ГЛАВА 14

ДВЕ НЕЗАВИСИМЫЕ ВЫБОРКИ

§ 1. АЛЬТЕРНАТИВЫ ОДНОРОДНОСТИ

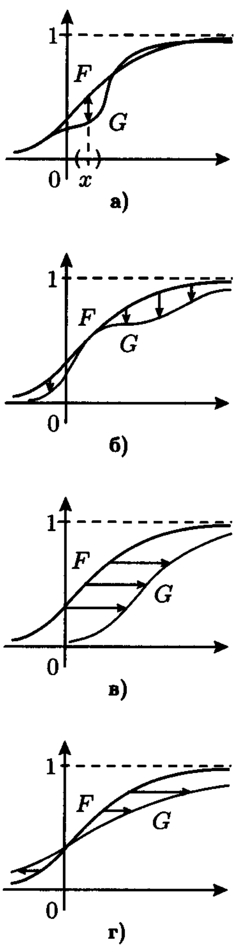

Рис. 1

Данные. Два набора наблюдений $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ будем рассматривать как реализовавшиеся значения случайных величин $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$.

На протяжении всей главы будем считать выполненными

Допущения

Д1. *Случайные величины* $X_1, \ldots, X_n$ *независимы и имеют общую функцию распределения* $F(x)$.

Д2. *Случайные величины* $Y_1, \ldots, Y_m$ *независимы и имеют общую функцию распределения* $G(x)$.

Д3. *Обе функции* F *и* G *неизвестны, но принадлежат множеству* Ω_c *всех непрерывных функций распределения.*

Нас будет интересовать

Гипотеза однородности

$$H_0 \colon G(x) = F(x) \text{ при всех } x.^{*)}$$

В качестве гипотез, конкурирующих с H_0, выделим следующие **альтернативы** (рис. 1):

а) *неоднородности* $H_1 \colon G(x) \neq F(x)$ при некотором x (а в силу непрерывности — и в некоторой окрестности точки x);

б) *доминирования* $H_2 \colon G(x) \leqslant F(x)$ при всех x, причем хотя бы для одного x неравенство строгое (говорят, что случайная величина Y_1 *стохастически больше* случайной величины X_1, поскольку $\mathbf{P}(Y_1 \geqslant x) \geqslant \mathbf{P}(X_1 \geqslant x)$ при каждом x);

в) *правого сдвига* $H_3 \colon G(x) = F(x - \theta)$, где параметр $\theta > 0$ (эта альтернатива — частный случай предыдущей);

г) *масштаба* $H_4 \colon G(x) = F(x/\theta)$, где $0 < \theta \neq 1$.

$^{*)}$ Формально H_0 представляет собой сложную непараметрическую гипотезу (см. § 1 гл. 13): в пространстве $\Omega_c \times \Omega_c$ она задает «диагональ» $\{(F,G) \colon G = F\}$.

Причины, по которым следует рассматривать конкурирующие гипотезы, отличные от H_1, таковы:

1) с практической точки зрения бывает важно уловить отклонения от H_0 только определенного вида (скажем, наличие систематического прироста у y_j по сравнению с x_i);

2) за счет сужения (по сравнению с H_1) множества пар распределений (F,G), составляющих альтернативное подмножество, обычно удается построить более эффективные (чувствительные) критерии, настроенные на обнаружение отклонений от H_0 конкретного вида.

Альтернатива доминирования H_2 встретится в § 3 и § 5. В гл. 15 приведены два полезных критерия, применяемых против альтернативы правого сдвига H_3. Методы анализа альтернативы масштаба H_4 (и ее обобщения, когда присутствует неизвестный «мешающий» параметр сдвига) изложены в гл. 24.

§ 2. ПРАВИЛЬНЫЙ ВЫБОР МОДЕЛИ

При проверке гипотезы однородности двух наборов данных $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ важно понять, с каким из *двух случаев* мы имеем дело: двумя реализациями независимых между собой выборок или парными повторными наблюдениями.

Примером первого случая может служить определение влияния удобрения на размер растений. Здесь $x_1, \ldots, x_n$ обозначают размеры растений на грядке, где удобрение не применялось, $y_1, \ldots, y_m$ — на соседней грядке, где оно применялось (см. пример 1 гл. 8). В этой ситуации можно предположить *независимость выборок* $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_m$. Формально это выражает допущение

Д4. *Все компоненты случайного вектора* $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ *независимы* (см. § 3 гл. 1).

Пример второго случая — исследование эффективности определенного воздействия (лекарства) на величину измеряемого показателя (скажем, артериального давления), где $x_1, \ldots, x_n$ — это значения показателя (у каждого из n наблюдаемых больных) *до воздействия*, а $y_1, \ldots, y_n$ — *после воздействия* ($m = n$). Для каждого фиксированного i ($i = 1, \ldots, n$) числам x_i и y_i в вероятностной модели Д1—Д3 соответствуют случайные величины X_i и Y_i, которые нельзя считать независимыми, так как x_i и y_i относятся к одному и тому же человеку.

Статистические методы, применимые ко второму случаю, рассматриваются в гл. 15. Конечно, их можно использовать и для независимых между собой выборок, отбросив, если $m \neq n$, лишние наблюдения в одной из реализаций (их надо отбирать случайно, скажем, с помощью таблицы Т1). Однако при этом игнорируется важная информация о совместной независимости, что снижает чувствительность методов по сравнению с критериями, рассматриваемыми в настоящей главе.

В свою очередь, использование приведенных в этой главе критериев для данных, относящихся ко второму случаю, представляет собой *грубую методическую ошибку*, нередко допускаемую неопытными прикладниками, которые пытаются проверить однородность своих наблюдений при помощи первого попавшегося метода.

Не все йогурты одинаково полезны.

Из телерекламы.

Рассмотрим три критерия проверки гипотезы однородности в предположении справедливости допущений Д1—Д4.

§ 3. КРИТЕРИЙ СМИРНОВА

Н. В. Смирнов (1900–1966), русский математик.

Для проверки гипотезы однородности H_0 против альтернативы неоднородности H_1 используется критерий Смирнова , статистикой которого служит величина

$$D_{n,m} = \sup_x \left| \widehat{F}_n(x) - \widehat{G}_m(x) \right|,$$

где $\widehat{F}_n(x) = \dfrac{1}{n} \sum\limits_{i=1}^{n} I_{\{X_i \leqslant x\}}$, $\widehat{G}_m(x) = \dfrac{1}{m} \sum\limits_{j=1}^{m} I_{\{Y_j \leqslant x\}}$,

т. е. $D_{n,m}$ — расстояние в равномерной метрике между эмпирическими функциями выборок (рис. 2).

Слишком большое расстояние противоречит гипотезе H_0. В [10, с. 350] приведена таблица критических значений $D_{n,m}$ для $n, m \leqslant 20$ и уровней значимости 1, 2, 5, 10%.

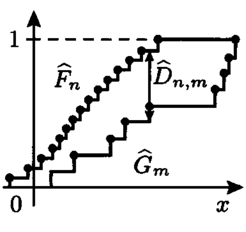

Рис. 2

Для нахождения значения статистики на реализациях $x_1, \ldots, x_n$ и $y_1, \ldots, y_m$ можно либо построить графики функций $\widehat{F}_n$ и $\widehat{G}_m$ и визуально определить их наибольшее расхождение, либо произвести вычисления на компьютере согласно формулам

$$D_{n,m} = \max\{D^+_{n,m}, D^-_{n,m}\},$$

где

$$D^+_{n,m} = \sup_x \left(\widehat{F}_n(x) - \widehat{G}_m(x)\right) = \max_{1 \leqslant i \leqslant n} \left\{ \frac{i}{n} - \widehat{G}_m(X_{(i)}) \right\},$$

$$D^-_{n,m} = \sup_x \left(\widehat{G}_m(x) - \widehat{F}_n(x)\right) = \max_{1 \leqslant j \leqslant m} \left\{ \frac{j}{m} - \widehat{F}_n(Y_{(j)}) \right\}.$$

Здесь $X_{(1)} \leqslant \ldots \leqslant X_{(n)}$ и $Y_{(1)} \leqslant \ldots \leqslant Y_{(m)}$ — упорядоченные по возрастанию элементы каждой из выборок.

Н. В. Смирнов в 1939 г. доказал, что если гипотеза H_0 верна, то при выполнении допущений Д1—Д4 имеет место сходимость

$$\mathbf{P}\left(\sqrt{nm/(n+m)}\, D_{n,m} \leqslant x\right) \to K(x) \text{ при } n,\, m \to \infty, \tag{1}$$

где $K(x)$ — функция распределения Колмогорова, определенная в § 2 гл. 12 (там же приведена небольшая таблица значений этой функции). Доказательство сходимости (1) при условии

Д5. *Размеры $n, m \to \infty$ так, что $n/(n+m) \to \gamma \in (0,1)$*

можно найти в [11, с. 428]. Контрпример в задаче 3 показывает, что условие непрерывности ДЗ необходимо. Данное приближение является довольно точным уже при $n, m \geqslant 20$ (см. [32, с. 108]).

Расстояние от пункта A до пункта B равно 1 км. Пусть n — скорость движения из A в B, а m — скорость движения на обратном пути. Тогда $2/(1/n + 1/m) = 2nm/(n + m)$ — средняя скорость. Эту величину называют *средним гармоническим* чисел n и m.

На рис. 3 изображена верхняя половина окружности с центром в точке O, построенная на диаметре PR длины $n + m$, $|PS| = n$, $|SR| = m$. Перпендикуляр ST к диаметру PR пересекает окружность в точке T. При этом $a = |OT| = (n+m)/2$ — *среднее арифметическое* n и m. Из подобия $\triangle PTS$ и $\triangle TRS$ следует, что $b = |ST| = \sqrt{nm}$ — *среднее геометрическое.*

Вопрос 1. Как на этом рисунке построить отрезок длины $c = 2nm/(n+m)$ так, чтобы стало очевидным неравенство $a \geqslant b \geqslant c$?

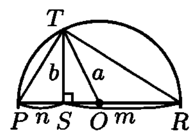

Рис. 3

Величина под корнем в формуле (1) представляет собой половину среднего гармонического n и m. Почему половину? Дело в том, что $\widehat{F}_n - \widehat{G}_m = (\widehat{F}_n - F) + (G - \widehat{G}_m)$, если $G = F$. При сложении независимых случайных величин их дисперсии складываются (П2). Поэтому для фиксированного x дисперсия отклонения $\widehat{F}_n(x) - \widehat{G}_m(x)$ при $m = n$ будет в 2 раза больше дисперсии отклонения $\widehat{F}_n(x) - F(x)$.

Замечание 1. В случае *альтернативы доминирования* (см. § 1) вместо критерия Смирнова надо применять *односторонний критерий*, основанный на следующей предельной теореме для определенной выше статистики $D^+_{n,m}$: при справедливости гипотезы H_0 для любого $x \geqslant 0$ имеет место сходимость

$$\mathbf{P}\left(\sqrt{nm/(n+m)}\, D^+_{n,m} \leqslant x\right) \to 1 - e^{-2x^2} \quad \text{при } n,\, m \to \infty. \tag{2}$$

(Для случая $m = n$ эта сходимость будет установлена в § 6.)

Согласно § 2 гл. 12 для правого «хвоста» распределения Колмогорова справедливо разложение

$$1 - K(x) = 2\left[e^{-2x^2} - e^{-8x^2} + e^{-18x^2} - \ldots\right].$$

Второй член заключенного в квадратные скобки ряда представляет собой четвертую степень его первого члена. Пренебрегая им и всеми последующими членами, из сравнения сходимостей (1) и (2) видим, что фактический уровень значимости (см. § 1 гл. 12) данного критерия примерно вдвое меньше, чем у критерия Смирнова.

§ 4. КРИТЕРИЙ РОЗЕНБЛАТТА

Для проверки гипотезы однородности H_0 двух выборок против *альтернативы неоднородности* H_1 (см. § 1) можно воспользоваться также критерием типа ω^2 из § 2 гл. 12. Статистика этого критерия задается формулой

$$\omega^2_{n,m} = \int\limits_{-\infty}^{\infty} [\widehat{F}_n(x) - \widehat{G}_m(x)]^2\, d\widehat{H}_{n+m}(x),$$

где $\widehat{H}_{n+m}(x) = \dfrac{n}{n+m}\,\widehat{F}_n(x) + \dfrac{m}{n+m}\,\widehat{G}_m(x)$ представляет собой эмпирическую функцию, построенную по объединенной выборке $(X_1,\ldots,X_n,Y_1,\ldots,Y_m)$.

Согласно [10, с. 86], статистика $\omega^2_{n,m}$ зависит лишь от порядковых номеров (*рангов*) выборочных элементов:

$$\omega^2_{n,m} = \frac{1}{nm}\left[1/6 + \frac{1}{m}\sum_{i=1}^{n}(R_i - i)^2 + \frac{1}{n}\sum_{j=1}^{m}(S_j - j)^2\right] - 2/3,$$

где R_i — ранг $X_{(i)}$, а S_j — ранг $Y_{(j)}$ в объединенном вариационном ряду (см. § 4 гл. 4).

Положим для краткости $Z = Z_{n,m} = \dfrac{nm}{n+m}\,\omega^2_{n,m}$. В 1952 г. М. Розенблатт доказал, что при условии справедливости гипотезы H_0 и выполнении допущений Д1—Д5 имеет место сходимость

$$\mathbf{P}(Z \leqslant x) \to A_1(x), \qquad (3)$$

где предельный закон A_1 тот же самый, что встречался в § 2 гл. 12. Математическое ожидание и дисперсия этого закона равны, соответственно, 1/6 и 1/45, в то время как

$$\mathbf{M}Z = \frac{1}{6}\left(1 + \frac{1}{n+m}\right),$$

$$\mathbf{D}Z = \frac{1}{45}\left(1 + \frac{1}{n+m}\right)\left[1 + \frac{1}{n+m} - \frac{3}{4}\left(\frac{1}{n} + \frac{1}{m}\right)\right].$$

Поэтому при вычислении приближенных критических значений рекомендуется вместо Z в формуле (3) использовать статистику $Z^* = (Z - \mathbf{M}Z)/\sqrt{45\,\mathbf{D}Z} + 1/6$.*) Это обеспечивает удовлетворительную точность приближения уже для $n,\ m \geqslant 7$.

§ 5. КРИТЕРИЙ РАНГОВЫХ СУММ УИЛКОКСОНА

Критерий ранговых сумм Уилкоксона применяется для проверки гипотезы однородности H_0 против *альтернативы доминирования* H_2 (см. § 1), в частности, — против альтернативы правого сдвига H_3.

*) Очевидно, $\mathbf{M}Z^* = 1/6$, $\mathbf{D}Z^* = 1/45$, причем из приведенных выше формул для $\mathbf{M}Z$ и $\mathbf{D}Z$ из (3) с учетом свойств сходимости (П5) следует, что $\mathbf{P}(Z^* \leqslant x) \to A_1(x)$.

Вычислим статистику V *критерия ранговых сумм Уилкоксона.*

1. Обозначим через S_j ранг порядковой статистики $Y_{(j)}$ ($j = 1,\ldots,m$) в вариационном ряду, построенном по объединенной выборке $(X_1,\ldots,X_n,Y_1,\ldots,Y_m)$ (рис. 4).
2. Положим $V = S_1 + \ldots + S_m$.

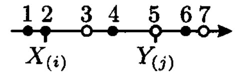

Рис. 4

Критерий, основанный на статистике V, был предложен Ф. Уилкоксоном в 1945 г. для выборок одинакового размера и распространен на случай $m \neq n$ Х. Манном и Д. Уитни в 1947 г.

Суть критерия сводится к следующему: если верна гипотеза H_0, то значения $Y_{(j)}$ должны быть рассеяны по всему вариационному ряду; напротив, достаточно большое значение V указывает на тенденцию преобладания Y_j над X_i, что свидетельствует в пользу справедливости гипотезы H_2. Таким образом, критическая область выбирается в виде луча $\{V > c\}$, где c — некоторая константа.

Малые выборки. Критические значения статистики V для $n, m \leqslant 25$ приведены в таблице [10, с. 357].

Большие выборки. Рассмотрим статистику

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I_{\{X_i < Y_j\}}. \qquad (4)$$

Вопрос 2. Верно ли, что все слагаемые $I_{\{X_i<Y_j\}}$ в сумме (4) независимы?

При отсутствии совпадений среди X_i и Y_j справедливо равенство (см. задачу 4)

$$U = V - m(m+1)/2, \qquad (5)$$

и, следовательно, критерии, основанные на V и U, эквивалентны. Предложенная Уилкоксоном ранговая форма V удобнее для вычислений. С другой стороны, с помощью *считающей формы* U, изученной Манном и Уитни, нетрудно установить (задача 5), что в случае справедливости гипотезы H_0 имеем:

$$\mathbf{M}U = nm/2, \quad \mathbf{D}U = nm(n+m+1)/12. \qquad (6)$$

В [86, с. 143] сообщается, что У. Краскел нашел статистику U в работе Г. Дехлера, опубликованной в Германии в 1914 г.

Когда гипотеза H_0 верна и выполнены условия Д1—Д5, имеет место сходимость

$$U^* = (U - \mathbf{M}U)/\sqrt{\mathbf{D}U} \xrightarrow{d} Z \sim \mathcal{N}(0,1). \qquad (7)$$

Доказательство этого результата можно найти в [86, с. 145].

Поправка. К сожалению, нормальное приближение (7) не обеспечивает достаточную точность при $n, m \leqslant 50$. Например, при $25 \leqslant n, m \leqslant 50$ (см. [88, с. 87]) в 40% случаев истинные критические точки для статистики V отличаются от точек, полученных на основе сходимости (7), более чем на 1. Существенно точнее следующая аппроксимация, предложенная Р. Иманом в 1976 г. Она использует полусумму нормальной и стьюдентовской квантилей. Положим $N = n + m$. Критическим α-значением статистики

$$\tilde{U}^* = \frac{1}{2} U^* \left[1 + \sqrt{(N-2)/\left(N-1-(U^*)^2\right)}\right] \qquad (8)$$

Численный пример применения критерия ранговых сумм Уилкоксона—Манна—Уитни содержится в задаче 2.*)

Комментарии

1. Как доказано в [86, с. 167], критерий ранговых сумм состоятелен против альтернативы доминирования H_2, в частности, против альтернативы правого сдвига H_3.

2. Распределение случайной величины $V = S_1 + \ldots + S_m$ можно найти, пользуясь тем, что при справедливости гипотезы H_0 вероятность каждого из C^m_{n+m} возможных сочетаний $S_1, \ldots, S_m$ (соответствующих расстановкам Y_j, $j = 1, \ldots, m$, по $n + m$ местам) одна и та же.

Вопрос 3.
Чему равна $\mathbf{P}(V \geqslant 8)$ для $n=3$ и $m=2$?

3. Покажем, как оценка $\widehat{\theta}$, определяемая равенством (10), связана со статистикой U. Ввиду формулы (4) при отсутствии совпадений, U равна числу положительных разностей $Y_j - X_i$. Естественной оценкой параметра θ будет такая величина θ', чтобы наборы $(Y_1' = Y_1 - \theta', \ldots, Y_m' = Y_m - \theta')$ и $(X_1, \ldots, X_n)$ выглядели как выборки из одного и того же закона. Для таких выборок распределение статистики U симметрично относительно среднего $nm/2$. Таким образом, приходим к следующему уравнению относительно θ':

$$\sum_{i=1}^{n}\sum_{j=1}^{m} I_{\{Y_j' - X_i > 0\}} = \sum_{i=1}^{n}\sum_{j=1}^{m} I_{\{Y_j - X_i > \theta'\}} = nm/2.$$

Когда величина θ' становится равной $\widehat{\theta}$ из формулы (10), происходит «перескок» через уровень $nm/2$.

4. Точный доверительный интервал для малых выборок строится с помощью метода 1 из § 3 гл. 11, примененного к

$$g(\boldsymbol{x}, \boldsymbol{y}, \theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} I_{\{y_j - x_i > \theta\}}.$$

Когда известно, что наблюдения имеют *нормальное* распределение (см. § 4 гл. 12), для проверки однородности можно использовать критерии из примера 1.

Пример 1. ***Однородность нормальных выборок.*** Проверим однородность двух *независимых* выборок $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$, где $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_j \sim \mathcal{N}(\mu_2, \sigma_2^2)$, причем все параметры μ_1, μ_2, σ_1, σ_2 неизвестны. Несмещенными оценками для дисперсий σ_1^2 и σ_2^2 служат

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{и} \quad S_2^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

(см. пример 3 гл. 6). В силу теоремы 1 гл. 11 $(n-1)\,S_1^2/\sigma_1^2 \sim \chi^2_{n-1}$, $(m-1)\,S_2^2/\sigma_2^2 \sim \chi^2_{m-1}$, причем S_1 не зависит от $\overline{X}$, а ввиду независимости выборок — также и от $\overline{Y}$. Это же верно и для S_2.

*) Обобщение критерия для многомерных данных см. в § 3 гл. 23.

Вопрос 4.
Что происходит с законом F_{k_1,k_2} при k_1, $k_2 \to \infty$?

Определение. Случайная величина ζ имеет F-распределение (Фишера—Снедекора) с k_1 и k_2 степенями свободы (обозначается $\zeta \sim F_{k_1,k_2}$), если

$$\zeta = \left(\frac{1}{k_1}\,\xi\right) \Big/ \left(\frac{1}{k_2}\,\eta\right), \quad \text{где } \xi \sim \chi^2_{k_1},\ \eta \sim \chi^2_{k_2},\ \xi \text{ и } \eta \text{ независимы.}$$

Критерий Фишера. Если верна гипотеза

$H'\colon \sigma_1 = \sigma_2, \mu_1$ и μ_2 — любые,

то в соответствии с приведенным выше определением статистика S_1^2/S_2^2 распределена по закону $F_{n-1,m-1}$. Ее критические значения можно найти в таблице Т5.

В случае, когда критерий Фишера не отвергает гипотезу H', для проверки однородности остается проверить гипотезу $H''\colon \mu_1 = \mu_2$.

Обозначим неизвестную общую дисперсию через σ^2 . Так как распределение хи-квадрат является частным случаем гамма-распределения ($\chi_k^2 \sim \Gamma(k/2,1/2)$), из леммы 1 гл. 4 вытекает, что

$\sigma^{-2}\left[(n-1)\,S_1^2 + (m-1)\,S_2^2\right] \sim \chi^2_{n+m-2}$.

Total (*англ.*) — общий.

Поскольку математическое ожидание закона χ^2_{n+m-2} равно $n+m-2$, статистика $S_{tot}^2 = \left[(n-1)\,S_1^2 + (m-1)\,S_2^2\right]/\,(n+m-2)$ несмещенно оценивает σ^2 по объединенной выборке.

При справедливости гипотезы H'' ввиду независимости выборок имеем: $\overline{X} - \overline{Y} \sim \mathcal{N}(0,(1/n+1/m)\,\sigma^2)$. При этом $\overline{X} - \overline{Y}$ (функция от $\overline{X}$ и $\overline{Y}$) не зависит от S_{tot} (функции от S_1 и S_2) в силу леммы о независимости из § 3 гл. 1. Отсюда согласно определению закона Стьюдента t_k с k степенями свободы (см. пример 4 гл. 11) имеем:

$$T = (\overline{X}-\overline{Y}) \Big/ \left(S_{tot}\sqrt{\frac{1}{n}+\frac{1}{m}}\right) = \sqrt{\frac{nm}{n+m}}\,(\overline{X}-\overline{Y}) \Big/ S_{tot} \sim t_{n+m-2}.$$

Это приводит к так называемому **критерию Стьюдента**, который позволяет проверить гипотезу H''. Критические значения статистики t_{n+m-2} даны в Т4.

Вопрос 5.
Какое распределение имеет статистика T^2?

Несмотря на то, что критерий Стьюдента оптимален для нормальных выборок, рассмотренная процедура проверки однородности имеет скорее теоретическое, чем практическое значение. Почему?

Во-первых, это объясняется тем, что критические значения статистики S_1^2/S_2^2 *существенно изменяются* даже при небольших возмущениях модели (см. в гл. 16 задачу 6 и замечание 2 при $k=2$).*)

Во-вторых, эффективность критерия Стьюдента быстро уменьшается при отклонении от строгой нормальности. (Относительная асимптотическая эффективность двух критериев при альтернативах правого сдвига определена, например, в [86, с. 76].) В частности,

*) Устойчивая ранговая альтернатива критерию Фишера, не предполагающая нормальности наблюдений, описывается в § 2 гл. 24.

эффективность критерия ранговых сумм Уилкоксона—Манна—Уитни по сравнению с критерием Стьюдента равна $E(F)$ (см. формулу (11)).

Рассмотрим для иллюстрации модель Тьюки смеси нормальных законов из примера 2 гл. 8 (при $\mu = 0$ и $\sigma = 1$), у которой функция распределения F выглядит так: $F_\varepsilon(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/3)$, где $\Phi(x)$ — функция распределения $\mathcal{N}(0,1)$, $0 \leqslant \varepsilon \leqslant 1$. Следующая таблица (из [86, с. 85]) показывает изменение эффективности $E(F_\varepsilon)$ в этой модели при небольшом утяжелении «хвостов».

ε	0	0,01	0,03	0,05	0,08	0,10	0,15
$E(F_\varepsilon)$	0,955	1,009	1,108	1,196	1,301	1,373	1,497

В силу теоремы 4 гл. 8 эффективность $E(F) = e_{W,\overline{X}}(F) \geqslant 0{,}864$ при всех $F \in \Omega_s$ и может быть сколь угодно велика.

Отметим также, что у критерия с $m \neq n$ по сравнению с критерием с $m' = n' = (n+m)/2$ эффективность уменьшается в $1/[4\gamma(1-\gamma)] > 1$ раз, $\gamma = n/(n+m)$ (см. [86, с. 171]), поэтому желательно брать выборки одинаковых размеров (если, конечно, есть такая возможность).

§ 6. ПРИНЦИП ОТРАЖЕНИЯ

Материал этого параграфа в основном заимствован из гл. III замечательной книги [81], которую автор настоятельно рекомендует прочитать заинтересовавшемуся читателю. В конце параграфа некоторые из полученных результатов будут использованы для решения двух задач из области проверки однородности выборок.

Рассмотрим случайное блуждание $S_n = \xi_1 + \ldots + \xi_n$, где независимые «шаги» ξ_i принимают значения $+1$ и -1 с одинаковой вероятностью $1/2$. *Траекторией (путем)* блуждания длины n будем называть ломаную, соединяющую точки плоскости с координатами (i, S_i), $i = 1, \ldots, n$. Каждый из 2^n возможных путей имеет одинаковую вероятность 2^{-n}.

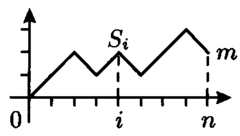

Рис. 5

Обозначим через $N_{n,m}$ количество путей, ведущих из точки $(0,0)$ в точку (n,m) (рис. 5). Пусть для такого пути k — это число шагов вверх ($\xi_i = +1$), l — число шагов вниз ($\xi_i = -1$). Тогда $k + l = n$ и $k - l = m$, откуда $k = (n+m)/2$. Расставить k «плюс единиц» по n местам можно C_n^k способами. Поэтому

$$N_{n,m} = C_n^{(n+m)/2}, \qquad (12)$$

где подразумевается, что биномиальный коэффициент равен 0, если $(n+m)/2$ не является целым числом между 0 и n.

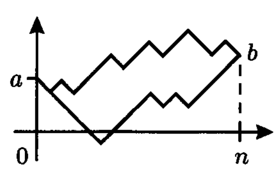

Рис. 6

Пусть a и b — положительные целые числа. Перенесем начальную ординату блуждания из 0 в a и потребуем, чтобы в момент n траектория приходила в точку с координатами (n, b) (рис. 6).

Глава 15

ПАРНЫЕ ПОВТОРНЫЕ НАБЛЮДЕНИЯ

§ 1. УТОЧНЕНИЕ МОДЕЛИ

Да вместе вы зачем?
Нельзя, чтобы случайно.

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

Методы этой главы предназначены для выявления неоднородности реализаций выборок $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$ одинакового размера, которые *нельзя считать независимыми* между собой (см. § 2 гл. 14).

Прежде всего, уточним статистическую модель из § 1 гл. 14 применительно к данной ситуации. Вычислим *приращения* $Z_i = Y_i - X_i$, $i = 1, \ldots, n$, и разложим каждое из них на две части: $Z_i = \theta + \varepsilon_i$, где θ — интересующий нас *эффект воздействия* — систематический сдвиг, который мы будем считать положительным, ε_i — *случайная ошибка*, включающая в себя влияние неучтенных факторов на Z_i.

В дополнение к допущениям Д1—Д3 из § 1 гл. 14 предположим, что выполняется условие

Д6. Случайные величины $\varepsilon_1, \ldots, \varepsilon_n$ *независимы и имеют непрерывные (вообще говоря, разные) распределения такие, что*

$$\mathsf{P}(\varepsilon_i \leqslant 0) = \mathsf{P}(\varepsilon_i \geqslant 0) = 1/2, \quad i = 1, \ldots, n.$$

Это означает, что равны нулю медианы функций распределения случайных величин ε_i (см. § 2 гл. 7).

Замечание 1. Предположения Д1—Д3 из § 1 гл. 14 не обеспечивают одинаковой распределенности ε_i. Действительно, пусть случайные величины X_1 и X_2 распределены по стандартному нормальному закону $\mathcal{N}(0,1)$ (см. § 2 гл. 3) и независимы. Положим $Y_1 = X_1 + X_2$ и $Y_2 = X_1 - X_2$. Нетрудно проверить, что Y_1 и Y_2 распределены по закону $\mathcal{N}(0,2)$ и независимы, так как $\mathbf{cov}(Y_1, Y_2) = 0$ (см. П9). Но $Z_1 = Y_1 - X_1 = X_2 \sim \mathcal{N}(0,1)$, $Z_2 = Y_2 - X_2 = X_1 - 2X_2 \sim \mathcal{N}(0,5)$. Отсюда $\varepsilon_1 = Z_1 - \theta \sim \mathcal{N}(-\theta,1)$, а $\varepsilon_2 = Z_2 - \theta \sim \mathcal{N}(-\theta,5)$. Кроме того, что ε_1 и ε_2 имеют разные распределения, они еще и зависимы, т. е. нарушается условие Д6: $\mathbf{cov}(\varepsilon_1, \varepsilon_2) = \mathbf{cov}(X_2, X_1) - 2\,\mathbf{cov}(X_2, X_2) = -2 \neq 0$.

Итак, пусть выполнено предположение Д6. Рассмотрим задачу проверки гипотезы $H_0'\colon \theta = 0$ против альтернативы $H_3'\colon \theta > 0$

(штрих указывает на то, что проверяемая гипотеза и сдвиговая альтернатива задаются не для пары законов (F, G) (см. § 1 гл. 14), а для распределений приращений Z_i). Для ее решения используем критерии знаков (§ 2) и знаковых рангов Уилкоксона (§ 3).*)

§ 2. КРИТЕРИЙ ЗНАКОВ

Выполним **следующие шаги**.

1) Зададим уровень значимости (см. § 1 гл. 12) — малую вероятность α ошибочно отвергнуть верную гипотезу H_0'.

2) Положим $U_i = I_{\{Z_i>0\}}$, $i = 1,\ldots,n$.

3) В качестве *статистики критерия знаков* возьмем сумму $S = U_1 + \ldots + U_n$ и подсчитаем ее значение s на реализациях $x_1,\ldots,x_n$ и $y_1,\ldots,y_n$.**)

Малые выборки. При $n \leqslant 15$ вычисляем фактический уровень значимости, определенный в § 1 гл. 12 (см. рис. 1):

$$\alpha_0 = \mathsf{P}_0(S \geqslant s) = 2^{-n} \sum_{i=s}^{n} C_n^i = 2^{-n} \sum_{i=0}^{n-s} C_n^i. \qquad (1)$$

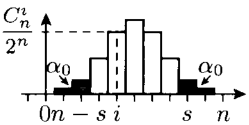

Рис. 1

Если $\alpha_0 \leqslant \alpha$, отвергаем гипотезу H_0', в противном случае — принимаем. В [10, с. 402] приведена таблица биномиальных коэффициентов, облегчающая вычисление α_0.

Большие выборки. Для расчета α_0 при $n > 15$ можно применить нормальную аппроксимацию распределения стандартизованной статистики

$$S^* = (S - \mathsf{M}S) \Big/ \sqrt{\mathsf{D}S} = (S - n/2) \Big/ \sqrt{n/4}\,.$$

Если гипотеза H_0' верна, то в соответствии с центральной предельной теоремой (П6) распределение величины S^* при $n \to \infty$ сходится к стандартному нормальному закону $\mathcal{N}(0,1)$ (см. § 2 гл. 3).

Пусть $x_{1-\alpha}$ — квантиль закона $\mathcal{N}(0,1)$ уровня $1-\alpha$ (см. § 3 гл. 7), s^* — наблюдаемое значение статистики S^*. Если $s^* \geqslant x_{1-\alpha}$, то отвергаем гипотезу H_0', в противном случае — принимаем.

Поправка. Можно значительно улучшить качество приближения дискретного биномиального распределения непрерывным нормальным законом за счет введения *поправки на непрерывность*. Рассмотрим «подправленную» статистику

$$\tilde{S}^* = (S - 0{,}5 - n/2) \Big/ \sqrt{n/4}\,. \qquad (2)$$

*) Их обобщения для многомерных данных приведены в § 2 гл. 23.

**) Если значение i-го приращения $z_i = y_i - x_i > 0$, то это отмечают знаком «+», если $z_i < 0$ — знаком «−». Отсюда происходит название критерия.

Как показывает рис. 2, сдвиг *влево* на 0,5 позволяет точнее аппроксимировать сумму площадей прямоугольников площадью под графиком *правого* «хвоста» нормальной плотности.

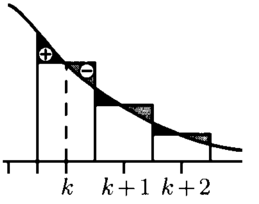

Рис. 2

Совпадения. Если среди значений Z_i встречаются нули, то их надо отбросить и соответственно уменьшить n до числа ненулевых значений Z_i.

Оценка параметра. Когда гипотеза H_0' отвергнута, принимается альтернатива H_3'. В этом случае представляет интерес величина сдвига θ. В качестве ее оценки $\widehat{\theta}$ можно взять *выборочную медиану приращений* $MED\{Z_i,\, i=1,\dots,n\}$ (см. § 2 гл. 7).

Доверительный интервал. Определим номер k_α как наибольшее число слагаемых, при котором

$$2^{-n}\sum_{i=0}^{k_\alpha} C_n^i \leqslant \alpha. \tag{3}$$

Тогда пара порядковых статистик $(Z_{(k_\alpha+1)}, Z_{(n-k_\alpha)})$ (см. § 4 гл. 4) образует доверительный интервал для θ с коэффициентом доверия $1-2\alpha$ (см. § 1 гл. 11). Для нахождения k_α можно также воспользоваться таблицей из [10, с. 353].

При большом n значение k_α с учетом поправки на непрерывность приближенно равно целой части числа

$n/2-0{,}5-x_{1-\alpha}\sqrt{n/4}$, где $x_{1-\alpha}$ — квантиль закона $\mathcal{N}(0,1)$.

Пример 1. ***Времена реакции* [80, с. 123].** Числа x_i и y_i в приведенной ниже таблице представляют собой времена реакции i-го испытуемого на световой и звуковой сигналы соответственно, $z_i = y_i - x_i$, $i = 1,\dots,12$.

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	176	163	152	155	156	178	160	164	169	155	122	144
y_i	168	215	172	200	191	197	183	174	176	155	115	163
z_i	−8	+52	+20	+45	+35	+19	+23	+10	+7	0	−7	+19

Поскольку $z_{10} = 0$, отбросим это наблюдение, уменьшив размер выборки до $n = 11$. Статистика знаков S имеет значение $s = 9$. По формуле (1) находим $\alpha_0 = (1+11+55)/2048 \approx 0{,}033$. Следовательно, на уровне значимости $\alpha \geqslant 3{,}3\%$ гипотеза $H_0'\colon \theta = 0$ отвергается.

Хотя в нашем случае $n < 15$, подсчитаем для сравнения значение статистики $\widetilde{S}^*$ по формуле (2). Получим 1,809. В таблице Т2 этому значению соответствует уровень значимости 3,5%. Упорядочив z_i по возрастанию, вычисляем оценку параметра сдвига $\widehat{\theta} = MED\{z_1,\dots,z_9,z_{11},z_{12}\} = 19$. Наконец, для $\alpha = 0{,}05$ из неравенства (3) находим $k_\alpha = 2$, что приводит к интервалу $(7,35)$ с коэффициентом доверия 90%.

Комментарии

1) Если потребовать, чтобы все величины ε_i в допущении Д6 имели одинаковое распределение, у которого нуль — единственная медиана, то в силу закона больших чисел (П6) критерий знаков будет состоятельным против альтернативы $H_3'\colon \theta > 0$.

2) В случае альтернативы $H_3''\colon \theta < 0$, очевидно, достаточно поменять местами выборки $X_1,\ldots,X_n$ и $Y_1,\ldots,Y_n$.

3) Покажем, как оценка MED параметра сдвига связана со статистикой S критерия знаков. Интуитивно понятно, что сдвиг разумно оценить такой величиной θ', чтобы набор $Z_i' = Z_i - \theta'$ $(i = 1,\ldots,n)$ выглядел как выборка из распределения с нулевой медианой. Для такой выборки S имеет биномиальное распределение (см. § 1 гл. 5), симметричное относительно своего математического ожидания $n/2$. Эти соображения приводят к следующему уравнению относительно θ':

$$\sum_{i=1}^{n} I_{\{Z_i'>0\}} = \sum_{i=1}^{n} I_{\{Z_i>\theta'\}} = n/2.$$

Когда величина θ' становится равной MED, происходит «перескок» через уровень $n/2$.

4) Нетрудно убедиться, что приведенный выше доверительный интервал получается в результате применения метода 1 из § 3 гл. 11 к функции $g(\boldsymbol{z},\theta) = \sum I_{\{z_i>\theta\}}$.

5) Ходжес и Леман показали (см. [88, с. 66]), что при оценивании сдвига с помощью MED следует использовать выборку четного размера $n = 2k$, поскольку выборочная медиана для выборки размера $2k+1$ имеет ту же самую точность.

§ 3. КРИТЕРИЙ ЗНАКОВЫХ РАНГОВ УИЛКОКСОНА

Пусть кроме допущения Д6 выполнено условие

Д7. *Случайные величины $\varepsilon_1,\ldots,\varepsilon_n$ имеют одинаковое распределение, симметричное относительно нуля:*

$F_{\varepsilon_1}(-x) = 1 - F_{\varepsilon_1}(x)$ *для всех x.*

Для проверки гипотезы H_0' против альтернативы H_3' (см. § 1) совершим **следующие шаги**.

1) Зададим уровень значимости критерия α (малую вероятность ошибочно отвергнуть верную гипотезу H_0').

2) Вычислим $Z_i = Y_i - X_i$, $i = 1,\ldots,n$, и упорядочим $|Z_1|,\ldots,|Z_n|$ по возрастанию. Пусть R_i обозначает *ранг* (порядковый номер) величины $|Z_i|$.

3) Положим $U_i = I_{\{Z_i>0\}}$, $i = 1,\ldots,n$.

4) В качестве *статистики критерия знаковых рангов* возьмем $T = R_1U_1 + \ldots + R_nU_n$ и подсчитаем ее значение t на реализациях $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$.

Малые выборки. При $n \leqslant 15$ отвергнем гипотезу H_0', если окажется, что $t \geqslant t_\alpha$, где критическое значение t_α берется из таблицы $A.4$ книги [88].

Большие выборки. Для $n > 15$ можно использовать стандартизированную статистику

$$T^* = \frac{T - \mathbf{M}T}{\sqrt{\mathbf{D}T}} = \frac{T - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24}}, \tag{4}$$

распределение которой сходится к $\mathcal{N}(0,1)$ при $n \to \infty$, если справедлива гипотеза H_0' и выполнены условия Д6—Д7 (задачи 4–6).

В случае, когда наблюдаемое значение этой статистики $t^* \geqslant x_{1-\alpha}$, где $x_{1-\alpha}$ — $(1-\alpha)$-квантиль закона $\mathcal{N}(0,1)$, гипотеза H_0' отвергается, иначе — принимается.

Поправка. В 1974 г. Р. Иман предложил следующую аппроксимацию, обеспечивающую значительное снижение относительной ошибки для критических значений. Она использует линейную комбинацию нормальной и стьюдентовской квантилей (см. [88, с. 47]). Положим

$$\tilde{t}^* = \frac{1}{2}\, t^* \left[1 + \sqrt{(n-1)/[n - (t^*)^2]}\right]. \tag{5}$$

С помощью таблиц Т2 и Т4 вычислим $z_\alpha = (x_{1-\alpha} + y_{1-\alpha})/2$, где $x_{1-\alpha}$ и $y_{1-\alpha}$ обозначают соответственно квантили уровня $(1-\alpha)$ закона $\mathcal{N}(0,1)$ и распределения Стьюдента с $(n-1)$ степенями свободы (см. § 2 гл. 11). Если $\tilde{t}^* \geqslant z_\alpha$, то гипотеза H_0' отвергается, иначе — принимается.

Совпадения. Если среди значений Z_i встречаются нули, то их надо отбросить и, соответственно, уменьшить n до числа ненулевых значений Z_i. Если среди ненулевых $|Z_i|$ есть равные, то для вычисления статистики T надо использовать средние ранги. В формуле (4) дисперсию $\mathbf{D}T$ следует заменить на

$$\frac{1}{24}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_{k=1}^{g} l_k(l_k^2 - 1)\right], \tag{6}$$

где g — число групп совпадений, l_k — количество элементов в k-й группе.*)

*) Не совпадающие с другими наблюдения считаются группой размера 1. Если совпадений нет вовсе, то сумма в выражении (6) пропадает.

Оценка параметра. Когда гипотеза H_0' отвергается, в качестве оценки параметра сдвига θ можно взять *медиану средних Уолша* (см. § 3 гл. 8)

$$W = MED\{(Z_i + Z_j)/2,\ 1 \leqslant i \leqslant j \leqslant n\}.$$

Доверительный интервал. Построение доверительного интервала для случая $n \leqslant 15$ описано в [88, с. 55]. При больших n пара порядковых статистик $(V_{(k_\alpha+1)}, V_{(M-k_\alpha)})$ образует приближенный доверительный интервал с коэффициентом доверия $1 - 2\alpha$. Здесь $V_{(1)} \leqslant \ldots \leqslant V_{(M)}$ — упорядоченные по возрастанию *средние Уолша* $(Z_i + Z_j)/2$ при $1 \leqslant i \leqslant j \leqslant n$ и $M = n(n+1)/2$; k_α — это целая часть числа

$$n(n+1)/4 - 0{,}5 - x_{1-\alpha}\sqrt{n(n+1)(2n+1)/24}, \qquad (7)$$

где $x_{1-\alpha}$ обозначает, как и ранее, $(1-\alpha)$-квантиль закона $\mathcal{N}(0,1)$, а 0,5 представляет собой поправку на непрерывность (см. § 2).

Проверка симметрии. Прежде чем применять критерий знаковых рангов, следует удостовериться в справедливости допущения Д7. Простой графический метод проверки основан на сходимости выборочных квантилей к теоретическим (см. § 3 гл. 7). Так как для теоретических квантилей z_p симметричного относительно медианы $z_{1/2}$ закона верно равенство $z_{1/2} - z_p = z_{1-p} - z_{1/2}$, то для порядковых статистик $Z_{(i)}$ можно ожидать выполнения соотношений

$$\xi_i = MED - Z_{(i)} \approx \eta_i = Z_{(n+1-i)} - MED, \quad i = 1, \ldots, [n/2],$$

(здесь $[\cdot]$ обозначает целую часть числа). Поэтому для выборки $Z_1, \ldots, Z_n$ из симметричного относительно медианы распределения точки плоскости (ξ_i, η_i) должны располагаться вблизи диагонали $y = x$.

Замечание 2. Условие *строгой симметрии* относительно медианы является почти столь же нереалистичным, как и предположение, что распределение величин Z_i в точности нормально. Как правило, надежно проверить симметрию можно лишь по выборке из нескольких сотен наблюдений. Асимптотический критерий Гупты для решения этой проблемы приведен в [88, с. 76]. Ссылки на другие критерии см. там же на с. 81.

Предположение о симметрии иногда оказывается справедливым в силу специфики получения наблюдений, приводящей к одинаковым вероятностям отклонения на произвольную величину от медианы как влево, так и вправо.

Симметрия распределения величин Z_i довольно естественно возникает в модели «контроль — обработка» (см. пример 1 гл. 8). Однако подчеркнем еще раз, что в случае совместной независимости выборок $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$ для проверки гипотезы однородности следует использовать не критерии знаков и знаковых рангов Уилкоксона, а методы, изложенные в гл. 14.

Пример 2. Для данных из примера 1 проверим гипотезу $H_0'\colon \theta = 0$ с помощью критерия знаковых рангов. После отбрасывания $Z_{10} = 0$ в выборке останется $n = 11$ наблюдений. Упорядочим их:

$Z_{(i)}$	-8	-7	7	10	19	19	20	23	35	45	52

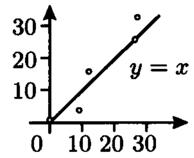

Рис. 3

Видим, что $MED = 19$. Для визуальной проверки симметрии построим точки (ξ_i,η_i), определенные выше. Проведем прямую $y = x$ (рис. 3). Хотя выборка слишком мала для уверенного заключения, построенная диаграмма, по-видимому, не противоречит допущению о симметрии распределения случайной величины Z_i.

Упорядочим по возрастанию величи́ны $|Z_i|$ и присвоим средние ранги совпадающим значениям:

$\lvert Z_i\rvert$	7	7	8	10	19	19	20	23	35	45	52
R_i	1,5	1,5	3	4	5,5	5,5	7	8	9	10	11
U_i	0	1	0	1	1	1	1	1	1	1	1

Согласно приведенной таблице статистика критерия знаковых рангов $T = \sum R_i U_i = 61{,}5$. Учитывая, что среди величин $|Z_i|$ есть две группы совпадений, по формуле (6) вычислим дисперсию $\mathbf{D}T = (11 \cdot 12 \cdot 23 - 3 - 3)/24 = 126{,}25$. Отсюда по формуле (4) для нормированной статистики T^* получаем значение $t^* \approx 2{,}54$. Положив $\alpha = 0{,}005$, из таблиц Т2 и Т4 (при $k = n - 1 = 10$) находим $z_\alpha = (2{,}576 + 3{,}169)/2 \approx 2{,}87$. Согласно формуле (5) имеем $\tilde{t}^* = 3{,}14$. Так как $3{,}14 \geqslant 2{,}87$, то гипотеза H_0' отвергается на уровне значимости 0,005.

С помощью компьютера вычисляем значение оценки параметра сдвига $W = MED\{(Z_i + Z_j)/2,\ i \leqslant j\} = 19{,}25$. На основе формулы (7) строим 90%-ный доверительный интервал $\big(V_{(15)},V_{(52)}\big) = (15/2, 31)$, который несколько у́же интервала $(7, 35)$, полученного ранее при применении критерия знаков к этим же данным.

Комментарии

1) Если в условии Д7 все ε_i имеют одинаковое *симметричное гладкое распределение* (см. § 1 гл. 8), то критерий знаковых рангов будет состоятельным против альтернативы $H_3'\colon \theta > 0$ (см. [86, с. 64]).

2) Покажем, что связь между статистикой T критерия знаковых рангов и медианой средних Уолша W аналогична рассмотренной ранее связи между статистикой S критерия знаков и выборочной медианой MED. Согласно задаче 1 при отсутствии нулевых значений и совпадений среди величин $|Z_i|$ статистика знаковых

Задачи

1. Установить формулу, связывающую статистики U и V критерия Уилкоксона — Манна — Уитни.

2. Доказать, что в случае однородности выборок $\mathsf{D}U = nm(n+m+1)/12$.

3. Найти предельное распределение статистики критерия Смирнова $\sqrt{nm/(n+m)}D_{n,m}$ для выборок из распределения Бернулли. (*Используйте центральную предельную теорему*.)

4. Найти, к какому предельному распределению сходится распределение Фишера при стремлении к бесконечности обеих степеней свободы.

5*. Вывести формулу для статистики критерия Розенблатта $\omega^2_{n,m}$, содержащую ранги R_i и S_j.