Однородность выборок

Не в совокупности ищи единства, но более — в единообразии разделения.

Козьма Прутков

Гипотеза однородности

<u>Данные</u>. 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 всего изложения будем считать выполняющимися

<u>Допущения</u>

Д1. $X_1, \ldots, X_n$ — *независимы и имеют общую ф. р.* $F(x)$.

Д2. $Y_1, \ldots, Y_m$ — *независимы и имеют общую ф. р.* $G(x)$.

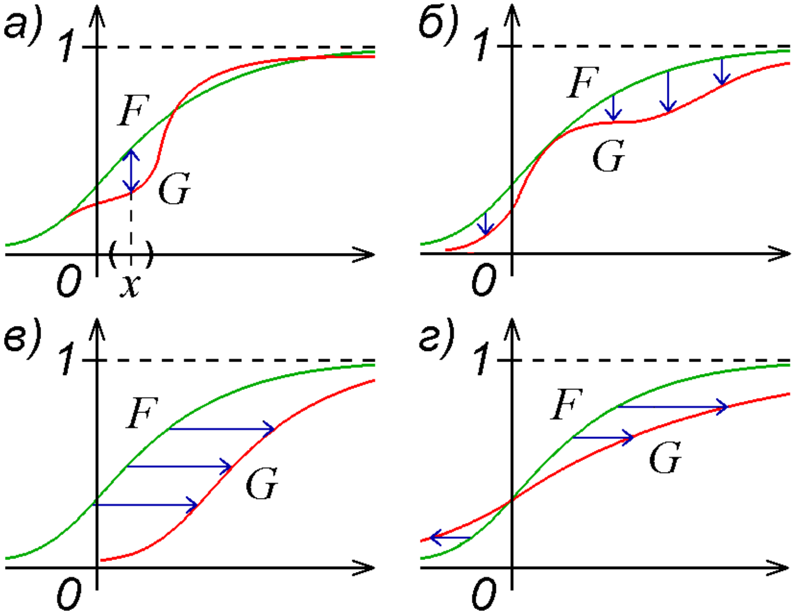
Д3. *Обе функции* F *и* G *неизвестны, но принадлежат множеству всех непрерывных функций распределения* Ω_c.

Нас будет интересовать

<u>Гипотеза однородности</u>

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

Альтернативы однородности



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

б) *доминирования* $H_2:\ G(x) \leqslant F(x)$ при всех x, причем хотя бы для одного x неравенство строгое;

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

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

Причины рассмотрения альтернатив

- С практической точки зрения бывает важно уловить отклонения от H_0 только определенного вида, скажем, наличие систематического прироста у элементов второй выборки относительно элементов первой выборки

- За счет сужения по сравнению с альтернативой неоднородности H_1 множества распределений, составляющих альтернативное подмножество, удается построить более эффективные (чувствительные) критерии, настроенные на обнаружение отклонений от H_0 конкретного вида

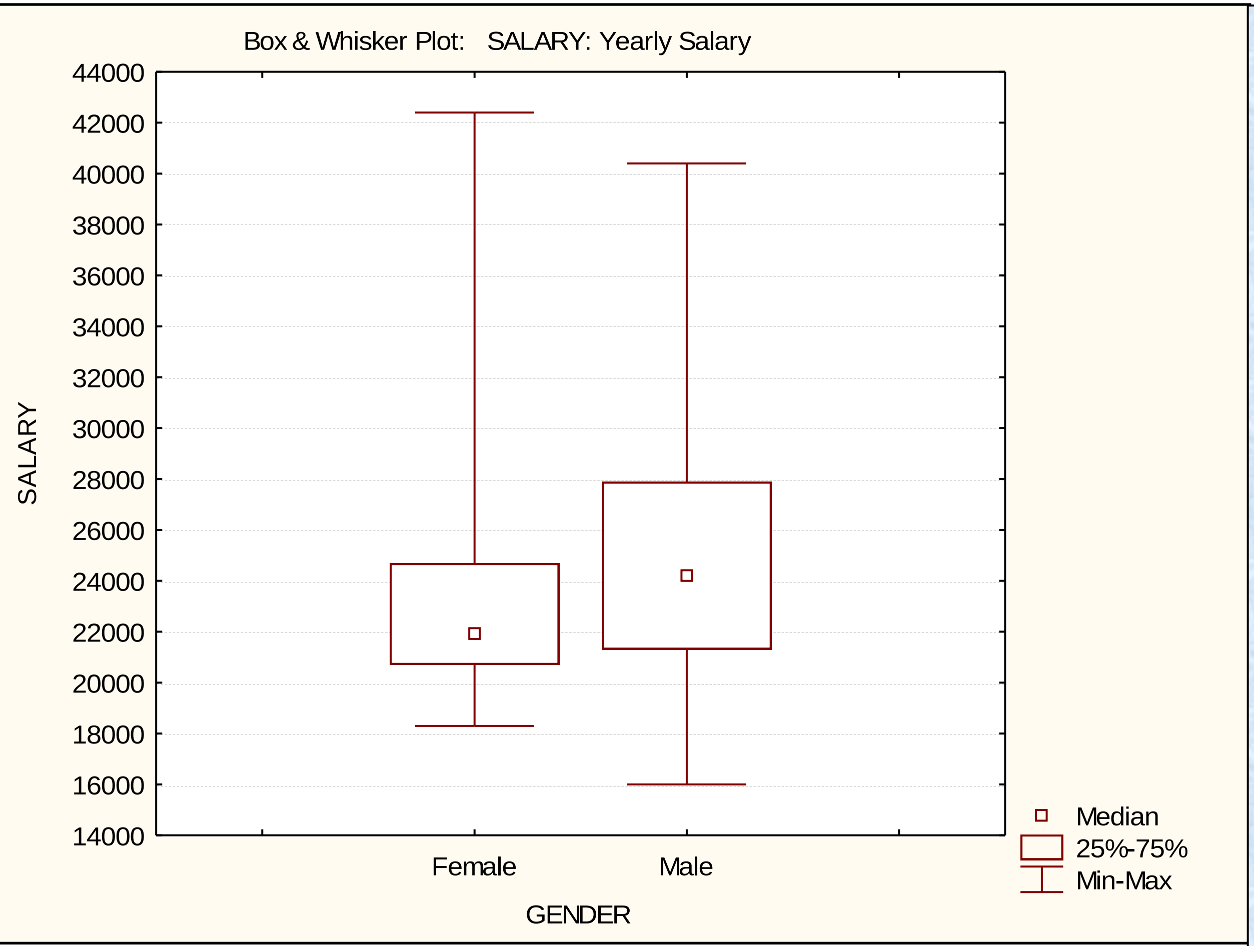
Правильный выбор модели

Две реализации независимых между собой выборок (групп)

При проверке гипотезы однородности двух наборов данных важно понять, с каким из *двух случаев* мы имеем дело

Парные повторные наблюдения («до» и «после»)

Сравнение зарплат мужчин и женщин



Выдвижение гипотезы

Видим, что в среднем зарплата мужчин немного выше. При этом типичный диапазон её изменения — межквартильный размах — вдвое шире.

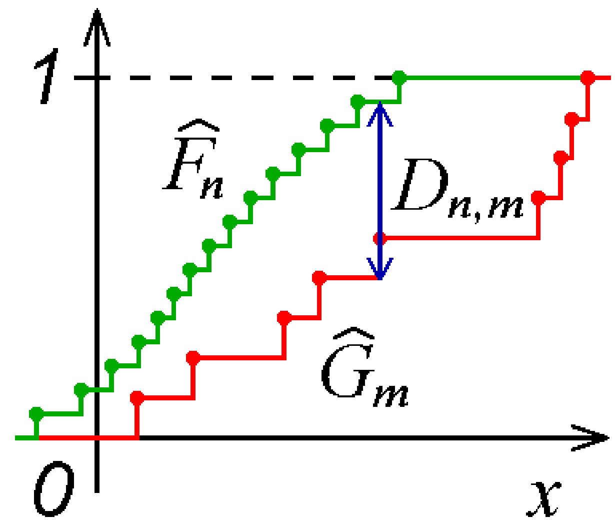
Для выяснения статистической значимости отличия выборок (групп) применяются непараметрические критерии

Смирнова, Розенблатта, Манна — Уитни.

Критерий Смирнова

Критерий предназначен для обнаружения наиболее общей *альтернативы неоднородности* H_1.

Он базируется на величине максимального различия между эмпирическими функциями выборок



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

$$\text{где } \widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leqslant x\}}, \quad \widehat{G}_m(x) = \frac{1}{m} \sum_{j=1}^{m} I_{\{Y_j \leqslant x\}},$$

т. е. $D_{n,m}$ — расстояние в равномерной метрике между эмпирическими функциями выборок. Слишком большое расстояние противоречит гипотезе H_0.

Предельная теорема для статистики критерия Смирнова

Предположим, что кроме допущений **Д1 – Д3**, выполнены ещё два допущения:

Д4. Все наблюдения $X_1, \ldots X_n, Y_1, \ldots, Y_m$ независимы.

Д5. Размеры выборок увеличиваются пропорционально: $n / (n + m) \to \alpha$, $0 < \alpha < 1$.

Если верна гипотеза однородности H_0, то при $n \to \infty$ и $m \to \infty$ имеет место сходимость

$$\mathbf{P}\left(\sqrt{nm/(n+m)}\, D_{n,m} \leqslant x\right) \to K(x).$$

Таким образом, предельным законом для статистики критерия Смирнова служит распределение Колмогорова. Поэтому данный критерий также называют *критерием Колмогорова — Смирнова.*

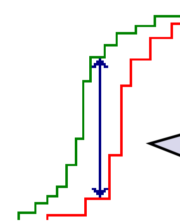
Критерий Розенблатта

Для проверки гипотезы однородности H_0 двух независимых выборок против альтернативы неоднородности H_1 можно воспользоваться также критерием типа ω^2. Статистика критерия задается формулой

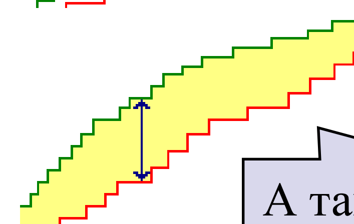
$$\omega^2_{n,m} = \int\limits_{-\infty}^{\infty} [\widehat{F}_n(x) - \widehat{G}_m(x)]^2\, d\widehat{H}_{n+m}(x),$$

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

Таким образом, статистика критерия Розенблатта измеряет расхождение между эмпирическими функциями в средне-квадратичном смысле — по существу, на основе величины площади области между эмпирическими функциями.



Такое отличие лучше обнаруживает критерий Смирнова



А такое — критерий Розенблатта

Ранговая форма статистики критерия Розенблатта

Статистику критерия Розенблатта можно также можно представить в виде

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

где R_i — *ранг* (номер по порядку возрастания значений в объединённой выборке) величины $X_{(i)}$, S_j — ранг величины $Y_{(j)}$.

М. Розенблатт в 1952 г. установил, что при выполнении допущений Д1-Д5

$$\mathbf{P}\left(\frac{nm}{n+m}\omega^2_{n,m} \le x\right) \to A_1(x) = 1 - \frac{1}{\pi}\sum_{j=1}^{\infty}(-1)^{j+1}\int_{(2j-1)^2\pi^2}^{4j^2\pi^2}\sqrt{\frac{-\sqrt{y}}{\sin(\sqrt{y})}}\frac{e^{-xy/2}}{y}dy$$

Реализация критерия Розенблатта на языке R

Критерий Розенблатта на языке R реализован функциями cvmts.test и cvmts.pval из пакета CvM2SL2Test (Cramer-von Mises Two Sample Test в метрике L2), вычисляющими, соответственно, статистику критерия и фактический уровень значимости (p-value).

Пакет CvM2SL2Test находится не в Repository (CRAN), а хранится в архиве. Оттуда его можно скачать, зайдя на сайт
https://cran.r-project.org/src/contrib/Archive/CvM2SL2Test/
и затем установить кнопкой Install, выбрав в поле ввода Install from: опцию Package Archive File (.zip, .tar.gz).

Для установки пакета CvM2SL2Test под Windows предварительно потребуется установить программу RTools.exe, которую можно скачать с сайта
https://cran.r-project.org/bin/windows/Rtools/

Альтернатива критерию Розенблатта

Альтернативой критерию Розенблатта является критерий, предложенный в статье L. Baringhaus and C. Franz (2004) On a new multivariate two-sample test, Journal of Multivariate Analysis, 88, p. 190-206, допускающий обобщение на случай многомерных выборок. Его статистикой служит

$$T_{n,m}=\frac{nm}{n+m}\left[\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\|X_i-Y_j\right\|-\frac{1}{2n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\left\|X_i-X_k\right\|-\frac{1}{2m^2}\sum_{j=1}^{m}\sum_{k=1}^{m}\left\|Y_j-Y_k\right\|\right]$$

В одномерном случае $T_{m,n}=\frac{nm}{n+m}\int_{-\infty}^{\infty}(\hat{F}_n(x)-\hat{G}_m(x))^2\,dx.$ Эта статистика была предложена Х. Крамером ещё в 1928 г. В отличие от критерия Розенблатта данный критерий не является свободным от вида непрерывного закона. Поэтому его p-value приходится вычислять методом bootstrap. На языке R он реализован функцией cramer.test из пакета cramer.

Практическое задание 1

В файле Prefix-ver.txt содержатся логарифмы частот встречаемости (признак LogFrequency) 985 голландских слов с приставкой ver. Также указан семантический класс слов (признак SemanticClass), имеющий две категории:
"opaque" — трудные для понимания, несоставные слова;
"transparent" — семантически простые, составные слова.

1) Импортируйте файл Prefix-ver.txt в Rstudio

2) Отберите слова класса "opaque", запишите их логарифмы частот в выборку x. Отберите слова класса "transparent", запишите их логарифмы частот в выборку y

3) Постройте гистограммы для выборок x и y (Похожи ли они?)

4) Постройте графики эмпирических функций распределения выборок x и y на одном рисунке: при построении 2-го графика задайте для команды plot аргументы add=TRUE, col=4

5) Установ пакет cramer, проверьте однородность выборок критериями Смирнова (ks.test) и Барингхауса — Франца (kramer.test)

Односторонний критерий Смирнова

Статистикой этого критерия служит

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

Теорема. При справедливости гипотезы H_0 и выполнении допущений Д1-Д5 для любого $x \geq 0$

$$\mathbf{P}\left(\sqrt{nm/(n+m)}\, D^{+}_{n,m} \leq x\right) \to 1 - e^{-2x^2}. \qquad (1)$$

Согласно определению функции Колмогорова для её правого «хвоста» выполняется разложение

$$1 - K(x) = 2(e^{-2x^2} - e^{-8x^2} + e^{-18x^2} - \ldots). \qquad (2)$$

Второй член заключённого в скобки ряда представляет собой четвёртую степень его первого члена. Пренебрегая им и всеми последующими членами, из сравнения формул (1) и (2) видим, что фактический уровень значимости одностороннего критерия Смирнова примерно вдвое меньше, чем у двустороннего.

Критерий Манна — Уитни

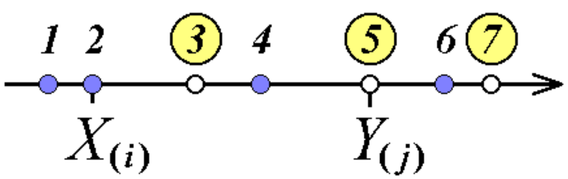
Критерий Манна — Уитни применяется для проверки гипотезы однородности H_0 против альтернативы доминирования H_2, в частности, — против альтернативы правого сдвига H_3.

Вычислим статистику V *критерия Манна — Уитни*.

1) Обозначим через S_j ранг порядковой статистики $Y_{(j)}$ ($j = 1, \ldots, m$) в вариационном ряду, построенном по объединенной выборке $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$.

2) Положим $V = S_1 + \ldots + S_m$.

Критерий, основанный на 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 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 H_2. Таким образом, критическая область выбирается в виде $\{V > c\}$, где c — некоторая константа.

Основная теоретическая проблема, решённая Манном и Уитни, заключалась в нахождении для заданной границы c соответствующего уровня значимости критерия α_c

Эквивалентные статистики

Несложно установить, что статистика V критерия Манна — Уитни и статистика U, определяемая формулой

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I_{\{X_i < Y_j\}},$$

У. Краскел нашел статистику U в работе Г. Дехлера, опубликованной в Германии в 1914 г.

связаны равенством

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

т. е. они отличаются на известную константу. Поэтому V и U эквивалентны в том смысле, что задают одинаковый критерий.

Аналогичным свойством обладает и статистика разности средних рангов

$$T = \frac{V}{m} - \frac{W}{n},$$

где $W = R_1 + \ldots + R_n$ — сумма рангов элементов X_i из первой выборки в объединённой выборке (поскольку, очевидно, что $V + W = (n+m)(n+m+1)/2.$)

Асимптотическая нормальность статистики критерия Манна — Уитни

Предложенная Уилкоксоном *ранговая форма* статистики критерия V более удобна для вычислений. В свою очередь, с помощью *считающей формы* U, изученной Манном и Уитни, нетрудно установить, что в случае справедливости гипотезы однородности выборок выполняются следующие равенства:

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

Если гипотеза однородности выборок H_0 верна, то при выполнении допущений **Д1 – Д5** имеет место сходимость распределения стандартизированной статистики U^* к стандартному нормальному закону:

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

Для альтернативы доминирования критической 5%-ной границей служит величина 1,645.

Практическое задание 2

1) Моделируйте две независимые выборки x и y размера 50:

а) выборка x должна быть равномерно распределена на отрезке [0, 1]: её функция распределения $F(x) = x$ на [0, 1]

б) выборка y должна иметь функцию распределения $G(x) = x^3$ на отрезке [0, 1] (используйте метод обратной функции)

2) Постройте на одном рисунке разноцветные графики эмпирических функций распределения выборок x и y

3) Проверьте выборки на однородность односторонними критериями Смирнова (ks.test) и Манна – Уитни (wilcox.test)

Предупреждение! Внимательно изучите описания функций ks.test и wilcox.test, чтобы правильно указать значение аргумента alternative для каждого из критериев.

Модель повторных наблюдений

Для выявления неоднородности реализаций двух зависимых выборок $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$ обычно используется следующая модель. Рассматриваются *приращения*

$$Z_i = Y_i - X_i, \quad i = 1, \ldots, n.$$

Каждое из них раскладывается 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.

Предполагается, что выполняется допущение

Сл. в. $\varepsilon_1, \ldots, \varepsilon_n$ — *независимы и имеют непрерывные* (вообще говоря, разные) *распределения такие, что*

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

При условии выполнения этого допущения проверяется гипотеза $H_0\colon\ \theta = 0$ против альтернативы $H_1\colon\ \theta > 0$.

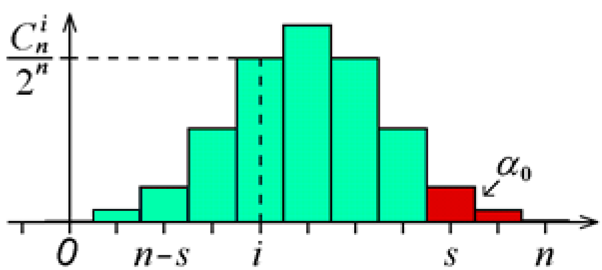
Критерий знаков

Критерий знаков предназен для выявления неоднородности реализаций выборок $X_1, \dots, X_n$ и $Y_1, \dots, Y_n$ одинакового размера, которые <u>нельзя считать независимыми</u> между собой.

1) Зададим уровень значимости — малую вероятность α ошибочно отвергнуть верную гипотезу однородности выборок.

2) Положим $Z_i = Y_i - X_i$ и $U_i = I_{\{Z_i > 0\}}$, $i = 1, \dots, n$.

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



Малые выборки. При $n \leqslant 15$ вычисляем фактический уровень значимости (p-level)

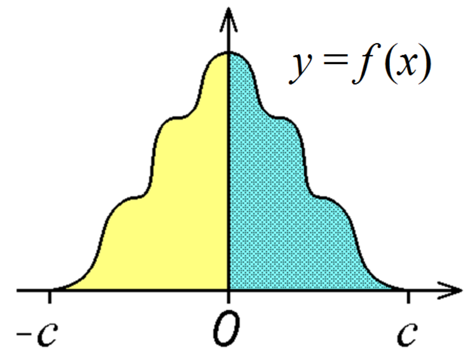
$$\alpha_0 = \mathbf{P}_0(S \geqslant s) = 2^{-n} \sum_{i=s}^{n} C_n^i.$$

Если $\alpha_0 \leqslant \alpha$, отвергаем гипотезу H_0.

Критерий знаковых рангов

Предположим, что приросты имеют вид $Z_i = Y_i - X_i = \theta + \varepsilon_i$ где ε_i — независимые случайные величины, распределение которых непрерывно и симметрично относительно нуля:

$$F(-x) = 1 - F(x) \quad \text{для всех } x.$$



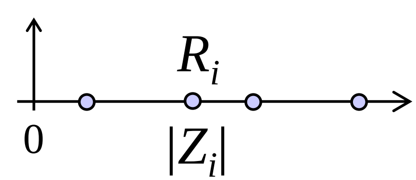
Плотность $f(x)$ — чётная функция

Тогда для проверки гипотезы

$$H_0\colon \theta = 0$$

можно применить критерий знаковых рангов, предложенный Уилкоксоном. Статистикой критерия служит взвешенная сумма

$$T = R_1 U_1 + \ldots + R_n U_n,$$



где в качестве *весов* при индикаторах U_i стоят ранги R_i абсолютных величин приростов Z_i.

Асимптотическая нормальность статистики критерия знаковых рангов

При выполнении допущений, указанных выше в описании критерия знаковых рангов, справедлива следующая

Теорема. При $n \to \infty$

$$T^* = \frac{T - \mathbf{M}T}{\sqrt{\mathbf{D}T}} = \frac{T - [n(n+1)/4]}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{d} Z \sim \mathcal{N}(0,1).$$

Для односторонней альтернативы $\{H_1\colon \theta > 0\}$ критической границей на уровне значимости 0,05 служит величина 1,645.

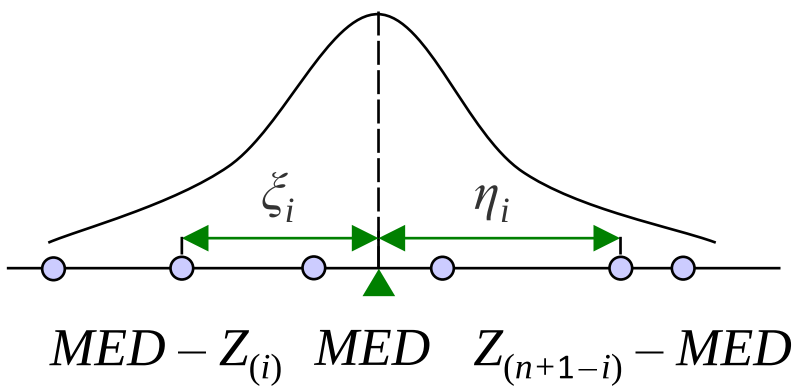
Визуальная проверка симметрии распределения

Одним из способов проверки симметрии распределения служит визуальный метод, состоящий в построении диаграммы рассеяния точек с координатами (ξ_i, η_i), где

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

Если распределение симметрично относительно медианы, то точки с координатами (ξ_i, η_i) должны располагаться вблизи прямой $y = x$.

Для проверки симметрии при $n \geq 100$ можно использовать асимптотические критерий Гаствирта, основанный на отличии $\overline{X}$ и MED, или критерий Орлова (см. след. слайд).



Асимптотический критерий Орлова для проверки симметрии

Проверить симметрию распределения относительно 0 можно с помощью **критерия Орлова** (критерий типа омега-квадрат).

Статистикой критерия Орлова служит

$$R = \sum_{i=1}^{n} (1 - \hat{F}_n(Z_i) - \hat{F}_n(-Z_i))^2 .$$

Видим, что R — сумма квадратов нарушений условия симметрии $F(-x) = 1 - F(x)$ в точках Z_i, правда не для самой F, а для её оценки $\widehat{F}_n$.

А. И. Орлов в 1972 г. установил, что при выполнении предположения о симметрии распределение статистики R сходится при $n \to \infty$ к некоторому предельному закону. Приведём три квантили этого закона: $x_{0,9} = 1{,}2$, $x_{0,95} = 1{,}66$, $x_{0,99} = 2{,}8$.

Известны и другие асимптотические критерии проверки симметрии, например,

Hollander M., Wolfe D., Chicken E. Nonparametric Statistical Methods, 3rd Edition, 2013, p. 94.

Практическое задание 3

В файле Pressure.txt приводятся для 15 пациентов данные о систолическом и диастолическом давлении крови до принятия и спустя 2 часа после принятия 25 мг каптоприла. Является ли снижение *систолического* давления статистически значимым?

1) Импортируйте файл Pressure.txt и запишите значения признаков SistBefore и SistAfter в векторы x и y соответственно

2) Вычислите приросты z, исключите нулевые приросты, если они будут, для оставшихся приростов постройте гистограмму

3) Проверьте гипотезу отсутствия эффекта от каптоприла критериями знаков и знаковых рангов Уилкоксона:
для 1-го критерия вычислите p-value с помощью функции pbinom,
для 2-го критерия используйте функцию wilcox.test

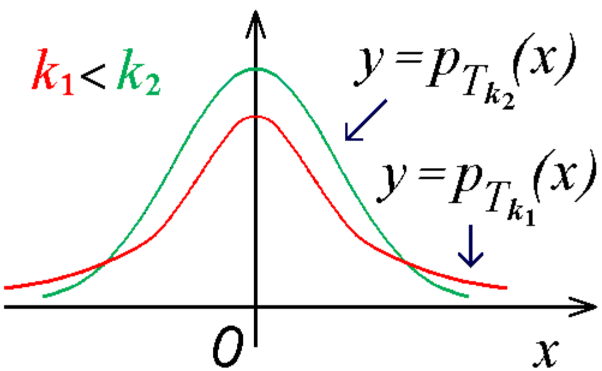
4) Постройте диаграмму для проверки симметрии приростов (используйте цикл для подсчёта ξ_i и η_i). Можно ли считать распределение приростов симметричным, т. е. законно ли в данном случае использовать критерий знаковых рангов?

Критерий Стьюдента

Критерий Стьюдента (или t-тест) позволяет проверять гипотезу H_0 о равенстве средних μ_1 и μ_2 двух независимых нормальных выборок:

$$X_i \sim \mathcal{N}(\mu_1, \sigma^2)\ (i = 1, \ldots, n) \ \text{ и } \ Y_j \sim \mathcal{N}(\mu_2, \sigma^2)\ (j = 1, \ldots, m),$$

где дисперсия σ^2 предполагается одинаковой, но неизвестной.



Известно, что при справедливости гипотезы H_0 статистика

$$T = \sqrt{\frac{nm}{n+m}}\left(\overline{X} - \overline{Y}\right)\Big/ S$$

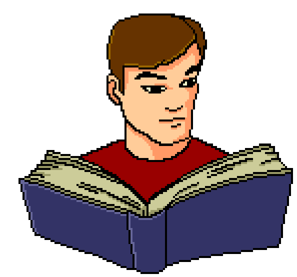
имеет распределение Стьюдента с $n+m-2$ степенями свободы, обычно обозначаемое как t_{n+m-2}. Здесь

$$S^2 = \left[\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{j=1}^{m}(Y_j - \overline{Y})^2\right]\Big/ (n+m-2).$$

Происхождение названия критерия

Распределение Стьюдента t_k впервые появилось в статье Уильяма Д. Госсета (1908), который в то время работал в Дублине на пивоваренном заводе Гиннеса.

Условия контракта не позволяли ему публиковать результаты исследований под собственным именем. Госсет выбрал скромный псевдоним «Student».



Почему не следует применять критерий Стьюдента на практике

- Главным недостатком критерия является допущение о строгой нормальности распределения элементов выборок, которое обычно не выполняется для реальных данных. Для надёжной проверки сложной гипотезы нормальности требуются несколько сотен наблюдений, которых может и не быть в распоряжении исследователя.
- Установлено, что малейшее отклонение распределения от нормального закона приводит к очень быстрому снижению эффективности (чувствительности) критерия Стьюдента (см. следующий слайд).
- Другим условием применимости критерия является равенство дисперсий. Классический F-критерий, предназначенный для проверки этого равенства, весьма чувствителен к утяжелению «хвостов» распределения. Правда, имеются его альтернативы — критерии Ливиня и Брауна — Форсайта. Однако они основаны на абсолютных отклонениях, поэтому тоже боятся «выбросов».
- Есть ранговые критерии, например, критерий Манна — Уитни, которые даже на нормальном законе уступают критерию Стьюдента в эффективности лишь 4,5%, а выиграть могут сколь угодно много на распределениях с тяжёлыми «хвостами».

Неробастность критерия Стьюдента

Робастность (robust (англ.) — крепкий, надёжный, устойчивый) — свойство критерия или оценки, означающее незначительное уменьшение эффективности при малом изменении («возмущении») модели.

Рассмотрим для иллюстрации *модель Тьюки* смеси двух нормальных распределений с математическим ожиданием 0 и стандартными отклонениями 1 и 3 соответственно. Функция распределения $F_\varepsilon(x)$ смеси имеет вид

$$F_\varepsilon(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/3),$$

где $\Phi(x)$ — функция распределения $\mathcal{N}(0,1)$, $0 \leqslant \varepsilon \leqslant 1$. Следующая таблица показывает изменение эффективности $E(F_\varepsilon)$ в этой модели при небольшом утяжелении «хвостов».

Эффективность критерия Манна — Уитни относительно критерия Стьюдента

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

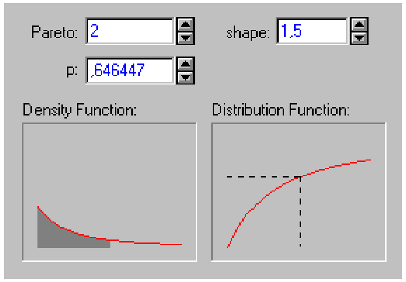
Известно также, что для всех гладких симметричных распределений F

$$E(F) \geq 0{,}864,$$

причём эффективность $E(F)$ может быть сколь угодно велика.

Распределение Парето

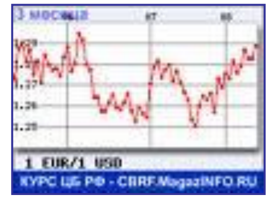
Говорят, что случайная величина X имеет (стандартное) распределение Парето с параметром $\alpha > 0$, если



$$P\{X \leqslant x\} = \begin{cases} 1 - x^{-\alpha}, & \text{если } x \geqslant 1, \\ 0, & \text{если } x < 1. \end{cases}$$

Распределение получило свое название в честь Вилфредо Парето, инженера, ставшего экономистом. Парето получил данное распределение примерно в 1895 году в результате анализа статистических данных о налогах на доход в Англии за 1843 год. Он был поражен тем фактом, что доля индивидуумов, чей доход в два раза превосходил средний, была существенно больше, чем следовало из предположения о нормальности распределения. Изучая данные о доходах во всей Европе в конце 19-го столетия, Парето предложил в качестве приближения для α значение 1,5, которое интерпретировалось им как индикатор заметного неравенства людей в доходах.

Распределения с тяжёлыми «хвостами» в финансовой математике



Б. Мандельброт заметил, что предположение о нормальности распределения цен на хлопок не согласуется с реальными данными. По этой причине он рекомендовал использовать для описания финансового рынка распределения с более тяжелыми «хвостами», чем у нормального закона. В одном из исследований по данному вопросу Э. Ф. Фама, изучая распределения цен акций, использовал закон Парето и получил для параметра α этого закона оценку 1,8.

Преимущества ранговых методов

- Инвариантность относительно формы распределения наблюдений (требуется только непрерывность функции распределения)
- Малая потеря в эффективности ($\leq$ 5%) по сравнению с критериями, опирающимися на предположение о том, что известен (с точностью до параметров) закон распределения наблюдений (обычно — нормальный)
- Хорошая защищённость от «выбросов» (выделяющихся значений) благодаря замене наблюдений на их ранги
- Применимость к малым выборкам (из 5 – 15 наблюдений)

См. книгу Hollander M., Wolfe D., Chicken E. «Nonparametric Statistical Methods», Third Edition, и сопровождающую её библиотеку NSM3 языка R.

Правильный выбор критерия

Главное в теме

1) Необходимо различать, с чем мы имеем дело: реализациями независимых между собой выборок или парными повторными наблюдениями

2) Важно знать, против каких именно альтернатив могут быть использованы те или иные критерии

3) Перед вычислением уровней значимости следует построить диаграммы размахов для обнаружения «выбросов» и гистограммы для выявления асимметрии

4) Критерий Розенблатта часто оказывается более чувствительным, чем критерий Смирнова, но не всегда

5) Критерий знаковых рангов обычно более эффективен, чем критерий знаков, но перед его применением следует проверить симметрию распределения приростов

6) Критерий Стьюдента можно использовать только как *вспомогательный инструмент,* на практике его вполне может заменить критерий Манна — Уитни


Проблема обнаружения неоднородности

Домашнее задание

1) Моделируйте две независимые выборки размера $n = 50$ из законов $N(0,1)$ и $N(0,9)$ соответственно.

2) Проверьте гипотезу однородности выборок критериями Смирнова и Барингхауса — Франца

3) Проверьте гипотезу однородности выборок критерием Манна – Уитни

4) Постройте графики эмпирических функций распределения выборок

5) Объясните, почему критерий Манна – Уитни не обнаруживает различия распределений выборок