_X^2)$, а выборка $\mathbb{Y}_n$ распределению $\mathcal{N}(m_Y;\sigma_Y^2)$;

2) дисперсии σ_X^2 и σ_Y^2 одинаковы $\sigma_X^2=\sigma_Y^2=\sigma^2$ и неизвестны;

3) выборки независимы.

Гипотеза об однородности в рамках этой модели имеет вид (8.4):

$$H_0:\ \theta = m_Y - m_X = 0.$$

Статистика критерия Стьюдента вычисляется следующим образом:

$$T(\mathbb{X}_m,\mathbb{Y}_n) = T(\mathbb{Z}_N) = \frac{\overline{Y}-\overline{X}}{S_N\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}, \tag{8.6}$$

где

$$S_N^2 = \frac{1}{N-2}\left[\sum_{i=1}^{m}(X_i-\overline{X})^2+\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right]$$

— оценка неизвестной дисперсии σ^2 по объединенной выборке $\mathbb{Z}_N=[X_1,\ldots,X_m,Y_1,\ldots,Y_n]^T$ объема $N=m+n$.

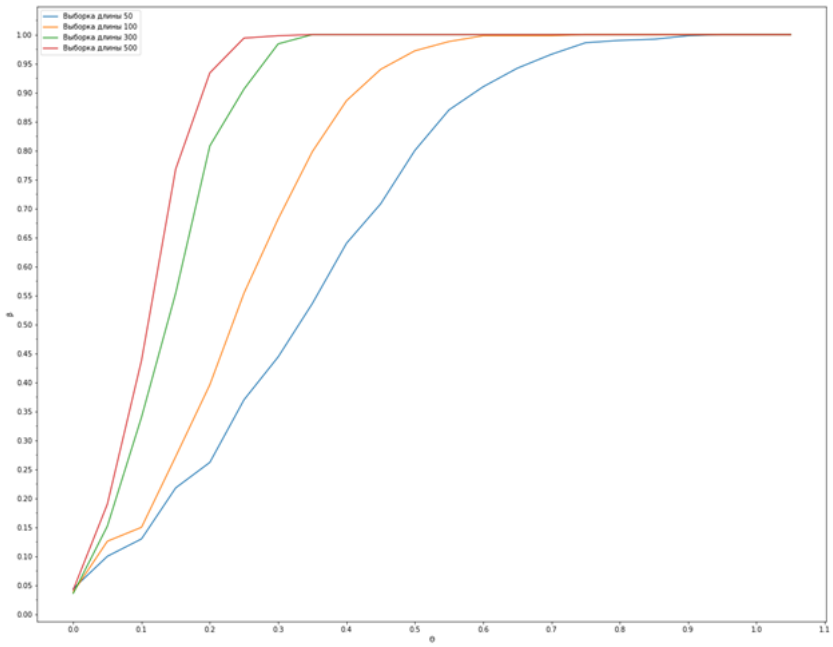
При справедливости гипотезы H_0 статистика $T(\mathbb{Z}_N)$ имеет распределение Стьюдента с $r=N-2$ степенями свободы. В зависимости от типа задачи имеет смысла рассматривать альтернативные гипотезы H_A различного вида. Критически области уровня значимости $\alpha\in(0;1)$ критерия Стьюдента, соответствующие различным H_A, приведены в табл. 8.1, где $t_\gamma(r)$ — квантиль уровня γ распределения Стьюдента.

H_A	Критические области для $T(\mathbb{Z}_N)$
$\theta<0$	$(-\infty;t_\alpha(r))$
$\theta>0$	$(t_{1-\alpha}(r);+\infty)$
$\theta\neq 0$	$(-\infty;t_{\frac{\alpha}{2}}(r))\cup(t_{1-\frac{\alpha}{2}}(r);+\infty)$

Таблица 8.1

Заметим, что при $N^*=\min(m,n)\to\infty$ статистика (8.6) асимптотически нормальна.

Ниже приведены графики, иллюстрирующие состоятельность критерия Стьюдента для проверки гипотезы о равенстве средних двух гауссовских величин против альтернативы сдвига. Генерируются пары гауссовских выборок заданного объёма. Первая выборка имеет стандартное гауссовское распределение, вторая - гауссовское распределение с математическим ожиданием θ и дисперсией 1. Рассматриваются значения $\theta=0;0{,}001\cdot k, r=1,...,1000$. Для каждого фиксированного объёма выборки при каждом значении θ моделирование проводится 1000 раз. На уровне значимости 0.05 для каждой смоделированной пары выборок проводится проверка гипотезы $H_0: m_1=m_2$ о равенстве средних против альтернативы $H_1: m_1<m_2$ с помощью критерия Стьюдента. На оси X - значения параметра сдвига θ, на оси Y - частота принятия альтернативной гипотезы. Графики для выборок объёма 50; 100; 300; 500 нарисованы соответственно синим, оранжевым, зелёным и красным цветом. Пример разработан студентом ОПИ ФКН Евдокимовым Максимом.



Критерий Вилкоксона

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует неизвестному непрерывному распределению $F(t)$, а выборка $\mathbb{Y}_n$ распределению $F(t-\theta)$;

2) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ независимы.

Критерий Вилкоксона позволяет проверить в модели (8.3) гипотезу вида (8.4)

$$H_0\colon\ \theta = 0.$$

Статистика критерия имеет вид

$$T(\mathbb{Z}_N) = W_{m,n} = \sum_{j=1}^{n} R_j, \tag{8.7}$$

где R_j — ранг элемента Y_j в объединенной выборке $\mathbb{Z}_N = [X_1,\ldots,X_m,Y_1,\ldots,Y_n]^T$ объема $N = m+n$.

Для выборок небольшого объема можно найти точное распределение статистики $W_{m,n}$ при справедливости H_0 вида (8.4). Квантили этого распределения табулированы, например, в книге Большев Л.Н., Смирнов Н.В. «Таблицы математической статистики» (М.: Наука, 1983) для $1 \leqslant n \leqslant m \leqslant 25$. Далее в задаче 3.1 мы укажем алгоритм вычисления квантилей статистики $W_{m,n}$ при справедливости H_0.

Можно показать, что при справедливости H_0 для любых m и n

$$\mathsf{E}\{W_{m,n}\} = \frac{n}{2}(N+1), \quad \mathsf{D}\{W_{m,n}\} = \frac{m\cdot n}{12}(N+1),$$

а стандартизованная статистика

$$W^*_{m,n} = \frac{W_{m,n} - \mathsf{E}\{W_{m,n}\}}{\sqrt{\mathsf{D}\{W_{m,n}\}}}$$

при $N^* = \min(m,n) \to \infty$ асимптотически нормальна.

Если в выборке имеются связки, то при вычислении статистики $W^*_{m,n}$ следует заменить дисперсию $\mathsf{D}\{W_{m,n}\}$ на выражение

$$\widetilde{\mathsf{D}}\{W_{m,n}\} = \mathsf{D}\{W_{m,n}\} - \frac{m\cdot n \sum\limits_{k=1}^{l} t_k(t_k^2-1)}{12N(N-1)},$$

где l — количество связок в выборке $\mathbb{Z}_N$, а t_k — размер k-й связки, $k = 1,\ldots,l$.

Важно отметить, что распределение статистик $W_{m,n}$ и $W^*_{m,n}$ критерия Вилкоксона при справедливости H_0 не зависит от распределения $F(t)$. Критерии, статистики которых обладают таким свойством, принято называть *свободными от распределения*.

Отметим также, что для применения критерия Вилкоксона не требуется выполнения условия $\mathsf{E}\{X\} < \infty$.

Критические области уровня значимости α критерия Вилкоксона, основанного на статистике $W^*_{m,n}$, соответствующие различным альтернативам H_A, указаны в табл. 3.2. Через u_γ обозначена квантиль уровня γ распределения $\mathcal{N}(0;1)$.

H_A	**Критические области для $W^*_{m,n}$**
$\theta < 0$	$(-\infty; u_\alpha]$
$\theta > 0$	$(u_{1-\alpha}; +\infty)$
$\theta \neq 0$	$(-\infty; u_{\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)$

Таблица 8.2

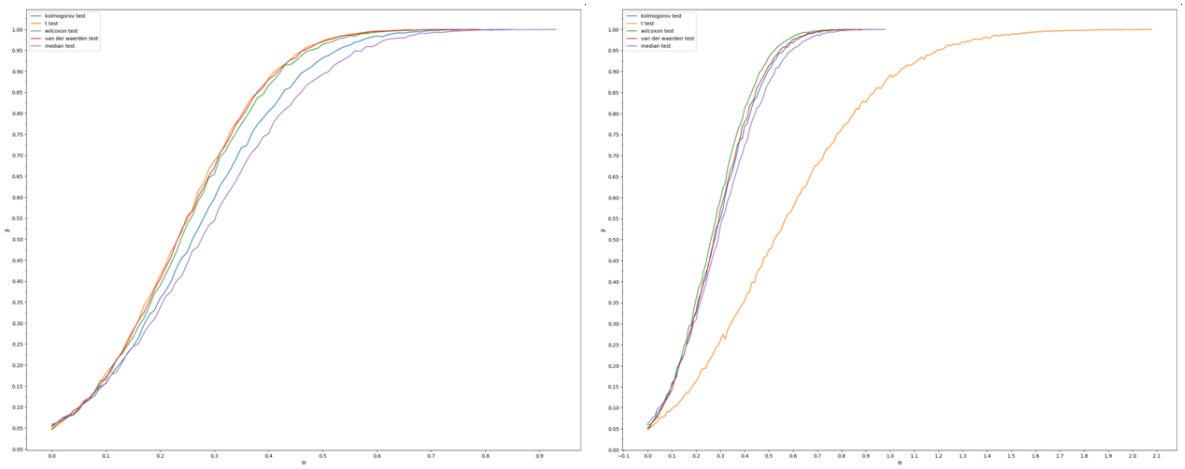
Критерии Стьюдента и Вилкоксона при выполнении указанных выше предположений являются состоятельными для альтернатив вида $H_1\colon\ \theta < 0$, $H_2\colon\ \theta > 0$, $H_3\colon\ \theta \neq 0$.

Ниже представлены графики эмпирических функций мощности для нескольких критериев, проверяющих гипотезу об однородности двух выборок против альтернативы сдвига. На левом графике первая выборка имеет стандартное гауссовское распределение, вторая - гауссовское распределение с математическим ожиданием θ и дисперсией 1. Рассматриваются значения $\theta = 0; 0{,}001 \cdot k, r = 1,...,1000$. Для объёма выборки 100 при каждом значении θ моделирование проводится 1000 раз. На уровне значимости 0.05 для каждой смоделированной пары выборок проводится проверка гипотезы $H_0 : m_1 = m_2$ о равенстве средних против альтернативы $H_1 : m_1 < m_2$ с помощью критерия Стьюдента, критерия Вилкоксона и критерия Колмогорова-Смирнова (о нём будет следующая лекция) . На оси X - значения параметра сдвига θ, на оси Y - частота принятия альтернативной гипотезы. Полученная в результате моделирования эмпирическая функция мощности критерия Стьюдента нарисована оранжевым цветом, критерия Вилкоксона - зелёным цветом, критерия Колмогорова-Смирнова - голубым цветом. Моделирование показывает, что в гауссовском случае критерий Стьюдента имеет самую высокую мощность среди рассмотренных критериев. Важно отметить, что критерий Вилкоксона проигрывает критерию Стьюдента незначительно.

На правом графике представлены эмпирические мощности критериев Стьюдента, Вилкоксона и Колмогорова-Смирнова для выборок, имеющих распределение Тьюки с функцией распределения

$$F(x) = (1-\gamma)\Phi(x) + \gamma\Phi(\frac{x}{\sigma})$$

с параметрами $\gamma = 0.05$ и $\sigma = 5$. Это распределение описывает ситуацию, когда в гауссовскую выборку попадает 5 процентов "засорений". Отметим, что в этой ситуации критерий Стьюдента существенно проигрывает (т.е. является менее мощным) критерию Вилкоксона и критерию Колмогорова-Смирнова.



Задача 8.1. Построить функцию распределения статистики Вилкоксона $W_{m,n}$ при справедливости гипотезы H_0 вида (8.3) – (8.4) для $m = 4$, $n = 2$. Найти квантили распределения статистики $W_{4,2}$ уровня 0,9 и 0,1.

Р е ш е н и е: При справедливости гипотезы H_0 вида (8.3)–(8.4) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ являются однородными, и вероятности появления любого набора рангов игреков $(R_1,\ldots,R_n)$ в объединенной выборке $\mathbb{Z}_N = [X_1,\ldots,X_m,Y_1,\ldots,Y_n]^T$ объема $N = m+n$ одинаковы. Поскольку существует C_N^n различных способов размещения элементов $Y_1,\ldots,Y_n$ среди $n+m$ элементов объединенной выборки, то вероятность появления любого набора рангов $(R_1,\ldots,R_n)$ будет равна $\dfrac{1}{C_N^n}$. Выпишем все возможные комбинации рангов игреков (R_1,R_2) для случая $m=4$, $n=2$ и вычислим соответствующие им значения статистики Вилкоксона

$$W_{4,2} = \sum_{i=1}^{2} R_i.$$

Таких комбинаций будет $C_6^2 = 15$ (табл. 8.3).

Таблица 8.3

(R_1,R_2)	(1;2)	(1;3)	(1;4)	(1;5)	(1;6)	(2;3)	(2;4)	(2;5)
$W_{4,2}$	3	4	5	6	7	5	6	7
(R_1,R_2)	(2;6)	(3;4)	(3;5)	(3;6)	(4;5)	(4;6)	(5;6)	
$W_{4,2}$	8	7	8	9	9	10	11	

Ряд распределения статистики $W_{4,2}$ при справедливости гипотезы H_0 имеет вид:

Таблица 8.4

$W_{4,2}$	3	4	5	6	7	8	9	10	11
P	1/15	1/15	2/15	2/15	3/15	2/15	2/15	1/15	1/15

По определению квантилью уровня γ, где $\gamma \in (0;1)$, распределения $F(t)$ называется такое число z_γ, что

$$z_\gamma = \min\{t\colon F(t) \geqslant \gamma\}.$$

Зная функцию распределения $F_W(t)$ статистики $W_{4,2}$, найдем квантили $W_{0,1}(4,2)$ уровня 0,1 и $W_{0,9}(4,2)$ уровня 0,9.

Поскольку в данном случае

$$F_W(3) = \frac{1}{15} \approx 0{,}067, \qquad F_W(4) = \frac{2}{15} \approx 0{,}13,$$

имеем

$$W_{0,1}(4,2) = \min\{t\colon F_W(t) \geqslant 0{,}1\} = 4.$$

Аналогично,

$$F_W(9) = \frac{13}{15} \approx 0{,}87, \qquad F_W(10) = \frac{14}{15} \approx 0{,}93.$$

Значит,

$$W_{0,9}(4,2) = \min\{t\colon F_W(t) \geqslant 0{,}9\} = 10.$$

Заметим, что минимальное значение статистики $W_{m,n}$ соответствует ситуации, при которой все элементы второй выборки меньше всех элементов первой выборки, т.е.

$$(R_1,\ldots,R_n) = (1,\ldots,n),$$

и

$$\min W_{m,n} = 1+\ldots+n = \frac{n(n+1)}{2}.$$

Максимальное значение статистики $W_{m,n}$ соответствует ситуации, при которой все элементы второй выборки больше всех элементов первой выборки, т.е.

$$(R_1,\ldots,R_n) = (m+1,\ldots,n+m),$$

и

$$\max W_{m,n} = (m+1) + \ldots + (n+m) = \frac{n(m+2n+1)}{2}.$$

Распределение статистики $W_{m,n}$ при справедливости H_0 симметрично относительно своего среднего значения

$$\mathsf{E}\{W_{m,n}\} = \frac{n(m+n+1)}{2}.$$

Поэтому статистические таблицы содержат квантили либо высоких (не менее 0,5) либо низких (менее 0,5) уровней.Таким образом, если известна квантиль уровня γ распределения $W_{m,n}$, то, используя симметрию распределения $F_W(t)$, можно найти квантиль уровня $1-\gamma$ распределения статистики $W_{m,n}$ из соотношения

$$W_{1-\gamma}(m;n) - \mathsf{E}\{W_{m,n}\} = \mathsf{E}\{W_{m,n}\} - W_{\gamma}(m;n).$$

Этим свойством можно было воспользоваться и в данном примере. Зная $W_{0,9}(4;2) = 10$ и вычисляя

$$\mathsf{E}\{W_{4,2}\} = \frac{2(2+4+1)}{2} = 7,$$

найдем

$$W_{0,1}(4;2) = 2\mathsf{E}\{W_{4,2}\} - W_{0,9}(4;2) = 14 - 10 = 4.$$

Пример 8.1. Изучается влияние кобальта на увеличение массы тела кроликов. Опыт проводится на двух группах животных: контрольной и опытной. Возраст животных 1,5 – 2 месяца, исходная масса 500 – 600 грамм. Рацион одинаков, но опытной группе добавляют в пищу 0,06 грамм хлористого кобальта на 1 кг массы тела кролика.

Прибавка в весе составила:

Таблица 8.5

Контрольная группа	560	580	600	420	530	490	580	740
Опытная группа	692	700	621	640	561	680	630	

Влияет ли кобальтовая добавка на увеличение массы тела?

Р е ш е н и е: Наблюдения в данной задаче можно считать независимыми случайными величинами. Пусть $X_1,\ldots,X_8 \sim F(t)$, а $Y_1,\ldots,Y_7 \sim F(t-\theta)$.

Гипотеза $H_0 : \theta = 0$ описывает ситуацию, когда контрольная и опытная выборки одинаково распределены. Альтернативная гипотеза $H_A : \theta > 0$ соответствует ситуации, когда среднее значение второй (опытной) выборки больше среднего значения первой (контрольной) выборки. Проверим гипотезу $H_0 : \theta = 0$ против альтернативы $H_A : \theta > 0$, с помощью критерия Вилкоксона. Объединим выборки и проранжируем объединённую выборку.

Таблица 8.6

Ранг	1	2	3	4	5	6,5	6,5	8	9	10	11	12	13	14	15
Вес	420	490	530	560	<u>561</u>	580	580	600	<u>621</u>	<u>630</u>	<u>640</u>	<u>680</u>	<u>692</u>	<u>700</u>	740

Подчеркнутые значения соответствуют элементам второй выборки. Вычислим реализацию статистики Вилкоксона по формуле (8.7).

$$W_{8,7} = \sum_{i=1}^{7} R_i = 5+9+10+11+12+13+14 = 74 \tag{8.8}$$

Построим доверительную и критическую области.

В таблице квантилей распределения Вилкоксона найдем квантиль распределения $W_{m,n}$ для $m=8$ и $n=7$ уровня $1-0{,}047$:

$$W_{8,7;1-0{,}047}=71.$$

Доверительная $\left[\min W_{8,7};71\right)$ и критическая $\left(71;\max W_{8,7}\right]$ области имеют вид, изображенный на рис. 8.1. Если данных много $(n+m>20)$, то можно аппроксимировать распределение статистики W гауссовским распределением.

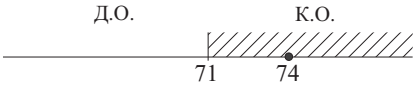


Рис 8.1

Вычислим $\mathsf{E}[W_{m,n}]$ и $D[W_{m,n}]$:

$$\mathsf{E}[W_{m,n}]=\frac{n\cdot(m+n+1)}{2}=\frac{7\cdot(7+8+1)}{2}=56 \tag{8.9}$$

$$D[W_{m,n}]=\frac{n\cdot m\cdot(m+n+1)}{12}=\frac{7\cdot 8\cdot(7+8+1)}{12}=\frac{224}{3} \tag{8.10}$$

Тогда

$$W^*=\frac{W_{m,n}-\mathsf{E}[W_{m,n}]}{\sqrt{D[W_{m,n}]}}=\frac{(74-56)}{8.64}=2.08$$

Так как $W^*|_{H_0}\sim N(0,1)$ и $Z_{1-\alpha}=Z_{0,95}=1{,}65$, то доверительная и критическая области выглядят следующим образом (см. рис. 8.2).

Видим, что вычисленные значения статистик попадают в критические области. Следовательно, на уровне значимости $\alpha=0.05$ гипотеза H_0 отвергается в пользу H_A.

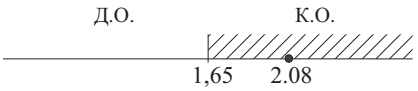


Рис 8.2

Решим теперь эту задачу с помощью критерия Стьюдента.

Предположим дополнительно (!), что:

$$X_1,\ldots,X_8\sim N\left(m_1,\sigma^2\right),\qquad \text{а}\qquad Y_1,\ldots,Y_7\sim N\left(m_2,\sigma^2\right). \tag{8.11}$$

Проверим гипотезу

$$H_0: m_1-m_2=0,\qquad \text{против}\qquad H_A: m_2-m_1>0.$$

Статистика Стьюдента вида 8.6 при справедливости H_0 имеет распределение Стьюдента $t(m+n-2)=t(13)$.

Вычислим $\overline{X}=562{,}5$, $\overline{Y}=646{,}29$ и выборочную дисперсию объединенной выборки

$$S^2=\frac{1}{13}\left(\sum_{i=1}^{8}\left(X_i-\overline{X}\right)^2+\sum_{i=1}^{7}\left(Y_i-\overline{Y}\right)^2\right)=\frac{1}{13}(60150+16905)=5927{,}31. \tag{8.12}$$

Тогда выборочное среднеквадратическое отклонение $S=77$, а реализация статистики

$$T=\frac{\overline{Y}-\overline{X}}{S\cdot\sqrt{\frac{1}{7}+\frac{1}{8}}}=\frac{646{,}29-562{,}5}{39}=2{,}15. \tag{8.13}$$

По таблице находим 0,95 квантиль распределения Стьюдента $t_{0{,}95;\,13}=1{,}77$.

Построим доверительную и критическую области. Доверительная область имеет вид $(-\infty;1{,}77)$, критическая область — $(1{,}77;+\infty)$.

Вычисленное значение статистики $t=2{,}15$ попадает в критическую область.Следовательно, на уровне значимости 0.05 основная гипотеза отвергается в пользу альтернативной.

Пример 8.2. Имеется набор данных «ирисы Фишера«. Эти данные собраны ботаником Эдгаром Андерсоном. Они включают длину и ширину чашелистиков, длину и ширину лепестков трёх видов ирисов (setosa, versocolor и виргинский). Выберем из этих данных случайным образом по 10 измерений длин чашелистиков цветов вида setosa и цветов вида versocolor. Длина (в мм) чашелистиков у цветов вида setosa:

$$5.1; 4.9; 4.7; 4.6; 5.0; 5.4; 4.6; 4.4; 4.8; 4.8.$$

Длина (в мм) чашелистиков у цветов вида versocolor:

$$5.7; 6.3; 4.9; 6.6; 5.2; 5.0; 5.9; 6.0; 5.6; 5.8.$$

Можно ли считать, опираясь на эти данные, что длина чашелистиков у цветов вида versocolor в среднем больше, чем у цветов вида setosa?

Пример 8.3. Имеются данные Федеральной службы государственной статистики о среднедушевых денежных доходах населения (рублей в месяц) 2008 г. по некоторым областям Центрального и Приволжского федеральных округов. Данные представлены в табл. 8.8.

Таблица 8.7

Центральный федеральный округ	**Доход, руб. в месяц**	**Приволжский федеральный округ**	**Доход, руб. в месяц**
Брянская область	10043	Республика Башкортостан	14253
Владимирская область	9596	Республика Марий Эл	7843
Воронежская область	10305	Удмуртская Республика	9581
Ивановская область	8354	Чувашская Республика	8594
Костромская область	9413	Пермский Край	16119
Московская область	19776	Кировская область	10112
Орловская область	9815	Пензенская область	10173
Рязанская область	11311	Ульяновская Область	9756
Тамбовская область	11253		
Тверская область	10856		
Тульская область	11389		

Выяснить, одинаковы ли в среднем среднедушевые доходы населения в этих округах. Уровень значимости считайте равным 0,05.

Р е ш е н и е: Пусть среднедушевые доходы по Центральному федеральному округу (ЦФО) являются выборкой $\mathbb{X}_m = [X_1, \ldots, X_m]^T$ объема $m = 11$, соответствующей распределению $F_X(t)$, а доходы по Приволжскому федеральному округу (ПФО) — выборкой $\mathbb{Y}_n = [Y_1, \ldots, Y_n]^T$ объема $n = 8$, соответствующей распределению $F_Y(t)$. В данной задаче естественно предположить, что неоднородность выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$ обусловлена различием средних значений случайной величины X (показатель среднедушевого дохода в ЦФО) и случайной величины Y (показатель среднедушевого дохода в ПФО), порождающих выборки $\mathbb{X}_m$

и $\mathbb{Y}_n$. Тогда $F_Y(t) = F_X(t-\theta)$. Для проверки гипотезы $H_0\colon\ \theta = 0$ об однородности выборок против альтернативы сдвига $H_A\colon\ \theta \neq 0$ можно применить критерий Вилкоксона.

Причина, по которой выбирается альтернативная гипотеза указанного вида, заключается в том, что мы не имеем априорной информации о том, что в каком-то из рассматриваемых нами округов показатели среднедушевых доходов должны быть больше или меньше, чем в другом округе.

Статистика критерия Вилкоксона имеет вид (8.7)

$$W_{m,n} = \sum_{j=1}^{n} R_j,$$

где R_j — ранг случайной величины Y_j в объединенной выборке $\mathbb{Z}_N = [X_1,\ldots,X_m,Y_1,\ldots,Y_n]^T$. Проведем ранжирование объединенной выборки и вычислим реализацию статистики Вилкоксона

$$W_{11,8} = 1+3+5+7+10+11+17+18 = 72.$$

Критическая область, соответствующая уровню значимости 0,05, имеет вид

$$\Big[\min W_{11,8}; W_{0,025}(11;8)\Big) \cup \Big(W_{0,975}(11,8); \max W_{11,8}\Big].$$

По таблице найдем квантиль $W_{0,025}(11,8) = 55$. Тогда (см. задачу 3.1) квантиль

$$W_{0,975}(11,8) = 2\mathbf{E}\left\{W_{11,8}\right\} - W_{0,025}(11,8) = \frac{2n(m+n+1)}{2} - 55 = 160 - 55 = 105.$$

Таким образом, критическая область критерия Вилкоксона, основанного на статистике $W_{m,n}$, имеет вид:

$$\Big[\min W_{11,8}; 55\Big) \cup \Big(105; \max W_{11,8}\Big]$$

Так как реализация статистики $W_{m,n}$ попадает в доверительную область, то гипотеза H_0 принимается на уровне значимости $\alpha = 0{,}05$. Предположим теперь, что наблюдаемые случайные величины соответствуют гауссовскому распределению. Такое предположение допустимо, так как каждый элемент выборки является выборочным средним большого количества случайных величин. Тогда задача может быть формализована следующим образом. Среднедушевые доходы по ЦФО $X_1,\ldots,X_m$ являются выборкой объема $m = 11$, соответствующей распределению $\mathcal{N}\left(m_X;\sigma_X^2\right)$, а среднедушевые доходы по ПФО $Y_1,\ldots,Y_n$ — выборкой объема $n = 8$, соответствующей распределению $\mathcal{N}\left(m_Y;\sigma_Y^2\right)$.

В рамках такой модели требуется проверить гипотезу

$$H_0\colon\ \theta = m_Y - m_X = 0$$

против альтернативы $H_A\colon\ \theta \neq 0$.

Так как дисперсии σ_X^2 и σ_Y^2 неизвестны, то сначала следует проверить гипотезу

$$H_0\colon\ \sigma_X^2 = \sigma_Y^2 = \sigma^2$$

против альтернативы

$$H_1\colon\ \sigma_X^2 \neq \sigma_Y^2.$$

Применим для этого критерий Фишера. Вычислим реализацию выборочных средних и выборочных несмещенных дисперсий:

$$\overline{X} = \frac{1}{11}\sum_{i=1}^{11} X_i = 11101{,}0; \quad \widetilde{S}_X^2 = \frac{1}{10}\sum_{i=1}^{11}\left(X_i - \overline{X}\right)^2 = (3025{,}4)^2;$$

$$\overline{Y} = \frac{1}{8}\sum_{i=1}^{8} Y_i = 10803{,}9; \quad \widetilde{S}_Y^2 = \frac{1}{7}\sum_{i=1}^{8}\left(Y_i - \overline{Y}\right)^2 = (2860{,}3)^2.$$

Так как $\widetilde{S}_X^2 > \widetilde{S}_Y^2$, то статистика Фишера будет иметь вид

$$T(\mathbb{X}_m, \mathbb{Y}_n) = F_{m,n} = \frac{\widetilde{S}_X^2}{\widetilde{S}_Y^2}.$$

Реализация статистики $F_{m,n} = 1{,}12$.

При справедливости гипотезы

$$H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$$

статистика
$F_{m,n}$ имеет распределение Фишера $F(10;7)$. Выберем уровень значимости $\alpha = 0{,}05$, тогда критическая область имеет вид

$$\left(f_{1-\frac{\alpha}{2}}(m-1;n-1);+\infty\right) = (f_{0{,}975}(10;7);+\infty),$$

где $f_{0{,}975}(10;7)$ квантиль распределения $F(10;7)$. По таблице находим, что $f_{0{,}975}(10;7) = 4{,}76$. Следовательно, реализация статистики $F_{m,n}$ попала в доверительную область, и гипотеза H_0 принимается на уровне значимости $\alpha = 0{,}05$.

Теперь предположения, требуемые для применения критерия Стьюдента со статистикой (8.6), выполнены. Вычислим реализацию статистики критерия Стьюдента

$$T(\mathbb{X}_m, \mathbb{Y}_n) = T(\mathbb{Z}_N) = \frac{\overline{Y} - \overline{X}}{S_N\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}},$$

где

$$S_N^2 = \frac{\left[\sum\limits_{i=1}^{m}(X_i - \overline{X})^2 + \sum\limits_{i=1}^{n}(Y_i - \overline{Y}_m)^2\right]}{m+n-2} = \frac{(m-1)\widetilde{S}_X^2 + (n-1)\widetilde{S}_Y^2}{m+n-2} =$$
$$= \frac{10\cdot(3025{,}4)^2 + 7\cdot(2860{,}3)^2}{11+8-2} = (2958{,}5)^2.$$

Окончательно получаем

$$T(\mathbb{Z}_N) = \frac{-297{,}13}{2958{,}5\sqrt{0{,}22}} = -0{,}216.$$

При справедливости H_0: $\theta = m_Y - m_X = 0$ статистика критерия Стьюдента (8.6) имеет распределение Стьюдента с $r = m+n-2 = 17$ степенями свободы. Так как альтернативная гипотеза имеет вид $H_1 : \theta \neq 0$, то критическая область уровня значимости α имеет вид

$$\left(-\infty; t_{\frac{\alpha}{2}}(r)\right) \cup \left(t_{1-\frac{\alpha}{2}}(r); +\infty\right).$$

Для $\alpha = 0{,}05$ по таблице находим $t_{0{,}975}(17) = 2{,}11$ и, учитывая то, что $t_{1-\alpha}(r) = -t_\alpha(r)$, имеем $t_{0{,}025}(17) = -2{,}11$. Так как реализация статистики попала в доверительную область, то принимается гипотеза H_0 на уровне значимости $\alpha = 0{,}05$, и можно считать, что среднедушевые доходы в этих федеральных округах в среднем одинаковы. ■

Пример 8.4. Пример от студентки Анны М. Объясните, почему задача Анны неверно формализована (т.е. критерии Вилкоксона и Стьюдента неприменимы в данной задаче)?

Имеются ежемесячные данные (сайт www.gks.ru) о средней розничной цене килограмма картофеля в Воронежской и Владимирской областях.

Таблица 8.8

Воронежская	20	20.71	21.36	22.13	23.31	23.99	25.16	25.30	25.31	26.61	26.89	26.98	27.44	27.63	27.68
Владимирская	21.16	21.27	22.46	22.74	23.34	24.07	24.45	25.32	25.67	26.36	26.8	27.61	28.27	28.37	28.86

Можно ли считать, что цена картофеля во Владимирской (нечернозёмной области) в среднем выше, чем в чернозёмной Воронежской области?

Р е ш е н и е: Анны М.

Пусть $X_1,\ldots,X_{15} \sim F(t)$, а $Y_1,\ldots,Y_{15} \sim F(t-\theta)$.

Гипотеза $H_0 : \theta = 0$ означает, что цена картофеля во Владимирской и Воронежской областях в среднем одинакова. Альтернативная гипотеза $H_A : \theta > 0$ соответствует ситуации, когда среднее значение цены во Владимирской области больше среднего значения среднего значения цены в Воронежской области. Проверим гипотезу $H_0 : \theta = 0$ против альтернативы $H_A : \theta > 0$, с помощью критерия Вилкоксона. Объединим выборки и проранжируем объединённую выборку.

$$W_{15,15} = \sum_{i=1}^{15} R_i = 3+4+7+8+10+12+13+17+18+19+21+25+28+29+30 = 244. \tag{8.14}$$

Вычислим $\mathsf{E}[W]$ и $D[W]$:

$$\mathsf{E}[W] = \frac{n\cdot(m+n+1)}{2} = \frac{15\cdot(15+15+1)}{2} = 232{,}5 \tag{8.15}$$

$$D[W] = \frac{n\cdot m\cdot(m+n+1)}{12} = \frac{15\cdot 15\cdot(15+15+1)}{12} = 581.25 \tag{8.16}$$

Тогда

$$W^* = \frac{W-\mathsf{E}[W]}{\sqrt{D[W]}} = \frac{(W-232.5)}{24.11} = 0.48$$

Реализация статистики попадает в доверительную область. Следовательно, не уровне значимости 0.05 нет оснований отвергать основную гипотезу.

Лекция 9

Проверка гипотезы об однородности против альтернатив растяжения/сжатия. Проверка гипотезы об однородности против альтернатив общего вида.

Критерий Фишера

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует распределению $\mathcal{N}(m_X;\sigma_X^2)$, а выборка $\mathbb{Y}_n$ распределению $\mathcal{N}(m_Y;\sigma_Y^2)$, причем параметры $m_X, m_Y, \sigma_X^2, \sigma_Y^2$ неизвестны;

2) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ независимы.

Критерий Фишера позволяет проверить гипотезу

$$H_0\colon \Delta = \frac{\sigma_Y}{\sigma_X} = 1. \tag{9.1}$$

Если последняя гипотеза верна, и при этом $m_X = m_Y = m$, то верна гипотеза H_0 вида (8.5). Мешающий параметр сдвига в этом случае совпадает с математическим ожиданием m.

Статистика критерия Фишера вычисляется следующим образом:

$$T(\mathbb{X}_m,\mathbb{Y}_n) = F_{n,m} = \frac{\dfrac{1}{n-1}\sum\limits_{i=1}^{n}(Y_i-\overline{Y})^2}{\dfrac{1}{m-1}\sum\limits_{i=1}^{m}(X_i-\overline{X})^2} = \frac{\widetilde{S}_Y^2}{\widetilde{S}_X^2}, \tag{9.2}$$

где $\widetilde{S}_X^2, \widetilde{S}_Y^2$ — несмещенные выборочные дисперсии, построенные по выборкам $\mathbb{X}_m$ и $\mathbb{Y}_n$ соответственно.

Если верна гипотеза H_0 вида (9.1), то статистика $F_{n,m}$ имеет распределение Фишера $F(n-1;m-1)$ с $n-1$ и $m-1$ степенями свободы.

В таблицах, как правило, представлены квантили распределения Фишера только высоких (не менее 0,5) уровней. Это обстоятельство связано с тем, что если случайная величина $T \sim F(n;m)$, то случайная величина $\dfrac{1}{T} \sim F(m;n)$. Можно показать, что число $\dfrac{1}{f_\beta(m;n)}$, где $f_\beta(m;n)$ — квантиль уровня β распределения $F(m;n)$, является квантилью уровня $1-\beta$ распределения $F(n;m)$.

При различных альтернативах процедуру проверки гипотезы H_0 вида (9.1) с помощью критерия Фишера целесообразно организовать следующим образом.

Если альтернативная гипотеза имеет вид

$$H_1\colon \frac{\sigma_Y}{\sigma_X} = \Delta < 1,$$

и $\widetilde{S}_Y^2 < \widetilde{S}_X^2$, то следует рассмотреть статистику $F_{m,n} = \dfrac{1}{F_{n,m}} = \dfrac{\widetilde{S}_X^2}{\widetilde{S}_Y^2}$.

Эта статистика при справедливости H_0 вида (9.1) имеет распределение $F(m-1;n-1)$, поэтому критическая область имеет вид: $(f_{1-\alpha}(m-1;n-1);+\infty)$, где α — уровень значимости критерия, а $f_{1-\alpha}(m-1;n-1)$ — квантиль уровня $(1-\alpha)$ распределения $F(m-1;n-1)$. Если же $\widetilde{S}_Y^2 > \widetilde{S}_X^2$, то принимается гипотеза H_0.

Если альтернативная гипотеза имеет вид

$$H_2\colon \frac{\sigma_Y}{\sigma_X} = \Delta > 1,$$

и $\widetilde{S}_Y^2 > \widetilde{S}_X^2$, то статистика $F_{m,n}$ имеет распределение $F(n-1;m-1)$, а критическая область критерия имеет вид: $(f_{1-\alpha}(n-1;m-1);+\infty)$, где $f_{1-\alpha}(n-1;m-1)$ — квантиль уровня $(1-\alpha)$ распределения $F(n-1;m-1)$. В случае когда $\widetilde{S}_Y^2 < \widetilde{S}_X^2$, принимается гипотеза H_0.

Если альтернативная гипотеза имеет вид

$$H_3\colon \frac{\sigma_Y}{\sigma_X} = \Delta \neq 1,$$

и $\widetilde{S}_Y^2 > \widetilde{S}_X^2$, то статистикой критерия Фишера будет $F_{m,n}$. Если же $\widetilde{S}_Y^2 < \widetilde{S}_X^2$, то статистикой критерия будет $F_{m,n} = \dfrac{1}{F_{n,m}}$. Соответствующие критические области будут иметь вид: $\left(f_{1-\frac{\alpha}{2}}(n-1;m-1);+\infty\right)$ в первом случае или $\left(f_{1-\frac{\alpha}{2}}(m-1;n-1);+\infty\right)$ во втором случае.

Квантили распределения Фишера табулированы, например, в книге Большев Л.Н., Смирнов Н.В. «Таблицы математической статистики» (М.: Наука, 1983)

Критерий Ансари–Брэдли

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует неизвестному непрерывному распределению $F(t-\mu)$, а выборка $\mathbb{Y}_n$ — распределению $F\left(\dfrac{t-\mu}{\Delta}\right)$, $\Delta > 0$, причем параметры μ и Δ неизвестны, и $F(\mu) = 0{,}5$;

2) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ независимы.

Замечание 9.1. Если случайная величина X, порождающая выборку $\mathbb{X}_m$, имеет распределение $F(t-\mu_1)$, а случайная величина Y, порождающая выборку $\mathbb{Y}_n$, имеет распределение $F\left(\dfrac{t-\mu_2}{\Delta}\right)$, и параметры μ_1 и μ_2 — неизвестны, то рекомендуется оценить параметры положения μ_1 и μ_2 выборочными медианами $\widehat{\mu}_X$ и $\widehat{\mu}_Y$. А затем построить преобразованные выборки $\widetilde{\mathbb{X}}_m = [X_1 - \widehat{\mu}_X, \ldots, X_m - \widehat{\mu}_X]^T$ и $\widetilde{\mathbb{Y}}_n = [Y_1 - \hat{\mu}_Y, \ldots, Y_n - \hat{\mu}_Y]^T$, и применить критерий Ансари–Брэдли к полученным выборкам $\widetilde{\mathbb{X}}_m$ и $\widetilde{\mathbb{Y}}_n$.

Гипотеза об однородности выборок имеет вид (8.5), т.е. $H_0\colon \Delta = 1$.

Статистика критерия Ансари–Брэдли имеет вид

$$A_{m,n} = \sum_{i=1}^{m} \left(\frac{N+1}{2} - \left| R_i - \frac{N+1}{2} \right| \right), \tag{9.3}$$

где R_i — ранг элемента X_i в объединенной выборке $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^T$ объема $N = m+n$.

Точное распределение статистики $A_{m,n}$ при справедливости H_0 табулировано для $2 \leqslant m \leqslant n$ при $m+n \leqslant 20$ (см., например, Холлендер М., Вульф Д. «Непараметрические методы статистики«.М.:Финансы и статистика,1983).

Известно, что при справедливости H_0 вида (8.5) и любых $m,n \geqslant 1$

$$\mathsf{E}\{A_{m,n}\} = \begin{cases} \dfrac{m(N+2)}{4}, & \text{если } N \text{ — четное,} \\ \dfrac{m(N+1)^2}{4N}, & \text{если } N \text{ — нечетное;} \end{cases}$$

$$\mathsf{D}\{A_{m,n}\} = \begin{cases} \dfrac{mn(N+2)(N-2)}{48(N-1)}, & \text{если } N \text{ — четное,} \\ \dfrac{mn(N^2+3)(N+1)}{48N^2}, & \text{если } N \text{ — нечетное;} \end{cases}$$

а стандартизованная статистика

$$A^*_{m,n} = \frac{A_{m,n} - \mathsf{E}\{A_{m,n}\}}{\sqrt{\mathsf{D}\{A_{m,n}\}}} \tag{9.4}$$

асимптотически нормальна при $N^* = \min(m,n) \to \infty$.

Если в выборке $\mathbb{Z}_N$ имеются связки, то дисперсию $\mathsf{D}\{A_{m,n}\}$ статистики $A_{m,n}$ следует заменить выражением

$$\widetilde{\mathsf{D}}\{A_{m,n}\} = \begin{cases} \dfrac{mn\left(16\sum\limits_{j=1}^{k} t_j R_j^2 - N(N+2)^2\right)}{16N(N-1)}, & \text{если } N \text{ — четное,} \\ \dfrac{mn\left(16N\sum\limits_{j=1}^{k} t_j R_j^2 - (N+1)^4\right)}{16N^2(N-1)}, & \text{если } N \text{ — нечетное,} \end{cases}$$

где k — количество связок в выборке $\mathbb{Z}_N$, t_j — размер j-й связки, R_j — средний ранг элементов j-й связки, $j = 1,\ldots,k$.

Отметим, что для применения критерия Ансари–Брэдли не требуется условия конечности дисперсии $\mathsf{D}\{X\} < \infty$. Критерий Ансари–Брэдли является свободным от распределения и состоятельным для альтернатив вида $H_1\colon \Delta < 1, H_2\colon \Delta > 1, H_3\colon \Delta \neq 1$.

Критические области критерия Ансари–Брэдли, основанного на статистике $A^*_{m,n}$ для этих альтернатив, указаны в табл. 9.2, где α — уровень значимости, u_γ — квантиль уровня γ стандартного гауссовского распределения.

H_A	**Критические области для** $A^*_{m,n}$
$\Delta < 1$	$(-\infty; u_\alpha)$
$\Delta > 1$	$(u_{1-\alpha}; +\infty)$
$\Delta \neq 1$	$(-\infty; u_{\frac{\alpha}{2}}) \cup (u_{1-\frac{\alpha}{2}}; +\infty)$

Таблица 9.1

Пример 9.1. Станок штампует детали, размер которых соответствует заданному нормативу, т.е. вероятность превышения и занижения нормативного размера одинакова. Технологи провели наладку станка для того, чтобы уменьшить отклонения размеров изготовленных деталей от размера, требуемого стандартом. До и после наладки случайным образом было выбрано по 11 деталей. Оказалось, что размер деталей, выбранных для наладки, составил (в мм):

52,4; 56,1; 48,6; 46,5; 46,0; 42,2; 48,8; 56,6; 59,8; 49,7; 51,6.

Размер деталей, изготовленных после наладки станка (в мм):

49,3; 47,7; 52,9; 48,3; 49,1; 46,4; 47,0; 52,0; 51,5; 51,2; 49,8.

Можно ли считать, опираясь на эти данные, что точность изготовления деталей увеличилась после наладки? Уровень значимости считать равным 0,05.

Р е ш е н и е: Пусть размеры деталей, проверенные до наладки станка, являются выборкой $\mathbb{X}_m = [X_1,\ldots,X_m]^T$ объема $m = 11$, порожденной непрерывной случайной величиной X, а размеры деталей, проверенные после наладки, выборкой $\mathbb{Y}_n = [Y_1,\ldots,Y_n]^T$ объема $n = 11$, порожденной непрерывной случайной величиной Y.

Поскольку размер деталей до и после наладки станка соответствует заданному нормативу, то это означает, что медианы случайных величин X и Y одинаковы $\mu_X = \mu_Y = \mu$, и параметр μ совпадает с нормативным размером. Так как наладка станка производится с целью уменьшения отклонения размеров изготовленных деталей от размера, требуемого

стандартом, то можно считать, что распределения случайных величин X и Y различаются лишь параметром масштаба Δ, т.е.

$$F_X(t) = F(t-\mu), \quad F_Y(t) = F\left(\frac{t-\mu}{\Delta}\right),$$

а $F(\mu) = 0{,}5$.

Тогда гипотеза $H_0\colon \Delta = 1$ будет означать, что выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ однородны, и наладка не привела к ожидаемому результату. В качестве альтернативной гипотезы в этой задаче следует выбрать $H_A\colon \Delta < 1$, так как справедливость этой альтернативы означает, что $\mathsf{D}\{X\} > \mathsf{D}\{Y\}$,т.е. точность изготовления деталей увеличилась.

Для проверки указанной гипотезы можно применить критерий Ансари–Брэдли со статистикой (9.3)

$$T(\mathbb{Z}_N) = A_{m,n} = \sum_{i=1}^{m}\left(\frac{N+1}{2} - \left|R_i - \frac{N+1}{2}\right|\right),$$

где R_i — ранг элемента X_i в объединенной выборке $\mathbb{Z}_N = [X_1,\ldots,X_m,Y_1,\ldots,Y_n]^T$ объема $N = m+n$.

Проранжировав реализацию объединенной выборки $\mathbb{Z}_N$, получим вектор искомых реализаций рангов

$$[r_1,\ldots,r_{11}]^T = [18,20,8,4,2,1,9,21,22,12,16]^T.$$

Реализация статистики

$$A_{11,11} = 53.$$

К сожалению, таблицы точного распределения статистики $A_{m,n}$ при справедливости H_0 составлены только для выборок объема $n+m \leqslant 20$, поэтому для построения критической области придется воспользоваться аппроксимацией.

При справедливости H_0 и $N^* = \min(m,n) \to \infty$, стандартизованная статистика $A^*_{m,n}$ вида (9.4) является асимптотически нормальной.

Так как $N = m+n = 22$ — четное число, то

$$\mathsf{E}\left\{A_{11,11}\right\} = \frac{m(N+2)}{4} = \frac{11\cdot 24}{4} = 66,$$

$$\mathsf{D}\left\{A_{11,11}\right\} = \frac{mn(N+2)(N-2)}{48(N-1)} = \frac{11\cdot 11\cdot 24\cdot 20}{48\cdot 21} = 57{,}62.$$

Следовательно, реализация стандартизованной статистики

$$A^*_{11,11} = \frac{53-66}{\sqrt{57{,}62}} = -1{,}71.$$

Критическая область $(-\infty;u_\alpha)$ для уровня значимости $\alpha = 0{,}05$ имеет вид $(-\infty;-1{,}65)$. Таким образом, реализация статистики попала в критическую область и гипотеза H_0 отвергается в пользу альтернативы H_A на уровне значимости 0,05 , т.е. точность изготовления деталей увеличилась после наладки станка. ■

Критерий Колмогорова—Смирнова

Пусть справедливы следующие предположения:

1) выборка $\mathbb{X}_m$ соответствует неизвестному непрерывному распределению $F_X(t)$, а выборка $\mathbb{Y}_n$ — непрерывному распределению $F_Y(t)$;

2) выборки $\mathbb{X}_m$ и $\mathbb{Y}_n$ независимы.

Критерий Колмогорова–Смирнова позволяет проверить гипотезу об однородности вида (8.1) против альтернативной гипотезы общего вида (8.2).

Статистика Колмогорова–Смирнова имеет вид:

$$D_{m,n} = \sup_{t\in\mathbb{R}^1} |\widehat{F}_{X,m}(t) - \widehat{F}_{Y,n}(t)|,$$

где $\widehat{F}_{X,m}(t)$ и $\widehat{F}_{Y,n}(t)$ — выборочные функции распределения, построенные по выборкам $\mathbb{X}_m$ и $\mathbb{Y}_n$ соответственно.

Так как выборочная функция распределения монотонна и изменяется в конечном числе точек, то

$$D_{m,n} = \max_{1 \leqslant i \leqslant m+n} \left| \widehat{F}_{X,m}(Z_i) - \widehat{F}_{Y,n}(Z_i) \right|, \tag{9.5}$$

где $\mathbb{Z}_N = [X_1, \ldots, X_m, Y_1, \ldots, Y_n]^T$ — объединенная выборка объема $N = m + n$.

Точное распределение статистики $D_{m,n}$ при справедливости гипотезы H_0 вида (8.1) табулировано, например, в книге Большев Л.Н., Смирнов Н.В. «Таблицы математической статистики» (М.: Наука, 1983) для $2 \leqslant m \leqslant n \leqslant 20$.

Если $N^* = \min\{m,n\} \to \infty$, то при справедливости H_0 статистика

$$D^*_{m,n} = \sqrt{\frac{nm}{n+m}} D_{m,n} \tag{9.6}$$

асимптотически имеет распределение Колмогорова с функцией распределения

$$K(t) = \sum_{k=-\infty}^{\infty} (-1)^k \exp\{-2k^2t^2\}.$$

Квантили распределения $K(t)$ также табулированы,например, в книге Большев Л.Н., Смирнов Н.В. «Таблицы математической статистики» (М.: Наука, 1983).

Критическая область уровня значимости α критерия Колмогорова–Смирнова, основанного на статистике (9.6), имеет вид: $(K_{1-\alpha}; +\infty)$, где $K_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения Колмогорова.

Пример 9.2. Известно, что одним из факторов риска сердечно-сосудистых заболеваний является склад психоэмоциональной сферы человека. Медики выделяют две основных модели поведения людей. Модель типа A характеризуется постоянным острым дефицитом времени и склонностью к соперничеству, модель типа B — спокойствием и размеренностью. Склонность к сердечно-сосудистым заболеваниям характерна для людей с моделью поведения типа A. Высказано предположение о том, что различие в типах поведения индивидуумов обусловлено их физиологическими различиями. Чтобы проверить это предположение , исследователи сравнили максимальные уровни концентрации гормонов роста в плазме крови у испытуемых различных типов поведения. Получены следующие результаты (в мг/мл). Испытуемые с моделью поведения типа A:

3,6; 2,6; 4,7; 8,0; 3,1; 8,8; 4,6; 5,8; 4,0; 4,6.

Испытуемые с моделью поведения типа B:

14,9; 16,6; 15,9; 5,3; 10,5; 16,2; 17,4; 8,5; 15,6; 5,4; 9,8.

Можно ли, опираясь на эти результаты исследования, считать предположение верным?

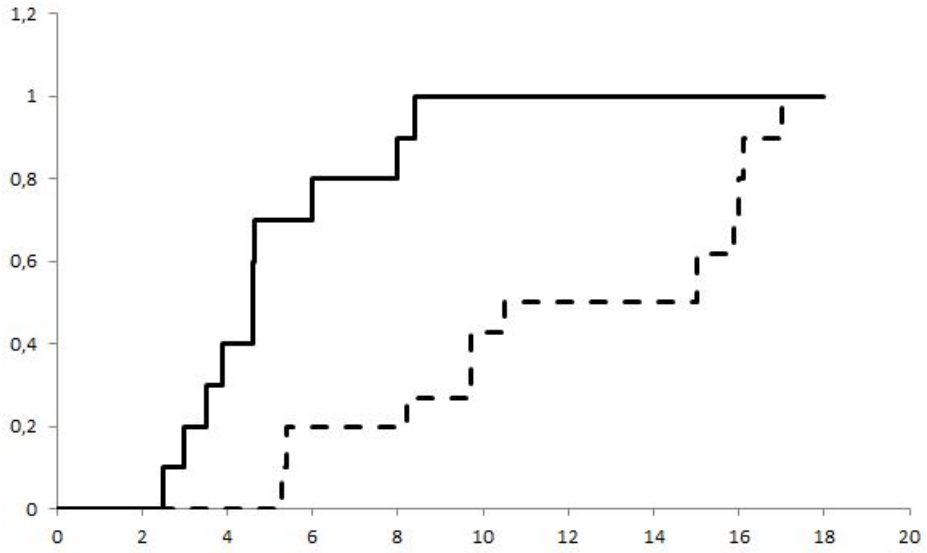
Р е ш е н и е: Пусть результаты измерений по группе A с поведением типа представляются выборкой $\mathbb{X}_m = [X_1, \ldots, X_m]^T$ объема $m = 10$, соответствующей непрерывному распределению $F_X(t)$, а результаты измерений по группе с поведением типа B — выборкой $\mathbb{Y}_n = [Y_1, \ldots, Y_n]^T$ объема $n = 11$, соответствующей непрерывному распределению $F_Y(t)$.

Проверим гипотезу об однородности выборок $\mathbb{X}_m$ и $\mathbb{Y}_n$. Так как медики не дают априорной информации о типе неоднородности, следует проверить гипотезу H_0 вида (8.1) против альтернативной гипотезы H_A общего вида (8.2). Для решения данной задачи можно применить критерий Колмогорова–Смирнова со статистикой (9.5)

$$D_{m,n} = \max_{1 \leqslant i \leqslant m+n} \left| \widehat{F}_{X,m}(Z_i) - \widehat{F}_{Y,n}(Z_i) \right|,$$

где $[Z_1,\ldots,Z_{m+n}]^T = [X_1,\ldots,X_m,Y_1,\ldots,Y_n]^T$ — объединенная выборка. Используя графический метод (см. рис.) или простой перебор, можно видеть, что максимальное расхождение между реализациями выборочных функций распределения $\widehat{F}_{x,10}(t)$ и $\widehat{F}_{y,11}(t)$ достигается в точке $t = 8{,}8$, и реализация статистики

$$D_{10,11} = \left|\widehat{F}_{x,10}(8{,}8) - \widehat{F}_{y,11}(8{,}8)\right| = 1 - \frac{3}{11} = \frac{8}{11}.$$



$-\ \widehat{F}_{x,10}(t); - - - \widehat{F}_{y,11}(t)$

Рис 9.1

Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид $(z_{0,95};1)$, где $z_{0,95} = \frac{6}{11}$ — квантиль уровня 0,95 распределения статистики $D_{10,11}$ при справедливости гипотезы H_0. Так как реализация статистики $D_{10,11}$ попадает в критическую область, то гипотеза H_0 отвергается в пользу альтернативы H_A на уровне значимости 0,05.

Если считать, что m и n достаточно велики, и использовать статистику (9.6) вида

$$D^*_{m,n} = \sqrt{\frac{nm}{n+m}} D_{m,n},$$

то критическая область будет иметь вид $(K_{0,95};\infty)$, где $K_{0,95}$ — квантиль уровня 0,95 распределения Колмогорова $K(t)$. Согласно таблицам, $K_{0,95} = 1{,}36$. Реализация статистики

$$D^*_{m,n} = \sqrt{\frac{10\cdot 11}{10+11}} \cdot \frac{8}{11} \approx 1{,}66$$

так же попадает в критическую область.

Таким образом, гипотеза H_0 об однородности вида (8.1) отвергается на уровне значимости $\alpha = 0{,}05$. Следовательно, можно считать, что различие типов поведения людей обусловлено их физиологическими различиями.

Лекция 10

Однофакторный дисперсионный анализ

В двух предыдущих лекциях были изучены критерии, предназначенные для выявления однородности (неоднородности) двух выборок, которые являлись измерениями однотипных показателей, полученных в результате различных «обработок». В частности, в одном из примеров лекции №8 мы рассмотрели показатели среднедушевых доходов населения в двух федеральных округах и, используя различные статистические критерии, показали, что среднедушевые доходы населения в Центральном и Приволжском округах можно считать в среднем одинаковыми. Это означает, что способ «обработки», под которым можно понимать экономические условия конкретного региона, не оказывает влияния на измеряемый показатель(среднедушевые доходы). Обобщим теперь эту задачу. Пусть имеется три и более выборок, соответствующих различным обработкам (округам). Требуется выяснить, равны ли средние значения измеряемого показателя для всех обработок. Задачу выявления однородности(или неоднородности) трех или большего числа выборок, которые могут различаться сдвигом, называют задачей дисперсионного анализа. В этой лекции будут рассмотрены классические и непараметрические ранговые критерии для решения задачи однофакторного дисперсионного анализа.

10.1 Теоретические положения

Пусть имеется k независимых выборок

$$Z_1 = [X_{11}, X_{21}, \ldots, X_{n_1 1}]^T, \ldots, Z_k = [X_{1k}, X_{2k}, \ldots, X_{n_k k}]^T,$$

порожденных случайными величинами $X_1, \ldots, X_k$ с распределениями $F(t-\theta_1), \ldots, F(t-\theta_k)$ соответственно. Требуется проверить гипотезу

$$H_0: \quad \theta_1 = \ldots = \theta_k = \theta \tag{10.1}$$

против альтернативы

$$H_A: \quad \exists i, j, \text{ такие что } \theta_i \neq \theta_j, i \neq j. \tag{10.2}$$

Справедливость гипотезы H_0 означает, что выборки $Z_1, \ldots, Z_k$ однородны, и объединенная выборка

$$\mathbb{Z}_N = [Z_1^T, \ldots, Z_n^T]^T = [X_{11}, \ldots, X_{n_1 1}, \ldots X_{1k}, \ldots, X_{n_k k}]$$

объема $N = n_1 + \ldots + n_k$ является однородной выборкой соответствующей распределению $F(t-\theta)$. Если же гипотеза H_0 нарушается, то это означает, что среди k рассматриваемых выборок найдутся выборки, распределения которых различаются сдвигом. Предполагается, что этот сдвиг вызван воздействием (влиянием) одной или нескольких переменных. Такие переменные называют *факторами*. Если предполагается наличие только одного фактора, то задача проверки гипотезы (10.1) называется *задачей однофакторного дисперсионного анализа*. При описании задач однофакторного анализа принято использовать следующие термины:

- *уровень фактора* (или способ обработки) — конкретная реализация фактора;

- *отклик* — значение измеряемой случайная величина.

Фактор может быть как количественной, так и качественной переменной. Однако при решении задачи однофакторного дисперсионного анализа должно быть выбрано конечное число k различных уровней фактора, при этом реализации $[x_{1j},\ldots,x_{n_j j}]^T$ выборок Z_j, $j=1,\ldots,k$, должны быть откликами, соответствующими j-му уровню фактора.

Отметим, что описанная задача проверки гипотезы $_0$ вида (10.1) является обобщением задачи проверки гипотезы об однородности двух выборок против альтернативы сдвига на случай $k>2$ выборок.

Таблица 10.1, в которой в первой строке записаны уровни факторов, а x_{ij}, $j=1,\ldots,k$, $i=1,\ldots,n_j$ есть реализации случайных величин X_{ij}, называется таблицей с одним входом или таблицей однофакторного анализа.

Таблица 10.1

1	2	$\cdots$	k
x_{11}	x_{22}	$\cdots$	x_{1k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$x_{n_1 1}$	$x_{n_2 2}$	$\cdots$	$x_{n_k k}$

Рассмотрим классический F-критерий и ранговые критерии Краскела—Уоллиса и Джокхиера для проверки гипотезы вида (10.1) .

10.2 Классический *F-критерий*

Для представления классического критерия удобно описать рассматриваемую задачу с помощью следующей статистической модели:

$$X_{ij}=\theta+\tau_j+\varepsilon_{ij},\quad j=1,\ldots,k,\quad i=1,\ldots,n_j, \tag{10.3}$$

где

θ — неизвестное математическое ожидание;

τ_j — неизвестные отклонения от общего среднего θ, вызванные изменениями уровня фактора (эффект j-й обработки), причем $\sum\limits_{j=1}^{k}\tau_j=0$;

ε_{ij} — независимые ненаблюдаемые погрешности соответствующие нормальному распределению $N(0;\sigma^2)$ с неизвестной дисперсией σ^2.

В рамках такой модели параметры $\theta+\tau_j$, $j=1,\ldots,k$ совпадают с параметрами θ_j, определёнными выше , и гипотеза (10.1) будет иметь вид

$$H_0:\tau_1=\tau_2=\cdots=\tau_k=0, \tag{10.4}$$

а альтернативная гипотеза (10.2) —

$$H_A:\exists j\quad\text{такое, что}\quad \tau_j\neq 0. \tag{10.5}$$

Статистика F-критерия для проверки гипотезы (10.4) имеет вид

$$T(\mathbb{Z}_N)=\frac{\dfrac{1}{k-1}\sum\limits_{j=1}^{k}n_j(\overline{X}_{\bullet j}-\overline{X}_N)^2}{\dfrac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij}-\overline{X}_{\bullet j})^2}, \tag{10.6}$$

где $\overline{X}_{\bullet j}=\dfrac{1}{n_j}\sum\limits_{i=1}^{n_j}X_{ij}$ — выборочное среднее, построенное по выборке $Z_j, j=1,\ldots,k$, а $\overline{X}_N=\dfrac{1}{N}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}X_{ij}$ — выборочное среднее, построенное по объединенной выборке $\mathbb{Z}_N=[Z_1^T,\ldots,Z_k^T]^T$ объема $N=n_1+\cdots+n_k$.

Поясним принцип построения такой статистики.

Рассмотрим сумму квадратов отклонений наблюдений от выборочного среднего $SS_{\text{общ.}}$:

$$SS_{\text{общ.}} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_N)^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_N \pm \bar{X}_{\bullet j})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\bullet j})^2 +$$

$$+\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_N-\bar{X}_{\bullet j})^2 + 2\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\bullet j})(\bar{X}_N-\bar{X}_{\bullet j})$$

Нетрудно показать, что последнее слагаемое равно 0.

Тогда

$$SS_{\text{общ.}} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\bullet j})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_N-\bar{X}_{\bullet j})^2 =$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\bullet j})^2 + \sum_{j=1}^{k} n_j(\bar{X}_N-\bar{X}_{\bullet j})^2.$$

Заметим, что первое слагаемое «отвечает» за разброс наблюдений внутри столбцов, второе – за разброс между столбцами. То есть, вариация первого слагаемого обусловлена вариацией случайной составляющей, второго слагаемого - изменением уровня фактора. Поэтому слагаемые принято обозначать:

$$SS_{\text{случ.}} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_{\bullet j})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}X_{ij}^2 - \sum_{j=1}^{k} n_j(\bar{X}_{\bullet j})^2$$

и

$$SS_{\text{ур. факт.}} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{X}_N-\bar{X}_{\bullet j})^2 = \sum_{j=1}^{k} n_j(\bar{X}_N-\bar{X}_{\bullet j})^2 = \sum_{j=1}^{k} n_j(\bar{X}_{\bullet j})^2 - N\bar{X}_N{}^2.$$

Таким образом, $SS_{\text{общ.}} = SS_{\text{случ.}} + SS_{\text{ур. факт.}}$.

Найдем распределения этих величин при справедливости гипотезы H_0.

$\frac{SS_{\text{общ.}}}{\sigma^2} \sim \chi^2(N-1)$, где $\mathrm{N} = \mathrm{n}_1 + \ldots + n_k$ – общее число наблюдений;

$$\frac{SS_{\text{случ.}}}{\sigma^2} \sim \chi^2(N-k).$$

В том случае, когда справедлива гипотеза H_0, $\frac{SS_{\text{ур. факт.}}}{\sigma^2} \sim \chi^2(k-1)$.

Таким образом, при справедливости H_0, имеем:

$$T(\mathbb{Z}_N) = \frac{\frac{1}{k-1}\cdot\frac{SS_{\text{ур. факт.}}}{\sigma^2}}{\frac{SS_{\text{случ.}}}{\sigma^2}\cdot\frac{1}{N-k}} = \frac{\frac{SS_{\text{ур. факт.}}}{(k-1)}}{\frac{SS_{\text{случ.}}}{(N-k)}}\Big|_{H_0} \sim \mathbf{F}(k-1,N-k).$$

При больших значениях отклонений τ_j величина $SS_{\text{ур. факт.}}$, которая стоит в числителе статистики (10.6), будет принимать большие значения. Следовательно, и статистика (10.6) при нарушении H_0 будет принимать большие значения. Таким образом, критическая область уровня значимости α для F-критерия будет иметь вид:

$$(f_{1-\alpha}(k-1;N-k);\infty),$$

где $f_{1-\alpha}(k-1;N-k)$ – квантиль уровня $1-\alpha$ распределения Фишера $F(k-1;N-k)$.

Если гипотеза H_0 вида (10.4) принимается, то выборки $Z_1,\ldots,Z_k$, полученные при различных значениях уровня фактора, однородны, и, следовательно, можно считать, что фактор не оказывает влияния на отклик. В этом случае задача исследования влияния фактора на отклик завершена. Если же гипотеза H_0 отвергнута, то это свидетельствует о том, что фактор оказывает влияние на отклик. В связи с этим возникает задача оценивания и сравнения средних значений случайных величин $X_1,\ldots,X_k$, порождающих выборки $Z_1,\ldots,Z_k$. Перейдём к решению этой задачи.

10.3 Доверительное оценивание параметров сдвига и контрастов

Обозначим

$$\theta_j = \theta + \tau_j, \quad j = 1,\ldots,k$$

математическое ожидание случайная величина X_j, порождающей выборку $Z_j = [X_{1j},\ldots,X_{n_j j}]$.

Тогда модель (10.3) будет иметь вид

$$X_{ij} = \theta_j + \varepsilon_{ij}, \quad j = 1,\ldots,k, \quad i = 1,\ldots,n_j. \tag{10.7}$$

Предположим, что $\varepsilon_{ij}, j = 1,\ldots,k, i = 1,\ldots,n_j$ имеют распределение $\mathbb{N}(0;\sigma^2)$ с неизвестной дисперсией σ^2.

Построим доверительные интервалы для параметров θ_j, $j = 1,\ldots,k$. Можно показать (см., например, пример 8.2 в Горяинова Е.Р., Панков А.Р.,Платонов Е.А. «Прикладные методы анализа статистических данных».- М.:Изд. дом ВШЭ,2012), что статистика

$$G(\mathbb{Z}_N,\theta_j) = \frac{\sqrt{n_j}(\overline{X}_{\bullet j} - \theta_j)}{\sqrt{\frac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2}}, \quad j = 1,\ldots,k \tag{10.8}$$

является центральной статистикой для θ_j и имеет распределение Стьюдента с $N-k$ степенями свободы, где $N = \sum\limits_{j=1}^{k} n_j$.

Тогда центральный доверительный интервал параметра θ_j, $j = 1,\ldots,k$ надежности $1-\alpha$ имеет вид:

$$P\left(\overline{X}_{\bullet j} - \frac{1}{\sqrt{n_j}}\sqrt{\widetilde{S}_N^2}\,t_{1-\frac{\alpha}{2},N-k} < \theta_j < \overline{X}_{\bullet j} + \frac{1}{\sqrt{n_j}}\sqrt{\widetilde{S}_N^2}\,t_{1-\frac{\alpha}{2},N-k}\right) = 1-\alpha, \tag{10.9}$$

где

$$\widetilde{S}_N^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2,$$

а $t_{1-\frac{\alpha}{2},N-k}$ — квантиль уровня $1-\frac{\alpha}{2}$ распределения Стьюдента с $N-k$ степенями свободы.

На практике при проведении сравнительного анализа бывает важно строить доверительные интервалы не только для средних значений θ_j, но и для разностей средних значений.

Определение 10.1. *Контрастом параметров* $\theta_j, j = 1,\ldots,k$ в модели (10.7) называется величина $\gamma = \sum\limits_{j=1}^{k} c_j\theta_j$, где c_j — константы, удовлетворяющие условию $\sum\limits_{j=1}^{k} c_j = 0$.

Например, если положить значения $c_l = 1$, $c_m = -1$ и $c_j = 0$ для $j \neq l$ и $j \neq m$, то контраст $\gamma = \theta_l - \theta_m$ представляет разность средних значений откликов, соответствующих l-му и m-му отклику уровням фактора.

Несмещенной оценкой контраста γ является статистика

$$\widehat{\gamma} = \sum_{j=1}^{k} c_j\widehat{\theta}_j = \sum_{j=1}^{k} c_j\overline{X}_{\bullet j} \tag{10.10}$$

Можно показать, что центральная статистика для γ имеет вид

$$G(\mathbb{Z}_N,\gamma) = \frac{\sum\limits_{j=1}^{k} c_j\overline{X}_{\bullet j} - \gamma}{\sqrt{\sum\limits_{j=1}^{k}\frac{c_j^2}{n_j}\left(\frac{1}{N-k}\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2\right)}} \tag{10.11}$$

и $G(\mathbb{Z}_N,\gamma)$ имеет распределение Стьюдента с $N-k$ степенями свободы.

Тогда, обозначая

$$\widetilde{S}_N^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\overline{X}_{\bullet j}),$$

получим

$$\mathsf{P}\Big(-t_{1-\frac{\alpha}{2},N-k} < G(\mathbb{Z}_N,\gamma) < t_{1-\frac{\alpha}{2},N-k}\Big) = 1-\alpha,$$

$$\mathsf{P}\left(\sum_{j=1}^{k} c_j\overline{X}_{\bullet j} - t_{1-\frac{\alpha}{2},N-k}\sqrt{\widetilde{S}_N^2\sum_{j=1}^{k}\frac{c_j^2}{n_j}} < \gamma < \sum_{j=1}^{k} c_j\overline{X}_{\bullet j} + t_{1-\frac{\alpha}{2},N-k}\sqrt{\widetilde{S}_N^2\sum_{j=1}^{k}\frac{c_j^2}{n_j}}\right) = 1-\alpha,$$

и центральный доверительный интервал контраста γ уровня надежности $1-\alpha$ имеет вид

$$\left(\sum_{j=1}^{k} c_j\overline{X}_{\bullet j} - t_{1-\frac{\alpha}{2},N-k}\sqrt{\widetilde{S}_N^2\sum_{j=1}^{k}\frac{c_j^2}{n_j}};\quad \sum_{j=1}^{k} c_j\overline{X}_{\bullet j} + t_{1-\frac{\alpha}{2},N-k}\sqrt{\widetilde{S}_N^2\sum_{j=1}^{k}\frac{c_j^2}{n_j}}\right). \tag{10.12}$$

10.4 Критерий Краскела—Уоллиса

Пусть имеется k независимых выборок

$$Z_1 = [X_{11},X_{21},\ldots,X_{n_1 1}]^T,\ldots,Z_k = [X_{1k},X_{2k},\ldots,X_{n_k k}]^T,$$

порожденных случайными величинами $X_1,\ldots,X_k$ с распределениями $F(t-\theta_1),\ldots,F(t-\theta_k)$ соответственно.

Пусть справедливы следующие предположения:

1) $F(t)$ — непрерывная функция распределения;

2)случайные величины $X_1,\ldots,X_k$, порождающие выборки $Z_1,\ldots,Z_k$ независимы.

Требуется проверить гипотезу

$$H_0: \quad \theta_1 = \ldots = \theta_k = \theta \tag{10.13}$$

против альтернативы

$$H_A: \quad \exists i,j, \text{ такие что } \theta_i \neq \theta_j, i \neq j. \tag{10.14}$$

Обозначим R_{ij} — ранг X_{ij} в объединенной выборке $\mathbb{Z}_N = [X_{11},\ldots,X_{n_1 1},\ldots,X_{1k},\ldots,X_{n_k k}]^T$, а $\overline{R}_{\bullet j} = \frac{1}{n_j}\sum\limits_{i=1}^{n_j} R_{ij}$ — средний ранг элементов выборки, соответствующий $j-$му уровню фактора, $j = 1,\ldots,k$.

Статистика критерия Краскела—Уоллиса имеет вид

$$T(\mathbb{Z}_N) = H = \frac{12}{N(N+1)}\sum_{j=1}^{k} n_j\left(\overline{R}_{\bullet j} - \frac{N+1}{2}\right)^2. \tag{10.15}$$

Для удобства вычислений можно использовать другую форму этой статистики:

$$H = \frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{1}{n_j}\left(\sum_{i=1}^{n_j} R_{ij}\right)^2 - 3(N+1). \tag{10.16}$$

Если в выборке $\mathbb{Z}_N$ имеются связки, то рекомендуется использовать модифицированную форму статистики H вида

$$H' = \frac{H}{(1-\frac{1}{(N^3-N)}\sum\limits_{i=1}^{g}(t_i^3-t_i))}, \tag{10.17}$$

где g — количество связок, а t_i — размер i-й связки.

Заметим, что для вычисления статистики (10.15) не обязательно знать количественные реализации откликов, достаточно иметь их совместную ранжировку. Поэтому табл. 10.1 однофакторного анализа заменим на табл. 10.2.

Таблица 10.2

1	2	$\cdots$	k
r_{11}	r_{22}	$\cdots$	r_{1k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$r_{n_1 1}$	$r_{n_2 2}$	$\cdots$	$r_{n_k k}$

Точные квантили статистики (10.15) при справедливости гипотезы H_0 вида (10.1) представлены, например, в Ликеш И.,Ляга Й «Основные таблицы математической статистики« для следующих значений

$$k = 3, 2 \leqslant n_1 \leqslant n_2 \leqslant n_3 \leqslant 8;$$

$$k = 4, 2 \leqslant n_1 \leqslant \ldots \leqslant n_4 \leqslant 4;$$

$$k = 5, 2 \leqslant n_1 \leqslant \ldots \leqslant n_5 \leqslant 3.$$

Статистика (10.15) при справедливости гипотезы H_0 вида (10.1) и $min\{n_1,\ldots,n_k\} \to \infty$ имеет распределение хи-квадрат с $r = k-1$ степенями свободы .

При нарушении гипотезы (10.1) расхождения между средним рангом $(N+1)/2$ объединенной выборки $\mathbb{Z}_N$ и средними рангами $\overline{R}_{\bullet j}$, $j = 1,\ldots,k$ столбцов, соответствующих j-м уровням фактора, будет большим. Поэтому статистика (10.15) в случае справедливости альтернативы (10.2) будет принимать большие значения, а критическая область уровня значимости α будет иметь вид $(\chi_{1-\alpha}(k-1);+\infty)$, где $\chi_{1-\alpha}(k-1)$-квантиль уровня $1-\alpha$ распределения хи-квадрат с $k-1$ степенями свободы.

10.5 Критерий Джонкхиера

Пусть справедливо следующее предположение:

$F(t)$ — непрерывная функция распределения.

Критерий Джонкхиера позволяет проверить гипотезу H_0 вида (10.1) против альтернативы

$$H_A : \theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_k, \tag{10.18}$$

где хотя бы одно из неравенств строгое.

Альтернативы такого вида принято называть упорядоченными. Упорядоченные альтернативы описывают ситуацию, при которой увеличение уровня фактора вызывает увеличение сдвига распределения соответствующей этому уровню выборки относительно распределения первой выборки.

Если имеется априорное предложение о том, что с увеличением уровня фактора средние значения случайных величин $X_1,\ldots,X_k$ уменьшаются, то следует перенумеровать столбцы табл. 10.2 в обратном порядке.

Введем обозначения

$$\varphi(y,z) = \begin{cases} 1, & \text{если } y < z, \\ 0{,}5, & \text{если } y = z, \\ 0, & \text{если } y > z; \end{cases}$$

$$U_{lm} = \sum_{i=1}^{n_l} \sum_{j=1}^{n_m} \varphi(X_{il},X_{jm}).$$

Статистика критерия Джонкхиера имеет вид

$$T(\mathbb{Z}_N) = J = \sum_{1 \leqslant l < m \leqslant k} U_{l,m}. \tag{10.19}$$

Заметим, что

$$U_{lm} = \sum_{i=1}^{n_l} \sum_{j=1}^{n_m} \varphi(X_{il},X_{jm}) = \sum_{i=1}^{n_l} \sum_{j=1}^{n_m} \varphi(R_{il},R_{jm}),$$

т.е. при вычислении реализации статистики Джонкхиера (10.19) можно использовать либо реализации выборок $z_1,\ldots,z_k$ либо реализации рангов объединенной выборки $\mathbb{Z}_N$.

Можно показать, что при справедливости гипотезы H_0 вида (10.1)

$$\mathsf{E}\{J\} = \frac{1}{4}\left[N^2 - \sum_{j=1}^{k} n_j^2\right]; \qquad \mathsf{D}\{J\} = \frac{1}{72}\left[N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right],$$

а стандартизированная статистика

$$J^* = \frac{J - \mathsf{E}\{J\}}{\sqrt{D\{J\}}}$$

при $\min(n_1,\ldots,n_k) \to \infty$ асимптотически нормальна.

Точные квантили распределения статистики J представлены,например, в книге Холлендер М., Вульф Д. «Непараметрические методы статистики« для следующих значений

$$k = 3, \quad 2 \leqslant n_1 \leqslant n_2 \leqslant n_3 \leqslant 8;$$

$$k = 4{,}5{,}6, \quad 2 \leqslant n_1 = \ldots = n_k \leqslant 6.$$

Критическая область уровня значимости α критерия Джонкхиера, основанного на статистике J^*, и соответствующая альтернативе (10.18), имеет вид $(u_{1-\alpha};+\infty)$, где $u_{1-\alpha}$ – квантиль уровня $1-\alpha$ распределения $\mathscr{N}(0;1)$.

Отметим, что при $k = 2$ статистика

$$J = W_{n_1,n_2} - \frac{n_2(n_2+1)}{2},$$

где W_{n_1,n_2} – статистика критерия Вилкоксона (8.7), вычисленная для выборок Z_1 и Z_2. Таким образом, статистика Джонкхиера (10.19) с точностью до известной константы представляется в виде суммы $\dfrac{k(k-1)}{2}$ статистик W_{n_1,n_m} , вычисленных для всех возможных $C_k^2 = \dfrac{k(k-1)}{2}$ пар выборок Z_l и Z_m, $1 \leqslant l < m \leqslant k$.

Для состоятельности критерия Джонкхиера против альтернатив вида (10.18) достаточно, чтобы $n_j \to \infty$ так, что

$$\frac{n_j}{N} \to \lambda_j, 0 < \lambda_j < 1, j = 1,\ldots,k.$$

Пример 10.1. Имеются данные (в руб.) Федеральной службы государственной статистики о среднедушевых месячных денежных

Таблица 10.3

Центральный федеральный округ	**Приволжский федеральный округ**	**Дальневосточный федеральный округ**
Брянская область 10043	Кировская область 10112	Камчатский край 19063
Владимирская область 9596	Республика Марий Эл 7843	Магаданская область 19703
Воронежская область 10305	Удмуртская республика 9581	Сахалинская область 24552
Ивановская область 8354	Пермский край 16119	Хабаровский край 15705
Костромская область 9413	Чувашская республика 8594	Еврейская автономная область 10877
Московская область 19776	Республика Башкортостан 14253	Чукотский автономный округ 32140
Орловская область 9815	Пензенская область 10173	
Рязанкская область 11311	Ульяновская область 9756	
Тамбовская область 11253		
Тверская область 10856		
Тульская область 11389		

доходах населения в 2008г. по некоторым областям Центрального, Приволжского и Дальновосточного федеральных округов. Данные представлены в табл.10.3. Можно ли считать, что средние значения среднедушевых месячных доходов населения одинаковы во всех трех округах?

Р е ш е н и е: Фактором, т.е. переменной, которая может оказывать влияние на измеряемую величину (среднедушевой доход), является федеральный округ. Фактор в данном случае имеет три уровня: 1 — «Центральный федеральный округ», 2 — «Приволжский федеральный округ», 3 — «Дальневосточный федеральный округ».

Имеющиеся данные представляются тремя выборками $Z_1 = [X_{11},\ldots,X_{n_1 1}]^T$ объема $n_1 = 11$, $Z_2 = [X_{12},\ldots,X_{n_2 2}]^T$ объема $n_2 = 8$ и $Z_3 = [X_{13},\ldots,X_{n_3 3}]$ объема $n_3 = 6$, которые соответствую непрерывным распределениям $F(t-\theta_1)$, $F(t-\theta_2)$ и $F(t-\theta_3)$.

Проверим гипотезу об однородности

$$H_0 : \theta_1 = \theta_2 = \theta_3$$

против альтернативы

$$H_1 : \exists \theta_i \neq \theta_j \quad \text{при } i \neq j.$$

Для проверки этой гипотезы можно применить критерий Краскела—Уоллиса. Статистика критерия имеет вид (10.15)

$$T(\mathbb{Z}_N) = H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j (\overline{R}_{\bullet j} - \frac{N+1}{2})^2.$$

Для вычисления реализации статистики составим табл. 10.4 реализаций рангов случайных величин X_{ij}, $j = 1,2,3$, $i = 1,..,n_j$.

Таблица 10.4

Уровни фактора	Реализация рангов										
1	9	6	12	2	4	23	8	16	15	13	17
2	18	1	5	3	20	10	11	7			
3	21	19	22	24	14	25					

Следовательно,

$$\overline{R}_{\bullet 1} = \frac{1}{11}\sum_{i=1}^{11} R_{i1} = 11{,}36, \quad \overline{R}_{\bullet 2} = \frac{1}{8}\sum_{i=1}^{8} R_{i2} = 9{,}38, \quad \overline{R}_{\bullet 3} = \frac{1}{6}\sum_{i=1}^{6} R_{i3} = 20{,}83$$

и реализация статистики критерия

$$T(z_n) = \frac{12}{25\cdot 26}(11(11{,}36-13)^2 + 8(9{,}38-13)^2 + 6(20{,}83-13)^2) \approx 9{,}281.$$

При справедливости гипотезы H_0 статистика (10.15) критерия Краскела—Уоллиса имеет распределение хи-квадрат с $r = k-1 = 2$ степенями свободы.

Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид

$$(\chi_{0{,}95}(2); +\infty) = (5{,}99; +\infty).$$

Реализация статистики попадает в критическую область, следовательно, гипотеза H_0 отвергается на уровне значимости 0,05. Таким образом, нельзя считать, что среднее значение показателей среднедушевых месячных доходов населения одинаковы во всех трех округах.

Если предположить, что рассматриваемые выборки соответствуют гауссовскому распределению, то имеющиеся данные можно описать моделью (10.3) и проверить гипотезу H_0 об однородности вида (10.4) против альтернативы H_A вида (10.5).

Для проверки гипотезы (10.4) применим $F-$критерий со статистикой (10.6):

$$T(\mathbb{Z}_N) = F = \frac{\dfrac{1}{k-1}\displaystyle\sum_{j=1}^{k} n_j(\overline{X}_{\bullet j} - \overline{X}_N)^2}{\dfrac{1}{N-k}\displaystyle\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2}.$$

Вычислим сначала реализации выборочных средних $\overline{X}_{\bullet 1}, \overline{X}_{\bullet 2}, \overline{X}_{\bullet 3}$ для выборок Z_1, Z_2, Z_3 и выборочного среднего $\overline{X}_N$ объединенной выборки $\mathbb{Z}_N$.

Имеем

$$\overline{X}_{\bullet 1} = 11101{,}0, \qquad \overline{X}_{\bullet 2} = 10803{,}9, \qquad \overline{X}_{\bullet 3} = 20340{,}0, \qquad \overline{X}_N = 13223{,}3.$$

Тогда

$$\frac{1}{k-1}SS_{\text{ур.факт.}} = \frac{1}{2}\sum_{j=1}^{3} n_j(\overline{X}_{\bullet j} - \overline{X}_N)^2 \approx 200129591$$

$$\frac{1}{N-k}SS_{\text{случ.}} = \frac{1}{22}\sum_{j=1}^{3}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2 = 19038613$$

и

$$T(z_N) = \frac{200129591}{19038613} \approx 10{,}51.$$

При справедливости гипотезы H_0 вида (10.4) статистика $T(\mathbb{Z}_N)$ имеет $F-$распределение $F(2;22)$.

Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид $(f_{0{,}95}(2;22); +\infty)$, где $f_{0{,}95}(2;22)$ – квантиль уровня 0,95 распределения $F(2;22)$. По таблице находим $f_{0{,}95}(2;22) = 3{,}44$.

Таким образом, реализация статистики $F-$критерия попадает в критическую область и гипотеза об однородности H_0 вида (10.4) отвергается на уровне значимости 0,05.

Пример 10.2. Пусть значение доходов X_{ij} из примера 10.1 описывается моделью

$$X_{ij} = \theta_j + \varepsilon_{ij}, \quad j = 1,2,3, \quad i = 1,\ldots,n_j,$$

где ε_{ij} соответствуют гауссовскому распределению с нулевым математическим ожиданием и неизвестной дисперсией σ^2,а $\theta_j \in \mathbb{R}^1, j = 1,2,3$ – неизвестные параметры. Постройте доверительный интервалы уровня надежности 0,95 для контрастов $\gamma_1, \gamma_2, \gamma_3$ параметров θ_j, где

$$\gamma_1 = \sum_{j=1}^{3} c_{1j}\theta_j = \theta_1 - \theta_2, \quad \text{т.е.} \quad c_{11} = 1, c_{12} = -1, c_{13} = 0;$$

$$\gamma_2 = \sum_{j=1}^{3} c_{2j}\theta_j = \theta_1 - \theta_3, \quad \text{т.е.} \quad c_{21} = 1, c_{22} = 0, c_{23} = -1;$$

$$\gamma_3 = \sum_{j=1}^{3} c_{3j}\theta_j = \theta_2 - \theta_3, \quad \text{т.е.} \quad c_{31} = 0, c_{32} = 1, c_{33} = -1.$$

Дайте содержательную трактовку контрастов $\gamma_1, \gamma_2, \gamma_3$.

Р е ш е н и е: Точечной оценкой параметра $\gamma_1 = \theta_1 - \theta_2$ будет согласно 10.10,

$$\widehat{\gamma}_1 = \sum_{j=1}^{3} c_{1j}\overline{X}_{\bullet j} = \overline{X}_{\bullet 1} - \overline{X}_{\bullet 2},$$

а доверительный интервал 10.12 уровня надежности 0,95 имеет вид

$$I_1(Z_N) = \left(\widehat{\gamma}_1 - t_{0,975,N-k}\sqrt{\widetilde{S}_N^2}\sum_{j=1}^{3}\frac{c_j^2}{n_j}; \qquad \widehat{\gamma}_1 + t_{0,975,N-k}\sqrt{\widetilde{S}_N^2}\sum_{j=1}^{3}\frac{c_j^2}{n_j} \right),$$

где

$$\widetilde{S}_N^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2.$$

Вычислим соответствующие реализации оценки контраста и доверительного интервала $I_1(Z_N)$. Поскольку (см. пример 10.1)

$$\overline{X}_{\bullet 1} = 11101{,}0, \qquad \overline{X}_{\bullet 2} = 10803{,}9, \qquad \overline{X}_{\bullet 3} = 20340{,}0,$$

то

$$\widehat{\gamma}_1 = 11101 - 10803{,}9 = 297{,}1,$$

$$\widetilde{S}_N^2 = \frac{1}{N-3}\sum_{j=1}^{3}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\bullet j})^2 = 19038613,$$

$$\sqrt{\sum_{j=1}^{3}\frac{c_j^2}{n_j}} = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{1}{11} + \frac{1}{8}} = 0{,}46.$$

По табл. находим квантиль $t_{0.975}(22) = 2{,}074$ и получаем реализацию доверительного интервала

$$I_1(z_{25}) = \Big(297{,}1 - 2{,}074 \cdot 2027{,}5; \quad 297{,}1 + 2{,}074 \cdot 2027{,}5\Big) = (-3907{,}6; 4501{,}8).$$

Аналогично

$$\widehat{\gamma}_2 = \sum_{j=1}^{3} c_{2j}\overline{X}_{\bullet j} = \overline{X}_{\bullet 1} - \overline{X}_{\bullet 3}; \qquad \widehat{\gamma}_3 = \sum_{j=1}^{3} c_{1j}\overline{X}_{\bullet j} = \overline{X}_{\bullet 2} - \overline{X}_{\bullet 3};$$

$$I_2(Z_N) = (\widehat{\gamma}_2 - t_{0,975}(N-k)\sqrt{\frac{1}{n_1} + \frac{1}{n_3}}\sqrt{\widetilde{S}_N^2}; \qquad \widehat{\gamma}_2 + t_{0,975}(N-k)\sqrt{\frac{1}{n_1} + \frac{1}{n_3}}\sqrt{\widetilde{S}_N^2});$$

$$I_3(Z_N) = (\widehat{\gamma}_3 - t_{0,975}(N-k)\sqrt{\frac{1}{n_2}+\frac{1}{n_3}}\sqrt{\widetilde{S}_N^2}\widehat{\gamma}_2 + t_{0,975}(N-k)\sqrt{\frac{1}{n_2}+\frac{1}{n_3}}\sqrt{\widetilde{S}_N^2}).$$

Вычисляем соответствующие реализации, получим:

$$\widehat{\gamma}_2 = 11101 + 20340 - 9239; \qquad \widehat{\gamma}_3 = 10803{,}9 - 20340 = -9536{,}1;$$

$$I_2(z_{25}) = (-13831{,}5; -4646{,}5); \qquad I_3(z_{25}) = (-14423{,}1; -4649{,}1).$$

Контраст γ_1 представляет разность средних значений случайных величин, порождающих выборки Z_1 и Z_2, контраст γ_2 — разность средних значений случайных величин, порождающих выборки Z_1 и Z_3, контраст γ_3 — разность средних значений случайных величин, порождающих выборки Z_2 и Z_3.

Важно отметить, что доверительный интервал для γ_1 включает значение ноль. Это означает, что на уровне доверия 0,95 можно считать, что разность параметров $\theta_1 - \theta_2 = \gamma_1$ равна нулю и гипотеза об однородности выборок Z_1 и Z_2 данной модели верна. Доверительные интервалы γ_2 и γ_3 не включают значения ноль, это означает, что средние значения θ_1 и θ_3, а так же θ_2 и θ_3 различаются. ■

Пример 10.3. После разрыва ахиллова сухожилия травмированному человеку необходимо сделать операцию и последующую иммобилизацию в течении 6 недель. Однако восстановление двигательных функций травмированной ноги требует длительного времени. Для сокращения восстановительного периода необходимо, по мнению врачей, пройти реабилитационный курс, включающий физиотерапевтическое лечение и занятие лечебной гимнастикой. Не все пациенты имеют силы и возможности пройти такой курс.

В табл. 10.5 представлены данные о времени (в неделях) восстановительного периода для трех групп успешно прооперированных пациентов примерно одинакового возраста и состояния здоровья. Пациенты первой группы прошли полный реабилитационный курс, пациенты второй группы получили только физиотерапевтическое лечение, а пациенты третьей группы целенаправленно не занимались реабилитацией.

Таблица 10.5

Группа 1	Группа 2	Группа 3
26	41	36
29	34	44
19	44	47
37	23	41
28	28	37
33	45	49
40	33	42
36	35	44
34	40	
31	54	

Можно ли считать, что указанные реабилитационные процедуры способствуют сокращению времени восстановления пациентов после травмы?

Р е ш е н и е: Фактором в данной задаче является наличие реабилитационного лечения. Уровни фактора: 1 — наличие полного реабилитационного курса лечения, 2 — частичное реабилитационное лечение, 3 — отсутствие реабилитационного лечения. Данные представлены тремя выборками

$$Z_j = [X_{1j}, \ldots, X_{n_j j}], \qquad j = 1,2,3$$

объемов $n_1 = 10, n_2 = 10$ и $n_3 = 8$, которые соответствуют неизвестным непрерывным распределениям $F(t - \theta_j)$.

Гипотеза

$$H_0 : \theta_1 = \theta_2 = \theta_3$$

означает, что выборки Z_1,Z_2 и Z_3 однородны, т.е. реабилитационное лечение не оказывает влияния на срок восстановления после травмы.

В качестве альтернативной гипотезы H_A можно выбрать и гипотезу общего вида

$$H_1 : \exists \theta_i \neq \theta_j \text{ при } i \neq j \ ,$$

и упорядоченную альтернативу

$$H_2 : \theta_1 \leqslant \theta_2 \leqslant \theta_3,$$

где хотя бы одно из неравенств строгое. Последняя альтернатива описывает ситуацию, когда более полное реабилитационное лечение обуславливает более быстрое восстановление.

Для проверки гипотезы H_0 против альтернативы H_1 можно использовать критерий Краскела—Уоллиса, для проверки H_0 против гипотезы H_2 — критерий Джонкхиера.

Чтобы вычислить статистику критерия Краскела—Уоллиса 10.15

$$T(\mathbb{Z}_N) = H = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(\overline{R}_{\bullet j} - \frac{N+1}{2} \right)^2$$

составим таблицу (табл. 10.6) реализаций рангов r_{ij} случайных величин X_{ij}, $j = 1,2,3$, $i = 1,\ldots,n_j$.

Таблица 10.6

Группа 1	Группа 2	Группа 3
3	19,5	13,5
6	10,5	23
1	23	26
15,5	2	19,5
4,5	4,5	15,5
8,5	25	27
17,5	8,5	21
13,5	12	23
10,5	17,5	
7	28	

Следовательно,

$$\overline{R}_{\bullet 1} = 8,7, \overline{R}_{\bullet 2} = 15,05, \overline{R}_{\bullet 3} = 21,06.$$

Тогда реализация статистики

$$H = \frac{12}{28 \cdot 29} \left[10(9,7 - 14,5)^2 + 10(15,05 - 14,5)^2 + 8(21,06 - 14,5)^2 \right] \approx 10,1.$$

Поскольку в выборке $\mathbb{Z}_N$ имеются связки, то следует использовать модифицированную форму 10.17 статистики H. Так как в выборке есть 8 связок, из которых 7 связок имеют размер 2, а одна — размер 3, то

$$H' = \frac{H}{1 - \frac{1}{N^3 - N} \sum\limits_{i=1}^{g} \left(t_i^3 - t_i \right)} = \frac{10,1}{1 - \frac{1}{28^3 - 28} \left(7(2^3 - 2) + (3^3 - 3) \right)} = 10,13.$$

При справедливости гипотезы H_0 статистика $T(\mathbb{Z}_N)$ имеет при $N \to \infty$ распределение хи-квадрат с $r = k - 1 = 2$ степенями свободы. Критическая область уровня значимости $\alpha = 0,05$ имеет вид $(\chi_{0,95}(2); +\infty)$, где $\chi_{0.95}(2)$ — квантиль уровня 0,95 распределения хи-квадрат с двумя степенями свободы.

По таблице находим $\chi_{0.95}(2) = 5,99$.

Таким образом реализация статистики попадает в критическую область и гипотеза H_0 отвергается в пользу альтернативы H_1 на уровне значимости 0,05.

Применим теперь критерий Джонкхиера, статистика которого 10.19 имеет вид

$$T(\mathbb{Z}_N) = J = U_{12} + U_{13} + U_{23},$$

где

$$U_{lm} = \sum_{i=1}^{n_l} \sum_{j=1}^{n_m} \varphi(X_{il}, X_{jm}).$$

Для вычисления реализации величины

$$U_{12} = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \varphi(X_{i1}, X_{j2})$$

необходимо сложить $n_1 \cdot n_2 = 10 \cdot 10$ значений $\varphi(x_{i1}, x_{j2})$.

Так, сравнивая каждое значение первого столбца с каждым значением второго столбца получим, $\varphi(x_{11}, x_{12}) = 1$, поскольку $x_{11} = 26 < x_{12} = 41$; $\varphi(x_{11}, x_{22}) = 1$, поскольку $x_{11} = 26 < x_{22} = 34$ и т.д. Таким образом, при $i = 1$

$$\sum_{j=1}^{10} \varphi(x_{11}, x_{j2}) = 9,$$

при $i = 2$

$$\sum_{j=1}^{10} \varphi(x_{21}, x_{j2}) = 8,$$

и т.д. для $i = 3, \ldots, 10$.

В итоге получаем

$$u_{12} = 9 + 8 + 10 + 5 + 8{,}5 + 7{,}5 + 4{,}5 + 5 + 6{,}5 + 8 = 72,$$

$$u_{13} = 8 + 8 + 8 + 7 + 8 + 8 + 6 + 7 + 8 + 8 = 76,$$

$$u_{23} = 5 + 8 + 3 + 8 + 8 + 2 + 8 + 8 + 6 + 0 = 56.$$

Тогда реализация статистики J

$$J = T(z_N) = 72 + 76 + 56 = 204.$$

Найдем математическое ожидание и дисперсию статистики J при справедливости гипотезы H_0

$$\mathsf{E}\{J\} = \frac{1}{4}\left(N^2 - \sum_{j=1}^{k} n_j^2\right) = \frac{1}{4}(28^2 - (100 + 100 + 64)) = 130,$$

$$D\{J\} = \frac{1}{72}\left(N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right) =$$

$$= \frac{1}{72}(28^2(2 \cdot 28 + 3) - (100 \cdot 23 + 100 \cdot 23 + 64 \cdot 19)) = 578{,}56.$$

Тогда реализация стандартизованной статистики J^*

$$J^* = T(z_N) = \frac{J - \mathsf{E}\{J\}}{\sqrt{D\{J\}}} = 3{,}07.$$

Статистика J^* при справедливости гипотезы H_0 асимптотически нормальна. Критическая область критерия уровня значимости $\alpha = 0{,}05$ имеет вид

$$(u_{1-\alpha}; +\infty) = (u_{0{,}95}; +\infty) = (1{,}65; +\infty).$$

Таким образом реализация статистики J^* попадает в критическую область, и гипотеза H_0 отвергается на уровне значимости 0,05 в пользу альтернативы H_2.

Важно отметить что полученная реализация статистики H', равная 10,13, совпадает с квантилью распределения хи-квадрат уровня $1 - \alpha \approx 0{,}99$, а реализация статистики J^*, равная 3,07, совпадает с квантилью распределения $\mathcal{N}(0;1)$ уровня $1 - \alpha \approx 0{,}999$. ■

Пример 10.4. Изучается влияние денежного стимулирования на производительность труда.

Шести однородным группам, по 5 человек, раздали задачи одинаковой сложности. Задачи были выданы каждому члену группы независимо от остальных. Группы различаются только по денежному вознаграждению за каждую решенную задачу. Величина вознаграждения зависит от номера группы: чем больше номер группы, тем больше вознаграждение. Каждой группе известна цена вознаграждения за решенную задачу.

Влияет ли вознаграждение на количество решенных задач?

В таблице представлено количество решенных задач каждым членом группы.

группа 1	группа 2	группа 3	группа 4	группа 5	группа 6
10	8	12	12	24	19
11	10	17	15	16	18
9	16	14	16	22	27
13	13	9	16	18	25
7	12	16	19	20	24

Лекция 11

Анализ статистической зависимости

11.1 Теоретические положения

Пусть выборки $\mathbb{X}_n = \{X_1,\ldots,X_n\}^T$ и $\mathbb{Y}_n = \{Y_1,\ldots,Y_n\}^T$ порождены случайными величинами X и Y соответственно. Предполагается, что $\mathbb{X}_n$ и $\mathbb{Y}_n$ получены в процессе совместного наблюдения за X и Y, а именно: если x_k — реализация случайной величины X_k, $k = 1,\ldots,n$ (т. е. реализация случайной величины X в k-ом опыте), то y_k — реализация случайной величины Y в этом же опыте. Обозначим $\mathbb{W}_n = \{W_1,\ldots,W_n\}^T$ — двумерную выборку с элементами $W_k = (X_k,Y_k)$, $k = 1,\ldots,n$. Обозначим через $F_X(x)$ и $F_Y(y)$ функции распределения случайных величин X и Y соответственно, а через $F_W(x,y)$ — функцию распределения случайного вектора $W = \{X,Y\}^T$, компонентами которого являются изучаемые случайные величины X и Y.

Определение 11.1. Статистическая гипотеза вида

$$H_0 : F_W(x,y) = F_X(x)F_Y(y),\ \forall x,y \in \mathbb{R}^1 \tag{11.1}$$

называется гипотезой о независимости случайных величин X и Y.

Проверка гипотезы о независимости двух номинальных признаков

При проведении социологических и психологических исследований часто приходится иметь дело с экспериментальными данными, которые не только не имеют количественного выражения, но и не могут быть упорядочены. Например, пол, профессия, принадлежность к политической партии, отношение к какой-либо религиозной конфессии и т.д. Такие категоризованные данные принято называть признаками, измеренными в номинальной шкале. Рассмотрим задачу выявления статистической зависимости (независимости) между признаками, измеренными в номинальной шкале, и укажем меры, описывающие силу связи между ними.

Пусть признак A измеряется в номинальной шкале и имеет категории (градации) $A_1,\ldots,A_m$, а признак B, измеренный в номинальной шкале, имеет категории $B_1,\ldots,B_k$. Например, признак A — цвет глаз человека, а признак B — пол человека. Тогда признак A имеет категории: A_1 — карий, A_2 — зеленый, A_3 — серый, A_4 — голубой, признак B категории: B_1 — мужской, B_2 — женский.

Определим следующие случайные события:

$A_i = \{$ признак A имеет i-ую категорию$\}$, $i = 1,\ldots,m$,

$B_j = \{$ признак B имеет j-ую категорию$\}$, $j = 1,\ldots,k$.

Введем обозначения

$$p_{i\bullet} = \mathsf{P}(A = A_i),\quad p_{\bullet j} = \mathsf{P}(B = B_j),\quad p_{ij} = \mathsf{P}(A = A_i, B = B_j),\quad i = 1,\ldots,m,\quad j = 1,\ldots,k.$$

Определение 11.2. Признаки A и B, измеренные в номинальной шкале, называются *независимыми*, если выполняются равенства

$$p_{ij} = p_{\bullet i} p_{\bullet j}, \qquad \forall i = 1,\ldots,m, \quad j = 1,\ldots,k.$$

Тогда гипотезу о независимости двух номинальных переменных можно сформулировать следующим образом:

$$H_0 : p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \forall i = 1,...,m, \forall j = 1,...,k. \tag{11.2}$$

Альтернативная гипотеза H_A общего вида формулируется как:

$$H_A : \exists (i,j) \text{ при которых } p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j} \tag{11.3}$$

Для того, чтобы проверять гипотезу H_0, сначала нужно провести эксперимент и составить таблицу сопряженности признаков.

Пусть случайным образом выбрано n объектов, и у каждого из этих объектов измерен признак A и признак B. Результаты измерений удобно представить в виде таблицы 6.1 размера $m \times k$. В этой таблице число n_{ij} обозначает количество объектов, у которых признак A имеет категорию A_i и признак B имеет категорию B_j; число $n_{i\bullet} = \sum_{j=1}^{k} n_{ij}$ обозначает количество объектов, имеющих категорию A_i, $i = 1,\ldots,m$; число $n_{\bullet j} = \sum_{i=1}^{m} n_{ij}$, —количество объектов, имеющих категорию B_j, $j = 1,\ldots,k$. Сумма наблюдений по всем строкам и всем столбцам $\sum_{i=1}^{m} \sum_{j=1}^{k} n_{ij} = n$ совпадает с числом проведённых испытаний.

Такую таблицу принято называть *таблицей сопряженности признаков*.

Таблица 11.1

A \ B	B_1	$\cdots$	B_k	
A_1	n_{11}	$\cdots$	n_{1k}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_m	n_{m1}	$\cdots$	n_{mk}	$n_{m\bullet}$
	$n_{\bullet 1}$	$\cdots$	$n_{\bullet k}$	n

Оценим неизвестные вероятности p_{ij}, $p_{i\bullet}$ и $p_{\bullet j}$ их частотами. Обозначим частоту случайного события $A_i \cdot B_j$, $i = 1,\ldots,m$, $j = 1,\ldots,k$ через $\hat{p}_{ij} = \frac{n_{ij}}{n}$, частоту события A_i через $\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n}$, а частоту события B_j через $\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}$.Согласно теореме Бернулли, частота $\hat{p}$ некоторого события сходится по вероятности к вероятности p этого события. Таким образом, в нашей модели имеем, что

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \xrightarrow{P} p_{ij} \text{ при } n \to \infty,$$
$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n} \xrightarrow{P} p_{i\bullet} \text{ при } n \to \infty,$$
$$\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n} \xrightarrow{P} p_{\bullet j} \text{ при } n \to \infty.$$

Если события A_i и B_j независимы, то $p_{ij} = p_{i\bullet} p_{\bullet j}$, значит

$$\frac{n_{ij}}{n} \approx \frac{n_{i\bullet}}{n} \cdot \frac{n_{\bullet j}}{n} \qquad \forall i = 1,...,k, \quad \forall j = 1,...,m.$$

Величины n_{ij} принято называть наблюдаемыми частотами, а величины $\frac{n_{i\bullet} n_{\bullet j}}{n}$ – ожидаемыми частотами. То есть, если признаки A и B независимы, то все ожидаемые частоты не должны сильно отклоняться от соответствующих наблюдаемых частот. Точное утверждение этого факта дается теоремой Фишера–Пирсона.

Теорема 11.1 (Фишер–Пирсон)**.** *При сделанных выше предположениях и $n \to \infty$ статистика*

$$\widehat{\chi}^2 = n \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(\hat{p}_{ij} - \hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{\hat{p}_{i\bullet}\hat{p}_{\bullet j}} = n \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{n_{i\bullet} n_{\bullet j}}. \tag{11.4}$$

при справедливости гипотезы H_0 вида 11.2 имеет распределение χ^2 с числом степеней свободы равным $(k-1)(m-1)$.

Если гипотеза о независимости признаков нарушается, то статистика $\hat{\chi}^2$ будет принимать большие значения. Следовательно, критическая область уровня значимости α имеет вид

$$(\chi^2_{1-\alpha,((k-1)(m-1))};+\infty),$$

где $\chi^2_{1-\alpha,((k-1)(m-1))}$ — квантиль уровня $1-\alpha$ распределения хи-квадрат с $r=(m-1)(k-1)$ степенями свободы.

Построенный критерий для проверки гипотезы H_0 вида 11.2 о независимости признаков против альтернативной гипотезы H_A вида (11.3) называется критерием Пирсона.

Для удобства вычислений можно использовать другую форму записи статистики (11.4) вида

$$\widehat{\chi}^2 = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij})^2}{n_{i\bullet}n_{\bullet j}}-1\right). \tag{11.5}$$

Для случая $m=k=2$, часто встречающегося на практике, формула (11.4) принимает весьма простой вид:

$$\widehat{\chi}^2 = \frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}. \tag{11.6}$$

Задача 11.1. Исследовалась связь между удовлетворенностью образом жизни и материальным положением семьи. На вопрос об удовлетворённости образом жизни каждый респондент должен дать один из ответов — не удовлетворён, удовлетворён, на вопрос о материальном положении — имеет низкий уровень или имеет высокий уровень. Было опрошено 665 респондентов. Результаты опроса представлены в таблице.

A \ B	не удовлетворены	удовлетворены	
низкий уровень	156	68	224
высокий уровень	74	357	431
	230	425	665

Имеется ли зависимость между материальным положением семьи и удовлетворённостью образом жизни?

Р е ш е н и е: Имеется два признака: «материальное положение семьи» (признак А) и «удовлетворенность образом жизни» (признак В). Проверим гипотезу H_0 вида (11.2) о независимости признаков А и В против альтернативной гипотезы H_A вида (11.3).

Вычисляя статистику Пирсона по формуле (11.6), получим $\hat{\chi}^2 = 176.38$.

При справедливости гипотезы H_0 о независимости признаков

$$\hat{\chi}^2|_{H_0} \underset{\text{n}\to\infty}{\sim} \chi^2(1).$$

По таблице находим 0.95 – квантиль распределения хи-квадрат с одной степенью свободы $\chi^2_{0.95;1} = 3.84$. Критическая область имеет вид: $(3.84;\infty)$.

Так как значение статистики попало в критическую область, то гипотеза о независимости признаков отвергается.

Задача 11.2. Врач офтальмолог провёл 954 операции по восстановлению зрения у возрастных пациентов. Перед операцией проводилось обследование зрительного аппарата пациентов, в ходе которого проверялось 5 показателей. Каждый из показателей оценивался доктором по следующей шкале: 0 — неудовлетворительное состояние, 1 — пограничное состояние, 2 — соответствует норме. Затем общее состояние зрительного аппарата пациента оценивалось как сумма пяти показателей. Некоторые пациенты имели осложнения после операции. Требуется выяснить, имеется ли зависимость между наличием осложнения после операции и состоянием зрительного аппарата пациента. Результаты представлены в таблице.

A \ B	нет осложнения	есть осложнение	
оценка 0-1	129	14	143
оценка 2-10	807	4	811
	936	18	954

Р е ш е н и е: Имеется два признака: «состояние зрительного аппарата пациента» (признак А) и «наличие осложнения» (признак В). Проверим гипотезу H_0 вида (11.2) о независимости признаков А и В против альтернативной гипотезы H_A вида (11.3).

Вычисляя статистику Пирсона по формуле (11.6), получим $\hat{\chi}^2 = 51.84$.

При справедливости гипотезы H_0 о независимости признаков

$$\hat{\chi}^2|_{H_0} \underset{n\to\infty}{\sim} \chi^2(1).$$

По таблице находим 0.95 — квантиль распределения хи-квадрат с одной степенью свободы $\chi^2_{0.95;1} = 3.84$. Критическая область имеет вид: $(3.84;\infty)$. Так как значение статистики попало в критическую область, то гипотеза о независимости признаков отвергается.

Если после применения критерия хи-квадрат принимается гипотеза H_0 о независимости признаков, то исследование можно считать завершенным. Если гипотеза H_0 отвергнута, т.е. признаки А и В зависимы, то нужно исследовать силу связи между ними и каким-то образом характеризовать имеющуюся зависимость.

11.2 Меры связи признаков в номинальной шкале.

Коэффициенты контингенции и ассоциации Юла для таблиц сопряжённости 2×2.

Рассмотрим сначала таблицы сопряжённости 2×2. Обозначим для удобства в таблице сопряженности

$$n_{11} = a, \quad n_{12} = b, \quad n_{21} = c, \quad n_{22} = d.$$

Тогда таблица будет иметь вид:

A \ B	B	$\bar{B}$	
A	a	b	$a+b$
$\bar{A}$	c	d	$c+d$
	$a+c$	$b+d$	n

Определение 11.3. *Коэффициентом контингенции* называется величина

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(d+c)(a+c)(b+d)}}. \tag{11.7}$$

Заметим, что коэффициент контингенции может принимать значения от -1 до 1.

Рассмотрим следующие характерные ситуации.

Если $\Phi = 0$, т.е. $ad = bc$, то, согласно формуле (11.6), статистика $\hat{\chi}^2 = 0$, и признаки A и B независимы.

Если $\Phi = 1$, то таблица сопряжённости имеет вид:

A \ B	B	$\bar{B}$
A	a	0
$\bar{A}$	0	d

Если $\Phi = -1$, то таблица сопряжённости имеет вид:

A \ B	B	$\bar{B}$
A	0	b
$\bar{A}$	c	0

Поясним смысл коэффициента контингенции. Для этого прошкалируем номинальные переменные A и B. Будем считать, что переменная A принимает значение 1, если признак A присутствует, и значение 0, если признак A отсутствует (событие $\bar{A}$). Аналогично и для признака B. Получим таблицу двумерного распределения, в которой вместо вероятностей выписаны соответствующие частоты.

A \ B	1	0
1	$\frac{a}{a+b+c+d}$	$\frac{b}{a+b+c+d}$
0	$\frac{c}{a+b+c+d}$	$\frac{d}{a+b+c+d}$

Оценим сначала математическое ожидание и дисперсию случайных величин A и :

$$\hat{\mathsf{E}}A = 0 \cdot \frac{c+d}{n} + 1 \cdot \frac{a+b}{n} = \frac{a+b}{n};$$

$$\hat{\mathsf{D}}A = \hat{\mathsf{E}}(A^2) - (\hat{\mathsf{E}}A)^2 = \frac{a+b}{n}\left(1 - \frac{a+b}{n}\right);$$

$$\hat{\mathsf{E}}B = 0 \cdot \frac{b+d}{n} + 1 \cdot \frac{a+c}{n} = \frac{a+c}{n};$$

$$\hat{\mathsf{D}}B = \hat{\mathsf{E}}(B^2) - (\hat{\mathsf{E}}B)^2 = \frac{a+c}{n}\left(1 - \frac{a+c}{n}\right);$$

$$\hat{\mathsf{E}}(A \cdot B) = \frac{a}{n}.$$

Подставим полученные оценки в формулу для вычисления выборочного коэффициента корреляции между A и B

$$\hat{\rho}_{AB} = \frac{\hat{\mathsf{E}}(AB) - (\hat{\mathsf{E}}A)(\hat{\mathsf{E}}B)}{\sqrt{\hat{D}A \cdot \hat{D}B}}.$$

Проведя простые арифметические преобразования, получим, что

$$\hat{\rho}_{AB} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \Phi.$$

Таким образом, коэффициент контингенции является аналогом коэффициента корреляции и характеризует двухстороннюю связь между случайными величинами.

Определение 11.4. *Коэффициентом ассоциации Юла* называется величина

$$Q = \frac{ad - bc}{ad + bc}.$$

Можно показать, что коэффициент ассоциации Юла принимает значение от -1 до 1.

Если $Q = 0$, то $\hat{\chi}^2 = 0$, и признаки A и B независимы.

Значению $Q = 1$ соответствуют следующие ситуации:

A \ B	B	$\bar{B}$
A	a	b
$\bar{A}$	0	d

или

A \ B	B	$\bar{B}$
A	a	0
$\bar{A}$	c	d

Т.о. значение $Q = 1$ описывает одну из следующих двух ситуаций:

1) если признак A отсутствует, то признак B тоже отсутствует, но отсутствие признака B не влечёт отсутствие признака A или

2) наличие признака A означает, что признак B также присутствует, но наличие признака B не влечёт наличие признака A.

То есть, коэффициент ассоциации Q характеризует одностороннюю связь между признаками.

Охарактеризуйте самостоятельно ситуацию, при которой коэффициент ассоциации $Q = -1$.

Если $Q = -1$, то

A \ B	B	$\bar{B}$
A	a	b
$\bar{A}$	c	0

или

A \ B	B	$\bar{B}$
A	0	b
$\bar{A}$	c	d

Задача 11.3. Вернёмся к задачам 6.1 и 6.2, в которых было выявлено наличие зависимости между признаками, и охарактеризуем эту зависимость с помощью коэффициентов контингенции и ассоциации.

Так, в задаче 11.1, имеем $\Phi = 0.5214$; $Q = 0.834$.

В задаче 11.2 получим $\Phi = -0.244$; $Q = -0.91$.

Задача 11.4. Характерный пример, демонстрирующий существенное различие между коэффициентом контингенции Φ и коэффициентом ассоциации Q, основан на следующих реальных данных (Macmillan Publishing Company). Имеются сведения о влиянии прививки на холерную инфекцию. Среди 1630 человек, привитых от инфекции, заболело 5 человек; среди 1033 непривитых заболело 11 человек. Выясним сначала, имеется ли зависимость между наличием прививки (признак A) и заболеваемостью холерой (признак B). Составим таблицу сопряжённости:

A \ B	заболел	не заболел	
есть	5	1625	1630
нет	11	1022	1033
	16	2647	2663

Вычисляя статистику критерия хи-квадрат, получим $\chi^2 = 6.06$.

Значение статистики попадает в критическую область, и гипотеза H_0 вида (11.2) о независимости признаков A и B против альтернативной гипотезы H_A вида (11.3) отвергается на уровне значимости $\alpha = 0.05$.

Опишем имеющуюся зависимость с помощью коэффициентов контингенции Φ и ассоциации Q.

$\Phi = -0.048$, $Q = -0.555$.

Несмотря на выявленную зависимость, коэффициент Φ оказался практически нулевым. Это говорит об отсутствии двусторонней связи между признаками. Например, имея информацию об отсутствии у человека заболевания, нельзя судить о наличии у него прививки. Коэффициент ассоциации Q, равный -0.555, говорит о наличии между признаками A и B односторонней связи «средней» силы. В данном примере это можно трактовать так: заболевание холерой свидетельствует о том, что заболевший человек скорее не имел прививки, однако отсутствие заболевания не означает наличие прививки.

Меры связи, основанные на статистике хи-квадрат.

Рассмотрим теперь таблицы сопряжённости произвольного размера. Понятно, что большие значения статистики $\widehat{\chi}^2$ говорят о наличии зависимости между признаками A и B. Однако, непосредственно величина $\widehat{\chi}^2$ не позволяет судить о степени этой зависимости, так как величина $\widehat{\chi}^2 \to \infty$ при неограниченном возрастании n, если признаки A и B зависимы.

В качестве меры связи признаков A и B К. Пирсон предложил *коэффициент взаимной сопряженности (или коэффициент Пирсона)*

$$P = \sqrt{\frac{\widehat{\chi}^2}{\widehat{\chi}^2 + n}}. \tag{11.8}$$

Основанием для его введения послужил следующий факт. Если для двумерной гауссовской выборки $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$ провести разбиение наблюдаемой двумерной выборки на $m \times k$ непересекающихся прямоугольников

$$\Delta_{ij} = \Delta_{X,i} \times \Delta_{Y,i}, \quad i = 1,\ldots,m, \quad j = 1,\ldots,k,$$

то при возрастании m и k коэффициент

$$P^2 = \frac{\widehat{\chi}^2}{\widehat{\chi}^2 + n} \to \rho_{XY}^2,$$

где $\widehat{\chi}^2$ — статистика вида (11.4), а ρ_{XY} — коэффициент корреляции случайной величины X и Y, порождающих выборки $\mathbb{X}_n$ и $\mathbb{Y}_n$.

Однако, в отличие от ρ_{XY}, максимальное значение P равно $\sqrt{\frac{l-1}{l}} < 1$, где $l = \min(m,k)$. Чтобы устранить этот недостаток, Крамер ввел другую функцию от $\widehat{\chi}^2$

$$C = \sqrt{\frac{\widehat{\chi}^2}{n \cdot \min\{(m-1),(k-1)\}}}, \tag{11.9}$$

которая называется *коэффициентом Крамера*.

Значение коэффициента Крамера $C \in [0;1]$ и верхний предел $C = 1$ достигается тогда и только тогда, когда каждая строка (при $m \geqslant k$) или каждый столбец (при $m \leqslant k$) таблицы 9.3 содержит лишь один отличный от нуля элемент.

Значения коэффициентов P и C близкие к 1 говорят о сильной связи между признаками A и B. Если гипотеза о независимости признаков A и B отвергнута, то принято считать, что значения коэффициентов P и C в интервале $[0;0,3)$ говорят о слабой силе связи признаков A и B, значения в интервале $[0,3;0,7)$ — об умеренной силе связи и значения в интервале $[0,7;1]$ — о значительной силе связи.

Отметим, что коэффициенты P и V равны нулю в том случае, когда статистика $\hat{\chi}^2 = 0$. Значение этих коэффициентов близкие к единице говорят о сильной связи между переменными. Однако таблиц с критическими значениями этих коэффициентов нет, что не позволяет на заданном уровне значимости сделать вывод о наличии или отсутствии зависимости. Кроме того, эти коэффициенты не позволяют судить о характере выявленной зависимости.

Пример 11.1. Имеются следующие данные о специализации и поле 900 английских студентов

Специализация	**Пол**		
	М	Ж	
Искусствоведение	165	185	350
Естественные науки	168	92	260
Социально-экономические науки	115	105	220
Музыка	32	38	70
	480	420	900

Выяснить имеется ли зависимость между выбранной специальностью и полом студента. Оценить силу связи.

Р е ш е н и е: Каждый объект (респондент) в данной задаче характеризуется двумя признаками. Пусть признак A — выбранная специализация имеет градации, а признак B — пол

студента. Тогда A имеет градации: A_1 — искусствоведение, A_2 — естественные науки, A_3 — социально-экономические науки, A_4 — музыка; а B градации: B_1 — мужской, B_2 — женский.

Проверка гипотезы H_0 о независимости признаков A и B формулируется следующим образом:

$$H_0 : p_{ij} = p_{i\bullet} p_{\bullet j} \quad \text{для любых} \quad i = 1,\ldots,4, \quad j = 1,2.$$

Для проверки этой гипотезы используем критерий хи-квадрат, статистика которого имеет вид (11.4).

Согласно представленной таблице сопряженности, реализации ожидаемых частот принимают следующие значения

$$\frac{n_{1\bullet} n_{\bullet 1}}{n} = \frac{350 \cdot 480}{900} = 186{,}7, \quad \frac{n_{1\bullet} n_{\bullet 2}}{n} = \frac{350 \cdot 420}{900} = 163{,}3,$$

$$\frac{n_{2\bullet} n_{\bullet 1}}{n} = 138{,}7, \quad \frac{n_{2\bullet} n_{\bullet 2}}{n} = 121{,}3, \quad \frac{n_{3\bullet} n_{\bullet 1}}{n} = 117{,}3,$$

$$\frac{n_{3\bullet} n_{\bullet 2}}{n} = 102{,}7, \quad \frac{n_{4\bullet} n_{\bullet 1}}{n} = 37{,}3, \quad \frac{n_{4\bullet} n_{\bullet 2}}{n} = 32{,}7,$$

а реализация статистики

$$\widehat{\chi}^2 = 2{,}52 + 6{,}19 + 0{,}05 + 0{,}75 + 2{,}88 + 7{,}08 + 0{,}05 + 0{,}86 = 20{,}38.$$

При справедливости гипотезы H_0 статистика хи-квадрат имеет распределение хи-квадрат с $r = (k-1)(m-1) = 3$ степенями свободы. Выберем уровень значимости $\alpha = 0{,}05$, тогда критическая область имеет вид:

$$(\chi^2_{0{,}95}(3); +\infty) = (7{,}81; +\infty).$$

Реализация статистики попадает в критическую область. Следовательно, гипотеза о независимости признаков A (выбранная специализация) и B (пол студента) отвергается.

Оценим силу связи между признаками A и B с помощью коэффициентов Пирсона и Крамера:

$$P = 0{,}149, \quad C = 0{,}15.$$

Значения этих коэффициентов близки к нулю, что говорит о достаточно слабой силе выявленной связи.

11.3 Коэффициенты связи, основанные на прогнозе.

Меры прогноза Гутмана (λ-меры)

Пример 11.2. Сначала покажем принцип построения этих коэффициентов на примере. Пусть проведён опрос 655 респондентов и составлена следующая таблица сопряжённости признаков A и B, измеренных в номинальной шкале.

Признак A — «удовлетворённость образом жизни» имеет две градациии: неудовлетворён и удовлетворён. Признак B — «материальное положение семьи» имеет пять градаций: плохое, ниже среднего, среднее, выше среднего, высшее.

A \ B	плохое	ниже среднего	среднее	выше среднего	хорошее	
неудовлетворён	92	64	48	23	3	230
удовлетворён	22	46	136	148	73	425
	114	110	184	171	76	655

Опираясь на результаты опроса, представленные в этой таблице, мы хотим предсказать материальное положение (категорию признака B) наугад выбранного респондента. Согласно таблице, следует сделать прогноз о том, что респондент имеет «среднее» материальное положение, так как именно «среднее» материальное положение является модальной

(наиболее вероятной) категорией признакак B. Назовём такой прогноз первым прогнозом. Делая такой прогноз, мы хотим оценить вероятность ошибки прогноза:

$$\hat{p}_1 = 1 - \frac{184}{655} = 0.719,$$

При прогнозировании мы не учли признак A, который связан с признаком B. Понятно, что у респондентов не удовлетворённых своим образом жизни (признак A имеет категорию «неудовлетворён») модальная категория признака B будет «плохое материальное положение», а у респондентов, удовлетворённых своим образом жизни (т.е. A имеет категорию «удовлетворён»), модальной категорией признака B будет «выше среднего».

Таким образом, если при прогнозировании признака B, будет учитываться значение признака A, то оценка вероятности ошибки такого прогноза будет

$$1 - \frac{92+148}{655} = 1 - 0.367 = 0.633,$$

Тогда значение качества прогноза модальной категории признака B, обусловленное учётом совместного распределения признаков A и B, можно характеризовать следующим коэффициентом

$$\lambda_B = \frac{\text{вероятность ошибки первого прогноза} - \text{вероятность ошибки второго прогноза}}{\text{вероятность ошибки первого прогноза}}.$$

Заменяя неизвестные вероятности их оценками, получим:

В данном примере

$$\hat{\lambda}_B = \frac{0.719 - 0.633}{0.719} = 0.119.$$

Это означает, что прогноз модальной категории признака B (материальное положение) будет улучшен на 11.9%, если при прогнозировании будет учтено совместное распределение признаков A и B.

Аналогично для признака A. Прогнозируемое значение признака A - это его модальная категория («удовлетворён»), оценка вероятности ошибки такого прогноза равна

$$\hat{p}_1 = 1 - \frac{425}{655} = 0.35.$$

Оценка вероятности ошибки второго прогноза, при котором учтено значение признака B, будет

$$\hat{p}_2 = 1 - \frac{92+64+136+148+73}{655} = 1 - 0.783 = 0.217,$$

а мера прогноза для признака A

$$\hat{\lambda}_A = \frac{0.35 - 0.217}{0.35} = 0.38$$

Это означает, что прогноз модальной категории признака «удовлетворённость образом жизни» (признак A) будет улучшен на 38%, если при прогнозировании будет учтено значение признака B (материальное положение).

Перейдём к формальному представлению мер прогноза, которые называются λ-мерами Гутмана.

Пусть известно совместное распределение признаков A и B. Назовем модальным значением B такое событие B_j, для которого $p_{\bullet j} = \max\limits_{1 \leqslant l \leqslant k} p_{\bullet l}$.

В качестве прогноза номинального признака B с категориями $B_1, \ldots, B_k$ естественно выбрать такую категорию B_j, которая представляет собой модальный исход признака B. Вероятность ошибки такого прогноза (назовем его первым прогнозом) будет

$$p_1 = 1 - \max_{1 \leqslant l \leqslant k} p_{\bullet l}.$$

При определении первого прогноза не было учтено значение признака A. Если же признаки A и B зависимы, то естественно предположить, что прогноз модальной категории признака B может быть улучшен (т. е. вероятность ошибки прогноза будет уменьшена), если при прогнозировании будет учтено совместное распределение признаков A и B.

Пусть известно, что в результате проведенного испытания реализовалось событие A_i. Назовем вторым прогнозом признака B такую категорию B_j, для которой

$$p_{ij} = \max_{1 \leqslant l \leqslant k} p_{il}.$$

Вероятность ошибки второго прогноза составит

$$p_2 = 1 - \sum_{i=1}^{m} \max_{1 \leqslant l \leqslant k} p_{il}.$$

Гутманом была предложена мера прогноза λ_B равная относительному уменьшению вероятности ошибки предсказания модальной категории признака B при переходе от первого прогноза ко второму.

Мерой λ_B называется величина

$$\lambda_B = \frac{p_1 - p_2}{p_1} = \frac{\sum\limits_{i=1}^{m} \max\limits_{1 \leqslant l \leqslant k} p_{il} - \max\limits_{1 \leqslant l \leqslant k} p_{\bullet l}}{1 - \max\limits_{1 \leqslant l \leqslant k} p_{\bullet l}}. \tag{11.10}$$

Аналогично, мерой прогноза признака A называется

$$\lambda_A = \frac{\sum\limits_{j=1}^{k} \max\limits_{1 \leqslant l \leqslant m} p_{lj} - \max\limits_{1 \leqslant l \leqslant m} p_{l\bullet}}{1 - \max\limits_{1 \leqslant l \leqslant m} p_{l\bullet}}. \tag{11.11}$$

Мера λ_B (λ_A) характеризует улучшение качества прогноза модальной категории признака B (признака A), которое обусловлено учетом совместного распределения признаков A и B. Меры λ_B и λ_A принимают значения от 0 до 1 и имеют следующую интерпретацию. Величина $\lambda_B \cdot 100\%$ ($\lambda_A \cdot 100\%$) показывает на сколько процентов улучшится прогноз модальной категории признака B (признака A), если при прогнозировании будет учтено совместное распределение этих двух признаков.

Меры λ_B и λ_A асимметричны, так как при прогнозировании один из признаков рассматривается как причина, а другой как следствие. Если непонятно какой из признаков является причиной, а какой следствием, то в качестве меры прогноза рассматривают симметричную меру $\lambda = \frac{\lambda_A + \lambda_B}{2}$.

Пусть имеется таблица сопряженности признаков A и B (см. табл. 6.1). Оценим по этой выборке меру λ_B.

Оценкой вероятности ошибки первого прогноза является соответствующая частота

$$\widehat{p}_1 = 1 - \frac{1}{n} \max_{1 \leqslant l \leqslant k} n_{\bullet l},$$

а оценкой вероятности ошибки второго прогноза

$$\widehat{p}_2 = 1 - \frac{1}{n} \sum_{i=1}^{m} \max_{1 \leqslant l \leqslant k} n_{il}.$$

Тогда оценкой меры λ_B будет коэффициент

$$\widehat{\lambda}_B = \frac{\widehat{p}_1 - \widehat{p}_2}{\widehat{p}_1} = \frac{\sum\limits_{i=1}^{m} \max\limits_{1 \leqslant l \leqslant k} n_{il} - \max\limits_{1 \leqslant l \leqslant k} n_{\bullet l}}{n - \max\limits_{1 \leqslant l \leqslant k} n_{\bullet l}}.$$

Аналогично, оценкой меры λ_A будет коэффициент

$$\widehat{\lambda}_A = \frac{\sum\limits_{j=1}^{k} \max\limits_{1\leqslant l\leqslant m} n_{lj} - \max\limits_{1\leqslant l\leqslant m} n_{l\bullet}}{n - \max\limits_{1\leqslant l\leqslant m} n_{l\bullet}},$$

а оценкой меры λ коэффициент $\widehat{\lambda} = \frac{\widehat{\lambda}_A + \widehat{\lambda}_B}{2}$.

Для мер Гутмана λ_B и λ_A можно построить не только точечные оценки $\widehat{\lambda}_B$, $\widehat{\lambda}_A$, но и асимптотические доверительные интервалы (см., например, стр. 128-129 учебного пособия Горяинова Е.Р., Панков А.Р.,Платонов Е.А. «Прикладные методы анализа статистических данных».- М.:Изд. дом ВШЭ,2012)

Меры прогноза Гутмана (λ - меры) обладают следующими свойствами:

1. $0 \leqslant \lambda_B \leqslant 1$, $0 \leqslant \lambda_A \leqslant 1$;
2. $\hat{\lambda}_B = 1$ в том случае, когда в каждой строке таблицы сопряжённости есть только одна ненулевая частота; $\hat{\lambda}_A = 1$ в том случае, когда в каждом столбце таблицы сопряжённости есть только одна ненулевая частота. Случай $\hat{\lambda}_B = 1$ $(\hat{\lambda}_A = 1)$ означает точную прогнозируемость признака B (признака A) по известному значению признака A (признака B).
3. Недостатком λ-мер является их равенство нулю в случае, когда все максимальные значения по строкам находятся в одном столбце $(\hat{\lambda}_B = 0)$ или все максимальные значения по столбцам находятся в одной строке.

Пример 11.3. В 2009 г. центром исследования гражданского общества и некоммерческого сектора НИУ ВШЭ была сформирована репрезентативная выборка из 2000 респондентов. Среди ста вопросов анкеты были, в частности, такие:

какое из шести перечисленных описаний точнее всего соответствует материальному положению вашей семьи;

удовлетворены ли вы своим здоровьем.

На первый вопрос предлагались ответы:

1. денег не хватает даже на питание (категория A_1);
2. на питание денег хватает, но не хватает на покупку одежды и обуви (категория A_2);
3. на покупку одежды и обуви денег хватает, но не хватает на покупку бытовой техники (категория A_3);
4. денег вполне хватает на покупку крупной бытовой техники, но не можем купить новый автомобиль (категория A_4);
5. денег хватает на все, кроме таких дорогих приобретений, как квартира, дом (категория A_5);
6. материальных затруднений не испытываем, при необходимости могли бы приобрести квартиру, дом (категория A_6).

Ответы на второй вопрос: удовлетворен (категория B_1) и не удовлетворен (категория B_2). Результаты опроса представлены в таблице сопряженности признаков A (материальное положение семьи) и B (удовлетворенность состоянием своего здоровья).

A \ B	B_1	B_2	
A_1	83	154	237
A_2	278	354	632
A_3	470	299	769
A_4	204	76	280
A_5	46	20	66
A_6	13	3	16
	1094	906	2000

Оценить меры прогноза Гутмана $\widehat{\lambda}_B$ и $\widehat{\lambda}_A$.

Р е ш е н и е: Оценкой меры прогноза Гутмана λ_B является

$$\widehat{\lambda}_B = \frac{\sum\limits_{i=1}^{m} \max\limits_{1\leqslant j\leqslant k} n_{ij} - \max\limits_{1\leqslant j\leqslant k} n_{\bullet j}}{n - \max\limits_{1\leqslant j\leqslant k} n_{\bullet j}}, \quad k=2, m=6.$$

Согласно таблице сопряженности признаков, максимальное значение сумм по столбцам имеет первый столбец, т. е.

$$\max_{1\leqslant j\leqslant 2} n_{\bullet j} = n_{\bullet 1} = 1094,$$

а

$$\sum_{i=1}^{6} \max_{1\leqslant j\leqslant 2} n_{ij} = 154+354+470+204+46+13 = 1241.$$

Тогда реализация оценки

$$\widehat{\lambda}_B = \frac{1241-1094}{2000-1094} = 0{,}168.$$

Аналогично, оценка меры прогноза Гутмана λ_A есть

$$\widehat{\lambda}_A = \frac{\sum\limits_{j=1}^{k} \max\limits_{1\leqslant i\leqslant m} n_{ij} - \max\limits_{1\leqslant i\leqslant m} n_{i\bullet}}{n - \max\limits_{1\leqslant i\leqslant m} n_{i\bullet}}, \quad k=2, m=6.$$

По таблице сопряженности признаков находим

$$\max_{1\leqslant i\leqslant 6} n_{i\bullet} = n_{3\bullet} = 769,$$

а

$$\sum_{j=1}^{2} \max_{1\leqslant i\leqslant 6} n_{ij} = 470+354 = 824.$$

Реализация оценки

$$\widehat{\lambda}_A = \frac{824-769}{2000-769} = 0{,}045.$$

Оценка для симметричной меры прогноза λ будет

$$\widehat{\lambda} = \frac{\widehat{\lambda}_A + \widehat{\lambda}_B}{2} = 0{,}107.$$

Построенные оценки позволяют сказать, что прогноз модальной (наиболее вероятной) категории признака B (удовлетворенность состоянием своего здоровья) улучшится на 16,8%, если при прогнозировании будет учтено значение признака A (материальное положение семьи), а прогноз модальной категории признака A улучшится на 4,5%, если при прогнозировании будет учтено значение признака B.

Меры прогноза Гудмана–Краскела (τ-меры)*

Как мы указали выше, меры Гутмана показывают улучшение качества прогноза только модальной категории признака. Более продвинутые меры прогноза Гудмана–Краскела показывают улучшение качества прогноза любой категории признака. Меры по-прежнему определяются как относительное уменьшению вероятности ошибки предсказания категории признака при переходе от первого прогноза ко второму, но возможность появления различных категорий признаков оценивается иначе. А именно, для признака B:

$$\hat{p}_1 = 1 - \frac{\sum_j \frac{n_{\bullet j} n_{\bullet j}}{n}}{n}, \tag{11.12}$$

$$\hat{p}_2 = 1 - \frac{\sum_i \sum_j n_{ij} \frac{n_{ij}}{n_{i\bullet}}}{n}, \tag{11.13}$$

и

$$\hat{\tau}_B = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}_1} = \frac{n \sum_i \sum_j \frac{n_{ij}^2}{n_{i\bullet}} - \sum_j n_{\bullet j}^2}{n^2 - \sum_j n_{\bullet j}^2} \tag{11.14}$$

Аналогично для признака A

$$\hat{p}_1 = 1 - \frac{\sum_i \frac{n_{i\bullet} n_{i\bullet}}{n}}{n}, \tag{11.15}$$

$$\hat{p}_2 = 1 - \frac{\sum_j \sum_i n_{ij} \frac{n_{ij}}{n_{\bullet j}}}{n}, \tag{11.16}$$

и

$$\hat{\tau}_A = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}_1} = \frac{n \sum_j \sum_i \frac{n_{ij}^2}{n_{\bullet j}} - \sum_j n_{i\bullet}^2}{n^2 - \sum_i n_{i\bullet}^2} \tag{11.17}$$

Лекция 12

Исследование зависимости признаков, измеренных в порядковой шкале.

12.1 Критерий Спирмена

Пусть показатели (признаки) X и Y измеряются в порядковой шкале. В результате наблюдения над n объектами получим двумерную выборку размера n

$$(R_1,S_1),\ldots,(R_n,S_n),$$

где R_i — ранг X_i, а S_i — ранг Y_i. Необходимо выяснить, зависимы ли X и Y. Впервые решение задачи проверки независимости порядковых признаков было предложено психологом Ч. Спирменом.

В качестве меры расхождения X и Y Спирмен предложил использовать величину

$$S=\sum_{i=1}^{n}(R_i-S_i)^2.$$

Если

$$R_1=S_1,\ldots,R_n=S_n,$$

т.е. последовательность рангов X полностью совпадает с последовательностью рангов Y, то $S=0$.

Рассмотрим ситуацию, когда S достигает максимального значения.

$R_i:$	1	2	...	n
$S_i:$	n	$(n-1)$	...	1

Тогда

$$S=\sum_{i=1}^{n}((n-i+1)-i)^2=\frac{1}{3}(n^3-n).$$

Это значение соответствует ситуации полной противоположной связи признаков. Хотелось бы, чтобы коэффициент, измеряющий силу связи между признаками, принимал значения от -1 (при обратной связи) до 1 (при прямой связи) и не зависел от размера выборки.

С этой целью Спирмен преобразовал меру S и ввел следующий *коэффициент ранговой корреляции Спирмена*

$$\hat{\rho}_s=1-\frac{6\cdot\sum\limits_{i=1}^{n}(R_i-S_i)^2}{n^3-n},\tag{12.1}$$

который изменяется -1 до 1.

Покажем, что коэффициент Спримена является аналогом выборочного коэффициента корреляции

$$\hat{\rho}_{XY} = \frac{\sum\limits_{i=1}^{n}(X_i-\bar{X})\cdot(Y_i-\bar{Y})}{\sqrt{\sum\limits_{i=1}^{n}(X_i-\bar{X})^2\cdot\sum\limits_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

для порядковой шкалы.

От количественных переменных в формуле $\hat{\rho}_{XY}$ перейдем к порядковым. Тогда

$$\bar{R} = \bar{S} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}.$$

Подставляя эти значения в $\hat{\rho}_{XY}$, получим

$$\hat{\rho}_{xy} = \frac{\sum\limits_{i=1}^{n}(R_i-\bar{R})\cdot(S_i-\bar{S})}{\sqrt{\sum\limits_{i=1}^{n}(R_i-\bar{R})^2\cdot\sum\limits_{i=1}^{n}\left(S_i-\bar{S}\right)^2}} = \frac{\sum\limits_{i=1}^{n}\left(R_i-\frac{n+1}{2}\right)\left(S_i-\frac{n+1}{2}\right)}{\sum\limits_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2} = 1-\frac{6\sum\limits_{i=1}^{n}(R_i-S_i)^2}{n^3-n} = \hat{\rho}_S \tag{12.2}$$

При наличии связок в выборках в формулу (12.1) нужно внести следующую поправку

$$\widehat{\rho}_S(n) = 1-\frac{\sum\limits_{i=1}^{n}(R_i-S_i)^2}{\frac{1}{6}(n^3-n)-(u_1+u_2)}, \tag{12.3}$$

где

$$u_1 = \frac{1}{12}\sum_{k=1}^{q}(u_{1k}^3-u_{1k}),$$

$$u_2 = \frac{1}{12}\sum_{k=1}^{g}(u_{2k}^3-u_{2k}),$$

q — количество связок в выборке $\mathbb{X}_n$, u_{1k} — размер k-ой связки выборки $\mathbb{X}_n$, g — количество связок в выборке $\mathbb{Y}_n$, u_{2k} — размер k-ой связки выборки $\mathbb{Y}_n$.

При справедливости гипотезы о независимости признаков X и Y квантили распределения статистики Спирмена $\widehat{\rho}_S(n)$ табулированы для $4 \leqslant n \leqslant 100$.

Доказано, что при $n \to \infty$ и справедливости гипотезы о независимости статистика

$$\sqrt{n-1}\,\widehat{\rho}_S \sim N(0,1). \tag{12.4}$$

Вопрос об альтернативной гипотезе оставим пока(!) открытым.

12.2 Критерий Кендалла

Исследуется независимость компонент вектора $W = (X,Y)^T$, по результатам n наблюдений $(X_1,Y_1),\ldots,(X_n,Y_n)$.

Определение 12.1. Назовем *параметром согласованности* случайных величин X и Y, являющихся компонентами вектора W, величину

$$\tau_{XY} = 1-2\mathsf{P}\{(X_2-X_1)(Y_2-Y_1)<0\},$$

где (X_1,Y_1) и (X_2,Y_2) — независимые двумерные случайные величины, имеющие распределение $F_W(x,y)$.

Параметр τ_{XY}, согласно определению 12.1, может принимать значения от -1 до 1, причём значения ± 1 достигаются в том случае, когда

$$Y = \varphi(X),$$

где $\varphi(\cdot)$ — строго монотонная функция. Действительно, если $\varphi(\cdot)$ — строго возрастающая функция, то

$$\mathsf{P}\{(X_2 - X_1)(Y_2 - Y_1) < 0\} = \mathsf{P}\{(X_2 - X_1)(\varphi(X_2) - \varphi(X_1)) < 0\} = 0,$$

т. е. $\tau_{XY} = 1$.

Если $\varphi(\cdot)$ — монотонно убывающая функция, то

$$\mathsf{P}\{(X_2 - X_1)(\varphi(X_2) - \varphi(X_1)) < 0\} = 1,$$

т. е. $\tau_{XY} = -1$.

Замечание 12.1. Нетрудно показать, что в случае независимости случайных величин X и Y, параметр τ_{XY} равен нулю.

Однако существуют ситуации, когда случайные величины X и Y зависимы, а $\tau_{XY} = 0$.

Пример 12.1. Пусть случайная величина X имеет чётную плотность распределения $f(x)$, то есть $f(-x) = f(x)$ для любого $x \in R^1$, а случайная величина $Y = X^2$. Данные величины являются функционально зависимыми. Покажите, что параметр τ_{XY} равен нулю.

Оценим параметр согласованности τ_{XY}. Пусть (X_i,Y_i) и (X_j,Y_j) — элементы выборки $\mathbb{W}_n$, $1 \leqslant i \neq j \leqslant n$.

Определение 12.2. Пары (X_i,Y_i) и (X_j,Y_j) называются *согласованными*, если

$$\operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\} = 1,$$

и *несогласованными*, если

$$\operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\} = -1.$$

Из n наблюдений можно составить

$$C_n^2 = \frac{n!}{(n-2)! \cdot 2!} = \frac{n \cdot (n-1)}{2}$$

пар. Обозначим через K случайное число несогласованных пар среди всех

$$C_n^2 = \frac{n(n-1)}{2}$$

пар выборки $\mathbb{W}_n$, через Q — число всех согласованных пар, а через

$$S = Q - K$$

— меру порядка.

Если все пары несогласованы, т.е. $Q = 0$, то

$$S = \frac{-n(n-1)}{2}.$$

Если все пары согласованы, т.е. $K = 0$, то

$$S = \frac{n(n-1)}{2}.$$

Таким образом

$$-\frac{n(n-1)}{2} \leqslant S \leqslant \frac{n(n-1)}{2}.$$

Разделив величину S на её максимальное значение, получим

$$\frac{S}{\max S} = \frac{2\left(\left(\frac{n(n-1)}{2} - K\right) - K\right)}{n(n-1)} = 1 - \frac{4K}{n(n-1)}.$$

Отметим, что

$$\frac{S}{\max S} \in [-1,1].$$

Статистика

$$\widehat{\tau}_{XY} = 1 - \frac{4K}{n(n-1)} \tag{12.5}$$

называется *коэффициентом согласованности Кендалла*.

Нетрудно показать, что возможны и другие формы записи коэффициента Кендалла:

$$\widehat{\tau}_{XY} = \frac{2(Q-K)}{n(n-1)} = 2\frac{\sum\limits_{1\leqslant i<j\leqslant n} \operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\}}{n(n-1)} = 2\frac{\sum\limits_{1\leqslant i<j\leqslant n} \operatorname{sign}\{(R_i - R_j)(S_i - S_j)\}}{n(n-1)}, \tag{12.6}$$

где R_i, R_j — ранги элементов X_i, X_j в выборке $\mathbb{X}_n$, а S_i, S_j — ранги Y_i, Y_j в выборке $\mathbb{Y}_n$.

Если в выборках $\mathbb{X}_n$ и $\mathbb{Y}_n$ имеются связки, то при вычислении коэффициента $\widehat{\tau}_{XY}$ следует внести поправку

$$\widehat{\tau}_{XY} = \frac{\sum\limits_{1\leqslant i<j\leqslant n} \operatorname{sign}\{(X_i - X_j)(Y_i - Y_j)\}}{\sqrt{\frac{1}{2}n(n-1) - u_1}\sqrt{\frac{1}{2}n(n-1) - u_2}}, \tag{12.7}$$

где

$$u_1 = \frac{1}{2}\sum_{k=1}^{q} u_{1k}(u_{1k}-1), \qquad u_2 = \frac{1}{2}\sum_{k=1}^{g} u_{2k}(u_{2k}-1),$$

q — количество связок в выборке $\mathbb{X}_n$, u_{1k} — размер k-ой связки выборки $\mathbb{X}_n$, g — количество связок в выборке $\mathbb{Y}_n$, u_{2k} — размер k-ой связки выборки $\mathbb{Y}_n$.

Можно показать, что

$$\mathsf{E}\widehat{\tau}_{XY}(n) = \tau_{XY}, \tag{12.8}$$

т. е. коэффициент $\widehat{\tau}_{XY}$ является несмещенной оценкой параметра τ_{XY}.

Кендалл доказал, что при справедливости гипотезы о независимости случайных величин X и Y

$$\mathsf{E}\widehat{\tau}_{XY} = 0, \quad \mathsf{D}\widehat{\tau}_{XY} = \frac{4n+10}{9n(n-1)},$$

а нормированный коэффициент согласованности

$$\frac{\widehat{\tau}_{XY}}{\sqrt{\mathsf{D}\widehat{\tau}_{XY}}} \approx \frac{3\sqrt{n}\,\widehat{\tau}_{XY}}{2} \tag{12.9}$$

асимптотически имеет стандартное нормальное распределение, то есть

$$\frac{3\sqrt{n}\widehat{\tau}_{XY}}{2}\Big|_{H_0} \underset{n\to\infty}{\sim} N(0,1).$$

Квантили распределения статистики $\widehat{\tau}_{XY}$ для $4 \leqslant n \leqslant 40$ при справедливости гипотезы о независимости случайных величин X и Y табулированы. Критические области уровня значимости α критерия Кендалла, основанного на статистике (12.9), соответствующие различным альтернативам H_A, приведены в табл. 12.2.

Таблица 12.2

H_A	**Критические области для $T_\tau(\mathbb{W}_n)$**
$\tau < 0$	$(-\infty; u_\alpha)$
$\tau > 0$	$(u_{1-\alpha}; +\infty)$
$\tau \neq 0$	$\left(-\infty; u_{\frac{\alpha}{2}}\right) \cup \left(u_{1-\frac{\alpha}{2}}; +\infty\right)$

где u_γ — квантиль уровня γ стандартного гауссовского распределения.

Критерий Кендалла является состоятельным только для альтернатив вида $H_1 : \tau < 0$, $H_2 : \tau > 0$ и $H_3 : \tau \neq 0$. Критерий Спирмена также является состоятельным только для альтернатив $H_1 : \tau < 0$, $H_2 : \tau > 0$ и $H_3 : \tau \neq 0$.

Замечание 12.2. Критерии Кендалла и Спирмена, вообще говоря, не являются состоятельными против альтернатив общего вида

$$H_1 : \exists x,y \in \mathbb{R}^1, \quad \text{такие что } F_W(x,y) \neq F_X(x)F_Y(y). \tag{12.10}$$

Важно отметить, что известно асимптотическое распределение коэффициента согласованности Кендалла $\widehat{\tau}_{XY}$ и в том случае, когда гипотеза о том, что X и Y независимы, неверна. А именно, Кендалл показал, что при нарушении гипотезы о независимости признаков X и Y, статистика $\widehat{\tau}_{XY}$ имеет асимптотически нормальное распределение с математическим ожиданием $\mathsf{E}\widehat{\tau}_{XY} = \tau$ и дисперсией

$$\mathsf{D}\widehat{\tau}_{XY} \leqslant \frac{2}{n}(1-\tau^2).$$

Это обстоятельство позволяет построить асимптотический доверительный интервал для параметра согласованности τ.

Пример 12.2. Изучается взаимосвязь способностей детей к математике и к музыке. Учитель математики и учитель музыки протестировали 10 детей и провели ранжировку способностей по своему предмету.

математика	1	2	3	4	5	6	7	8	9	10
музыка	6	5	1	4	2	7	8	10	3	9

Можно ли считать, что способности к математике и к музыке связаны?

Р е ш е н и е: Проверим гипотезу

$$H_0 : \tau_{XY} = 1 - 2\mathsf{P}\{(X_2 - X_1)(Y_2 - Y_1) < 0\} = 0$$

против альтернативной гипотезы

$$H_A : \tau_{XY} = 1 - 2\mathsf{P}\{(X_2 - X_1)(Y_2 - Y_1) < 0\} \neq 0.$$

Пусть уровень значимости равен 0.05. Вычислим коэффициент Спирмена

$$S = \sum_{i=1}^{10} (R_i - S_i)^2 = 25 + 9 + 4 + 0 + 9 + 1 + 1 + 4 + 36 + 1 = 90,$$

$$\hat{\rho}_S = 1 - \frac{6 \cdot 90}{10 \cdot (100 - 1)} = 0.45.$$

Предполагая независимость признаков, по таблице найдём квантиль уровня 0.975 статистики Спирмена, вычисленной по 10 наблюдениям,

$$\rho_{10;0.975} = 0.648.$$

Тогда

$$\rho_{10;0.025} = -0.648.$$

Доверительная область критерия имеет вид $(-0.648;0.648)$, а критическая область —

$$[-1;-0.648]\bigcup[0.648;1].$$

Реализация статистики Спирмена принадлежит доверительной области. Следовательно, на уровне значимости $\alpha = 0.05$ основная гипотеза не отвергается.

Решим эту задачу, используя критерий Кендалла. Вычислим количество несогласованных пар

$$K = 5+4+0+2+0+1+1+2+0+0 = 15.$$

Тогда оценка коэффициента согласованности

$$\hat{\tau} = 1 - \frac{4K}{n(n-1)} = 1 - \frac{4\cdot 15}{10\cdot 9} = \frac{1}{3}.$$

по таблице найдём квантиль уровня 0.975 статистики Кендалла, вычисленной по 10 наблюдениям,

$$\tau_{10;0.975} = 0.511.$$

Тогда

$$\tau_{10;0.025} = -0.511.$$

Доверительная область критерия имеет вид $(-0.511;0.511)$, а критическая область —

$$[-1;-0.511]\bigcup[0.511;1].$$

Реализация статистики Кендалла принадлежит доверительной области. Следовательно, на уровне значимости $\alpha = 0.05$ основная гипотеза не отвергается.

Пример 12.3. Имеются данные о ВВП в паритетах покупательной способности (показатель X) и коэффициенте младенческой смертности в промиллях (показатель Y) по 15 странам.

i	**Страна**	x_i	y_i	i	**Страна**	x_i	y_i
1	Мозамбик	3,0	113	9	Бразилия	20	44
2	Чад	2,6	117	10	Греция	43,4	8
3	Бангладеш	5,1	79	11	Респ. Корея	42,4	10
4	Индия	5,2	68	12	Италия	73,7	7
5	Египет	14,2	56	13	Канада	78,3	6
6	Белорусия	15,6	13	14	США	100	8
7	Польша	20	14	15	Швейцария	95,9	6
8	Мексика	23,7	33				

Применяя критерии Спирмена и Кендалла, выяснить являются ли показатели X и Y зависимыми. Уровень значимости выбрать равным 0,05.

Р е ш е н и е: Пусть выборка $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}^T$, $n = 15$, порождена двумерным случайным вектором $W = (X,Y)^T$, имеющим некоторое непрерывное распределение $F_W(x,y)$. Проверим гипотезу H_0 о независимости случайных величин X и Y.

Применим критерий Спирмена. Для вычисления статистики критерия составим таблицу рангов

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
r_i	2	1	3	4	5	6	7,5	9	7,5	11	10	12	13	15	14
s_i	14	15	13	12	11	7	8	9	10	4,5	6	3	1,5	4,5	1,5

в которой r_i — реализация ранга элемента x_i в выборке $\{X_1,\ldots,X_{15}\}^T$, а s_i — реализация ранга элемента y_i в выборке $\{Y_1,\ldots,Y_{15}\}^T$.

Реализация статистики Спирмена (12.1)

$$\widehat{\rho}_S = 1 - \frac{6[(2-14)^2+(1-15)^2+\ldots+(14-1{,}5)^2]}{15^3-15} =$$

$$= 1 - \frac{6[144+196+\ldots+156{,}25]}{3360} = 1 - \frac{6\cdot 1085{,}5}{3360} = -0{,}938.$$

Так как в выборках имеются связки, то при вычислении $\widehat{\rho}_S$ следует внести поправку (12.3). В реализации выборки соответствующей случайной величине X есть одна связка размера $u_{11}=2$, а в выборке, соответствующей случайной величине Y — две связки размера $u_{21}=2$ и $u_{22}=2$. Тогда

$$u_1=\frac{1}{12}(2^3-2)=0{,}5, \qquad u_2=\frac{1}{12}(2^3-2)2=1,$$

а

$$\widehat{\rho}_S=1-\frac{\sum\limits_{i=1}^{n}(r_i-s_i)^2}{\frac{1}{6}(n^3-n)-(u_1+u_2)}=1-\frac{1085{,}5}{\frac{1}{6}3360-1{,}5}=-0{,}944.$$

Если в качестве альтернативной гипотезы выбрать H_A: $\tau\neq 0$, то критическая область будет иметь вид

$$\left[-1;z_{\frac{\alpha}{2},n}\right)\cup\left(z_{1-\frac{\alpha}{2},n};1\right],$$

где $z_{\frac{\alpha}{2},n}$, $z_{1-\frac{\alpha}{2},n}$ — квантили уровня $\frac{\alpha}{2}$ и $1-\frac{\alpha}{2}$ распределения коэффициента ранговой корреляции Спирмена при справедливости гипотезы H_0 о независимости для выборки объема n. По таблицам находим

$$z_{0{,}975,15}=0{,}521$$

и учитывая, что

$$z_{1-\frac{\alpha}{2},n}=-z_{\frac{\alpha}{2},n},$$

имеем

$$z_{0{,}025,15}=-0{,}521.$$

Тогда реализация статистики

$$\widehat{\rho}_S=-0{,}944$$

попадает в критическую область

$$[-1;-0{,}521)\cup(0{,}521;1].$$

Таким образом, гипотеза о независимости случайных величин X и Y отвергается на уровне значимости 0,05 в пользу альтернативы

$$H_A:\tau_{XY}\neq 0$$

о том, что случайные величины X и Y зависимы.

Если исследователь имеет целью проверить гипотезу о независимости против альтернативы об отрицательной связи между случайными величинами X и Y, то в качестве альтернативной гипотезы следует выбрать гипотезу

$$H_A:\tau_{XY}<0.$$

Тогда критическая область будет иметь вид

$$[-1;z_{\alpha,n})=[-1;z_{0{,}05,15})=[-1;-0{,}443).$$

Реализация статистики $\widehat{\rho}_S$ также попадает в указанную критическую область. Следовательно, на уровне значимости 0,05 гипотеза H_0 отвергается и принимается

$$H_A:\tau_{XY}<0.$$

Применим теперь критерий Кендалла. Поскольку в наблюдениях имеются связки, то для вычисления коэффициента $\widehat{\tau}_{XY}$ надо воспользоваться формулой (12.7).

$$\widehat{\tau}_{XY}=\frac{\sum\limits_{1\leqslant i<j\leqslant n}\operatorname{sign}\{(X_i-X_j)(Y_i-Y_j)\}}{\sqrt{\frac{1}{2}n(n-1)-u_1}\sqrt{\frac{1}{2}n(n-1)-u_2}}.$$

Для удобства вычислений реализацию двумерной выборки $(x_1,y_1),\ldots,(x_{15},y_{15})$ запишем таким образом, чтобы $x_1\leqslant x_2\leqslant\ldots\leqslant x_{15}$. Получим

i	2	1	3	4	5	6	7	8	9	10	11	12	13	15	14
x_i	2,6	3,0	5,1	5,2	14,2	15,6	20	20	23,7	42,4	43,4	73,7	78,3	95,9	100
y_i	117	113	79	68	56	13	14	44	33	10	8	7	6	6	8

В силу того, что реализация выборки «иксов» упорядочена по возрастанию, имеем

$$\sum_{1\leqslant i<j\leqslant n}\operatorname{sign}((x_i-x_j)(y_i-y_j))=\sum_{i=1}^{14}\sum_{j=i+1}^{15}\operatorname{sign}((x_i-x_j)(y_i-y_j))=\sum_{i=1}^{14}\sum_{j=i+1}^{15}\operatorname{sign}((-1)(y_i-y_j)).$$

Если $i=1$, а $j=2$, то

$$\operatorname{sign}((-1)(y_1-y_2))=\operatorname{sign}((-1)(117-113))=-1.$$

Проведя аналогичные вычисления для $i=1$ и $j=3,\ldots,15$, получаем, что

$$\sum_{j=2}^{15}\operatorname{sign}(-1)(y_1-y_j))=-14,$$

а

$$\sum_{i=1}^{14}\sum_{j=i+1}^{15}\operatorname{sign}((-1)(y_i-y_j))=-14-13-12-11-10-3-5-7-6-5-3-1+1+1=-88.$$

В выборке иксов есть одна связка размера два, т. е. $u_{11}=2$, а в выборке игреков две связки размера два, т. е. $u_{21}=2$, $u_{22}=2$. Тогда

$$u_1=\frac{1}{2}2\cdot 1=1,\qquad u_2=\left(\frac{1}{2}2\cdot 1\right)2=2.$$

Тогда реализация коэффициента согласованности

$$\widehat{\tau}_{XY}=\frac{-88}{\sqrt{\frac{15\cdot 14}{2}-1}\sqrt{\frac{15\cdot 14}{2}-2}}=-0{,}85.$$

Если проверяется гипотеза H_0 вида (11.1) о независимости случайных величин X и Y против альтернативы

$$H_A:\tau<0,$$

то критическая область имеет вид $[-1;z_{\alpha,n})$, где $z_{\alpha,n}$ — квантиль уровня α распределения коэффициента согласованности Кендалла при справедливости H_0 для выборки объема n. По таблицам находим ближайшую к $\alpha=0{,}05$ квантиль

$$z_{0{,}046,15}=-0{,}33.$$

Таким образом, гипотеза о независимости случайных величин X и Y отвергается на уровне значимости $\alpha=0{,}046$ в пользу альтернативы

$$H_A:\tau_{XY}<0.$$

Пример 12.4. У 12 школьников изучались две характеристики: оценки IQ, определённые с помощью шкалы интеллекта Стенфорда-Бине (показатель X) и успеваемость по химии, оцененная на основе теста из 35 вопросов (показатель Y). Данные внесены в следующую таблицу

X	122	112	110	120	103	126	113	114	106	108	128	109
Y	31	25	19	24	17	28	18	20	16	15	27	21

Можно ли считать, что успеваемость по химии выше у школьников с более высоким IQ?

12.3 Сравнительные свойства критерия Спирмена и Кендалла.

Соотношение между коэффициентом Спирмена и коэффициентом Кендалла даётся следующей теоремой.

Теорема 12.1. *Пусть выборка $\mathbb{X}_n$ упорядочена в порядке возрастания, а $T_1,\ldots,T_n$ — ранги соответствующих элементов $Y_1,\ldots,Y_n$.*

Обозначим

$$I(t)=\begin{cases}1,\ t>0,\\ 0,\ t\leqslant 0.\end{cases}$$

Тогда

$$\widehat{\tau}_{XY}=1-\frac{4}{n(n-1)}\sum_{i<j}I(T_i-T_j),$$

а коэффициент Спирмена равен

$$\widehat{\rho}_S=1-\frac{12}{n(n^2-1)}\sum_{i<j}(j-i)\cdot I(T_i-T_j). \tag{12.11}$$

Из теоремы следует, что за исключением ситуаций $\widehat{\rho}_S=\widehat{\tau}_{XY}=1$ или $\widehat{\rho}_S=\widehat{\tau}_{XY}=-1$, коэффициенты $\widehat{\rho}_S$ и $\widehat{\tau}_{XY}$, вообще говоря, не равны. А именно, в коэффициенте корреляции Спирмена наблюдениям, отстоящим друг от друга на большое расстояние,приписывается больший вес, чем в коэффициенте Кендалла. Однако, при справедливости гипотезы о независимости случайных величин X и Y, эти статистики сильно коррелированы, а именно коэффициент корреляции случайных величин $\hat{\rho}_S$ и $\hat{\tau}$ есть

$$\rho(\hat{\rho}_S,\hat{\tau})=\frac{2(n+1)}{\sqrt{2n(2n+5)}}.$$

Минимальное значение этого коэффициента корреляции, равное 0.98, достигается при $n=5$.

Лекция 13

Проверка гипотезы о независимости двух случайных величин, измеренных в количественной шкале.

13.1 Критерий, основанный на выборочном коэффициенте корреляции

Пусть справедливо предположение о том, что вектор $W = (X,Y)^T$ — гауссовский, причем $\mathsf{D}X > 0$ и $\mathsf{D}Y > 0$.

Тогда гипотеза H_0 вида (11.1) о независимости случайных величин X и Y эквивалентна гипотезе

$$H_0: \quad \rho_{XY} = 0, \tag{13.1}$$

где

$$\rho_{XY} = \frac{k_{XY}}{\sqrt{\mathsf{D}X\mathsf{D}Y}}$$

— коэффициент корреляции случайных величин X и Y, а k_{XY} — их ковариация.

Эквивалентность (11.1) и (13.1) следует из того, что компоненты гауссовского вектора X и Y независимы тогда и только тогда, когда X и Y некоррелированы.

Оценкой неизвестного коэффициента корреляции ρ_{XY} является *выборочный коэффициент корреляции*

$$\widehat{\rho}_{XY} = \frac{\widehat{k}_{XY}}{S_X S_Y}, \tag{13.2}$$

где S_X^2, S_Y^2 — выборочные дисперсии, построенные по выборкам $\mathbb{X}_n$ и $\mathbb{Y}_n$ соответственно, а

$$\widehat{k}_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

— выборочная ковариация, построенная по двумерной выборке $\mathbb{W}_n$.

Можно показать, что при выполнении сделанного предположения и справедливости гипотезы H_0 вида (13.1), статистика

$$T(\mathbb{W}_n) = \frac{\sqrt{n-2}\,\widehat{\rho}_{XY}}{\sqrt{1-\widehat{\rho}_{XY}^2}} \tag{13.3}$$

имеет распределение Стьюдента с $l = n-2$ степенями свободы.

При $n \to \infty$ и справедливости H_0 статистика вида

$$\widetilde{T}(\mathbb{W}_n) = \sqrt{n}\,\widehat{\rho}_{XY} \tag{13.4}$$

асимптотически нормальна.

Критические области уровня значимости α для критериев, основанных на статистиках (13.3) и (13.4), приведены в таблице 8.1, где $t_\gamma(l)$, u_γ — квантили уровня γ распределений Стьюдента и $\mathcal{N}(0;1)$ соответственно.

Таблица 8.1

H_A	Критические области для $T(\mathbb{W}_n)$	Критические области для $\widetilde{T}(\mathbb{W}_n)$
$\rho_{XY} < 0$	$(-\infty; t_\alpha(n-2))$	$(-\infty; u_\alpha)$
$\rho_{XY} > 0$	$(t_{1-\alpha}(n-2); +\infty)$	$(u_{1-\alpha}; +\infty)$
$\rho_{XY} \neq 0$	$\left(-\infty; t_{\frac{\alpha}{2}}(n-2)\right) \cup \left(t_{1-\frac{\alpha}{2}}(n-2); +\infty\right)$	$\left(-\infty; u_{\frac{\alpha}{2}}\right) \cup \left(u_{1-\frac{\alpha}{2}}; +\infty\right)$

Важно отметить, что если вектор $W = (X,Y)^T$ — гауссовский, то утверждение о том, что случайные величины X и Y зависимы, справедливо тогда и только тогда, когда $\rho_{XY} \neq 0$. Таким образом, в рассматриваемом случае альтернативная гипотеза общего вида

$$H_A: \quad \exists x,y \in \mathbb{R}^1 \text{ такие, что } F_W(x,y) \neq F_X(x)F_Y(y) \tag{13.5}$$

эквивалентна гипотезе

$$H_1: \quad \rho_{XY} \neq 0.$$

Критерии, основанные на статистиках (13.3) и (13.4), в гауссовском случае являются состоятельными против альтернативы (13.5). Если же распределение вектора W отлично от гауссовского, то эти критерии состоятельны против альтернатив вида

$$H_1: \quad \rho_{XY} \neq 0,$$

$$H_2: \quad \rho_{XY} < 0,$$

$$H_3: \quad \rho_{XY} > 0,$$

означающих коррелированность случайных величин X и Y.

Если гипотеза H_0 отвергнута, т. е. случайные величины X и Y зависимы, то коэффициент корреляции ρ_{XY} может служить характеристикой силы связи между X и Y. Чтобы построить доверительный интервал для ρ_{XY}, необходимо знать распределение $\widehat{\rho}_{XY}$ при справедливости гипотезы

$$H_1: \quad \rho_{XY} \neq 0.$$

Известно, что в этом случае статистика $\widehat{\rho}_{XY}$ асимптотически нормальна с

$$\mathsf{E}\widehat{\rho}_{XY} = \rho_{XY}\left(1 - \frac{1-\rho_{XY}^2}{2n}\right) + O(n^{-2}),$$

$$\mathsf{D}\widehat{\rho}_{XY} = \frac{(1-\rho_{XY}^2)^2}{n} + O(n^{-2}).$$

Тот факт, что дисперсия статистики $\widehat{\rho}_{XY}$ явно зависит от неизвестного параметра ρ_{XY} снижает точность асимптотического доверительного интервала (см. пример 13.1). Более того, если абсолютное значение ρ_{XY} близко к единице, то границы асимптотического доверительного интервала могут оказаться меньше -1 или больше 1.

В связи с этим Фишером было построено z-преобразование коэффициента выборочной корреляции

$$\widehat{z} = \operatorname{arcth}\widehat{\rho}_{XY} = \frac{1}{2}\ln\left(\frac{1+\widehat{\rho}_{XY}}{1-\widehat{\rho}_{XY}}\right),$$

распределение которого приближается к нормальному быстрее, чем распределение самой статистики $\widehat{\rho}_{XY}$, а дисперсия $\widehat{z}$ не зависит от ρ_{XY}.

Доказано, что

$$\mathsf{E}\widehat{z} = \frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}} + \frac{\rho_{XY}}{2(n-1)} + O(n^{-2}), \quad \mathsf{D}\widehat{z} = \frac{1}{n-3} + O(n^{-2}).$$

Тогда доверительный интервал параметра $z = \operatorname{arcth}(\rho_{XY})$ уровня надежности $1-\alpha$ имеет вид

$$I_1 = \left[\operatorname{arcth}\widehat{\rho}_{XY} - \frac{\widehat{\rho}_{XY}}{2(n-1)} - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}; \quad \operatorname{arcth}\widehat{\rho}_{XY} - \frac{\widehat{\rho}_{XY}}{2(n-1)} + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right] = [z_1; z_2],$$

а искомый доверительный интервал для коэффициента корреляции ρ_{XY} уровня надежности $1-\alpha$ вид

$$I_2 = [\mathrm{th}\, z_1; \mathrm{th}\, z_2],$$

где $u_{1-\frac{\alpha}{2}}$ — квантиль распределения $\mathcal{N}(0;1)$ уровня $1-\frac{\alpha}{2}$,

$$\mathrm{th}\, x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

— тангенс гиперболический величины x.

Отметим, что слагаемым $\frac{\rho_{XY}}{2(n-1)}$ в формуле для $\mathsf{E}\widehat{z}$ можно, вообще говоря, пренебрегать.

13.2 Критерий хи-квадрат

Критерий предназначен для проверки гипотезы H_0 вида (11.1) против альтернативы H_A общего вида (13.5). Предположим сначала, что случайный вектор $W = (X,Y)$ является дискретным и первая компонента X принимает конечное множество значений $\{a_1,\ldots,a_m\}$, а вторая компонента Y конечное множество значений $\{b_1,\ldots,b_k\}$.

Обозначим через n_{ij}, $i = 1,\ldots,m$; $j = 1,\ldots,k$ случайное число пар $\{X_l,Y_l\}$ элементов двумерной выборки $\mathbb{W}_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, реализации $\{x_l,y_l\}$ которых равны $\{a_i,b_j\}$. Пусть также

$$n_{i\bullet} = \sum_{j=1}^{k} n_{ij}, \quad i = 1,\ldots,m, \qquad n_{\bullet j} = \sum_{i=1}^{m} n_{ij}, \quad j = 1,\ldots,k.$$

Понятно, что $n_{i\bullet}$ — число элементов выборки $\mathbb{X}_n$, принявших значение a_i $i = 1,\ldots,m$, а $n_{\bullet j}$ — число элементов выборки $\mathbb{Y}_n$, принявших значение b_j, $j = 1,\ldots,k$. Заметим также, что

$$\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij} = n$$

по построению.

Обозначим через через

$$\hat{p}_{ij} = \frac{n_{ij}}{n}, \qquad i = 1,\ldots,m, \quad j = 1,\ldots,k$$

частоту появления соответствующих значений $\{a_i,b_j\}$ в выборке $\mathbb{W}_n$.

Аналогично,

$$\hat{p}_{i\bullet} = \frac{n_{i\bullet}}{n},$$

— частота появления значения a_i, $i = 1,\ldots,m$, а

$$\hat{p}_{\bullet j} = \frac{n_{\bullet j}}{n}$$

— частота появления значения b_j, $j = 1,\ldots,k$.

Если компоненты вектора $W = (X,Y)^T$ имеют другую структуру, например,случайные величины X и Y являются непрерывными, то следует провести предварительную группировку данных. Для этого область V_X всех возможных значений случайных величин X разбивается на $m > 1$ непересекающихся интервалов

$$\{\Delta_{X,i}, \quad i = 1,\ldots,m\} \quad \text{так, что} \quad \bigcup_{i=1}^{m} \Delta_{X,i} = V_X.$$

Аналогично, разобьем область V_Y всех возможных значений случайных величин Y на $k > 1$ непересекающихся интервалов

$$\{\Delta_{Y,j}, \quad j = 1,\ldots,k\}, \quad \bigcup_{j=1}^{k} \Delta_{Y,j} = V_Y.$$

При такой структуре n_{ij} — это случайное число пар $\{X_l,Y_l\}$ элементов двумерной выборки $\{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, реализации $\{x_l,y_l\}$ которых попали в прямоугольник

$$\Delta_{ij} = \Delta_{X,i} \times \Delta_{Y,j}, \qquad i = 1,\ldots,m, \quad j = 1,\ldots,k, \quad l = 1,\ldots,n.$$

Тогда $n_{i\bullet}$ — число элементов выборки $\mathbb{X}_n$ реализации которых попали в $\Delta_{X,i}$, $i = 1,\ldots,m$, а $n_{\bullet j}$ — число элементов выборки $\mathbb{Y}_n$ реализации которых попали в $\Delta_{Y,j}$, $j = 1,\ldots,k$.

Статистика критерия хи-квадрат имеет вид

$$\widehat{\chi}^2 = T(\mathbb{W}_n) = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(\hat{p}_{ij} - \hat{p}_{i\bullet}\hat{p}_{\bullet j})^2}{\hat{p}_{i\bullet}\hat{p}_{\bullet j}} = n\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{n_{i\bullet}n_{\bullet j}}. \tag{13.6}$$

Статистику (13.6) принято называть *статистикой хи-квадрат или статистикой Пирсона*.

Согласно теореме Фишера–Пирсона асимптотическое распределение статистики (13.6) при справедливости гипотезы H_0 вида (11.1) есть распределение хи-квадрат с

$$r = (m-1)(k-1)$$

степенями свободы, т. е. при $n \to \infty$

$$\widehat{\chi}^2 \sim \chi^2(r).$$

Большие расхождения между частотами$\hat{p}_{ij}$ и соответствующими им произведениями частот $\hat{p}_{i\bullet}\,\hat{p}_{\bullet j}$ говорят о нарушении гипотезы H_0 вида (11.1). Таким образом, в пользу альтернативной гипотезы (13.5) свидетельствуют большие значения статистики (13.6).

Следовательно, критическая область уровеня значимости α для данного критерия имеет вид

$$(\chi^2_{1-\alpha}(r), +\infty),$$

где $\chi^2_{1-\alpha}(r)$ — квантиль уровня $1-\alpha$ распределения хи-квадрат с r степенями свободы.

Для удобства вычислений можно использовать другую форму записи статистики (13.6) вида

$$\widehat{\chi}^2 = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(n_{ij})^2}{n_{i\bullet}n_{\bullet j}} - 1\right). \tag{13.7}$$

Отметим, что критерий хи-квадрат является состоятельным для альтернатив общего вида (13.5).

Пример 13.1. Изучается зависимость между показателем X (ВВП в паритетах покупательной способности) и показателем Y (коэффициент младенческой смертности в промиллях). По имеющимся данным этих показателей для 30 стран была вычислена реализация коэффициента выборочной корреляции $\widehat{\rho}_{XY} = -0{,}734$. Предполагая, что наблюдаемая выборка $(X_1,Y_1),\ldots,(X_{30},Y_{30})$, порожденная случайным вектором $W = (X,Y)^T$, соответствует двумерному гауссовскому распределению, сделать вывод о зависимости (независимости) случайных величин X и Y. Построить асимптотический доверительный интервал уровня надежности 0,95 для коэффициента корреляции ρ_{XY} случайных величин X и Y. Построить также доверительный интервал уровня надежности 0,95 для ρ_{XY}, используя преобразование Фишера.

Р е ш е н и е: Если двумерная выборка $\{(X_i,Y_i),\, i = 1,\ldots,n\}$, где X_i — ВВП, а Y_i — коэффициент младенческой смертности в i-ой стране, порождена двумерным гауссовским вектором вектором $W = (X,Y)^T$, то гипотеза о независимости СВ X и Y вида (11.1) эквивалентна гипотезе вида (13.1) некоррелированности этих случайных величин. В данной задаче естественной альтернативой гипотезе H_0 вида (13.1) является гипотеза H_A: $\rho_{XY} < 0$ об отрицательной коррелированности показателей X и Y. Для проверки H_0 используется критерий, основанный на выборочном коэффициенте корреляции.

Статистика (13.3) этого критерия

$$T_r(\mathbb{W}_n) = \frac{\sqrt{n-2}\,\widehat{\rho}_{XY}}{\sqrt{1-\widehat{\rho}_{XY}^2}}$$

имеет при справедливости H_0 распределение Стьюдента с $n-2$ степенями свободы.

Реализация этой статистики

$$T_r(\mathbb{W}_n) = \frac{-0{,}734\sqrt{28}}{\sqrt{1-0{,}734^2}} = -5{,}72.$$

Критическая область критерия уровня значимости $\alpha = 0{,}05$ при альтернативе

$$H_A: \quad \rho_{XY} < 0$$

имеет вид

$$(-\infty; t_{0{,}05}(n-2)].$$

По таблицам находим квантиль уровня 0,05 распределения Стьюдента с $n-2=28$ степенями свободы

$$t_{0{,}05}(28) = -t_{0{,}95}(28) = -1{,}701.$$

Так как реализация статистики попадает в критическую область $(-\infty; -1{,}701]$, то на уровне значимости 0,05 гипотеза о независимости ВВП и коэффициента младенческой смертности отвергается в пользу того, что эти показатели отрицательно коррелированы.

Будем считать, что объем выборки $n = 30$ достаточно велик, и построим асимптотический доверительный интервал для коэффициента корреляции ρ_{XY} уровня надежности 0,95. Так как случайные величины $\widehat{\rho}_{XY}(n)$ имеет асимптотическое гауссовское распределение с параметрами

$$\mathsf{E}\widehat{\rho}_{XY}(n) = \rho_{XY} - \frac{\rho_{XY}(1-\rho_{XY}^2)}{2n}, \quad \mathsf{D}\widehat{\rho}_{XY}(n) = \frac{(1-\rho_{XY}^2)^2}{n},$$

то при больших n

$$\mathsf{P}\left\{-u_{0{,}975} < \frac{\widehat{\rho}_{XY} - \mathsf{E}\widehat{\rho}_{XY}}{\sqrt{\mathsf{D}\widehat{\rho}_{XY}}} < u_{0{,}975}\right\} = 0{,}95. \tag{13.8}$$

Поскольку $\mathsf{D}\widehat{\rho}_{XY}$ и смещение оценки ρ_{XY} равное $\frac{\rho_{XY}(1-\rho_{XY}^2)}{2n}$ зависят от неизвестного значения ρ_{XY}, то в формуле (13.8) значение ρ_{XY} в выражениях $\mathsf{D}\widehat{\rho}_{XY}$ и $\frac{\rho_{XY}(1-\rho_{XY}^2)}{2n}$ заменяется выборочным коэффициентом корреляции $\widehat{\rho}_{XY}$. Из-за этого обстоятельства построенный асимптотический доверительный интервал будет иметь надежность отличную от 0,95. Проведем преобразование в (13.8) так, чтобы получить интервал для величины r_{XY}:

$$\mathsf{P}\left\{-u_{0{,}975} < \frac{\widehat{\rho}_{XY} - \left(\rho_{XY} - \frac{\widehat{\rho}_{XY}(1-\widehat{\rho}_{XY}^2)}{2n}\right)}{\frac{1}{\sqrt{n}}\left(1-\widehat{\rho}_{XY}^2\right)} < u_{0{,}975}\right\} = 0{,}95;$$

$$\mathsf{P}\left\{\frac{-u_{0{,}975}(1-\widehat{\rho}_{XY}^2)}{\sqrt{n}} - \frac{\widehat{\rho}_{XY}(1-\widehat{\rho}_{XY}^2)}{2n} - \widehat{\rho}_{XY} < -\rho_{XY} < \right.$$

$$\left. < \frac{u_{0{,}975}(1-\widehat{\rho}_{XY}^2)}{\sqrt{n}} - \frac{\widehat{\rho}_{XY}(1-\widehat{\rho}_{XY}^2)}{2n} - \widehat{\rho}_{XY}\right\} = 0{,}95;$$

$$\mathsf{P}\left\{\widehat{\rho}_{XY} + \frac{\widehat{\rho}_{XY}(1-\widehat{\rho}_{XY}^2)}{2n} - \frac{u_{0{,}975}(1-\widehat{\rho}_{XY}^2)}{\sqrt{n}} < \rho_{XY} < \right.$$

$$\left. < \widehat{\rho}_{XY} + \frac{\widehat{\rho}_{XY}(1-\widehat{\rho}_{XY}^2)}{2n} + \frac{u_{0{,}975}(1-\widehat{\rho}_{XY}^2)}{\sqrt{n}}\right\} = 0{,}95.$$

Подставляя имеющуюся реализацию $\widehat{\rho}_{XY}$ в последнюю формулу, получим

$$\mathsf{P}\left\{-0{,}734 + \frac{-0{,}734(1-0{,}734^2)}{2\cdot 30} - \frac{1{,}96(1-0{,}734^2)}{\sqrt{30}} < \rho_{XY} < \right.$$

$$\left. < -0{,}734 + \frac{-0{,}734(1-0{,}734^2)}{2\cdot 30} + \frac{1{,}96(1-0{,}734^2)}{\sqrt{30}}\right\} = 0{,}95;$$

$$\mathsf{P}\{-0{,}734-0{,}006-0{,}165<\rho_{XY}<-0{,}734-0{,}006+0{,}165\}=0{,}95.$$

Итак,

$$\mathsf{P}\{-0{,}905<\rho_{XY}<-0{,}575\}=0{,}95.$$

Построим теперь асимптотический доверительный интервал надежности 0,95 для ρ_{XY}, используя z-преобразование Фишера. Асимптотический доверительный интервал надежности 0,95 для величины

$$z=\frac{1}{2}\ln\frac{1+\rho_{XY}}{1-\rho_{XY}}$$

имеет вид

$$\mathsf{P}\left\{\widehat{z}-\frac{\widehat{\rho}_{XY}(n)}{2(n-1)}-\frac{u_{0,975}}{\sqrt{n-3}}<z<\widehat{z}-\frac{\widehat{\rho}_{XY}(n)}{2(n-1)}+\frac{u_{0,975}}{\sqrt{n-3}}\right\}=0{,}95, \tag{13.9}$$

где

$$\widehat{z}=\frac{1}{2}\ln\frac{1+\widehat{\rho}_{XY}}{1-\widehat{\rho}_{XY}}.$$

Подставляя в формулу (13.9) реализацию статистики $\widehat{z}=-0{,}937$ и $\widehat{\rho}_{XY}(n)=-0{,}734$, получим

$$\mathsf{P}\{-0{,}937+0{,}013-0{,}377<z<-0{,}937+0{,}013+0{,}377\}=\mathsf{P}\{-1{,}301<z<-0{,}544\}=0{,}95.$$

Поскольку

$$\operatorname{th}z=\frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}=\rho_{XY},$$

то

$$\mathsf{P}\{\operatorname{th}(-1{,}301)<\rho_{XY}<\operatorname{th}(-0{,}544)\}=\mathsf{P}\{-0{,}860<\rho_{XY}<-0{,}496\}=0{,}95.$$

Таким образом, асимптотический доверительный интервал уровня надежности 0,95 для ρ_{XY} есть $(-0{,}860;-0{,}496)$.

Пример 13.2. Имеются сведения о возрасте X (в годах) и среднемесячной заработной плате Y (в тыс. рублей) 30 сотрудников некоторой организации.

X	19	20	22	24	28	30	31	32	34	37
Y	8,5	9,2	11,2	10,4	16,2	17,4	14,3	24,9	22,8	20,8
X	39	40	41	43	45	46	47	48	50	51
Y	34,1	30,4	28,8	33,8	35,3	34,4	32,3	29,4	31,8	30,3
X	53	54	55	58	60	62	65	68	70	72
Y	28,7	31,9	25,5	19,9	22,3	20,6	18,3	14,7	14,1	15,0

Применяя критерий Спирмена и критерий хи-квадрат проверить гипотезу о том, что возраст сотрудника и величина его заработной платы независимы. Уровень значимости считать равным 0,05. Прокомментировать полученные результаты.

Р е ш е н и е: Пусть данные об X и Y представляются двумерной выборкой $\mathbb{W}_n$ объема $n=30$ с элементами $(X_1,Y_1),\ldots,(X_n,Y_n)$, соответствующей неизвестному непрерывному распределению $F_W(x,y)$, где $W=(X,Y)^T$.

Проверим гипотезу H_0 вида (11.1) о независимости случайных величин X и Y, используя критерий Спирмена. Для того, чтобы вычислить коэффициент ранговой корреляции (12.1), составим таблицу рангов, в которой r_i — реализации рангов X_i в выборке $\{X_1,\ldots,X_n\}^T$, а s_i — реализации рангов Y_i в выборке $\{Y_1,\ldots,Y_n\}^T$, $i=1,\ldots,n$.

r_i	1	2	3	4	5	6	7	8	9	10
s_i	1	2	4	3	9	10	6	17	16	14
r_i	11	12	13	14	15	16	17	18	19	20
s_i	28	23	20	27	30	29	26	21	24	22
r_i	21	22	23	24	25	26	27	28	29	30
s_i	19	25	18	12	15	13	11	7	5	8

Вычислив

$$\sum_{i=1}^{30}(r_i - s_i)^2 = 0+0+1+1+4^2+\ldots+25^2+22^2 = 3540,$$

найдем реализацию статистики

$$\widehat{\rho}_S = 1 - \frac{6\cdot 3540}{30^3-30} = 0{,}291.$$

Критическая область критерия Спирмена уровня значимости $\alpha = 0{,}05$ при альтернативе

$$H_A: \quad \tau \neq 0$$

имеет вид

$$[-1;-0{,}362)\cup(0{,}362;1],$$

где значения $-0{,}362$ и $0{,}362$ — квантили уровня $0{,}025$ и $0{,}975$, соответственно, распределения коэффициента ранговой корреляции при справедливости гипотезы H_0. Так как реализация статистики $\widehat{\rho}_{XY}$ не попала в критическую область, то на уровне значимости $0{,}05$ принимается гипотеза о независимости случайных величин X и Y.

Применим критерий хи-квадрат. Разобьем множество значений случайных величин X на $m = 3$ непересекающихся интервала:

$$\Delta_{X,1} = [0;35), \quad \Delta_{X,2} = [35;55), \quad \Delta_{X,3} = [55;100],$$

а множество значений случайных величин Y на $k = 2$ интервала

$$\Delta_{Y,1} = [0;21000), \quad \Delta_{Y,2} = [21000;+\infty).$$

В прямоугольник $\Delta_{X,1}\times\Delta_{Y,1}$ попало 7 реализаций выборки $\mathbb{W}_{30}$, т. е. $n_{11} = 7$. Далее,

$$n_{12} = 2, \quad n_{21} = 1, \quad n_{22} = 12, \quad n_{31} = 6, \quad n_{32} = 2.$$

Соответственно,

$$n_{1\bullet} = \sum_{j=1}^{2} n_{1j} = 9,$$

$n_{2\bullet} = 13$, $n_{3\bullet} = 8$,

$$n_{\bullet 1} = \sum_{i=1}^{3} n_{i1} = 14,$$

$n_{\bullet 2} = 16$.

Вычислим реализацию статистики хи-квадрат вида (13.7)

$$\widehat{\chi}^2 = 30\left[\frac{7^2}{14\cdot 9}+\frac{2^2}{9\cdot 19}+\frac{1^2}{13\cdot 14}+\frac{12^2}{16\cdot 13}+\frac{6^2}{8\cdot 14}+\frac{2^2}{16\cdot 8}-1\right] = 14{,}03.$$

При справедливости гипотезы H_0 статистика (13.7) имеет распределение хи-квадрат с

$$r = (m-1)(k-1) = 2$$

степенями свободы. Критическая область уровня значимости $\alpha = 0{,}05$ имеет вид

$$(\chi^2_{0{,}95}(2);+\infty),$$

где $\chi^2_{0{,}95}(2) = 5{,}99$ — квантиль уровня $0{,}95$ распределения χ^2 с двумя степенями свободы. Таким образом, реализация статистики критерия попала в критическую область. Следовательно, на уровне значимости $0{,}05$ гипотеза о независимости случайных величин X и Y отвергается.

При применении критериев хи-квадрат и Спирмена для проверки гипотезы о независимости случайных величин X и Y, были получены разные выводы о справедливости этой гипотезы. Этот факт можно объяснить следующим образом.

Критерий хи-квадрат является состоятельным для альтернативы вида (13.5), т. е. этот критерий способен “улавливать” зависимость любого типа. Критерий Спирмена является состоятельным только для альтернатив вида

$$\tau \neq 0, \quad \tau < 0, \quad \tau > 0,$$

т. е. способен обнаружить лишь монотонную зависимость между случайными величинами. В данном примере монотонной зависимости между возрастом и величиной заработной платы человека нет. Однако представленные данные позволяют заметить следующие тенденции: самые высокие заработки имеют сотрудники среднего возраста, а относительно невысокие — сотрудники молодого и пожилого возраста.

Лекция 14

Исследование зависимости между несколькими случайными величинами

Как было показано ранее, коэффициент корреляции ρ_{XY} может быть использован в качестве меры, описывающей силу связи между двумя случайными величинами X и Y. Однако встречаются ситуации, когда коррелированность двух случайных величин является лишь отражением того факта, что обе они коррелированы с некоторой третьей величиной или совокупностью величин. В этом случае говорят о наличии "ложной" корреляции. Для того, чтобы выяснить является ли наблюдаемая коррелированность "ложной", требуется устранить влияние третьих величин. С этой целью рассматривают условное распределение двух случайных величин при фиксированных значениях третьих величин, и определяют частный коэффициент корреляции.

14.1 Частные коэффициенты корреляции

Пусть $\xi = \{\xi_1, \dots \xi_l\}^T$ — невырожденный гауссовский вектор с корреляционной матрицей $\mathbb{R}_\xi = \{\rho_{ij}\}$, $i,j = 1,\dots,l$, где ρ_{ij} — коэффициент корреляции между ξ_i и ξ_j.

Определение 14.1. *Частным коэффициентом корреляции СВ* ξ_1 *и* ξ_2 при фиксированных значениях $\xi_3,\dots,\xi_l$ называется

$$\rho_{12;3,\dots,l} = \frac{-\mathbb{R}_{12}}{\sqrt{\mathbb{R}_{11}\mathbb{R}_{22}}}, \tag{14.1}$$

где $\mathbb{R}_{ij}$ — алгебраическое дополнение элемента ρ_{ij} матрицы $\mathbb{R}_\xi$.

Аналогично определяются частные коэффициенты корреляции между любыми двумя компонентами вектора ξ при фиксированных значениях других компонент.

Нетрудно показать (см. пример 14.1), что при $l = 3$ частным коэффициентом корреляции СВ ξ_1 и ξ_2 при фиксированной ξ_3 будет

$$\rho_{12;3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}.$$

По различию между $\rho_{12;3,\dots,l}$ и $\rho_{\xi_1\xi_2}$ можно судить о том, зависимы ли ξ_1 и ξ_2 между собой или зависимость между ними есть следствие зависимости каждой из них от величин $\xi_3,\dots,\xi_l$. Так если корреляция между ξ_1 и ξ_2 уменьшается, когда зафиксированы $\xi_3,\dots\xi_l$, то это означает, что зависимость между ξ_1 и ξ_2 частично обусловлена воздействием величин $\xi_3,\dots\xi_l$. Если частная корреляция равна нулю или мала, то зависимость между ξ_1 и ξ_2 целиком обусловлена воздействием СВ $\xi_3,\dots\xi_l$. Наоборот, если частная корреляция больше парной $\rho_{\xi_1\xi_2}$, то величины $\xi_3,\dots\xi_l$ ослабляют связь между ξ_1 и ξ_2.

Пусть имеется l выборок

$$X_1 = \{X_{11},\dots,X_{n1}\}^T,\dots,X_l = = \{X_{1l},\dots,X_{nl}\}^T,$$

порожденных СВ $\xi_1,\ldots,\xi_l$ соответственно. Предполагается, что $X_1,\ldots,X_l$ получены в процессе совместного наблюдения за $\xi_1,\ldots\xi_l$, а именно: если x_{k1} — реализация X_{k1} в k-ом опыте, то $x_{k2},\ldots,x_{kl}$ — реализации соответственно $X_{k2},\ldots,X_{kl}$ в k-ом опыте, $k=1,\ldots,n$.

Обозначим через $\widehat{\rho}_{ij}$ — выборочный коэффициент корреляции СВ ξ_i и ξ_j, построенный по выборкам X_i и X_j, а через $\widehat{\mathbb{R}}_\xi$ — симметричную матрицу размера $l\times l$, элементами которой являются $\widehat{\rho}_{ij}$, $i,j=1,\ldots,l$, причем $\widehat{\rho}_{ii}=1$ для любого $i=1,\ldots,l$.

Оценками неизвестных частных коэффициентов корреляции служат выборочные частные коэффициенты корреляции, которые получаются путем замены в формуле (14.1), коэффициентов корреляции ρ_{ij} на соответствующие оценки $\widehat{\rho}_{ij}$.

Известно, что распределение выборочного частного коэфициента корреляции $\widehat{\rho}_{12;3,\ldots,l}$, вычисленного по выборке объема n, такое же, как у выборочного коэффициента корреляции $\widehat{\rho}_{12}$, основанного на $n-d$ наблюдениях, где $d=l-2$ — количество фиксированных переменных. Таким образом, имеется возможность для построения доверительных интервалов частного коэффициента корреляции и проверки гипотезы о равенстве его нулю.

14.2 Множественный коэффициент корреляции

Определение 14.2. *Множественым коэффициентом корреляции* между случайной величиной ξ_1 и совокупностью СВ $\xi_2,\ldots\xi_l$ называется величина

$$R_{1(2\ldots l)}=\sqrt{1-\frac{\det\mathbb{R}_\xi}{\mathbb{R}_{11}}}. \tag{14.2}$$

Можно показать, что $R^2_{1(2\ldots l)}$ выражается через частные коэффициенты корреляции следующим образом

$$R^2_{1(2\ldots l)}=1-\left(1-\rho^2_{12}\right)\left(1-\rho^2_{13;2}\right)\cdot\ldots\cdot\left(1-\rho^2_{1l;2,3,\ldots,l-1}\right). \tag{14.3}$$

Используя соотношение (14.3), покажем, что множественный коэффициент корреляции характеризует зависимость между компонентами вектора ξ. Обозначим через $I^{(j)}$ — подмножество элементов из множества $\{2,\ldots,l\}$, которое не содержит элемент j, а через $\rho_{1j;I^{(j)}}$ — частный коэффициент корреляции между ξ_1 и ξ_j при фиксированных $\xi_{i_1},\ldots,\xi_{i_m}$, где индексы $i_1,\ldots,i_m\in I^{(j)}$.

Если $R_{1(2\ldots l)}=0$, то, как следует из (14.3), для любого $j=2,\ldots,l$ и любого множества индексов $I^{(j)}$ частные коэффициеты $\rho_{1j;I^{(j)}}=0$. Последнее означает, что случайная величина ξ_1 некоррелирована со всеми остальными компонентами вектора ξ. Если же $R_{1(2\ldots l)}=1$, то по крайней мере один из частных коэффициентов корреляции $\rho_{1j;I^{(j)}}$ должен быть равен единице. Это означает, что случайная величина ξ_1 является линейной функцией от СВ $\xi_2,\ldots,\xi_l$.

Отметим также еще некоторые важные свойства множественного коэффициента корреляции:

1. $0\leqslant R_{1(2\ldots l)}\leqslant 1$;
2. $R_{1(2\ldots l)}\geqslant\left|\rho_{1j;I^{(j)}}\right|$;
3. $R^2_{1(2)}=\rho^2_{12}$;
4. $R^2_{1(2)}\leqslant R^2_{1(2,3)}\leqslant\ldots\leqslant R^2_{1(2\ldots l)}$.

Последнее свойство означает, что коэффициент множественной корреляции нельзя уменьшить путем расширения множества СВ, относительно которых измеряется зависимость случайной величины ξ_1.

Рассмотрим оценку $\widehat{R}_{1(2\ldots l)}$ неизвестного коэффициета $R_{1(2\ldots l)}$, которая получается заменой матрицы $\mathbb{R}_\xi$ в формуле (14.3) матрицей $\widehat{\mathbb{R}}_\xi$.

Фишер доказал, что при справедливости гипотезы

$$H_0 : R_{1(2\ldots l)} = 0 \tag{14.4}$$

статистика

$$T_n(X_1,\ldots,X_l) = \widehat{F} = \frac{\frac{1}{l-1}\widehat{R}^2_{1(2\ldots l)}}{\frac{1}{n-l}\left(1-\widehat{R}^2_{1(2\ldots l)}\right)} \tag{14.5}$$

имеет F-распределение $F(l-1;n-l)$.

Критическая область уровня значимости α критерия со статистикой (14.5) для проверки гипотезы (14.4) против альтернативы

$$H_1 : \quad R_{1(2\ldots l)} > 0$$

имеет вид

$$(f_{1-\alpha}(l-1;n-l);+\infty),$$

где $f_{1-\alpha}(l-1;n-l)$ — квантиль распределения $F(l-1;n-l)$ уровня $1-\alpha$.

14.3 Коэффициент конкордации Кендалла

Пусть имеется m ранжировок

$$\{R_{11},\ldots,R_{n1}\},\ldots,\{R_{1m},\ldots,R_{nm}\},$$

построенных по выборкам

$$X_1 = \{X_{11},\ldots,X_{n1}\}^T,\ldots,X_m = \{X_{1m},\ldots,X_{nm}\}^T,$$

порожденных непрерывными СВ $\xi_1,\ldots,\xi_m$ соответственно. Предполагается, что выборки получены в результате совместного наблюдения за СВ $\xi_1,\ldots,\xi_m$.

Требуется проверить гипотезу о независимости СВ $\xi_1,\ldots,\xi_m$ вида

$$H_0 : \quad F_\xi(x_1,\ldots,x_m) = F_{\xi_1}(x_1)\cdot\ldots\cdot F_{\xi_m}(x_m) \quad \forall x_1,\ldots,x_m \in \mathbb{R}^1, \tag{14.6}$$

где $F_\xi(x_1,\ldots,x_m)$ — функция распределения случайного вектора $\xi = (\xi_1,\ldots,\xi_m)^T$, а $F_{\xi_i}(x_i)$ — функции распределения СВ ξ_i, $i = 1,\ldots,m$.

Коэффициентом конкордации Кендалла называется

$$T_n(X_1,\ldots,X_m) = \widehat{W}_n(m) = \frac{12}{m^2(n^3-n)}\sum_{i=1}^{n}\left[\sum_{j=1}^{m}R_{ij} - \frac{m(n+1)}{2}\right]^2. \tag{14.7}$$

Для $3 \leqslant n \leqslant 7$ и $3 \leqslant m \leqslant 20$ есть таблицы квантилей распределения статистики $\widehat{W}_n(m)$ при справедливости гипотезы H_0 вида (14.6).

При справедливости H_0 вида (14.6) и достаточно большом m статистика

$$m(n-1)\widehat{W}_n(m) \tag{14.8}$$

имеет распределение хи-квадрат ${\chi_r}^2$ с $r = n-1$ степенями свободы. Критическая область уровня значимости α критерия со статистикой (14.8) имеет вид

$$(k_{1-\alpha}(n-1);+\infty).$$

Отметим свойства коэффициента конкордации:

1. $0 \leqslant \widehat{W}_n(m) \leqslant 1$;
2. $\widehat{W}_n(m) = 1$ тогда и только тогда, когда все m ранжировок совпадают;

3. Пусть $\overline{\rho}(m)$ — среднее арифметическое значений ранговых коэффициентов корреляции Спирмена, вычисленных по всем C_m^2 парам ранжировок выборок $X_1,\ldots,X_m$.

   Тогда $\overline{\rho}(m)=\frac{m\widehat{W}_n(m)-1}{m-1}$.

Критерий, основанный на коэффициенте конкордации Кендалла, используется в задачах исследования согласованности экспертной группы. Пусть имеется n объектов, характеризующихся некоторым качественным показателем, и группа из m экспертов, способных оценить данный показатель. Каждый из m экспертов производит ранжирование всех объектов по рассматриваемому показателю. Обозначим R_{ij} — ранг присвоенный i-му объекту j-м экспертом, $i=1,\ldots,n,\ j=1,\ldots,m$. Тогда коэффициент конкордации $\widehat{W}_n(m)$ может служить оценкой степени согласованности суждений экспертов. Действительно, согластно свойству 2), коэффициент $\widehat{W}_n(m)$ принимает свое максимальное значение в том случае, когда ранжировки всех экспертов совпадают. Если же различия между ранжировками экспертов велики, то суммы рангов, присвоенные каждому из n объектов $\sum_{j=1}^m R_{ij}$, $i=1,\ldots,n$ будут близки к среднему значению суммы рангов всех экспертов равному

$$\frac{1}{n}\left(m\sum_{i=1}^{n} i\right)=\frac{m(n+1)}{2},$$

а коэффициент $\widehat{W}_n(m)$ близок к нулю.

Таким образом, если реализация статистики $\widehat{W}_n(m)$ близка к нулю, то говорят, что суждения экспертов не характеризуется общностью предпочтений, т. е. экспертная группа рассогласована. Если же реализация $\widehat{W}_n(m)$ близка к единице, то это свидетельствует в пользу того, что экспертная группа обладает единой системой предпочтений, т. е. является согласованной.

Пример 14.1. В некоторой области Англии исследовалось влияние погоды на урожай. Рассматривалось три показателя: урожай сена в центнерах на акр (X_1), весеннее количество осадков в дюймах (X_2) и накопленная за весну температура выше 42°F (X_3). По данным двадцатилетних наблюдений были вычислены реализации выборочных коэффициентов корреляции:

$$\widehat{\rho}_{X_1X_2}=0{,}8,\quad \widehat{\rho}_{X_1X_3}=-0{,}4,\quad \widehat{\rho}_{X_2X_3}=-0{,}56.$$

Оценить частные коэффициенты корреляции $\rho_{12;3}$, $\rho_{13;2}$ и $\rho_{23;1}$. Прокомментировать полученный результат.

Р е ш е н и е: Пусть $\mathbb{R}_X$ — корреляционная матрица вектора $X=(X_1,X_2,X_3)^T$. Пользуясь определением 14.1, найдем частные коэффициенты корреляции $\rho_{12;3}$, $\rho_{12;2}$ и $\rho_{23;1}$:

$$\rho_{12;3}=\frac{-\mathbb{R}_{12}}{\sqrt{\mathbb{R}_{11}\mathbb{R}_{22}}}=\frac{-(\rho_{12}-\rho_{13}\rho_{23})(-1)^{1+2}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{13}^2)}}=\frac{\rho_{12}-\rho_{13}\rho_{23}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{13}^2)}},$$

$$\rho_{13;2}=\frac{-\mathbb{R}_{13}}{\sqrt{\mathbb{R}_{11}\mathbb{R}_{33}}}=\frac{-(\rho_{12}\rho_{23}-\rho_{13})(-1)^{1+3}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{12}^2)}}=\frac{\rho_{13}-\rho_{12}\rho_{23}}{\sqrt{(1-\rho_{23}^2)(1-\rho_{12}^2)}},$$

$$\rho_{23;1}=\frac{-\mathbb{R}_{23}}{\sqrt{\mathbb{R}_{22}\mathbb{R}_{33}}}=\frac{-(\rho_{23}-\rho_{12}\rho_{13})(-1)^{2+3}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}}=\frac{\rho_{23}-\rho_{12}\rho_{13}}{\sqrt{(1-\rho_{12}^2)(1-\rho_{13}^2)}}.$$

Заменяя в полученных формулах неизвестные коэффициенты корреляции ρ_{ij}, $1\leqslant i<j\leqslant 3$ их выборочными оценками $\widehat{\rho}_{ij}$, получаем оценки частных коэффициентов корреляции

$$\widehat{\rho}_{12;3}=\frac{\widehat{\rho}_{12}-\widehat{\rho}_{13}\widehat{\rho}_{23}}{\sqrt{(1-\widehat{\rho}_{13}^2)(1-\widehat{\rho}_{23}^2)}}=\frac{0{,}8-(-0{,}4)(-0{,}56)}{\sqrt{(1-0{,}4^2)(1-0{,}56^2)}}=0{,}759,$$

$$\widehat{\rho}_{13;2} = \frac{\widehat{\rho}_{13} - \widehat{\rho}_{12}\widehat{\rho}_{23}}{\sqrt{\left(1-\widehat{\rho}_{12}^2\right)\left(1-\widehat{\rho}_{23}^2\right)}} = \frac{-0{,}4 - 0{,}8(-0{,}56)}{\sqrt{(1-0{,}8^2)(1-0{,}56^2)}} = 0{,}097,$$

$$\widehat{\rho}_{23;1} = \frac{\widehat{\rho}_{23} - \widehat{\rho}_{12}\widehat{\rho}_{13}}{\sqrt{\left(1-\widehat{\rho}_{12}^2\right)\left(1-\widehat{\rho}_{13}^2\right)}} = \frac{-0{,}56 - 0{,}8(-0{,}4)}{\sqrt{(1-0{,}8^2)(1-0{,}4^2)}} = -0{,}436.$$

Оценки парных коэффициентов корреляции показывают, что урожайность X_1 и количество осадков X_2 имеют положительную корреляцию, а урожайность X_1 и накопленные температуры X_3 — отрицательную. Последнее можно интепретировать как неблагоприятное влияние высоких температур на урожай. Однако, оценка частного коэффициента корреляции $\widehat{\rho}_{13;2}$ между урожайностью и накопленными температурами при фиксированном количестве осадков оказывается положительной. Это означает, что существует положительная связь между урожаем и температурой, когда влияние осадков устранено. Отрицательное значение $\widehat{\rho}_{13}$ является следствием воздействия фактора осадков.

Пример 14.2. Для исследования зависимости между производительностью труда (показатель X_1), возрастом работника (показатель X_2) и производственным стажем (показатель X_3) была составлена случайная выборка из 100 рабочих одной специальности, и для каждого рабочего были указаны показатели X_1, X_2 и X_3. На основании этих данных были вычислены оценки парных коэффициентов корреляции: $\widehat{\rho}_{12} = 0.2$, $\widehat{\rho}_{13} = 0.41$, $\widehat{\rho}_{23} = 0.82$.

Вычислите частный коэффициент корреляции между производительностью труда и возрастом рабочего при условии, что стаж рабочего фиксирован. Проверьте гипотезу о том, что частный коэффициент корреляции $\rho_{12;3} = 0$. Прокомментируйте полученный результат.

Р е ш е н и е: Вычислим оценку частного коэффициента корреляции

$$\widehat{\rho}_{12;3} = \frac{\widehat{\rho}_{12} - \widehat{\rho}_{13}\widehat{\rho}_{23}}{\sqrt{\left(1-\widehat{\rho}_{13}^2\right)\left(1-\widehat{\rho}_{23}^2\right)}} = \frac{0{,}2 - (0{,}41)(0{,}82)}{\sqrt{(1-0{,}41^2)(1-0{,}86^2)}} = -0{,}26,$$

Проверим гипотезу $H_0 : \rho_{12;3} = 0$. В данной задаче естественной альтернативой гипотезе H_0 является гипотеза H_A: $\rho_{12;3} < 0$ об отрицательной коррелированности показателей X_1 и X_2 при фиксированном значении показателя X_3. Для проверки H_0 используется критерий, основанный на выборочном частном коэффициенте корреляции.

Статистика этого критерия имеет вид

$$T = \frac{\sqrt{n-2-d}\,\widehat{\rho}_{12;3}}{\sqrt{1-\widehat{r}_{12;3}^2}},$$

где d есть количество зафиксированных показателей. Последняя статистика имеет при справедливости H_0 распределение Стьюдента с $n-2-d$ степенями свободы. В данной задаче количество зафиксированных показателей d равно 1, так как зафиксирован только один показатель X_3. Вычисляя реализацию статистики, получим

$$t = \frac{\sqrt{97}(-0{,}26)}{\sqrt{1-(-0{,}26)^2}} = -2.65,$$

Пример 14.3. Автосалон предоставил сведения о продажной цене ξ_1, ширине ξ_2, длине ξ_3 и массе ξ_4 автомобиля. За последний месяц было продано 34 автомобиля. На основании этих данных вычислены выборочные коэффициенты корреляции:

$$\widehat{\rho}_{12} = 0{,}33, \quad \widehat{\rho}_{13} = 0{,}16, \quad \widehat{\rho}_{14} = 0{,}53, \quad \widehat{\rho}_{23} = 0{,}71, \quad \widehat{\rho}_{24} = 0{,}72, \quad \widehat{\rho}_{34} = 0{,}63.$$

Оценить множественный коэффициент корреляции $R_{1(2,3,4)}$ между продажной ценой автомобиля и совокупностью его трех технических характеристик, описывающих длину, высоту и массу. Проверить гипотезу о том, что $R_{1(2,3,4)} = 0$, предполагая, что данные имеют гауссовское распределение. Прокомментировать полученный результат.

Р е ш е н и е: Пусть выборки

$$X_1 = \{X_{11},\ldots,X_{n1}\}^T, \quad X_2 = \{X_{12},\ldots,X_{n2}\}^T, \quad X_3 = \{X_{13},\ldots,X_{n3}\}^T, \quad X_4 = \{X_{14},\ldots,X_{n4}\}^T$$

объема $n = 34$ порождены СВ ξ_1, ξ_2, ξ_3 и ξ_4 соответственно, которые являются компонентами гауссовского случайного вектора $\xi = (\xi_1,\ldots,\xi_4)^T$.

Оценкой множественного коэффициента корреляции между СВ ξ_1 и совокупностью ξ_2, ξ_3, ξ_4 является

$$\widehat{R}_{1(2\ldots4)} = \sqrt{1 - \frac{\det\widehat{\mathbb{R}}_\xi}{\widehat{\mathbb{R}}_{11}}},$$

где $\widehat{\mathbb{R}}_\xi$ — матрица составленная из выборочных коэффициентов корреляции $\widehat{r}_{ij}$, $i,j = 1,\ldots,4$, $\widehat{\mathbb{R}}_{11}$ — алгебраическое дополнение элемента $(1,1)$ матрицы $\widehat{\mathbb{R}}_\xi$.

По условию реализации этих величин имеют вид

$$\widehat{\mathbb{R}} = \begin{bmatrix} 1 & 0{,}33 & 0{,}16 & 0{,}53 \\ 0{,}33 & 1 & 0{,}71 & 0{,}72 \\ 0{,}016 & 0{,}71 & 1 & 0{,}63 \\ 0{,}53 & 0{,}72 & 0{,}63 & 1 \end{bmatrix}, \qquad \widehat{\mathbb{R}}_{11} = \begin{vmatrix} 1 & 0{,}71 & 0{,}72 \\ 0{,}71 & 1 & 0{,}63 \\ 0{,}72 & 0{,}63 & 1 \end{vmatrix} = 0{,}22,$$

а $\det\widehat{\mathbb{R}} = 0{,}15$.

Тогда

$$\widehat{R}^2_{1(2,3,4)} = 1 - \frac{0{,}15}{0{,}22} = 0{,}32,$$

а $\widehat{R}_{1(2,3,4)} = 0{,}57$.

Проверим гипотезу

$$H_0: \quad R_{1(2,3,4)} = 0.$$

При справедливости H_0 статистика (14.5) будет иметь F-распределение с $l-1=3$ и $n-l=30$ степенями свободы. Критическая область уровня значимости $\alpha = 0{,}05$ критерия со статистикой (14.5) имеет вид

$$(2{,}922;+\infty),$$

где 2,922 есть квантиль уровня 0,95 распределения $F(3;30)$.

Реализация статистики (14.5)

$$\widehat{F} = \frac{\frac{1}{3}0{,}32}{\frac{1}{30}(1-0{,}32)} = 4{,}706.$$

попадает в критическую область. Следовательно, на уровне значимости 0,05 гипотеза H_0 отвергается. То есть, можно считать, что имеется статистическая связь между продажной ценой автомобиля и совокупностью таких его технических характеристик, как длина, ширина и масса.

Пример 14.4. Три квалифицированных эксперта (A, B и C) проранжировали в порядке предпочтения семь представленных бизнес-проектов. Результаты представлены в таблице.

	1	2	3	4	5	6	7
A	1	4	2	5	3	7	6
B	2	1	3	4	5	6	7
C	2	1	4	5	3	7	6

Можно ли считать, что данная экспертная группа обладает общей системой предпочтений?

Р е ш е н и е: Пусть случайная величина ξ_j, $j = 1,2,3$ — оценка качества представленных бизнес-проектов, согласно оценке j-го эксперта, а

$$\{R_{11},\ldots,R_{n1}\}, \quad \{R_{12},\ldots,R_{n2}\}, \quad \{R_{13},\ldots,R_{n3}\}, \quad n = 7$$

— ранжировки выборок

$$X_1 = \{X_{11},\ldots,X_{n1}\}^T, \quad X_2 = \{X_{12},\ldots,X_{n2}\}^T, \quad X_1 = \{X_{13},\ldots,X_{n3}\}^T,$$

порожденных СВ ξ_1, ξ_2 и ξ_3.

Из справедливости гипотезы вида (14.6) о независимости СВ ξ_1, ξ_2 и ξ_3 будет следовать, в частности, тот факт, что при каждой ранжировке выборки X_1 вероятность появления любого набора рангов выборки X_2 или выборки X_3 будет равна $\frac{1}{n!}$. То есть, мнения экспертов рассогласованы.

Проверим гипотезу вида (14.6), используя критерий основанный на коэффициенте конкордации Кендалла (14.7)

$$\widehat{W}_n(m) = \frac{12}{m^2(n^3-n)} \sum_{i=1}^{n} \left[\sum_{j=1}^{m} R_{ij} - \frac{m(n+1)}{2} \right]^2.$$

Для вычисления реализации статистики $\widehat{W}_n(m)$ вычислим сначала реализации $r_{i\bullet}$ статистик

$$\sum_{j=1}^{m} R_{ij}, \qquad m = 3, \quad i = 1,\ldots,7.$$

Имеем

$$r_{1\bullet} = \sum_{j=1}^{3} r_{ij} = 1+2+2 = 5,$$

$$r_{2\bullet} = 4+1+1 = 6,$$

$$r_{3\bullet} = 2+3+4 = 9,$$

$$r_{4\bullet} = 14, \quad r_{5\bullet} = 11, \quad r_{6\bullet} = 20, \quad r_{7\bullet} = 19.$$

Тогда реализация коэффициента конкордации

$$\widehat{W}_n(3) = \frac{12}{3^2(7^3-7)} \left((5-12)^2 + \ldots + (19-12)^2 \right) = \frac{12}{9(7^3-7)} \cdot 212 = 0{,}84.$$

Реализация коэффициента конкордации близка к единице. Однако для проверки гипотезы (14.6) следует указать критическую область. В книге Кобзарь А.И. «Прикладная математическая статистика« (М.:Физматлит, 2006) представлены квантили уровня 0,95 и 0,99 статистики

$$S = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} R_{ij} - \frac{m(n+1)}{2} \right)^2$$

при справедливости гипотезы H_0 вида (14.6). Так для $n = 7$ и $m = 3$ квантиль уровня 0,95 статистики S равна 157,3.

Реализация статистики S, равная 212, попадает в критическую область. Следовательно, гипотеза о независимости СВ ξ_1, ξ_2, ξ_3 отвергается на уровне значимости 0,05. Таким образом, можно считать на уровне значимости 0,05, что группа экспертов согласована, т. е. обладает единой системой предпочтений.

Лекция 15

Линейный регрессионный анализ

Пусть имеется переменная Y, описывающая некоторый экономический, технический или какой-либо показатель. Предполагается, что Y зависит от некоторых величин $X_1,\ldots,X_p$,то есть изменение величин $X_1,\ldots,X_p$ вызывает вполне определённое изменение Y. В таких случаях переменную Y называют откликом или зависимой переменной, а переменные $X_1,\ldots,X_p$ — факторами, влияющими на отклик, или регрессорами или объясняющими независимыми переменными. Построение или восстановление функциональных зависимостей между откликом и факторами относят к задачам регрессионного анализа.

Формализуем эту задачу. Пусть проведено n измерений, и при каждом заданном значении факторных переменных $X_{i1}\ldots,X_{ip}$, $i=1,\ldots,n$ получено значение отклика Y_i. Пусть переменные Y и $X_1,\ldots,X_p$ связаны соотношением

$$Y_i = f(X_{i1},\ldots X_{ip})+\varepsilon_i, \qquad i=1,\ldots,n, \tag{15.1}$$

где ε_i — некоторая ненаблюдаемая случайная величина. Случайное слагаемое ε_i выражает либо внутренне присущую отклику изменчивость, либо влияние на него факторов, не учтённых в соотношении (15.1), либо и то и другое вместе. Иногда ε_i называют ошибкой эксперимента, связывая её присутствие с несовершенством метода измерения Y.

Предполагается, что случайные величины $\varepsilon_1\ldots,\varepsilon_n$ статистически независимы, одинаково распределены и имеют нулевое математическое ожидание.

15.1 Предположения о регрессионной функции

Для того, чтобы задача о подборе функции f была осмысленной, следует определить набор допустимых функций $f(X)$. Как правило предполагается, что множество допустимых функций является параметрическим семейством$f(X,\theta)$, где $\theta\in\Theta\subset R^{p+1}$ — параметр семейства. Тогда соотношение (15.1) можно переписать в виде

$$Y_i = f(X_{i1},\ldots X_{ip},\theta)+\varepsilon_i, \qquad i=1,\ldots,n. \tag{15.2}$$

Теперь задача восстановления зависимости между Y и $X_1,\ldots,X_p$ эквивалентна задаче оценивания неизвестных коэффициентов $\theta\in\Theta\subset\mathbb{R}^{p+1}$. Если функция $f(X,\theta)$ линейна по θ, то задача оценивания называется задачей линейного регрессионного анализа. Приведём несколько наиболее распространённых примеров функций $f(X,\theta)$ линейных по θ:

1) $f(X_1,...,X_p,\theta)=\theta_0+\sum_{i=1}^{p}\theta_iX_i$. В частности, модель $f(X,\theta)=\theta_0+\theta_1X$ называется простой парной линейной моделью;

2) $f(X_1,...,X_p,\theta)=\theta_0+\sum_{i=1}^{p}\theta_iX_i+\sum_{i=1}^{p}\theta_{i+p}X_i^2$;

3) модель вида $f(X_1,...,X_p,\theta)=\theta_0X_1^{\theta_1}\cdots X_p^{\theta_p}$ выглядит как нелинейная, однако её можно линеаризовать. Прологарифмировав правую и левую части, получим функцию $\ln f(X_1,...X_p,\theta)=\ln\theta_0+\theta_1\ln X_1+\cdots+\theta_p\ln X_p$, которая уже является линейной по θ.

Пример нелинейной по θ функции: $f(X_1,X_2,\theta)=\frac{\theta_0}{\theta_1X_1+\theta_2X_2}$.

Далее будем обсуждать только задачу линейного регрессионного анализа.

Выбор конкретного вида функциональной зависимости $f(X,\theta)$ основан на следующих соображениях:

1) физические соображения;

2) геометрические соображения;

3) использование нескольких подходящих, по мнению исследователя, моделей с последующей проверкой этих моделей на адекватность статистическими методами;

4) аппроксимировать функцию полиномом достаточно высокой степени,а затем правильность выбора порядка полинома проверить соответствующими статистическими методами.

15.2 Метод наименьших квадратов (МНК)

Существуют различные методы оценивания параметров в регрессионной модели. Одним из наиболее известных и распространённых является *метод наименьших квадратов (МНК)*. Идея этого метода основана на минимизации функции потерь $S(\theta)$ вида

$$S(\theta) = \sum_{i=1}^{n} (Y_i - f(X_{i1}, \dots X_{ip}, \theta))^2 \to \min_{\theta}. \tag{15.3}$$

Определение 15.1. Вектор $\hat{\theta}$ называется *оценкой метода наименьших квадратов (МНК-оценкой)* вектора параметров θ в регрессионной модели вида (15.2), если

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (Y_i - f(X_{i1}, \dots X_{ip}, \theta))^2. \tag{15.4}$$

Запишем уравнение регрессии в матричном виде. Введём следующие обозначения

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_p \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & & & \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Тогда уравнение регрессии примет вид

$$Y = X\Theta + \varepsilon. \tag{15.5}$$

Будем предполагать, что справедливы следующие предположения:

$$\mathsf{E}\varepsilon = 0, \qquad K_\varepsilon = \sigma^2 I.$$

и

$$\det(X^T X)^{-1} \neq 0.$$

Если последнее предположение справедливо, то то задача нахождения минимума по θ для функции

$$S(\theta) = (Y - X\Theta)^T (Y - X\Theta)$$

имеет единственное решение

$$\hat{\theta} = (X^T X)^{-1} X^T Y. \tag{15.6}$$

В частности, если рассматривать простую линейную регрессию вида

$$Y_i = \theta_0 + \theta_1 X_i + \varepsilon_i, \qquad i = 1, \dots, n, \tag{15.7}$$

то МНК-оценка для параметров θ_0 и θ_1 имеет вид

$$\hat{\theta}_0 = \overline{Y} - \hat{\theta}_1 \overline{X}, \qquad \hat{\theta}_1 = \frac{\sum\limits_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2} = \hat{\rho}_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}. \tag{15.8}$$

Неизвестную дисперсию случайных погрешностей можно оценить следующим образом:

$$\hat{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - f(X_{i1},\dots X_{ip},\hat{\theta}))^2. \tag{15.9}$$

Обозначим $\hat{Y}_i = f(X_{i1},\dots X_{ip},\hat{\theta})$ оценку значения переменной Y в i-ом опыте, а $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ — оценку ошибки $\hat{\varepsilon}_i$ в i-ом опыте.

Нетрудно проверить, что общую (total) сумму квадратов (SS) отклонений Y_i от их выборочного среднего $\overline{Y}$ представляется в виде

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{SS_{\text{общ., total=TSS}}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{SS_{\text{ост. = residual=RSS}}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}_{SS_{\text{регр.=explained=}ESS}}$$

Таким образом разброс зависимой переменной Y около выборочного среднего $\overline{Y}$ равен сумме разброса, «объясняемого регрессией» $SS_{\text{регр.}} = ESS$, и разброса, который объяснить регрессией не удалось $SS_{\text{ост.}} = RSS$.

Определение 15.2. Величина

$$\hat{R}^2_{YX} = 1 - \frac{\sum\limits_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2}.$$

называется множественным коэффициентом детерминации регрессии порядка $p \geqslant 1$.

Коэффициент детерминации показывает в какой степени рассматриваемая регрессионная модель порядка $p \geqslant 1$ лучше объясняет величину Y, чем тривиальная модель $Y = \theta_0 + \varepsilon$, (то есть Y не зависит от объясняющих переменных). Значения $\hat{R}^2_{YX}$ близкие к 1 говорят о том, что рассматриваемая регрессия достаточно хорошо объясняет изменения переменной Y. Наоборот, значения $\hat{R}^2_{YX}$ близкие к нулю свидетельствуют о том, что регерессионная модель плохо описывает переменную Y. Заметим также, что близость коэффициента детерминации к 1 ещё не означает, что модель правильно описывает зависимость между Y и $X_1,\dots,X_p$.

Чтобы коэффициент детерминации не возрастал с увеличением количества регрессоров p, рассматривают скорректированный несмещённый коэффициент детерминации

$$\hat{R}^2_{\text{скор.}} = 1 - \frac{\frac{1}{n-p-1}\sum\limits_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n-1}\sum\limits_{i=1}^{n}(Y_i - \overline{Y})^2}.$$

Нетрудно показать, что скорректированный и нескорректированный коэффициент R^2 связанны следующим соотношением

$$\hat{R}^2_{\text{скор.}} = 1 - (1 - \hat{R}^2)\frac{n-1}{n-p-1}.$$

Лекция 16

Статистические свойства МНК-оценок

1. *МНК-оценка является несмещенной, то есть* $E\hat{\theta} = \theta$

   Действительно,

$$E((X^TX)^{-1}X^T(X\theta+\varepsilon)) = (X^TX)^{-1}(X^TX)\theta + E\varepsilon = \theta.$$

2. *Ковариационная матрица МНК-оценки имеет вид*

$$K_{\hat{\theta}} = \sigma_\varepsilon^2 (X^TX)^{-1}.$$

   Покажем это:

$$E(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T = E((X^TX)^{-1}X^T\varepsilon)((X^TX)^{-1}X^T\varepsilon)^T = \\ = E(X^TX)^{-1}\varepsilon\varepsilon^TX\left((X^TX)^{-1}\right)^T = (X^TX)^{-1}E\varepsilon\varepsilon^T = \sigma_\varepsilon^2(X^TX)^{-1}.$$

   В частности, для модели простой парной регрессии вида (15.7) имеем, что

$$D\hat{\theta}_0 = \sigma_\varepsilon^2 \frac{\sum_{i=1}^n X_i^2}{n\sum_{i=1}^n (X_i-\overline{X})^2}; D\hat{\theta}_1 = \sigma_\varepsilon^2 \frac{1}{\sum_{i=1}^n (X_i-\overline{X})^2}.$$

3. *Оценка дисперсии* $\hat{\sigma_\varepsilon^2} = \dfrac{1}{n-(p+1)}\sum_{i=1}^n (Y_i - f(X_{i1},\dots X_{ip},\hat{\theta}))^2$ *является несмещённой, то есть* $M\hat{\sigma_\varepsilon^2} = \sigma_\varepsilon^2$.

4. *При большом количестве наблюдений асимптотическое распределение МНК оценки имеет вид*

$$\hat{\theta} \sim N\left(\theta, \sigma_\varepsilon^2(X^TX)^{-1}\right).$$

5. ***Теорема Гаусса-Маркова.*** *Пусть для модели (15.5) выполнены предположения:*

$$\mathsf{E}\varepsilon = 0, \qquad K_\varepsilon = \sigma^2 I$$

   и детерминированная матрица X *имеет максимальный ранг* p*. Тогда МНК-оценка* $\hat{\theta} = (X^TX)^{-1}X^TY$ *является наилучшей (в смысле наименьшей дисперсии) оценкой в классе линейных (по* Y*) несмещённых оценок.*

Замечание 16.1. В англоязычной литературе такие оценки имеют аббревиатуру BLUE (Best Linear Unbiased Estimator).

Замечание 16.2. Если оценивается вектор параметров θ, то наилучшей в смысле наименьшей дисперсии оценкой называют такую оценку $\hat{\theta}$, для которой матрица $K_{\tilde{\theta}} - K_{\hat{\theta}}$ будет неотрицательно определённой при любых несмещённых оценках $\tilde{\theta}$.

Лекция 17

Статистические свойства МНК-оценок в гауссовских моделях

Дополнительные свойства при предположении о гауссовости шумов, т.е. $\varepsilon \sim N(0,\sigma^2 I)$

1. $\hat{\theta} \sim N\left(\theta, \sigma_\varepsilon^2 (X^T X)^{-1}\right)$
2. $\frac{\Sigma \hat{\varepsilon_i^2}}{\sigma^2} \sim \chi^2(n-(p+1))$
3. Если c — неслучайный вектор размерности $p+1$, то
$$D(c^T \hat{\theta}) \leqslant D(c^T \widetilde{\theta}),$$
где $\widetilde{\theta}$ — любая несмещенная оценка θ.
4. МНК оценка совпадает с ОМП.

При справедливости предположения о гауссовости шумов возможно решить следующие задачи проверки адекватности регресионной модели

1) Проверка значимости уравнения множественной регрессии.

Пусть
$$Y_i = \theta_0 + \theta_1 X_{i1} + \ldots + \theta_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n.$$
Проверим
$$H_0 : \theta_1 = \ldots = \theta_p = 0, \qquad (\text{т.е. } Y_i = \theta_0 + \varepsilon_i)$$
против
$$H_A : \exists i : \theta_i \neq 0.$$
Пусть $\hat{\theta}_0, \ldots, \hat{\theta}_p$ — МНК-оценки соответствующих параметров регрессии, обозначим
$$\hat{Y}_i = \hat{\theta}_0 + \sum_{l=1}^{p} X_{il} \cdot \hat{\theta}_l.$$
Представив общую (total) сумму квадратов (SS) отклонений Y_i от их выборочного среднего $\overline{Y}$ в виде
$$\underbrace{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}_{SS_{\text{общ., total=TSS}}} = \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{SS_{\text{ост. = residual=RSS}}} + \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2}_{SS_{\text{регр.=explained}}}$$
и разделив все части равенства на σ^2, получим, что
$$\frac{SS_{\text{общ.}}}{\sigma^2} \sim \chi^2(n-1),$$

$$\frac{SS_{\text{сл.}}}{\sigma^2} \sim \chi^2(n-p-1),$$

а при справедливости гипотезы H_0 и

$$\frac{SS_{\text{регр.}}}{\sigma^2} \sim \chi^2(p).$$

Заметим, что число степеней свободы в правой части совпадает с числом степеней свободы в левой части:

$$(n-1) = (n-p-1) + p.$$

Тогда при справедливости гипотезы H_0 статистика

$$\hat{F} = \left. \frac{\frac{1}{p} SS_{\text{регр.}}}{\frac{1}{n-p-1} SS_{\text{случ.}}} \right|_{H_0} \sim F(p, n-p-1)$$

Можно показать, что F-статистика выражается и через коэффициент детерминации следующим образом

$$\hat{F} = \left. \frac{\frac{1}{p} \hat{R}^2}{\frac{1}{n-p-1}(1-\hat{R}^2)} \right|_{H_0} \sim F(p, n-p-1).$$

Если реализация F-статистики попадает в доверительную область, то гипотеза H_0 принимается на уровне значимости α, и рассматриваемую регрессионную модель следует признать незначимой. Если реализация F-статистики попадает в критическую область, то гипотеза H_0 отвергается на уровне значимости α,то есть хотя бы один из регрессионных параметров $\theta_1, \ldots, \theta_p$ отличен от нуля.

2) Проверка значимости коэффицентов регрессии (построение доверительного интервала параметров регрессии).

Проверим гипотезу

$$H_0 : \theta_{k-1} = 0,$$

где k любое целое число из множества $\{1, \ldots, p+1\}$ против альтернативы

$$H_A : \theta_{k-1} \neq 0.$$

Согласно свойству МНК-оценки

$$\hat{\theta} \sim N\left(\theta; \sigma_\varepsilon^2 (X^T X)^{-1}\right).$$

Тогда

$$\frac{\hat{\theta}_{k-1} - \theta_{k-1}}{\sigma_\varepsilon \sqrt{C_{kk}}} \sim N(0,1),$$

где C_{kk} — элемент (k,k) матрицы $(X^T X)^{-1}$. Так как σ_ε^2 неизвестно, то его оценивают:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum\limits_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-p-1}.$$

Тогда t-статистика

$$\frac{\hat{\theta}_{k-1} - \theta_{k-1}}{\hat{\sigma}_\varepsilon \sqrt{C_{kk}}} \sim t(n-p-1)$$

имеет распределение Стьюдента с $n-p-1$ степенями свободы. И при справедливости гипотезы $H_0 : \theta_{k-1} = 0$ t-статистика имеет вид

$$\frac{\hat{\theta}_{k-1}}{\hat{\sigma}_\varepsilon \sqrt{C_{kk}}}. \tag{17.1}$$

Следовательно, с вероятностью $1-\alpha$ при справедливости H_0

$$\left|\frac{\hat{\theta}_{k-1}}{\hat{\sigma}_\varepsilon\sqrt{C_{kk}}}\right| < t_{1-\alpha/2,n-p-1},$$

где $t_{1-\alpha/2,n-p-1}$-квантиль уровня $1-\alpha/2$ распределения Стьюдента с $n-p-1$ степенями свободы.

3) Рассмотим две модели линейной регрессии.

1. Первая модель имеет вид

$$Y_i = \theta_0 + \theta_1 X_{i1} + \ldots + \theta_p X_{ip} + \varepsilon_i, \qquad i = 1,\ldots,n.$$

Пусть $\hat{\theta}_0,\ldots,\hat{\theta}_p$ — МНК-оценки параметров $\theta_0,\ldots,\theta_p$. Найдем остаточную сумму квадратов

$$S_1^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

2. Вторая (“урезанная”) модель имеет вид

$$Y_i = \theta_0 + \theta_1 X_{i1} + \ldots + \theta_q X_{iq} + \varepsilon_i, \qquad q < p.$$

Пусть $\hat{\theta}_0,\ldots,\hat{\theta}_q$ — МНК-оценки параметров $\theta_0,\ldots,\theta_q$ (заметим, что эти оценки могут не совпадать с оценками $\hat{\theta}_0,\ldots,\hat{\theta}_q$ первой модели). Найдем остаточную сумму квадратов

$$S_2^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{q}\hat{\theta}_j X_{ij})^2.$$

Проверим гипотезу

$$H_0 : \theta_{q+1} = \ldots = \theta_p = 0.$$

При справедливости гипотезы H_0 F-статистика вида

$$F = \left.\frac{\frac{1}{p-q}(S_2^2 - S_1^2)}{\frac{1}{n-p-1}S_1^2}\right|_{H_0} \sim F(p-q,n-p-1)$$

имеет распределение Фишера с $p-q$ и $n-p-1$ степенями свободы.

Если реализация F-статистики попадает в доверительную область, то на уровне значимости α принимается гипотеза H_0, т.е. более адекватной является вторая (урезанная) модель.

Этот критерий, в частности, позволяет для зависимости

$$Y_i = \theta_0 + \theta_1 X_i + \ldots + \theta_p X_i^p + \varepsilon_i$$

наилучшим образом подобрать степень полинома.

4) Построение доверительного интервала для линии регрессии в простой парной модели.

МНК-оценкой регрессионной функции в модели (15.7) является

$$\hat{f}(X_0) = \hat{\theta}_0 + \hat{\theta}_1 X_0$$

. Нетрудно показать, что математическое ожидание функции $f(X)$ в точке X_0 равно

$$E\hat{f}(X_0) = \theta_0 + \theta_1 X_0,$$

а дисперсия

$$D\hat{f}(X_0) = D(\bar{Y} - \hat{\theta}_1\bar{X} - \hat{\theta}_1 X_0) =$$

$$= \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2} \right).$$

Дисперсия погрешностей σ_ε^2 неизвестна, но её можно оценить (см. формулу (15.9)). Тогда центральная статистика

$$\frac{f(X_0) - \hat{\theta}_0 - \hat{\theta}_1 X_0}{\hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}}}.$$

будет иметь распределение Стьюдента с $n-2$ степенями свободы, а доверительный интервал уровня надёжности $1-\alpha$ для функции регрессии в любой фиксированной точке X_0 будет иметь вид:

$$\left(\hat{\theta}_0 - \hat{\theta}_1 X_0 - t_{1-\frac{\alpha}{2},n-2} \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}}; \hat{\theta}_0 - \hat{\theta}_1 X_0 + t_{1-\frac{\alpha}{2},n-2} \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}} \right)$$

Отметим, что самый короткий интервал будет в точке $X_0 = \bar{X}$.

5) Построение доверительного интервала для отклика в простой парной модели. Согласно модели (15.7) зависимая переменная имеет вид

$$Y = \theta_0 + \theta_1 X + \varepsilon.$$

МНК-оценкой для отклика Y в точке $X = X_0$

$$\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X_0.$$

Найдём математическое ожидание и дисперсию случайной величины $\hat{Y}(X_0) + \varepsilon$.

$$E(\hat{Y}(X_0) + \varepsilon) = \theta_0 + \theta_1 X_0,$$

а дисперсия

$$D(\hat{Y}(X_0) + \varepsilon) = \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2} \right).$$

Тогда нетрудно показать, что центральная статистика для отклика Y в точке X_0

$$\frac{Y(X_0) - \hat{\theta}_0 - \hat{\theta}_1 X_0}{\hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}}} \sim t(n-2).$$

Следовательно, доверительный интервал уровня надёжности $1-\alpha$ для отклика Y в любой фиксированной точке X_0 будет иметь вид:

$$\left(\hat{\theta}_0 - \hat{\theta}_1 X_0 - t_{1-\frac{\alpha}{2},n-2} \hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}}; \hat{\theta}_0 - \hat{\theta}_1 X_0 + t_{1-\frac{\alpha}{2},n-2} \hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2}} \right)$$

Пример 17.1. Имеются данные (см. пример выше) о ВВП (переменная Y) и коэффициенте младенческой смертности (переменная X) для 15 стран. Было доказано, что эти показатели отрицательно коррелированы, т.е. между ними имеется зависимость. Рассмотрим модель простой линейной регрессии вида

$$Y_i = \theta_0 + \theta_1 X_i + \varepsilon_i, \qquad i = 1,\ldots,15, \tag{17.2}$$

и оценим неизвестные параметры θ_0 и θ_1 методом наименьших квадратов.

Р е ш е н и е: Согласно формуле (15.8) , получим $\hat{\theta}_0 = 60.62, \hat{\theta}_1 = -0.629$, т.е.

$$\hat{Y}_i = 60.62 - 0.629 X_i.$$

Проверим гипотезу $H_0 : \theta_1 = 0$ о значимости коэффициента θ_1. Реализация соответствующей t-статистики вида (17.1) равна -3.707. Полученное значение соответствует квантили уровня 0.0015 распределения Стьюдента с 13-ю степенями свободы. Таким образом, на уровне значимости $\alpha = 0.003$ гипотеза $H_0 : \theta_1 = 0$ отвергается.

Коэффициент детерминации данной регрессии $\hat{R}^2_{YX} = 0.51$. Это означает, что только 51% вариации ВВП объясняется данной моделью регрессии. Поэтому попытаемся включить в модель ещё один фактор $X_2 = X^2$ и рассмотреть линейную регрессионную модель вида

$$Y_i = \theta_0 + \theta_1 X_i + \theta_2 X_i^2 + \varepsilon_i, \qquad i = 1,\ldots,15.$$

МНК-оценка вектора параметров $\theta = (\theta_0, \theta_1, \theta_2)$, вычисленная по формуле (15.6) равна (75,59;-1,81;0,01). Коэффициент детерминации этой модели вырос до значения 0.64. Однако параметр θ_2 оказался статистически незначимым, так как t-статистика вида (17.1), соответствующая критерию, проверяющему гипотезу $H_0 : \theta_2 = 0$, оказалась равной 2.07. Это значение соответствует квантили уровня 0.97 распределения Стьюдента с 12-ю степенями свободы. Таким образом, на уровне значимости $\alpha = 0.05$ следует принять гипотезу $H_0 : \theta_2 = 0$, и исключить переменную $X_2 = X^2$ из уравнения регрессии.

Попытаемся подобрать другую функцию, описывающую зависимость между ВВП и коэффициентом младенческой смертности, используя геометрические соображения. График наблюдений, представленный на рис. 17.1, наводит на мысль о том, что тип функциональной зависимости между Y и X описывается гиперболической функцией, т.е.

$$Y_i = \theta_0 + \frac{\theta_1}{X_i} + \varepsilon_i, \qquad i = 1,\ldots,15.$$

Оценив параметры этой модели по МНК, получим $\hat{\theta}_0 = -1.35, \hat{\theta}_1 = 516.76$. Отметим, что коэффициент θ_1 в этой модели является статистически значимым, так как соответствующая t-статистика вида (17.1) равна 7.57. Коэффициент детерминации этой модели $\hat{R}^2_{YX} = 0.82$. Таким образом, последняя модель объясняет 82% вариации ВВП и гораздо лучше описывает зависимость между ВВП и коэффициентом младенческой смертности, чем модель (18.4). Теперь можно сделать вывод о том, что из трёх рассмотренных моделей наиболее адекватной является последняя модель вида

$$Y_i = -1.35 + \frac{516.76}{X_i}.$$

График этой функции представлен на рис. 17.2.

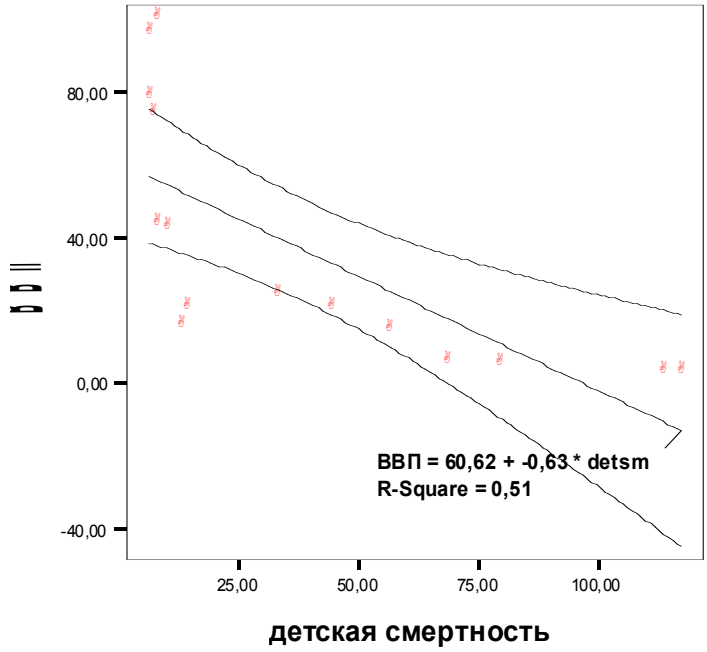


Рис 17.1

Таблица 1.

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	60,618	9,213		6,579	,000	40,714	80,522
детская смертность	-,629	,170	-,717	-3,707	,003	-,996	-,262

Таблица 2.

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	75,591	10,952		6,902	,000	51,728	99,454
детская смертность	-1,807	,588	-2,059	-3,071	,010	-3,089	-,525
dsmsq	,010	,005	1,389	2,072	,060	-,001	,021

Таблица 3.

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	-1,351	6,214		-,217	,831	-14,776	12,074
1/(детская смертность)	516,758	66,619	,907	7,757	,000	372,837	660,679

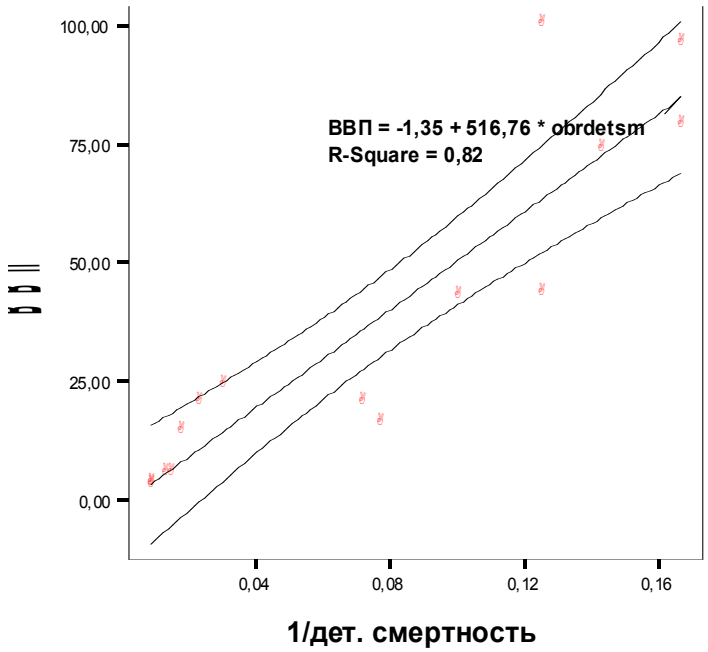

Linear Regression with
95,00% Mean Prediction Interval
ВВП = -1,35 + 516,76 * obrdetsm
R-Square = 0,82
ВВП
100,00
75,00
50,00
25,00
0,00
0,04
0,08
0,12
0,16
1/дет. смертность

Рис 17.2

Лекция 18

Регрессионные модели с переменной структурой

18.1 Фиктивные переменные (dummy variables)

При анализе реальных данных нередко возникают ситуации, когда влияние некоторой качественной переменной изменяет взаимосвязь между факторами и откликом, и модели с постоянной структурой уже не являются достаточно точными для описания имеющихся закономерностей. В таких случаях прибегают к построению моделей с переменной структурой. Эти модели строят с помощью включения в регрессионное уравнение фиктивных переменных (dummy variables). Эти переменные обычно являются дихотомическими, т.е. принимают только два значения: 0 или 1. Например, пусть заработная плата Y зависит от нескольких объясняющих количественных показателей X_1,...,X_p, и мы рассматриваем регрессионную модель

$$Y_i = \theta_0 + \theta_1 X_{i1} + \ldots + \theta_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Затем мы хотим включить в рассмотрение такой фактор, как наличие или отсутствие высшего образования. Для этого введём новую бинарную переменную d, которая будет принимать значение $d_i = 1$, если i-е наблюдение соответствует респонденту с высшим образованием и $d_i = 0$ в иных случаях. Тогда новая модель будет иметь вид:

$$Y_i = \theta_0 + \theta_1 X_{i1} + \ldots + \theta_p X_{ip} + \delta d_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

В рамках такой модели мы считаем, что средняя заработная плата для людей без высшего образования равна $\theta_0 + \theta_1 X_{i1} + \ldots + \theta_p X_{ip}$, а для людей с высшим образованием равна $\theta_0 + \theta_1 X_{i1} + \ldots + \theta_p X_{ip} + \delta$. Т.е. величина δ интерпретируется как среднее изменение зарплаты при переходе из одной категории (без высшего образования) в другую (с высшим образованием) при неизменных значениях остальных показателей.

Для оценивания параметров последней регрессионной модели можно применить МНК и затем проверить гипотезу $H_0 : \delta = 0$. Если гипотеза H_0 принимается, то это говорит о том, что различия в зарплате между категориями несущественно.

Если включаемый в рассмотрение качественный показатель имеет не два, а большее число значений, то в принципе можно было бы ввести дискретную фиктивную переменную, принимающую такое же количество значений. Однако так практически никогда не делают, поскольку интерпретация соответствующих коэффициентов становится затруднительной. В этих ситуациях используют несколько бинарных переменных. Типичным примером является исследование сезонных колебаний. Например, пусть Y_i - потребление некоторого продукта в месяц i, и есть основания считать, что потребление зависит от времени года. Для выявления сезонности вводят три бинарные переменные d_1,d_2,d_3 :

$$d_{i1} = \{1, \text{если месяц i -зимний}, 0 - \text{в остальных случаях}$$

$$d_{i2} = \{1, \text{если месяц i -весенний}, 0 - \text{в остальных случаях}$$

$$d_{i3} = \{1, \text{если месяц i -летний}, 0 - \text{в остальных случаях},$$

и оценивают параметры регрессионного уравнения

$$Y_i = \theta_0 + \theta_1 d_{i1} + \theta_2 d_{i2} + \theta_3 d_{i3} + \varepsilon_i, \qquad i = 1, \ldots, n.$$

Отметим, что бинарная переменная d_4, относящаяся к осени, не вводится. Введение такой переменной приведёт к тому, что для любого i будет выполняться равенство $d_{i1} + d_{i2} + d_{i3} + d_{i4} = 1$. Последнее означает линейную зависимость регрессоров и невозможность получения МНК-оценок. В рамках рассмотренной модели получаем, что среднемесячный объём потребления равен θ_0 для осенних месяцев, $\theta_0 + \theta_1$ - для зимних, $\theta_0 + \theta_2$ - для весенних, $\theta_0 + \theta_3$ - для летних. Коэффициенты при фиктивных переменных показывают отличие среднего значения текущего сезона от эталонного (в данном случае - осеннего).

Ещё одно важное приложение фиктивных переменных - построение кусочно-линейных моделей. Такие модели применяются для исследования структурных изменений. Например, пусть Y_t - объём производства некоторого предприятия в момент времени $t = 1,...,n$, X_t - размер основных фондов этого предприятия в момент времени $t = 1,...,n$, и $Y_t = \theta_0 + \theta_1 X_t + \varepsilon_t$. Предполагается, что в момент времени $t = t_0$ произошла структурная перестройка,и линии регрессии до и после момента времени t_0 различаются. При этом линия регрессии должна быть непрерывной. Для построения модели, описывающей такое явление, введём бинарную переменную

$$\delta_t = \begin{cases} 0, & t \leqslant t_0, \\ 1, & t > t_0 \end{cases}$$

и рассмотрим уравнение

$$Y_t = \theta_0 + \theta_1 X_t + \theta_2 (X_t - X_{t_0}) \delta_t + \varepsilon_t.$$

Проверяя гипотезу $H_0 : \theta_2 = 0$, мы тестируем отсутствие структурного сдвига в точке t_0.

18.2 Критерий Чоу (Chow)

Рассмотрим следующую задачу. Пусть имеется два набора данных, в первом наборе (назовём его подвыборкой A) имеется n наблюдений, во втором наборе (назовём его подвыборкой B) - m наблюдений. Например, зависимый показатель Y - заработная плата, а регрессоры $X_i, i = 1,...,p$ - возраст, стаж, уровень образования и т.п. Но первый набор данных относится к женщинам, а второй - к мужчинам. По набору данных A мы строим регрессионную модель

$$Y_i = \widetilde{\theta}_0 + \tilde{\theta}_1 X_{i1} + \ldots + \tilde{\theta}_p X_{ip} + \tilde{\varepsilon}_i, \qquad i = 1, \ldots, n.$$

и по набору данных B модель

$$Y_i = \breve{\theta}_0 + \breve{\theta}_1 X_{i1} + \ldots + \breve{\theta}_p X_{ip} + \breve{\varepsilon}_i, \qquad i = n, \ldots, n+m.$$

Вопрос, который интересует исследователя: верно ли, что эти две модели совпадают? Такая задача сводится к проверке гипотезы

$$H_0 : \widetilde{\theta}_0 = \breve{\theta}_0, \tilde{\theta}_1 = \breve{\theta}_1, ..., \tilde{\theta}_p = \breve{\theta}_p.$$

Статистика критерия строится следующим образом. Рассмотрим объединённую выборку из $n+m$ наблюдений и построим МНК-оценки параметров регрессии

$$Y_i = \theta_0 + \theta_1 X_{i1} + \ldots + \theta_p X_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n+m.$$

Отметим, что такую регрессию можно назвать регрессией с ограничениями (понятно, что ограничениями здесь являются условия $\widetilde{\theta}_0 = \breve{\theta}_0, \tilde{\theta}_1 = \breve{\theta}_1, ..., \tilde{\theta}_p = \breve{\theta}_p, \tilde{\sigma} = \breve{\sigma}$). Оценив параметры модели, найдём $SS_{\text{сл.}} = \sum_{i=1}^{n+m} (Y_i - \hat{Y}_i)^2$. Далее оценим по отдельности параметры

модели A и параметры модели B. Обозначим соответствующие суммы квадратов остатков как $SS^{\text{A}}_{\text{сл.}}$ и $SS^{\text{B}}_{\text{сл.}}$. Показателем улучшением качества оценивания из-за снятия ограничений (т.е. из-за разделения выборки на две подвыборки A и B) может служить величина

$$SS_{\text{сл.}} - SS^{\text{A}}_{\text{сл.}} - SS^{\text{B}}_{\text{сл.}}.$$

В гауссовском случае имеем, что

$$\frac{SS^{\text{A}}_{\text{сл.}}}{\sigma^2_\varepsilon} \sim \chi^2(n-p-1)$$

$$\frac{SS^{\text{B}}_{\text{сл.}}}{\sigma^2_\varepsilon} \sim \chi^2(m-p-1).$$

При справедливости гипотезы H_0 с учётом наложенных $p+1$ ограничений получим, что случайная величина

$$\frac{SS_{\text{сл.}} - SS^{\text{A}}_{\text{сл.}} - SS^{\text{B}}_{\text{сл.}}}{\sigma^2_\varepsilon} \sim \chi^2(p+1).$$

Тогда при справедливости гипотезы H_0 статистика

$$\frac{\frac{1}{p+1}(SS_{\text{сл.}} - (SS^{\text{A}}_{\text{сл.}} + SS^{\text{B}}_{\text{сл.}}))}{\frac{1}{m+n-2p-2}(SS^{\text{A}}_{\text{сл.}} + SS_{\text{сл.B}})} \sim F(p+1,n+m-2p-2)$$

имеет распределение Фишера с $p+1$ и $n+m-2p-2$ степенями свободы. Таким образом, если реализация статистики окажется больше, чем квантиль уровня $1-\alpha$ распределения $F(p+1,n+m-2p-2)$, то основную гипотезу на уровне значимости α следует отвергнуть, и считать, что подвыборки A и B описываются разными уравнениями регрессии.

18.3 Нарушения предпосылок линейной модели: гетероскедастичность, модель с коррелированными остатками

18.4 Гетероскедастичность.

Одним из нарушений в классической модели линейной регрессии является нарушение условия постоянства дисперсий некоррелированных остатков. Такое явление называется гетероскедастичностью. Например, если исследуется зависимость прибыли предприятия от размеров основных фондов, то естественно ожидать, что для крупных предприятий колебание прибыли будет выше, чем для малых. В этих случаях ковариационная матрица вектора погрешностей ε является диагональной и имеет вид: $K_\varepsilon = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix}$.

Если бы дисперсии $\sigma_i^2, i = 1,...,n$, были известны, то для оценивания параметров можно применить взвешенный МНК. Взвешенный МНК состоит в минимизации по θ функции потерь вида

$$S(\theta) = \sum_{i=1}^{n} (\frac{1}{\sigma_i}(Y_i - f(X_{i1},\dots X_{ip},\theta)))^2 \to \min_\theta. \tag{18.1}$$

Взвешенной МНК-оценкой называют $\hat{\theta}$, при котором достигается минимум функции $S(\theta)$.

Идея этого метода заключается в следующем. При использовании обычного МНК, мы минимизируем функцию

$$S(\theta) = \sum_{i=1}^{n} (Y_i - f(X_{i1},\dots X_{ip},\theta))^2 \to \min_\theta, \tag{18.2}$$

в которую слагаемые вносят, вообще говоря, разный вклад из-за различных дисперсий. "Взвешивая"каждое наблюдение с помощью коэффициента $\frac{1}{\sigma_i}$, мы устраняем эту неоднородность. Отметим, что больший вес здесь будут иметь наблюдения с меньшей дисперсией. Запишем явный вид взвешенной МНК-оценки. Для этого введём следующие обозначения.

Пусть $V^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\sigma_n^2} \end{pmatrix}$, $X^* = V^{-\frac{1}{2}}X, Y^* = V^{-\frac{1}{2}}Y$.

Тогда взвешенная МНК-оценка имеет вид

$$\hat{\theta} = ((V^{-\frac{1}{2}}X)^T(V^{-\frac{1}{2}}X))^{-1}(V^{-\frac{1}{2}}X)^T(V^{-\frac{1}{2}}Y) = (X^{*T}X^*)^{-1}X^{*T}Y^*. \tag{18.3}$$

Асимптотическое распределение такой оценки

$$\hat{\theta} \sim N\Big(\theta, \sigma^2(X^TV^{-1}X^T)^{-1}\Big).$$

Однако в реальности дисперсии $\sigma_i^2, i = 1,...,n$, как правило, неизвестны. Поскольку количество неизвестных параметров равно числу наблюдений n, то без дополнительных ограничений на структуру ковариационной матрицы K_ε не удастся получить приемлемые оценки дисперсий. Ознакомимся лишь с некоторыми моделями структур ковариационной матрицы.

Тест Голдфелда-Квандта (Golgfeld-Quandt).

Данный тест проверки наличия гетероскедастичности является состоятельным в случаях, когда имеется наличие монотонной связи дисперсии остатков с одним из регрессоров. Проверяется основная гипотеза о гомоскедастичности $H_0 : \sigma_1 = ... = \sigma_n$ против $H_1 : \sigma_i = \sigma X_{ij}, i = 1,...,n$, где $\sigma > 0$, а X_{ij} -значение j-го регрессора в i-м наблюдении.

Алгоритм построения статистики критерия следующий:

1) упорядочить данные по убыванию того регрессора, относительно которого имеется подозрение на гетероскедастичность;

2) исключить в этом упорядочении d средних наблюдений;

3) построить две независимые линии регрессии по первым $n/2-d/2$ и по последним $n/2-d/2$ наблюдениям и оценить вектор остатков $\hat{\varepsilon}_1$ для первой модели и вектор остатков $\hat{\varepsilon}_2$ для второй модели;

4) вычислить статистику $F = \frac{\hat{\varepsilon}_1^T\hat{\varepsilon}_1}{\hat{\varepsilon}_2^T\hat{\varepsilon}_2}$.

При справедливости гипотезы H_0 статистика F имеет F-распределение с $(n/2-d/2-p; n/2-d/2-p)$ степенями свободы, где p- число оцениваемых параметров. Большие значения этой статистики говорят о нарушении гомоскедастичности, и критическая область имеет вид: $(f_{n/2-d/2-p;n/2-d/2-p;1-\alpha}, \infty)$.

Замечание 18.1. Количество исключаемых наблюдений d не должно быть слишком малым и не должно быть слишком большим. Обычно d составляет примерно четверть от общего количества наблюдений. Необходимо также, чтобы $n/2-d/2 > p$.

Если гипотеза H_0 отвергнута тестом Голдфелда-Квандта, то следует провести следующую коррекцию на гетероскедастичность. Разделить $i-$е $i = 1,...,n$ уравнение на X_{ij}, получить новые переменные $X_{ik}^* = \frac{X_{ik}}{X_{ij}}$ и $Y_i^* = \frac{Y_i}{X_{ij}}$, затем построить взвешенную МНК-оценку (18.3).

Замечание 18.2. Если в исходной модели первый столбец матрицы X состоял из единиц, то оценка свободного члена в новой (преобразованной) модели будет при переменной X_{ij}, а оценка коэффициента при X_{ij}^* - оценкой свободного члена в исходной модели.

Пример 18.1. Продемонстрируем работу теста Голдфелда-Квандта на следующем примере. Пример разработан студенткой ФЭН Бурковой Анастасией.

Имеются данные о цене (показатель Y) и площади (показатель X) 300 квартир. Предполагается, что справедлива модель

$$Y_i = \theta_0 + \theta_1 X_i + \varepsilon_i, \qquad i = 1,\ldots,300. \tag{18.4}$$

Данные и МНК-оценка регрессии (18.4) представлены на рис. 18.1.

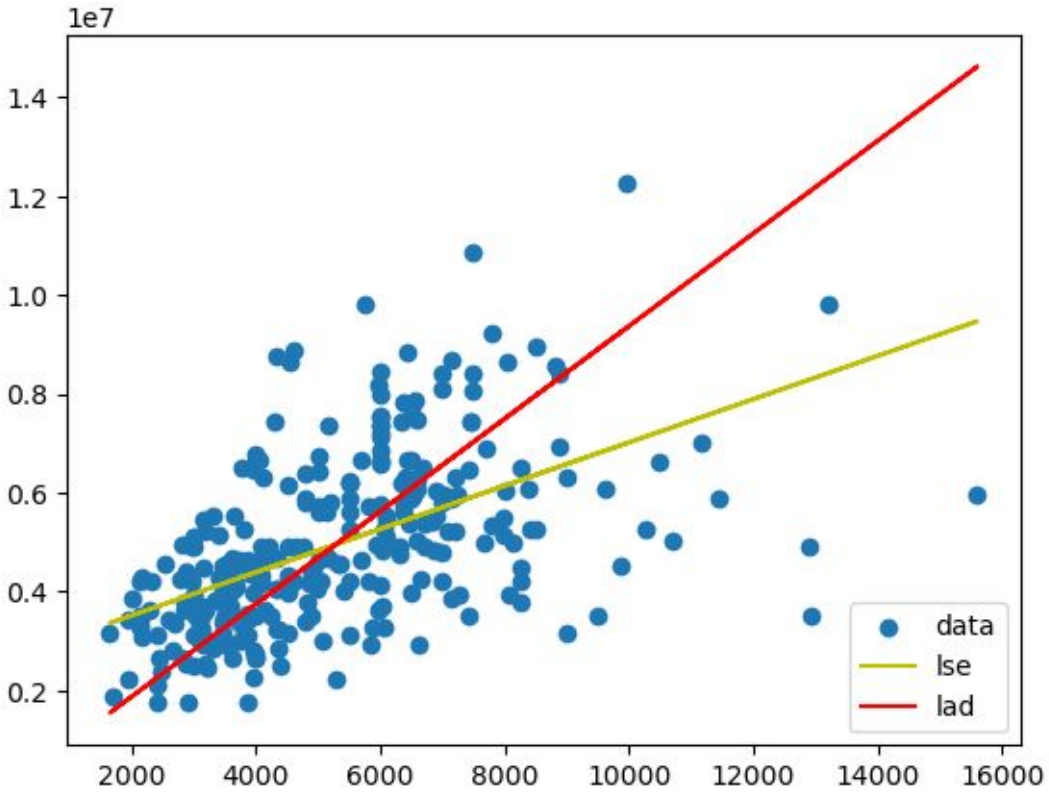


Рис 18.1

Проверим гипотезу о гомоскедастичности $H_0 : \sigma_1 = ... = \sigma_n$ против $H_1 : \sigma_i = \sigma X_i, i =$ 1,...,n, где $\sigma > 0$, а X_i -площадь i-й квартиры. Упорядочим данные по убыванию показателя X и удалим 74 средних наблюдения. Построим МНК-оценки регрессии по первым 113 наблюдениям и по последним 113 наблюдениям. Графики линий первой и второй регрессии представлены на рис. 18.2.

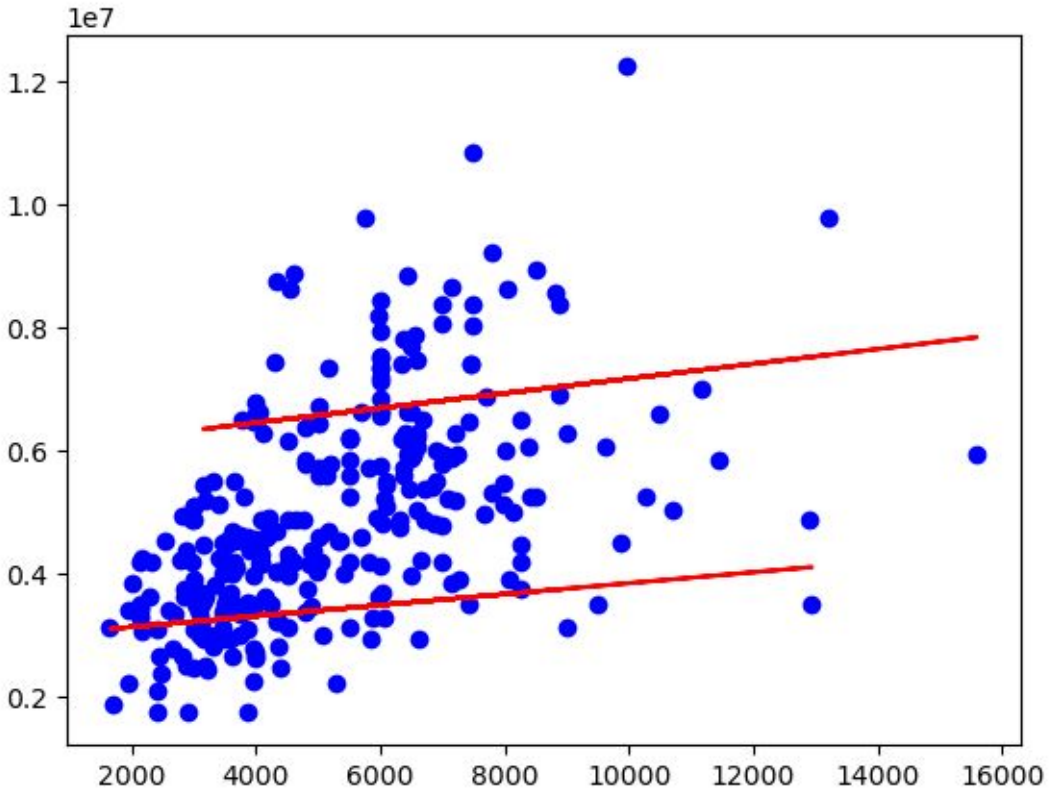


Рис 18.2

Оценим вектор остатков $\hat{\varepsilon}_1$ для первой модели и вектор остатков $\hat{\varepsilon}_2$ для второй модели. Вычислим реализацию статистики $F = \frac{\hat{\varepsilon}_1^T \hat{\varepsilon}_1}{\hat{\varepsilon}_2^T \hat{\varepsilon}_2} = 4{,}77$.

При справедливости гипотезы о гомоскедастичности статистика F должна иметь F-распределение со $(111;111)$ степенями свободы. Критическая область имеет вид: $(f_{111;111;0.95}, \infty) = (1{,}37, \infty)$. Поскольку реализация статистики попала в критическую область, то гипотезу о гомоскедастичности следует (на уровне значимости 0.05) отвергнуть в пользу альтернативы о том, что дисперсии погрешностей обратно пропорциональны значению регрессора X (площади соответствующей квартиры). Следовательно, требуется провести корректировку на гетероскедастичность. Для этого разделим i-е уравнение регрессии на X_i и применим МНК для оценивания параметров регрессии

$$Y_i^* = \frac{\theta_0}{X_i} + \theta_1 + \varepsilon_i,$$

где $Y_i^* = \frac{Y_i}{X_i}$. МНК-оценки последней модели совпадают с оценками взвешенного МНК (18.1) с весами $1/X_i$. Графики линий регрессии, полученные с помощью МНК (зелёная линия) и взвешенного МНК (красная линия) показаны на (см. рис. 18.3).

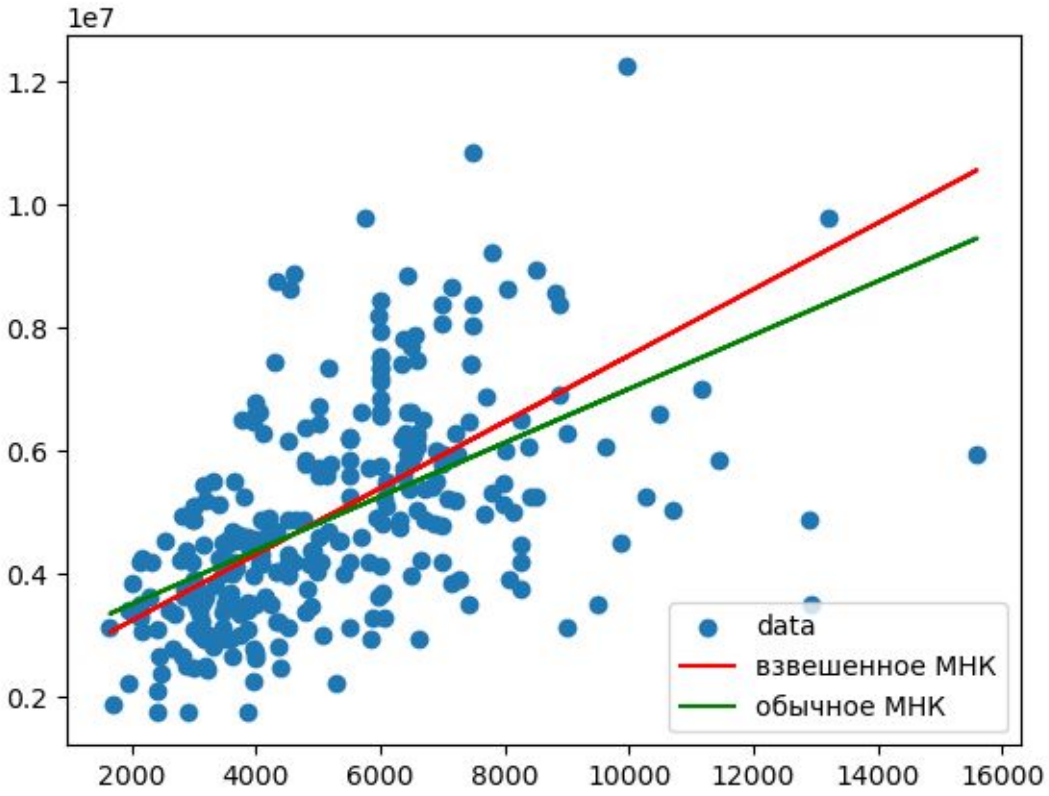


Рис 18.3

Тест Бреуша-Пагана (Breusch-Pagan).

Этот тест применяется в тех случаях, когда предполагается, что дисперсии σ_i^2 зависят от некоторых дополнительных переменных. А именно, альтернативная гипотеза имеет вид: $H_1 : \sigma_i^2 = \alpha_0 + \sum_{j=1}^{k} \alpha_j z_{ij}, i = 1,...,n$, где $(\alpha_0, \alpha_1,...,\alpha_k)$, а $(z_{i1},...,z_{ik})$ -вектор наблюдаемых независимых переменных в i-м наблюдении.

Алгоритм построения статистики критерия следующий:

1) Построить обычную регрессию и оценить вектор остатков $\hat{\varepsilon}_1,...,\hat{\varepsilon}_n$;

2)Оценить дисперсию остатков $\hat{\sigma_\varepsilon^2} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon_i^2}$;

3)Построить регрессию $\frac{\hat{\varepsilon_i^2}}{\hat{\sigma_\varepsilon^2}} = \alpha_0 + \sum_{j=1}^{k} \alpha_j z_{ij} + u_i$ и найти для неё $SS_{\text{регр.}}$ (вариацию зависимой переменной, объяснённую регрессией).

Доказано, что при справедливости гипотезы о гомоскедастичности $H_0 : \sigma_1 = ... = \sigma_n$, статистика $\frac{SS_{\text{регр.}}}{2}$ асимптотически будет иметь распределение $\chi^2(k)$ хи-квадрат с k степенями свободы.

Если гетероскедастичность выявлена с помощью теста Бреуша-Пагана, то рекомендуется провести следующую коррекцию. Для оценивания параметров регрессионной модели использовать взвешенный МНК, выбирая в качестве весов значения

$$\left(\hat{\alpha}_0 + \sum_{j=1}^{k} \hat{\alpha}_j z_{ij}\right)^{-1/2}.$$

Замечание 18.3. Может оказаться, что для некоторых наблюдений весовой коэффициент будет отрицательным. В случае, если число таких наблюдений невелико, то их можно проигнорировать. В противном случае предлагается использовать мультипликативную форму

$$\sigma_i^2 = \exp(\alpha_0 + \sum_{j=1}^{k} \alpha_j z_{ij}).$$

Тест Уайта (White)

Достаточно часто наличие гетероскедастичности обусловлено тем, что дисперсии ошибок некоторым достаточно сложным образом зависят от регрессоров. Уайт предположил,

что гетероскедастичность должна каким-то образом отражаться в остатках обычной регрессии исходной модели и предложил метод проверки гипотезы H_0 о гомоскедастичности без каких-либо предположений относительно структуры гетероскедастичности. Алгоритм построения статистики критерия следующий:

1) Оцениваются параметры модели с помощью МНК и находятся оценки остатков $\hat{\varepsilon}_1,...,\hat{\varepsilon}_n$;

2) Строится новая регрессия с зависимой переменной $\hat{\varepsilon}^2$. Входными переменными будут $X_1,...,X_p$, $X_1^2,...,X_p^2$, все попарные произведения X_iX_j и константа.

3)Вычисляется коэффициент детерминации R^2 для второй регрессии.

Если верна гипотеза H_0 о гомоскедастичности, то статистика NR^2 асимптотически будет иметь распределение $\chi^2(N-1)$ хи-квадрат с $N-1$ степенями свободы,где N- количество регрессоров второй модели.

Тест Уайта является универсальным, однако в случае отклонения гипотезы H_0 этот критерий не даёт указания на форму гетероскедастичности. Поэтому единственной возможностью коррекции гетероскедастичности является коррекция ковариационной матрицы ошибок.

Как было показано (см. свойство 2) МНК-оценок) ковариационную матрицу МНК-оценки можно записать

$$E(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T = E((X^TX)^{-1}X^T\varepsilon)((X^TX)^{-1}X^T\varepsilon)^T = E(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1} =$$
$$= (X^TX)^{-1}X^TK_\varepsilon X(X^TX)^{-1} = n(X^TX)^{-1}\left(\frac{1}{n}X^TK_\varepsilon X\right)(X^TX)^{-1}.$$

Запишем матрицу $X^TK_\varepsilon X$ следующим образом. Обозначим через $x_i^T, i=1,...,n$ вектор-строки размера $p+1$ матрицы X. Тогда матрица

$$X^TK_\varepsilon X = \sum_{i=1}^{n} x_ix_i^T\sigma_i^2.$$

Уайт показал, что оценка

$$\hat{K}_{\hat{\theta}} = n(X^TX)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_ix_i^T\hat{\varepsilon}_i^2\right)(X^TX)^{-1}$$

является состоятельной оценкой ковариационной матрицы $K_{\hat{\theta}}$.

Лекция 19

Краткий обзор методов оценивания параметров линейной регрессионной модели

Помимо уже рассмотренного метода наименьших квадратов для оценивания параметров линейной регрессионной модели используются и другие методы, краткий обзор которых приведён ниже.

Метод наименьших модулей (МНМ)

Оценкой метода наименьших модулей в модели (15.5) называется такое значение $\hat{\theta}$, при котором достигается минимум следующей функции потерь:

$$S(\theta) = \sum_{i=1}^{n} |Y_i - f(X_{i1}, \ldots, X_{ip}, \theta)|.$$

Минимум этой функции находится с помощью методов линейного программирования. Но можно рассмотреть эту задачу как частный случай метода взвешенного МНК с весами

$$W_i = (\operatorname{sign} \varepsilon_i) \cdot \varepsilon_i,$$

где $\varepsilon_i = Y_i - f(X_{i1}, \ldots, X_{ip}, \theta)$

Асимптотическое распределение МНМ оценки

$$\hat{\theta} \sim N\Big(\theta, \frac{1}{4f_\varepsilon^2(0)}(X^T X)^{-1}\Big).$$

Ранговая R-оценка

Ранговой оценкой параметров в модели (15.5) называется такое значение $\hat{\theta}$, при котором достигается минимум следующей функции потерь:

$$S(\theta) = \sum_{i=1}^{n} \Big(R(Y_i - f(X_{i1}, \ldots, X_{ip}, \theta)) - \frac{n+1}{2}\Big)(Y_i - f(X_{i1}, \ldots, X_{ip}, \theta),$$

где

$$R(Y_i - f(X_{i1}, \ldots, X_{ip}, \theta)$$

— ранг

$$Y_i - f(X_{i1}, \ldots, X_{ip}, \theta)$$

среди

$$Y_1 - f(X_{11}, \ldots, X_{np}, \theta), \cdots, Y_n - f(X_{n1}, \ldots, X_{np}, \theta).$$

Асимптотическое распределение R-оценки

$$\hat{\theta} \sim N\Big(\theta, \frac{1}{12(\int f_\varepsilon^2(x)dx)^2}(X^T X)^{-1}\Big).$$

Класс М-оценок

Пусть функция потерь имеет вид

$$S(\theta) = \sum_{i=1}^{n} \rho(Y_i - f(X_{i1}, \ldots, X_{ip}, \theta)),$$

где $\rho(X)$ — выпуклая функция.

Если в качестве функции $\rho(X)$ выбрать функцию $\rho(X) = X^2$, то получится МНК-оценка, если $\rho(X) = |X|$, то М-оценка совпадает с МНМ-оценкой, если $\rho(X) = a(R(X))X$, то М-оценка совпадает с R-оценкой.

Пусть $\psi(X) = \rho'(X)$. Тогда М-оценка $\hat{\theta}$ является решением системы уравнений

$$\sum_{i=1}^{n} \psi(Y_i - f(X_{i1}, \ldots, X_{ip}, \theta))X_{ie} = 0, \qquad e = 1, \ldots, p.$$

Асимптотическое распределение М-оценки

$$\hat{\theta} \sim N\Big(\theta, \frac{M(\psi^2(\varepsilon))}{(M(\psi'(\varepsilon)))^2}(X^T X)^{-1}\Big).$$