Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное
образовательное учреждение высшего образования
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

А. В. Бородина
Р. С. Некрасова

СТАТИСТИЧЕСКИЕ КРИТЕРИИ В АНАЛИЗЕ ДАННЫХ

Учебное пособие для обучающихся по направлениям подготовки бакалавриата «Математика», «Прикладная математика и информатика», «Программная инженерия», «Информационные системы и технологии»

Петрозаводск
Издательство ПетрГУ
2023

УДК 519.22
ББК 22.172
Б833

Печатается по решению редакционно-издательского совета
Петрозаводского государственного университета

Р е ц е н з е н т ы :
А. С. Румянцев, кандидат физико-математических наук,
доцент кафедры прикладной математики и кибернетики
Петрозаводского государственного университета;
О. В. Лукашенко, кандидат физико-математических наук, научный
сотрудник лаб. математической кибернетики Института прикладных
математических исследований Карельского научного центра РАН

Бородина, Александра Валентиновна.

Б833 Статистические критерии в анализе данных : учебное пособие для обучающихся по направлениям подготовки бакалавриата «Математика», «Прикладная математика и информатика», «Программная инженерия», «Информационные системы и технологии» / А. В. Бородина, Р. С. Некрасова ; М-во науки и высш. образования Рос. Федерации, Федер. гос. бюджет. образоват. учреждение высш. образования Петрозав. гос. ун-т. – Петрозаводск : Издательство ПетрГУ, 2023. – 45 с.
ISBN 978-5-8021-4040-6

Одной из основных задач статистической обработки данных является анализ результатов экспериментов на основе статистических критериев. В настоящем издании в доступной сжатой форме представлены базовые теоретические результаты по проверке статистических гипотез, рассмотрены основные виды статистических критериев и их примеры. Описание алгоритмов вычисления статистик дополнено актуальными реализациями на языке Python, предложены способы визуализации и структурирования данных. Пособие также содержит рекомендации к лабораторным занятиям и самостоятельной работе по курсу «Прикладная статистика».

УДК 519.22
ББК 22.172

ISBN 978-5-8021-4040-6

Оглавление

Введение

В общем случае анализ данных как область математики включает в себя извлечение знаний из экспериментальных (выборочных) данных, а также преобразование и моделирование данных с целью извлечения полезной информации и принятия решений. Анализ данных имеет множество подходов, затрагивающих широкий спектр областей. Представленный материал сконцентрирован на обработке результатов случайных экспериментов посредством статистического анализа. В частности, в настоящем учебном пособии подробно рассмотрен механизм проверки статистических гипотез. По сути, проверка гипотез – это многоступенчатая процедура, которая на основании данных частной выборки и при помощи методов теории вероятностей позволяет сделать вывод об обоснованности гипотезы – предложении о вероятностных свойствах рассматриваемой (неизвестной) случайной величины. Другими словами, это способ проверить, действительны ли результаты, полученные на выборке, и для всех элементов генеральной совокупности .

Первая глава пособия посвящена теоретическому аспекту проверки статистических гипотез, содержит подробное описание механизма проверки гипотез и интерпретацию таких базовых понятий, как статистический критерий, уровень значимости, критическая область, надежность критерия и т. д. Материал снабжен иллюстративными примерами. Авторы в первую очередь опирались на источники [1–3]. Также стоит порекомендовать книги [5–7].

Вторая глава содержит описание основных видов статистических критериев и частные примеры критериев. Более того, в материал включено описание алгоритмов вычисления статистик, дополненное актуальными реализациями на языке Python, предложены способы визуализации и структурирования данных. Авторы в большой степени полагались на материал, представленный в [1; 4; 6; 8]. Стоит отметить, что данные источники также

могут быть полезны обучающимися для самостоятельного рассмотрения.

Настоящее издание в первую очередь рассчитано на студентов 4-го курса, обучающихся по направлениям подготовки бакалавриата «Математика», «Прикладная математика и информатика», «Информационные системы и технологи», «Программная инженерия», но может быть рекомендовано для студентов, магистров и аспирантов других технических направлений.

Глава 1

Проверка статистических гипотез

1.1 Статистические гипотезы и критерии

Одной из базовых задача статистики (наряду с оцениванием неизвестных параметров распределений) является проверка статистических гипотез. Под *статистической гипотезой* (гипотезой) понимается любое предположение о виде и свойствах распределения случайных величин.

Пусть

- $X = (X_1, X_2, \dots)$ – некоторая выборка,
- $\mathcal{X}$ – выборочное пространство ($X \in \mathcal{X}$),
- $\mathcal{F} = \{F\}$ – совокупность априори допустимых распределений элементов выборки X,
- F_X – неизвестное истинное распределение данных рассматриваемой выборки.

По сути, элементы выборки представляют из себя независимые копии некоторой случайной величины (с. в.) с функцией распределения (ф. р.) $F_X \in \mathcal{F}$.

В общем случае задача проверки гипотез формулируется следующим образом:

1) выделяется некоторое подмножество ф. р. $\mathcal{F}_0 \subseteq \mathcal{F}$ допустимых распределений;

2) по данным выборки X проверяется истинность утверждения

$$H_0 : F_X \in \mathcal{F}_0.$$

В этом случае H_0 есть *основная (нулевая) гипотеза*. Далее определим альтернативные распределения $\mathcal{F}_1 = \mathcal{F} \backslash \mathcal{F}_0$. Утверждение

$$H_1 : F_X \in \mathcal{F}_1$$

называется *альтернативной (конкурирующей) гипотезой*. Таким образом, задача состоит в проверке гипотезы H_0 против альтернативной гипотезы H_1. (Иногда утверждается, что H_0 проверяется внутри общей гипотезы $H : F_X \in \mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$.)

Если подмножество $\mathcal{F}_0\,(\mathcal{F}_1)$ состоит из одного элемента, то гипотеза $H_0\,(H_1)$ называется *простой*, в противном случае – сложной.

Правило, согласно которому на основе наблюдений за X однозначно решается принять основную гипотезу H_0 как истинную или как ложную (т. е. конкурирующая гипотеза H_1 принимается как истинная) называется *статистическим критерием* (критерием).

Заметим, что простые гипотезы точно определяют распределение F_X, а сложные ограничивают лишь некоторое подмножество ф. р. – класс допустимых распределений, которому, по предположению, принадлежит истинное распределение данных. В более общем случае $\mathcal{F}$ может представлять из себя все возможные распределения на выборочном пространстве $\mathcal{X}$. Тогда альтернативная гипотеза H_1 не конкретизируется, и речь идет о согласии данных выборки X с гипотезой H_0. Соответствующие критерии называются *критериями согласия*, более подробно о них речь пойдет в разделе 2.1.

Если класс допустимых распределений задан некоторым параметрическим семейством, т. е.

$$\mathcal{F} = \{F(x,\, \theta),\, \theta \in \Theta\},$$

то соответствующие гипотезы, называемые *параметрическими*, имеют вид

$$H_0 : \theta \in \Theta_0 \subseteq \Theta, \quad H_1 : \theta \in \Theta_1 = \Theta \backslash \Theta_0$$

и могут быть проверены на основе *параметрических критериев*. По сути, параметрические гипотезы представляют из себя предположения об области изменения параметров заданной ф. р. F_X, а непараметрические – предположения о виде F_X.

Выбор критерия должен быть основан на специфике проверяемой гипотезы: в непараметрическом случае в ходе проверки гипотезы должны быть обнаружены любые отклонения H_0, в параметрическом – критерий должен быть направлен лишь на обнаружение конкретных (в рамках рассматриваемой параметрической модели $\mathcal{F}$) отклонений от основной гипотезы.

1.2 Критическая область и область принятия критерия

Статистический критерий – это правило, которое для данной реализации $(X_1,\, X_2, \dots)$ выборки X должно приводить к одному из двух решений: принять гипотезу H_0 или отклонить ее в пользу альтернативной H_1. Таким образом, каждому критерию соответствует некоторое разбиение выборочного пространства $\mathcal{X}$ на два множества $\mathcal{X}_0$ и $\mathcal{X}_1$ таких, что

$$\mathcal{X}_0 \cap \mathcal{X}_1 = \emptyset,\ \mathcal{X}_0 \cup \mathcal{X}_1 = \mathcal{X}.$$

Причем $\mathcal{X}_0$ состоит из элементов, для которых гипотеза H_0 принимается, а $\mathcal{X}_1$ – из тех, для которых H_0 отвергается (т. е. принимается H_1).

Таким образом, $\mathcal{X}_0$ – это *область принятия гипотезы* H_0, а $\mathcal{X}_1$ область ее отклонения, называемая *критической областью*. Любой критерий проверки гипотезы H_0 однозначно задается соответствующей критической областью, т. о.

$$H_0 \text{ отвергается} \Leftrightarrow X \in \mathcal{X}_1.$$

Выбор критической области для конкретной статистической задачи осуществляется в соответствии с *общим принципом принятия решения:* если в эксперименте наблюдается маловероятное при справедливости гипотезы H_0 событие, то считается, что H_0 не согласуется с данными (противоречит им), в этом случае она отвергается. В противном случае считается, что данные не противоречат H_0 (согласуются с основной гипотезой), и H_0 принимается.

В соответствии с данным принципом критическая область выбирается таким образом, чтобы обеспечить малое значение вероятности попадания выборки X в эту область при условии выполнения основной гипотезы. На практике при построении критерия верхняя граница такой вероятности заранее задается малым числом α, называемым *уровнем значимости*. Таким образом, критическая область определяется в соответствии с условием

$$\mathsf{P}\big(X \in \mathcal{X}_1 \big| H_0\big) \le \alpha. \tag{1.1}$$

Очевидно, что условие (1.1) неоднозначно определяет критическую область, чтобы устранить эту неопределенность, вводятся понятия *ошибок критерия*.

Следуя некоторому критерию, можно прийти к одному из трех результатов:

- верное решение;
- *ошибка перового рода* – отвергается гипотеза H_0, в то время как она верна («ложноположительный результат»);
- *ошибка второго рода* – принимается гипотеза H_0, когда она является ложной («ложноотрицательный результат»).

Вероятность попасть в критическую область при заданном уровне значимости, при условии что верна альтернативная гипотеза (т. е. вероятность не допустить ошибку второго рода) называется *мощностью* статистического критерия и обозначается β:

$$\beta = \mathsf{P}\big(X \in \mathcal{X}_1 \big| H_1\big), \tag{1.2}$$

тогда $1 - \beta$ определяет вероятность ошибки второго рода.

Вероятность попасть в область принятия критерия при условии выполнения гипотезы H_0 называется *надежностью* статистического критерия и обозначается γ:

$$\gamma = \mathsf{P}\big(X \in \mathcal{X}_0 \big| H_0\big). \tag{1.3}$$

При построении критерия стоит стремиться к тому, чтобы свести к минимуму вероятности ошибок обоих типов. Однако сложность заключается в том, что сумма $(1 - \beta) + \alpha$ не может быть сколь угодно мала.

Пример 1. *По итогам медицинского исследования на основе значения некоторого количественного показателя делается вывод о подозрении на аппендицит. Если значение показателя превышает заданное значение a, пациенту ставится диагноз, и будет назначена операция. В этом случае существует опасность поставить диагноз здоровому пациенту (ложноположительный результат) либо не назначить операцию пациенту с аппендицитом (ложноотрицательный результат). Задача заключается в определении наиболее подходящего значения порога a.*

Были проведены исследования показателей заведомо здоровых пациентов и пациентов с аппендицитом. Результаты предоставлены на рис. 1.1.

Рис. 1.1. Исследование показателей пациентов

Рис. 1.1 иллюстрирует, что увеличение a снижает вероятность допущения ошибки первого рода (уменьшает уровень значимости), однако влечет за собой более частые ошибки второго рода. На практике при выборе уровня значимости α принято исходить их условий конкретной задачи. В частности, в примере 1 ошибка второго рода (отказ от операции при аппендиците) влечет за собой летальный исход для пациента, а ошибка первого рода – неоправданную операцию. В данном случае ложноположительный диагноз предпочтительнее ложноотрицательного, следовательно, стоит выбирать меньшее значение порога a.

1.3 Критерий отношения правдоподобия

Заметим, что малые значения уровня значимости сужают критическую область, в то время как мощность критерия β, близкая к 1 (обеспечивающая малую вероятность ошибки второго рода), сужает область принятия критерия. Данное явление представляет из себя своего рода принцип неопределенности при проверке гипотез.

Вопрос оптимального выбора критической области при заданном уровне значимости помогает решить критерий отношения правдоподобия.

Выборка $X = (X_1, X_2, \ldots, X_n)$ может быть интерпретирована как с. в. с n-мерной плотностью распределения $f_X(y)$, $y = (y_1, \ldots, y_n) \in \mathcal{R}^n$ (в данном разделе будем рассматривать непрерывный случай).

Ввиду независимости наблюдений выполнено

$$f_X(y) = \Pi_{i=1}^n f_{X_i}(y_i),$$

где f_{X_i} – маргинальная плотность i-й компоненты X.

Функцией правдоподобия L выборки объема n называется плотность выборочного вектора $X = (X_1, \ldots, X_n)$ при фиксированном значении переменных $y_1, y_2, \ldots, y_n$:

$$L(y_1, \ldots, y_n) = \Pi_{i=1}^n f_{X_i}(y_i).$$

Определим $f_X^{(0)}$ ($f_X^{(1)}$) – плотность распределения X, если выполнена гипотеза H_0 (H_1). Заметим, что соответствующая функция правдоподобия есть вероятность получения некой выборки при выполнении той или иной гипотезы

$$L_0(y_1, \ldots, y_n) = \Pi_{i=1}^n f_{X_i}^{(0)}, \quad L_1(y_1, \ldots, y_n) = \Pi_{i=1}^n f_{X_i}^{(1)}.$$

Далее определим статистику

$$\Lambda := \Lambda(x_1, \ldots, x_n) = \frac{L_1(x_1, \ldots, x_n)}{L_0(x_1, \ldots, x_n)},$$

и минимальное пороговое значение C для отношения так, что H_0 отвергается при $\Lambda > C$, тогда справедлив следующий результат.

Лемма Неймана – Пирсона. *Среди всех критериев заданного уровня значимости α, проверяющих две простых гипотезы H_0 и H_1, критерий отношения правдоподобия*

$$\Lambda > C$$

является наиболее мощным.

Доказательство. Пусть $\mathcal{X}^*$ – область допустимых значений критерия отношения правдоподобия, а $\mathcal{X}_1^* \in \mathcal{X}^*$ – его критическая область на уровне значимости α^*. Мощность данного критерия обозначим β^*. Для *некоторого другого критерия*, проверяющего те же гипотезы H_0 и H_1, но на уровне

значимости α введем соответствующие характеристики $\mathcal{X}_1 \in \mathcal{X}$ и β. Тогда

$$\begin{aligned} \alpha^* &= \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}_1^* | H_0\big), \\ \alpha &= \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}_1 | H_0\big). \end{aligned}$$

Поскольку выбор критической области однозначно определяет критерий, имеем

$$\mathcal{X}_1^* = \mathcal{X}^* \setminus (\mathcal{X}^* \cap \mathcal{X}), \quad \mathcal{X}_1 = \mathcal{X} \setminus (\mathcal{X}^* \cap \mathcal{X}).$$

Следовательно,

$$\begin{aligned} 1-\beta^* &= \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}^* \cup \mathcal{X} \setminus \mathcal{X}_1^* | H_1\big) = \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}_1 | H_1\big) \\ &+ \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}^* \cap \mathcal{X} | H_1\big); \\ 1-\beta &= \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}_1^* | H_1\big) + \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}^* \cap \mathcal{X} | H_1\big). \end{aligned}$$

Заметим,

$$\mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}_1 | H_j\big) = \int\cdots\int_{\mathcal{X}_1} f_{X_1}^{(j)}(y_1)\cdots f_{X_n}^{(j)}(y_n) dy_1 \cdots dy_n, \quad j = 0,\, 1.$$

Обозначим:

$$f_X^{(j)}(y)dy := f_{X_1}^{(j)}(y_1)\cdots f_{X_n}^{(j)}(y_n) dy_1 \cdots dy_n, \quad j = 0,\, 1.$$

Тогда по условию теоремы при любом фиксированном $y = (y_1, \ldots, y_n)$ гипотеза H_0 отвергается, если

$$\frac{f_X^{(1)}(y)dy}{f_X^{(0)}(y)dy} > C. \tag{1.4}$$

Имеем

$$\begin{aligned} \alpha^* &= \int_{\mathcal{X}_1^*} f_X^{(0)}(y)dy, \\ \alpha &= \int_{\mathcal{X}_1} f_X^{(0)}(y)dy, \\ 1-\beta^* &= \int_{\mathcal{X}_1} f_X^{(1)}(y)dy + \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}^* \cap \mathcal{X} | H_1\big), \\ 1-\beta &= \int_{\mathcal{X}_1^*} f_X^{(1)}(y)dy + \mathsf{P}\big((X_1,\ldots,X_n) \in \mathcal{X}^* \cap \mathcal{X} | H_1\big). \end{aligned}$$

В области $\mathcal{X}_1$ по критерию отношения правдоподобия гипотеза H_0 не отвергается, следовательно, условие (1.4) нарушено и

$$1-\beta^* \quad \leq \quad \alpha C + \mathsf{P}\big((X_1,\ldots,X_n)\in \mathcal{X}^*\cap\mathcal{X}|H_1\big).$$

В области $\mathcal{X}_1^*$ по критерию отношения правдоподобия гипотеза H_0 отвергается, следовательно, условие (1.4) выполнено и

$$1-\beta \quad > \quad \alpha^* C + \mathsf{P}\big((X_1,\ldots,X_n)\in \mathcal{X}^*\cap\mathcal{X}|H_1\big).$$

Таким образом, при равных уровнях значимости $\alpha = \alpha^*$ вероятности ошибок второго рода соотносятся как:

$$1-\beta^* \leq \alpha C + \mathsf{P}\big((X_1,\ldots,X_n)\in \mathcal{X}^*\cap\mathcal{X}|H_1\big) < 1-\beta.$$

Отсюда следует соотношение мощностей

$$\beta^* > \beta,$$

и утверждение леммы является доказанным.

При заданном уровне значимости α среди всех критериев, проверяющих основную гипотезу H_0 против конкурирующей гипотезы H_1, наилучшим является критерий с наибольшей мощностью. Однако на практике найти такой критерий не всегда удается и приходится ограничиваться более умеренными требованиями. В частности, свойством *несмещенности* , которое предполагает, что наряду с (1.1) выполнено

$$\mathsf{P}\big((X_1,\ldots,X_n)\in \mathcal{X}_1|H_1\big) > \alpha.$$

Для выборок большого объема n критерий должен обладать свойством *состоятельности*

$$\mathsf{P}\big((X_1,\ldots,X_n)\in \mathcal{X}_1|H_1\big) \to 1, \quad n\to\infty. \tag{1.5}$$

Свойство (1.5) означает, что при большом числе наблюдений, в случае выполнения альтернативной гипотезы, верное решение будет приниматься с вероятностью, близкой к единице.

Заметим, что часто при решении прикладных задач выбор (задание) значения уровня значимости связан с практической стороной вопроса. В частности, минимизация вероятностей ошибки первого/второго рода может быть связана с ростом затрат или большим потерям. Как правило, выбирают одно из стандартных значений α: $0,05;\ 0,01;\ 0,05$.

1.4 Статистика критерия и p-value

На практике при исследовании статистических гипотез попадание значений рассматриваемой выборки в заданную критическую область проверяется с помощью вспомогательной одномерной функции от элементов выборки $S = S(X) = S(X_1, \ldots, X_n)$, называемой *статистикой критерия*. Статистика S "измеряет" отклонение эмпирических данных от соответствующих (гипотезе H_0) гипотетических значений, распределение статистики при выполнении гипотезы H_0 должно быть известно. Если

$$W = \{S(X_1, \ldots, X_n);\ (X_1, \ldots, X_n) \in \mathcal{X}\}$$

есть пространство значений статистики S, то критическая область $W_1 \subset W$ рассматриваемого критерия при заданном уровне значимости α задается как

$$\mathsf{P}\big(S(X_1, \ldots, X_n) \in W_1 \big| H_0\big) \leq \alpha. \tag{1.6}$$

Таким образом, в терминах выбранной статистики $S(X)$ *правило принятия решения* имеет вид:

$$H_0 \text{ отвергается} \iff S(X) \in W_1, \tag{1.7}$$

где при заданном уровне значимости α критическая область W_1 удовлетворяет (1.6).

Отметим, что решающим моментом для расчета критерия (для выполнения (1.6)) является проблема отыскания статистики S, соответствующей гипотезе H_0.

Поскольку статистика критерия есть одномерная случайная величина, все ее возможные значения принадлежат некоторому интервалу. Поэтому критическая область и область принятия гипотезы $W_0 = W \backslash W_1$ также являются интервалами и, следовательно, существуют точки, которые их разделяют.

Границы, отделяющие критическую область от области принятия гипотезы, называют *критическими точками*.

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области. *Правосторонней* называют критическую область вида

$$W_1 = [k_\alpha, \infty),$$

где $k_\alpha > 0$ – критическая точка. (Очевидно, что *левосторонняя* критическая область определяется как $W_1 = (-\infty,\ k_\alpha]$.) Если

$$W_1 = (-\infty,\ -k_\alpha^{(1)}] \cup [k_\alpha^{(2)},\ \infty),$$

то рассматривается *двусторонняя* критическая область.

В частности, если в случае двусторонней критической области критические точки симметричны относительно нуля $k_\alpha := k_\alpha^{(1)} = k_\alpha^{(2)}$, то при $|S| < k_\alpha$ гипотеза H_0 принимается.

Альтернативным способом для принятия решения является вычисление так называемого *достигаемого уровня значимости* p-value. Обозначим s наблюдаемое значение статистики $S(X)$, вычисленное по имеющейся реализации выборки X.

Уровень p-value определяется как вероятность того, что статистика S принимает значение, столь же экстремальное или более экстремальное, чем наблюдаемое значение s, при условии, что гипотеза H_0 верна, т. е. для левосторонней критической области

$$\text{p-value} = \mathsf{P}\big(S \le s \big| H_0\big);$$

для правосторонней критической области

$$\text{p-value} = \mathsf{P}\big(S \ge s \big| H_0\big);$$

для двустронней критической области

$$\text{p-value} = \min\{2P\big(S \le s \big| H_0\big), 2P\big(S \ge s \big| H_0\big)\}.$$

Таким образом, p-value – это наименьший уровень значимости, который привел бы к отклонению гипотезы H_0. Чем меньше значение p-value, тем больше оснований для отклонения гипотезы H_0:

- при p-value $< 0,10$ *предположительно* следует отклонить H_0;
- при p-value $< 0,05$ необходимо *скорее всего* отклонить H_0;
- при p-value $< 0,01$ имеются *веские основания* отклонить H_0.

Другими словами, гипотеза принимается, когда не были получены убедительные доказательства для ее отклонения. Тогда *правило принятия решения* на основе p-value эквивалентно правилу (1.7) и имеет вид:

$$H_0 \text{ отвергается} \iff \text{p-value} \le \alpha. \tag{1.8}$$

В общем случае проверка гипотез состоит из набора стандартных шагов, представленных ниже (алгоритм проверки гипотез):

1) сформулировать основную и альтернативную гипотезы $\{H_0, H_1\}$.

2) вычислить тестовую статистику критерия $S(X)$.

3) для заданного уровня значимости α построить критическую область W_1 либо вычислить значение p-value.

4) принять или отклонить гипотезу H_0 по отношению к альтернативной гипотезе H_1.

Пример 2. Артериальное давление

Пусть систолическое артериальное давление у мужчин в возрасте 35-44 лет является нормально распределенной случайной величиной с математическим ожиданием 127 *и дисперсией* 49.

Имеется выборка объема $n = 101$ *с данными измерений верхнего давления у мужчин, болеющих диабетом:*

$$X_1, \ldots, X_n \sim \mathbb{N}(\mu, 49),$$

выборочное среднее значение для которой равно

$$\overline{x}_n := \frac{1}{n} \sum_{i=1}^{n} x_i = 130.$$

Является ли данный факт достаточным основанием полагать, что в среднем у диабетиков систолическое давление выше, чем у здоровых людей? Другими словами, насколько вероятно, что любая выборка из 101 *диабетика дает среднее значение, превышающее* 127*?*

Для ответа на вопрос рассмотрим шаги алгоритма проверки гипотез. Основная гипотеза $H_0 : \mu = 127$, альтернативная $H_1 : \mu > 127$. В качестве статистики S рассмотрим случайную величину

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, 49/101).$$

```
In [6]: # Подключение библиотек
        from scipy import stats as st
        import math

In [7]: # Нормализация с.в.
        (130-127)/math.sqrt(49/101)

Out[7]: 4.307089551908954

In [8]: # p-value
        1-st.norm.cdf(_)

Out[8]: 8.270832654755367e-06
```

Рис. 1.2. Python: вычисления для примера 2

Критическая область будет правосторонней, т. к. H_0 отклоняется при значениях, превышающих 127. Наблюдаемое значение статистики имеет вид

$$s = \overline{x}_n,$$

равно 130, тогда:

$$\begin{aligned} \text{p-value} &= \mathsf{P}(\overline{X}_n \geq 130 | \mu = 127) = \mathsf{P}\Big(\frac{\overline{X}_n - 127}{\sqrt{49/101}} \geq \frac{130 - 127}{\sqrt{49/101}}\Big) = \\ &= 1 - \mathsf{P}(Z < 4,31) < 0,01, \end{aligned}$$

где $Z \sim N(0,1)$. Вычисления приводятся на рис. 1.2. Таким образом, есть веские основания считать, что давление людей, болеющих диабетом, отличается от давления здоровых людей.

Пример 3. Смещенная монета

В результате эксперимента с подбрасыванием монеты в выборке из $n = 100$ *элементов оказалось* 60 *«орлов». Достаточно ли оснований считать, что монета имеет смещение в сторону грани с «орлом»?*

Для ответа на вопрос рассмотрим стандартные шаги алгоритма проверки гипотез. Пусть p – неизвестная вероятность выпадения «орла», тогда общее число «орлов» в 100 экспериментах имеет биномиальное распределение с параметрами $X \sim Bin(100, p)$.

```
In [1]: # Подключение библиотек
        from scipy import stats as st

In [2]: # p-value
        1-st.binom.cdf(59,100,1/2)

Out[2]: 0.02844396682049044
```

Рис. 1.3. Python: вычисления для примера 3

Необходимо проверить альтернативную гипотезу $H_1 : p > 1/2$ при нулевой гипотезе $H_0 : p = 1/2$. В качестве статистики можно рассматривать саму случайную величину X. Наблюдаемое значение статистики равно 60, тогда p-value для правосторонней критической области можно вычислить следующим образом:

$$\begin{aligned} \text{p-value} &= \mathsf{P}(X \geq 60 | p = 1/2) = \sum_{k=60}^{100} \binom{100}{k} \left(\frac{1}{2}\right)^{100} = \\ &= 1 - \mathsf{P}(X \leq 59) < 0,05, \end{aligned}$$

см. вычисления на рис. 1.3.

Глава 2

Примеры статистических критериев

2.1 Критерии согласия

Критерии согласия помогают решить классическую задачу о виде распределения наблюдаемых значений. В простейшем случае задача формулируется следующим образом: дана выборка $(X_1, X_2, \ldots, X_n)$ из некоторого распределения с.в. ξ с *неизвестной* функцией распределения F_ξ. Выдвигается простая гипотеза

$$H_0 : F_\xi(x) = F(x),$$

где функция $F(x)$ задана. Отметим, что альтернативные распределения выборки никак не конкретизируются, изначально может быть известен только тип распределения F_ξ (абсолютно непрерывный/ дискретный), обычно определяемый исходя из условий предлагаемой задачи. Таким образом, в данном случае альтернативная гипотеза H_1 задается как $\overline{H}_0$, т. е. «не H_0», и при проверке речь идет о согласии выборочных данных с основной гипотезой.

2.1.1 Критерий Колмогорова

Наиболее известным примером критерия согласия является критерий Колмогорова, который применим только в случаях, если неизвестная функция F_ξ *непрерывна*. В качестве неизвестной функции берется эмпирическая функция $\overline{F}_n$, построенная по выборке X. Статистика критерия Колмогорова определяется как

$$D_n(X) = \sup_{-\infty<x<\infty} \left|\overline{F}_n(x) - F(x)\right|, \tag{2.1}$$

т. е. представляется собой максимальное отклонение эмпирической функции распределения $\overline{F}_n$ от известной теоретической F (гипотетической, определяемой гипотезой H_0).

```
In [2]: import seaborn as sns; sns.set()
        import pandas as pd
        data=[58.7, 57.3, 52.7, 52.5, 50.0, 51.8, 48.8, 49.4, 56.3, 55.4, 49.8, 48.4, 50.3, 47.8, 49.5, 50.3, 57.5, 61.3, 49.6,
              53.0, 50.2, 50.7, 55.7, 55.0, 54.8, 56.2, 54.5, 55.8, 56.9, 54.9, 56.5, 57.3, 57.6, 57.8, 58.3, 56.6, 58.6, 59.1,
              58.5, 58.7, 55.9, 58.5, 56.0, 56.1, 58.2, 59.0, 58.5, 60.3, 59.0, 60.1, 60.7, 61.0, 60.1, 60.9, 62.3, 62.6, 61.9,
              62.8, 57.3, 59.4, 56.9, 56.8, 55.3, 51.4, 53.8, 55.0, 53.8, 51.5, 51.2, 51.1, 52.7, 49.8, 52.3, 51.8, 51.3, 53.2,
              51.8, 52.1, 52.4, 51.3, 52.0, 52.3, 52.6, 53.5, 53.7, 53.6]

        data_array=np.array(data)
        fig=plt.figure(figsize=(15,10))
        plt.xlabel('X',fontsize=18)
        plt.xticks(fontsize=18)
        plt.yticks(fontsize=18)
        plt.hist(data_array, bins=10, density=True)
        ser = pd.Series(data_array)
        ser.plot(kind='kde')   # kde = Kernel Density Estimation plot
        plt.ylabel('Оценка плотности распределения', fontsize=18)
        plt.show()
```

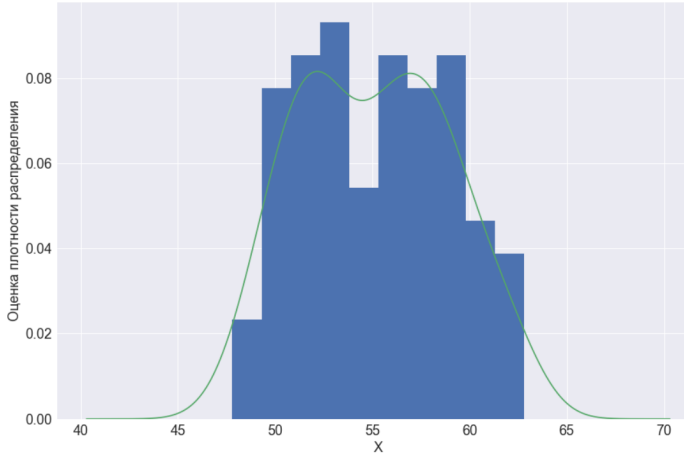

Рис. 2.1. Python: оценка плотности распределения

Отметим, что при условии справедливости гипотезы H_0 функция $\overline{F}_n(x)$ для любого x является оптимальной несмещенной оценкой F. Более того, эта оценка состоятельна, т. е. сходится с F при $n \to \infty$.Таким образом, если H_0 истинна, то при больших n значение статистики D_n не должно существенно отклоняться от 0. Следовательно, большие значения D_n интерпретируются как свидетельство против проверяемой гипотезы H_0,

тогда для данного критерия целесообразно задавать правостороннюю критическую область $W_1 = [k_\alpha, \infty)$. Критическая точка k_α рассчитывается на основании теоремы Колмогорова (см. [1]) следующим образом: если n достаточно велико (хотя бы $n \geq 20$), то положив $k_\alpha = \lambda_\alpha/\sqrt{n}$, где $K(\lambda_\alpha) = 1 - \alpha$, а

$$K(z) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2 z^2}$$

есть функция Колмогорова. Тогда

$$\mathsf{P}\big(D_n \in W_1 \big| H_0\big) = \mathsf{P}\big(\sqrt{n} D_n \geq \lambda_\alpha \big| H_0\big) \approx 1 - K(\lambda_\alpha) = \alpha.$$

Таким образом, критерий согласия Колмогорова формулируется следующим образом: если $n \geq 20$, при выбранном уровне значимости α число $K(\lambda_\alpha)$ определено соотношением $K(K(\lambda_\alpha)) = 1 - \alpha$, то

$$H_0 \text{ отвергается } \iff \sqrt{n} D_n \geq \lambda_\alpha.$$

Заметим, что распределение статистики D_n не зависит от вида функции $F_0(y)$, более того данный критерий является *асимптотическим*, т.е. работает только для больших выборок.

```
In [3]: # параметры распределения
        mu=data_array.mean()
        sigma=data_array.std(ddof=1)
        mu, sigma

Out[3]: (55.0918604651163, 3.8314554174093005)

In [4]: # нормализация
        normed_data=(data-mu)/sigma

In [5]: # Тест Колмогорова
        from scipy.stats import kstest

        # масштабируем данные
        normed_data=(data-mu)/sigma
        kstest(normed_data,'norm')

Out[5]: KstestResult(statistic=0.09423978586104687, pvalue=0.4080417993454337)
```

Рис. 2.2. Python: реализация теста Колмогорова

Для практических вычислений статистики D_n удобно использовать следующую эквивалентную (2.1) формулу.

$$D_n = \max_{1 \leq i \leq n} \Big[\Big|F_n(x_i) - \frac{2i-1}{2n}\Big| + \frac{1}{2n}\Big],$$

где x_i – элемент вариационного ряда для исходной выборки.

Пример 1. *Для выборки объема $n = 86$ проверим данные на соответствие нормальному распределению, реализовав тест Колмогорова.*

Для заданного набора данных построим график оценки функции плотности, см. рис. 2.1.

Применим тест Колмогорова, предварительно выполнив нормализацию данных (используя соответствующие оценки для матожидания μ и среднеквадратического отклонения σ) для сравнения со стандартным нормальным распределением $\mathbb{N}(0, 1)$. Необходимые вычисления см. на рис. 2.2. Тестовая статистика равна $0,094$, а значение p-value = $0,408$. Таким образом, гипотеза H_0 о соответствии нормальному распределению принимается на уровне $\alpha = 0,01$.

```
In [8]: # Построение qqplot
        %matplotlib notebook
        import statsmodels.api as sm
        import matplotlib.pyplot as plt

        qqfig = sm.qqplot(normed_data, line='45')
        plt.xticks(fontsize=18)
        plt.yticks(fontsize=18)
        plt.ylabel('Выборочные квантили',fontsize=18)
        plt.xlabel('Квантили $N(0,1)$',fontsize=18)
        plt.show()
```

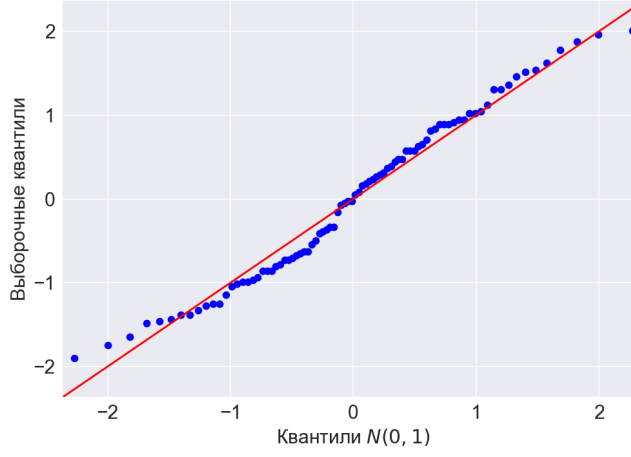

Рис. 2.3. Python: QQplot

Для визуального сравнения двух распределений часто используют так называемый график **квантиль-квантиль** QQplot. Точка (x, y) на графике QQplot соответствует одному из квантилей второго распределения (y-квантиль), построенному против того же квантиля первого распределения (x-квантиль). Две выборки имеют схожие распределения если все точки квантилей лежат на прямой или вблизи нее под углом 45 градусов к оси x. График QQplot полезен для визуального анализа значений данных, попадающих на хвосты распределения. Для построения графика средствами языка python можно использовать библиотеку statsmodels.

На рисунке 2.3 данные рассматриваемой выборки представлены параметрической кривой, определеяемой квантилями стандартного нормального распределения. Выводы теста Колмогорова хорошо согласуются с графиком.

Дополнительно для сравнения эмпирической и теоретической функций распределения иногда используют график типа **вероятность-вероятность** PPplot.

2.1.2 Критерий Пирсона χ^2

Одним из наиболее универсальных методов проверки статистических гипотез является так называемый *метод* χ^2. В общем случае метод работает с дискретными данными, однако любые данные можно свести к дискретным методом группировки, т. е. перейти от исходной (непрерывной) выборки к частотам попадания ее элементов в соответствующие подмножества группировки. Таким образом, рассматриваемый метод применим к данным любой природы, в том числе многомерным (в отличии от рассмотренного ранее критерия Колмогорова, подходящего только для анализа выборки из непрерывного одномерного распределения).

Рассмотрим *дискретную* с.в. ξ, принимающую значения $1,\ 2, \ldots,\ N$ с соответствующими вероятностями $p_1,\ p_2, \ldots,\ p_N$. Произведено n независимых наблюдений за ξ, т. е. получена выборка $(X_1, \ldots,\ X_n)$. Определим частоты

$$\nu_j = \sum_{i=1}^{n} \mathbb{I}(X_i = j), \quad j = 1, \ldots,\ N.$$

По сути, в методе χ^2 проверяется простая гипотеза о соответствии наблюдаемых частот $\nu = (\nu_1, \ldots,\ \nu_N)$ заданному вероятностному вектору $p = (p_1, \ldots, p_N)$. Если учесть, что по частотам $\nu = (\nu_1, \ldots,\ \nu_N)$ строится

эмпирическая (дискретная) функция распределения $\overline{F}_n(x)$ – оценка неизвестной ф.р. F_ξ, а по заданному вектору $p = (p_1, \ldots, p_N)$ – известная ф.р. $F(x)$, то основная гипотеза примет привычный вид

$$H_0 : F_\xi(x) = F(x).$$

В 1900 г. К. Пирсоном в качестве меры отклонения эмпирических данных (относительных частот ν_j/n) от гипотетических (заданных гипотезой H_0) значений p_j была предложена мера хи-квадрат:

$$\hat{X}_n^2 = \sum_{j=1}^{N} \frac{(\nu_j - n \cdot p_j)^2}{n \cdot p_j}. \tag{2.2}$$

На данной тестовой статисте строится критерий χ^2 Пирсона. Идея критерия базируется на следующих выводах: при условии справедливости гипотезы H_0 относительная частота ν_j/n есть состоятельная оценка вероятности p_j. В этом случае можно предположить, что при большом объеме выборки n абсолютные разности $|\nu_j/n - p_j|$ должны быть малы, следовательно, и значение статистики $\hat{X}_n^2$ не должно быть слишком большим. Тогда естественно задать правостороннюю критическую область, т. е.

$$\mathsf{P}\big(\hat{X}_n^2 > k_\alpha \big| H_0\big) = \alpha.$$

Далее возникает вопрос выбора значения критической точки k_α. Можно показать (см. [1]), что с ростом n распределение статистики $\hat{X}_n^2$ не зависит от основной гипотезы H_0 и сходится к распределению хи-квадрат с $(N-1)$ степенями свободы: $\chi^2(N-1)$, т. е.

$$\mathsf{P}\big(\hat{X}_n^2 > x \big| H_0\big) \approx 1 - F_{\chi^2(N-1)}(x), \quad n \to \infty,$$

где $F_{\chi^2(N-1)}(x)$ – ф. р. с. в. $\chi^2(N-1)$. Следовательно, критическая точка k_α находится из соотношения

$$1 - F_{\chi^2(N-1)}(k_\alpha) = \alpha,$$

тогда

$$k_\alpha = F^{-1}_{\chi^2(N-1)}(1-\alpha) =: \chi^2_{1-\alpha,\, N-1}.$$

Здесь $\chi^2_{1-\alpha,\, N-1}$ есть квантиль (аргумент функции распределения) с.в. с распределением хи-квадрат с $N-1$ степенями свободы, на уровне значимости α.

Таким образом, классический критерий χ^2 имеет следующий вид: пусть объем выборки n и наблюдаемые частоты $\nu = (\nu_1, \ldots, \nu_N)$ удовлетворяют условиям $n \geq 50$, $\nu_j \geq 5$, $j = 1, \ldots, N$. При заданном уровне значимости α

$$H_0 \text{ отвергается } \iff \hat{X}_n^2 \geq \chi^2_{1-\alpha,\, N-1}.$$

На практике при анализе выборки из непрерывного распределения критерий Пирсона применяется следующим образом: интервал между минимальным и максимальным элементами выборки делится на N интервалов равной длины, ν_j – количество элементов в j-м интервале; теоретическая p_j – вероятность попадания в j-й интервал (заданное значение), причем для применимости критерия необходимо $np_j \geq 5$.

Стоит отметить, что недостатком представленного метода является то, что при группировке данных может происходить некоторая потеря информации. Кроме того, возникает вопрос оптимального выбора числа интервалов N. Однако бесспорными преимуществами данного критерия является его простота, наглядность и универсальность (нет необходимости учитывать точные значения наблюдений.)

2.2 Критерии однородности

Одной из важных прикладных задач статистики является проверка однородности исходных данных, т. е. гипотезы о том, что закон распределения наблюдаемой случайной величины оставался неизменных в течение всего эксперимента. Формализация данной задачи имеет следующий вид. Пусть имеются две *независимые выборки* $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$, описывающие один и тот же процесс, явление, но полученные, к примеру, в разное время или при разных условиях. Возникает вопрос: являются ли выборки однородными, т. е. полученными из одного и того же распределения, или закон (функция) распределения изменился. Если выборки однородны, то они могут быть объединены в общую, более информативную выборку объема $n + m$.

Будем полагать, что $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ есть выборки из распределений с. в. ξ и η соответственно. Тогда гипотеза об однородности выборок принимает вид

$$H_0 : F_\xi = F_\eta$$

при конкурирующей гипотезе $H_1 : F_\xi \neq F_\eta$. Заметим, что функции распределения F_ξ и F_η, вообще говоря, неизвестны.

В общем случае возможно рассматривать произвольное конечно число независимых выборок.

2.2.1 Критерий Смирнова

Данный критерий, иногда называемый критерием Колмогорова – Смирнова, является одним из первых критериев однородности в задаче о двух выборках. Тестовая статистика имеет вид

$$D_{n,\,m} = \sup_{-\infty<x<\infty} \left|\overline{F}_n(x) - \overline{F}_m(x)\right|,$$

где $\overline{F}_n(x)$ и $\overline{F}_m(x)$ есть эмпирические функции распределения, построенные по выборкам $(X_1, \ldots, X_n)$ и $(Y_1, \ldots, Y_m)$ соответственно.

По аналогии с критерием согласия Колмогорова можно предположить, что для принятия гипотезы об однородности значение статистики $D_{n,\,m}$ не должно существенно отклоняться от 0. Следовательно, разумно выбрать правостороннюю критическую область. На основании теоремы Смирнова можно показать, что

$$\mathsf{P}\big(D_{n,\,m} > k_\alpha \big| H_0\big) = \alpha,$$

где

$$k_\alpha = \sqrt{\frac{1}{n} + \frac{1}{m}}\lambda_\alpha, \quad K(\lambda_\alpha) = 1 - \alpha,$$

а K есть функция Колмогорова.

$$H_0 \text{ отвергается } \iff \chi^2_n \geq \chi^2_{1-\alpha,\,N-1}.$$

Данный критерий является асимптотическим, т. е. при больших $n,\ m$

$$H_0 \text{ отвергается } \iff D_{n,\,m} > k_\alpha.$$

Важно заменить, что критерий Смирнова не зависит от конкретного вида функций F_ξ и F_η, важно лишь выполнения свойства *непрерывности*. В случае дискретного распределения для проверки однородности целесообразно применять *аналог критерия согласия Пирсона* χ^2, заменив в статистике известную функцию $F(x)$ на эмпирическую $\overline{F}_m$.

На практике значение статистики удобно вычислять по следующей формуле:

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

где

$$D^{-}_{n,\,m} = \max_{1\le i\le n}\big|\frac{i}{n}-\overline{F}_m(x_i)\big| = \max_{1\le j\le m}\big|\overline{F}_n(y_j)-\frac{j-1}{m}\big|,$$
$$D^{+}_{n,\,m} = \max_{1\le j\le m}\big|\frac{j}{m}-\overline{F}_n(y_j)\big| = \max_{1\le i\le n}\big|\overline{F}_m(x_i)-\frac{i-1}{n}\big|,$$

а $x_1 \le x_2 \le \cdots \le x_n$ и $y_1 \le y_2 \le \cdots \le y_m$ есть вариационные ряды, построенные по выборкам $(X_1,\ldots,X_n)$ и $(Y_1,\ldots,Y_m)$, соответственно.

2.2.2 Критерий Пирсона χ^2

Для группированных данных проверка гипотезы об однородности может быть выполнена с использованием критерия Пирсона. Далее рассмотрим случай двух серий испытаний по $m_i, i = 1, 2$ наблюдений в каждой. В общем случае с помощью данного критерия можно исследовать произвольное конечное число выборок (подробнее см. в [1]). Пусть число групп (например, значений какого-либо признака) конечно и равно N. Обозначим частоты для каждой выборки и их вероятности соответственно

$$(n_{1,1},\ldots,n_{1,N});\quad \overline{p}_1 = (p_{1,1},\ldots,p_{1,N})$$
$$(n_{2,1},\ldots,n_{2,N});\quad \overline{p}_2 = (p_{2,1},\ldots,p_{2,N}).$$

Тогда основная гипотеза однородности может быть записана в виде

$$H_0 : \overline{p}_1 = \overline{p}_2 = p = (p_1,\ldots,p_N),$$

где p – некоторый неизвестный вектор вероятностей и $\sum_{i=1}^{N} p_i = 1$.

Статистика критерия строится следующим образом:

$$\hat{X}^2_n = m_1 m_2 \textstyle\sum_{j=1}^{N} \dfrac{1}{n_{1,j}+n_{2,j}}\Big(\dfrac{n_{1,j}}{m_1}-\dfrac{n_{2,j}}{m_2}\Big)^2 =$$
$$= \frac{1}{w(1-w)}\Big(\sum_{j=1}^{N} w_j n_{1,j} - m_1 w\Big),$$

где $w = m_1/(m_1+m_2)$, $w_j = n_{1,j}/(n_{1,j}+n_{2,j})$.

Критическая область W_1 строится на основе равенства для заданного уровня значимости α

$$\mathbb{P}(\hat{X}^2_n > \chi^2_{1-\alpha,N-1} | H_0 \text{ верна}) = \alpha.$$

Тогда при больших значениях m_1, m_2 можно применять следующее правило:

$$H_0 \text{ отвергается } \iff \hat{X}_n^2 > \chi^2_{1-\alpha,N-1}.$$

2.2.3 Критерий знаков

Данный критерий не требует больших вычислений. Заметим, что основным его недостатком является неэкономное использование информации, содержавшейся в результатах наблюдений. В связи с этим критерий знаков рекомендовано использовать на стадии предварительного анализа и только для выборок *одинакового объема*.

Рассматриваются выборки $(x_1, \ldots, x_n)$ и $(y_1, \ldots, y_n)$ из наблюдения за с. в. ξ и η соответственно. Основная гипотеза имеет вид

$$H_0 : \mathsf{P}(\xi < \eta) = \mathsf{P}(\xi > \eta).$$

Убираем пары $x_i = y_i$, остается $m \leq n$ пар. Определим $z_i = x_i - y_i$, $i = 1, \ldots, m$. Заметим $z_i \neq 0$. Далее введем в рассмотрение

$$u = \sum_{i=1}^{m} \mathbb{I}(z_i > 0).$$

Можно показать, что с. в. u имеет биномиальное распределение $\mathbb{B}(m,\, 0.5)$, более того

$$\frac{u - \mathbb{E}u}{\sqrt{\mathbb{D}u}} \sim \mathbb{N}(0,\, 1)$$

при больших m. Пусть $p = \mathsf{P}(\xi < \eta)$. Фактически необходимо проверить гипотезу $H_0 :\ p = 0,5$.

Статистика критерия строится следующим образом:

$$b = \frac{1}{2^n} \sum_{i=1}^{u} C_n^i.$$

Критическая область зависит от вида конкурирующей гипотезы.
При $H_1 : p \neq 0,5$

$$H_0 \text{ отвергается } \iff b \notin [\alpha/2,\, 1 - \alpha/2];$$

при $H_1 : p < 0,5$

$$H_0 \text{ отвергается } \iff b < \alpha;$$

при $H_1 : p > 0,5$

$$H_0 \text{ отвергается } \iff b > 1 - \alpha.$$

Также заметим, что данный критерий применим только в случае *связных выборок*, т. е. элементы $x_i,\ y_i$ соответствуют одному и тому же объекту, но измерения сделаны в разные моменты (например, до и после обработки).

2.2.4 Критерий Манна – Уитни

Данный критерий, называемый иногда критерием Вилкоксона (Вилкоксона – Манна – Уитни) является примером *рангового статистического критерия*. Иной раз представленная для обработки информация задается не числовыми значениями, а отношением порядка. Например, больше/меньше, лучше/хуже. Часто данный подход имеет место при анализе естественных систем: психологические исследования, анализ социологических опросов и т. п. В таких случаях наблюдения принято ранжировать, т. е. упрочивать по степени их предпочтения. Номер места, которое занимает наблюдение в упорядоченном ряду, называют *рангом* данного наблюдения. Критерии, применяемые для анализа ранговых наблюдений, и называются ранговыми критериями. Заметим, что теория ранговых критериев в полной мере изложена в [1].

В качестве примера рангового критерия рассмотрим так называемый U-тест. Данный метод был предложен Вилкоксоном (F. Wilcoxon) в 1945 г. для выборок одинакового объема и развит Манном и Уитни (H. B. Mann, D. Whitney) в 1947 г. для выборок разного объема.

Рассмотрим объединенную выборку $(x_1, \ldots, x_n,\ y_1, \ldots, y_m)$ из n наблюдений за с.в. ξ и m наблюдений за с. в. η. Далее построим ее вариационный ряд и вычислим ранги (порядковые номера в вариационном ряду)

$$r(x_i),\ r(y_j),\ i = 1, \ldots, n,\ j = 1, \ldots, m$$

для всех элементов обеих выборок. Затем считаем характеристики

$$R_1 = \sum_{i=1}^{n} r(x_i); \quad U_1 = nm + \frac{n(n+1)}{2} - R_1, \tag{2.3}$$

$$R_2 = \sum_{j=1}^{m} r(y_j); \quad U_2 = nm + \frac{m(m+1)}{2} - R_2. \tag{2.4}$$

Итоговая тестовая статистика для критерия Манна – Уитни имеет вид

$$U = U(n,\, m) = \min(U_1,\, U_2). \tag{2.5}$$

Отметим, что если определить U как случайную величину на основе случайного вектора $(X_1, \ldots, X_n,\, Y_1, \ldots, Y_m)$, где $X_i,\, Y_j$ – независимые реализации некоторых с. в. $\xi,\, \eta$ соответственно, то можно показать, что для любых непрерывных распределений $F_\xi,\, F_\eta$

$$\mathbb{M}U = nm\mathsf{P}\big(\xi < \eta\big) = nm \int_{-\infty}^{\infty} F_\xi(x) f_\eta(x) dx = anm,$$

где a – некоторый параметр. Следовательно, при нулевой гипотезе $F_\xi = F_\eta$ получаем $\mathbb{M}U = nm/2$. Таким образом, с позиции статистики U альтернативы характеризуются величиной параметра a: $a < 1/2$, $a > 1/2$ или $a \neq 1/2$.

На практике для расчета критической точки принято опираться на следующий результат: можно показать, что в случае выполнения гипотезы H_0 и при $n,\, m \to \infty$ статистика $U(n,\, m)$ слабо сходится к нормальной с. в. следующего вида

$$\mathbb{N}\Big(\frac{nm}{2},\, \frac{nm(n+m+1)}{12}\Big).$$

Заметим, что данное предельное распределение с хорошим приближением можно использовать уже при $n,\, m \geq 4, n + m \geq 20$.

Критическая точка для данного критерия определяется как

$$t_\alpha(n,\, m) = \frac{nm}{2} - \sqrt{\frac{nm(n+m+1)}{12}} t_\alpha, \text{ где } \quad \mathsf{P}(\mathbb{N}(0,\, 1) < -t_\alpha) = \alpha.$$

Также стоит отметить, что критическая область имеет различный вид в зависимости от целей, для которых строится критерий. Если необходимо получить критерий, состоятельный против альтернатив $(F_\xi,\, F_\eta)$ с $a < 1/2$, то критическая область задается в виде $\{U(m,\, n) \leq t_\alpha(m,\, n)\}$ и

$$H_0 \text{ отвергается } \iff U(n,\, m) \leq t_\alpha(m,\, n).$$

При $a > 1/2$ рассматриваем перестановку F_ξ и F_η. При $a \neq 1/2$ задается двусторонняя критическая область:

$$H_0 \text{ отвергается } \iff \Big|U(n,\, m) - \frac{nm}{2}\Big| > \sqrt{\frac{nm(n+m+1)}{12}} t_{\alpha/2}.$$

Пример 2. *С помощью U-тест проверим стохастическую упорядоченность двух выборок.*

Будем рассматривать основную гипотезу $H_0 : \mathbb{P}(\xi > \eta) = 0,5$ относительно альтернативы $H_1 : \mathbb{P}(\xi > \eta) > 0,5$. Другими словами, проверим предположение, что с.в. ξ скорее *стохастически больше* с.в. η, чем эквивалентна.

Рассмотрим две выборки объема $n = 100$, для визуализации данных будем использовать график типа stripplot из библиотеки seaborn, см. рисунок 2.4.

```
In [2]: import seaborn as sns; sns.set()
        import matplotlib.pyplot as plt
        import pandas as pd

        sample1 = [75, 80, 80, 73, 81, 81, 83, 81, 78, 73, 78, 70, 74, 72, 71, 53, 67, 53, 58, 58, 69, 73, 67, 59, 66, 71, 64, 67,
                   72, 72, 72, 70, 72, 68, 74, 71, 63, 66, 74, 72, 71, 53, 67, 53, 58, 70, 72, 68, 74, 71, 63, 67, 53, 58, 58, 69,
                   73, 72, 71, 53, 67, 53, 58, 70, 80, 73, 81, 81, 83, 81, 78, 73, 78, 70, 58, 69, 73, 67, 74, 71, 63, 66, 74, 72,
                   71, 53, 67, 53, 58, 70, 72,  53, 67, 53, 58, 70, 80, 73, 81, 85]
        sample2 = [69, 68, 68, 74, 69, 65, 65, 63, 59, 54, 54, 48, 50, 50, 51, 48, 31, 46, 34, 28, 65, 68, 72, 64, 44, 56, 55, 49,
                   54, 57, 43, 40, 39, 41, 59, 51, 46, 49, 48, 31, 46, 34, 28, 65, 68, 69, 65, 65, 43, 40, 39, 41, 56, 55, 59, 54,
                   54, 48, 50, 50, 51, 48, 31, 46, 34, 28, 65, 68, 72, 64, 44, 61, 59, 54, 54, 48, 50, 50, 51, 48, 31, 54, 57, 43,
                   40, 39, 41, 59, 51, 46, 49, 48, 69, 65, 65, 63, 59, 54, 54, 48]

        dsample={"sample1": pd.Series(sample1), "sample2": pd.Series(sample2)}
        df=pd.DataFrame(dsample)
        sns.set(style='whitegrid')
        fig=plt.figure(figsize=(15,10))

        plt.xticks(fontsize=18)
        plt.yticks(fontsize=18)
        plt.ylabel('Значения элементов выборок',fontsize=18)
        sns.stripplot(data=df, linewidth=1, size=7, palette="Set1")
        plt.show()
```

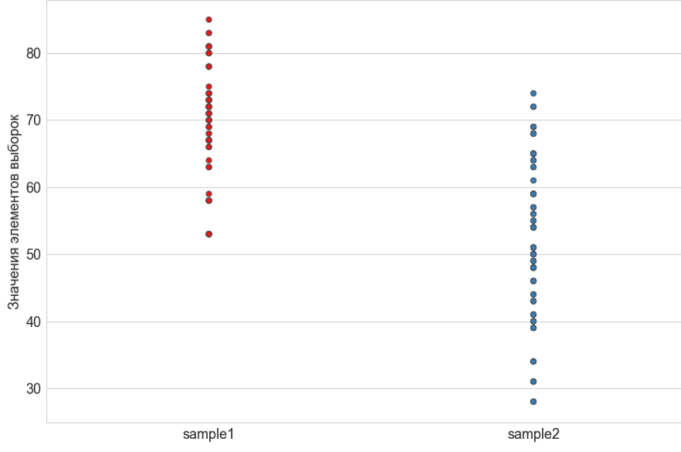

Рис. 2.4. Python: визуализация данных выборок

На рисунке 2.5. показаны несколько вариантов реализации сравнения выборок.

```
In [3]: # U-test

        st.mannwhitneyu(sample1, sample2, alternative='greater')

Out[3]: MannwhitneyuResult(statistic=8807.0, pvalue=6.567413297306493e-21)

In [4]: st.mannwhitneyu(sample1, sample2, alternative='less')

Out[4]: MannwhitneyuResult(statistic=8807.0, pvalue=1.0)

In [5]: u, p_value = st.mannwhitneyu(sample2, sample1, alternative='less')
        print("two-sample U-test p-value", p_value)

        two-sample U-test p-value 6.567413297306493e-21
```

Рис. 2.5. Python: U-тест для двух выборок

Первый вариант с использованием альтернативы greater. Тестовая статистика равна 8807, p-value = $6,56 \cdot 10^{-21}$, что существенно меньше уровня значимости $\alpha = 0,01$, а следовательно, имеются веские основания для отклонения гипотезы H_0 в пользу H_1. Второй вариант с альтернативой less выдает значение p-value = 1. Это значит, что оснований отвергать гипотезу H_0 в пользу альтернативной гипотезы $H_1 : \mathbb{P}(\xi > \eta) < 0,5$ нет. Наконец, в третьем варианте при смене порядка выборок и альтернативе less снова p-value = $6,56 \cdot 10^{-21}$, т. е. принимается альтернативная гипотеза $H_1 : \mathbb{P}(\eta > \xi) < 0,5$.

Таким образом имеются веские основания считать, что с.в. ξ стохастически больше с.в. η. На рисунке 2.6. представлены графики эмпирических функций распределения, построенные с помощью для выборок sample1 и sample2. При этом график для sample2 расположен выше, что снова согласуется с результатами U-теста.

2.2.5 Критерий серий

В ряде задач особый интерес представляют альтернативные гипотезы вида $H_1 : F_\xi(x) > F_\eta(x)$. В этом случае говорят, что с.в. η *стохастически больше*, чем ξ: при каждом x с.в. η с большей вероятностью превосходит x, чем с. в. ξ. Примером простого критерия, хорошо улавливающего подобные

```
In [7]: # Empirical CDF

from statsmodels.distributions.empirical_distribution import ECDF
#fig=plt.figure(figsize=(30,10))
ecdf_1=ECDF(sample1)
ecdf_2=ECDF(sample2)

fig=plt.figure(figsize=(15,10))
plt.step(ecdf_1.x, ecdf_1.y, label='Sample1 ECDF')
plt.step(ecdf_2.x, ecdf_2.y, label='Sample2 ECDF')
plt.legend(fontsize=18)
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.ylabel('Эмпирическая функция распределения',fontsize=18)
plt.xlabel('$x$',fontsize=18)
plt.show()
```

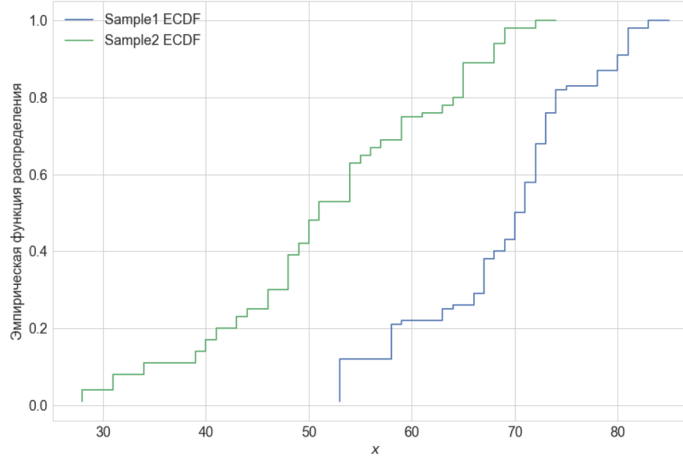

Рис. 2.6. Python: Эмпирические функции распределения двух выборок

изменения, является критерий *Вальда – Вольфовица*, называемый также критерием серий.

Для объединенной выборки $(x_1, \ldots, x_n, y_1, \ldots, y_m)$ строится общий вариационный ряд. Далее элементам первой выборки ставится в соответствие «+», элементам второй выборки «-». Последовательность знаков одного вида называется *серией*. В итоге получаем некоторую последовательность из «+» и «-». Заметим, что общее чисто таких последовательностей есть C_{n+m}^n. Можно предположить, что при выполнении гипотезы $H_0 : F_\xi(x) = F_\eta(x)$ элементы обеих выборок должны вести себя одинаково, т. е. все возможные последовательности из плюсов и минусов равновероятны. При выполнении гипотезы H_1 ожидается, что с повышенной вероятностью будут наблюдаться последовательности, в которых элементы одного вида имеют тенденцию смещаться к какому-нибудь краю

последовательности.

Вычисляются $n_1,\ n_2$ – количество элементов в сериях, N – общее число *серий*. Очевидно $n_1 + n_2 = n + m$ и $N \leq n + m$. Заметим, что статистика N количественно характеризует степень перемешивания в получаемой последовательности из «+» и «-».

Статистика для критерия серий вычисляется по следующей формуле:

$$Z = \frac{N - \frac{2n_1n_2}{n_1+n_2} - 1}{\sqrt{\frac{2n_1n_2(2n_1n_2-(n_1+n_2))}{(n_1+n_2)^2(n_1+n_2-1)}}}$$

и далее сравнивается с квантилем стандартизованного нормального распределения соответствующего порядка. В частности, при $H_1 :\ F_\xi(x) > F_\eta(x)$:

$$k_\alpha = F^{-1}_{\mathbb{N}(0,\,1)}\big(1-\alpha\big);$$

при $H_1 :\ F_\xi(x) \neq F_\eta(x)$:

$$k_\alpha = F^{-1}_{\mathbb{N}(0,\,1)}\big(\frac{1-\alpha}{2}\big).$$

Вывод следующий:

$$H_0 \text{ отвергается} \Longleftrightarrow |Z| \geq k_\alpha.$$

Критерий серий является примером асимптотического критерия, применим для выборок большого объема $n,\ m \to \infty$.

2.2.6 Критерий Краскела – Уоллиса

Данный критерий, известный также как H-критерий, предназначен для проверки равенства медиан *нескольких выборок* и является многомерным обобщением критерия Вилкоксона – Манна – Уитни.

Критерий Краскела – Уоллиса относится к ранговым критериям, поэтому он инвариантен по отношению к любому монотонному преобразованию шкалы измерения. Рассматриваются k независимых выборок:

$$\begin{gathered}(x_1^{(1)}, \ldots, x_{n_1}^{(1)}),\\ \ldots\ldots\ldots\ldots,\\ (x_1^{(k)}, \ldots, x_{n_k}^{(k)})\end{gathered}$$

из неизвестных непрерывных распределений с ф.р. $F_1(x), \ldots, F_k(x)$. Проверяется гипотеза

$$H_0 : F_1(x) = F_2(x) = \cdots = F_{k-1}(x) = F_k(x)$$

при альтернативной гипотезе о том, что все выборки не являются стохастически эквивалентными.

Далее строится объединенная выборка объема $n := n_1 + \cdots + n_k$, элементы которой ранжируются аналогично тому, как ранжировались элементы двух выборок при проверке по критерию Манна – Уитни. Пусть R_j – средний ранг выборки $j = 1, \ldots, k$, тогда

$$R = \frac{1}{k} \sum_{j=1}^{k} R_j$$

есть средний ранг в объединенной выборке. Статистика критерия вычисляется по формуле

$$K = \sum_{j=1}^{k} n_j \big(\frac{R_j}{n_j} - R \big)^2 = \sum_{j=1}^{k} \frac{R_j^2}{n_j} - \frac{n(n+1)^2}{4}.$$

Далее можно показать, что величина

$$\frac{12}{n(n+1)} K$$

при больших n имеет распределение хи-квадрат с $k-1$ степенями свободы.

Таким образом,

$$H_0 \text{ отвергается } \iff \frac{12}{n(n+1)} K \geq \chi^2_{1-\alpha,\, k-1}.$$

2.2.7 Медианный критерий

Данный критерий является непараметрическим и относится к классу ранговых критериев сдвига. То есть проверяет гипотезу о том, что распределения двух выборок имеют одинаковую форму и отличаются только сдвигом на константу.

Далее строится общий вариационный ряд, вычисляется *медиана объединенной выборки*. В каждой из k выборок определяется количество элементов строго меньше медианы n_j^- и строго больше медианы n_j^+. Элементы, равные медиане, исключаются из выборки, объем выборки при этом уменьшается.

Затем вычисляются *ожидаемые количества элементов* больших и меньших медианы $\hat{n}_j$, $j = 1, \ldots, k$. Если выполнена основная гипотеза, то можно ожидать, что около половины элементов каждой выборки будут меньше общей выборочной медианы и около половины – больше. Тогда

$$\hat{n}_j = \frac{n_j}{2}, \tag{2.6}$$

где n_j – объем выборки j, заметим, что в случае, если из j-й выборки были исключены значения, равные медиане общей выборки, вместо n_j в формулу (2.6) включаем уменьшенный объем.

Статистика критерия вычисляется по формуле

$$M = \sum_i \frac{(n_i^-)^2}{\hat{n}_i} + \sum_i \frac{(n_i^+)^2}{\hat{n}_i} - n,$$

где $n = n_1 + \cdots + n_k$ – объем объединенной выборки.

Критическая точка – квантиль распределения хи-квадрат с $k-1$ степенями свободы.

$$H_0 \text{ отвергается } \iff M > \chi^2_{1-\alpha,\, k-1}.$$

2.3 Критерии независимости

Критерии независимости служат для проверки гипотезы о независимости двух (или более) случайных величин. Данные представляются из себя наблюдения за двумерной с. в. $(X,\, Y)$ с неизвестной функцией распределения $F(x,\, y) = \mathsf{P}(X \le x,\, Y \le y)$. Требуется проверить гипотезу

$$H_0 :\, F(x,\, y) = F_X(x) F_Y(y), \tag{2.7}$$

где $F_X(x) = \mathsf{P}(X \le x)$, $F_Y = \mathsf{P}(Y \le y)$. Отметим, что если гипотеза H_0 выполнена, то с.в. X и Y по определению независимы. Следует отметить, что на практике различают разные типы зависимости между с. в.

Самый сильный вид зависимости между с. в. – это *функциональная зависимость*, т. е. $Y = f(X)$, где f – некоторая функция.

Если с. в. X и Y зависят от набора случайных факторов, то наличие общих случайных факторов свидетельствует о *статистической зависимости*, когда изменение одной с. в. влечет изменение распределения другой с.в. Кроме этого, с. в. X и Y могут *коррелировать*. На отсутствие корреляционной зависимости указывает нулевая ковариация и нулевой коэффициент корреляции:

$$Cov(X,Y) = 0, \quad \rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\mathbb{D}X}\sqrt{\mathbb{D}Y}} = 0.$$

Если $|\rho(X,Y)| = 1$, то с.в. X и Y линейно зависимы, что является частным случаем функциональной зависимости, а значение коэффициента корреляции отражает степень линейной зависимости. В общем случае коэффициент корреляции может быть «нечувствителен» к наличию других типов зависимости между с. в. Понятия независимости и некоррелированности с. в. не эквивалентны и справедливо соотношение:

$$\text{Независимость} \Longrightarrow \text{Некоррелированность}.$$

Обратное утверждение неверно, за исключением случая, когда (X,Y) имеет совместное нормальное распределение. Более подробно про типы зависимости между с. в. см. в [3], Глава 12. Корреляционный анализ.

Далее рассмотрим статистические критерии проверки независимости двух выборок как на основе классического определения независимости так и на основе исследования коэффициентов корреляции.

2.3.1 Критерий независимости χ^2

Рассмотрим задачу проверки независимости двух с. в. X, Y (или признаков) на основе методики Пирсона, где основная гипотеза имеет вид (2.7), а альтернативная

$$H_1 : F(x,\, y) \neq F_X(x)F_Y(y).$$

Пусть имеются реализации двух с. в. X и Y

$$x = (x_1, \ldots, x_n), \quad y = (y_1, \ldots, y_n), \tag{2.8}$$

а пары значений (x_i, y_i), $i = 1, \ldots, n$ представляют собой реализации двумерной с. в. (X,Y).

Выполним группировку данных по каждой компоненте отдельно на s и k интервалов соответственно

$$\delta_x^1, \dots, \delta_x^s, \quad \delta_y^1, \dots, \delta_y^k.$$

Обозначим $n_{i,j}$ частоту попадания двумерной с. в. в прямоугольник $\delta_x^i \times \delta_y^j$, т. е. $n_{i,j}$ – это количество пар реализаций, у которых первая компонента принадлежит интервалу δ_x^i, а вторая δ_y^j, при этом

$$\sum_{i=1}^{s}\sum_{j=1}^{k} n_{i,j} = n, \quad N_i = \sum_{j=1}^{k} n_{i,j}, \quad M_j = \sum_{i=1}^{s} n_{i,j}.$$

Построим таблицу сопряженности в следующем виде.

Таблица 2.1 Таблица сопряженности

X	Y				$\sum$
	δ_y^1	δ_y^2	$\dots$	δ_y^k	
δ_x^1	$n_{1,1}$	$n_{1,2}$	$\dots$	$n_{1,k}$	N_1
δ_x^2	$n_{2,1}$	$n_{2,2}$	$\dots$	$n_{2,k}$	N_2
δ_x^s	$n_{s,1}$	$n_{s,2}$	$\dots$	$n_{s,k}$	N_s
$\sum$	M_1	M_2	$\dots$	M_k	n

Тестовая статистика критерия независимости Пирсона вычисляется следующим образом:

$$\hat{X}_n^2 = n\sum_{i=1}^{s}\sum_{j=1}^{k} \frac{1}{N_i M_j}\big(n_{i,j} - \frac{N_i M_j}{n}\big)^2 = n\big(\sum_{i=1}^{s}\sum_{j=1}^{k} \frac{(n_{i,j})^2}{N_i M_j} - 1\big)$$

и имеет распределение хи-квадрат с $(s-1)(k-1)$ степенями свободы.

Критическая область определяется из условия

$$\mathbb{P}(\hat{X}_n^2 > \chi^2_{1-\alpha,\,(s-1)(k-1)} | H_0 \text{ верна }) = \alpha,$$

и асимптотический (при больших n) критерий независимости при заданном уровне значимости α имеет вид:

$$H_0 \text{ отвергается } \iff \hat{X}_n^2 > \chi^2_{1-\alpha,\,(s-1)(k-1)}.$$

В задачах на выявление зависимости между признаками, имеющими альтернативную структуру (т. е. все значения разбиваются на две взаимоисключающие группы), критерий проверки гипотезы независимости H_0 можно построить (см. подробнее [1] Глава 4.4. Гипотеза независимости) основываясь на статистике $Z_n = \sqrt{n}\widehat{\rho}_{XY}$, связанной с выборочным коэффициентом корреляции Пирсона

$$\widehat{\rho}_{XY} = \frac{S_{12}}{S_1 S_2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

где $\overline{x}, \overline{y}$ – выборочные средние, а S_1, S_2 – выборочные среднеквадратические отклонения X и Y соответственно, S_{12} – выборочная ковариация. При этом $s = k = 2$, и статистика имеет вид

$$Z_n = \sqrt{n}\Big(\frac{n_{1,1}}{N_1} - \frac{n_{2,1}}{N_2}\Big)\sqrt{\frac{N_1 N_2}{M_1 M_2}}.$$

В этом случае общий критерий независимости χ^2 проверяет гипотезу $H_0 : \rho(X, Y) = 0$ против двусторонней альтернативы $H_1 : \rho(X, Y) \neq 0$, или, другими словами, проверяет *значимость* коэффициента корреляции. Если H_0 принимается, то коэффициент корреляции *не значим.*

Критическая область является двусторонней

$$W_1 = \{|Z_n| > t_{\alpha/2}\} = \{\hat{X}_n^2 > \chi^2_{1-\alpha,1}\},$$

где $t_{\alpha/2} = \phi^{-1}(1 - \alpha/2)$ –это $(1 - \alpha/2)$ – квантиль стандартного нормального распределения и справедливо $t^2_{\alpha/2} = \chi^2_{1-\alpha,1}$, тогда

$$H_0 \text{ отвергается } \Longleftrightarrow \hat{X}_n^2 > \chi^2_{1-\alpha,\,1}.$$

При $H_1 : \rho(X, Y) > 0$

$$H_0 \text{ отвергается } \Longleftrightarrow Z_n > t_\alpha = \phi^{-1}(1 - \alpha).$$

При $H_1 : \rho(X, Y) < 0$

$$H_0 \text{ отвергается } \Longleftrightarrow Z_n < -t_\alpha = \phi^{-1}(\alpha).$$

Из свойств коэффициента корреляции следует, что $\rho(X, Y)$ является показателем тесноты линейной зависимости между с.в. X и Y при их совместном нормальном распределении. В этой связи выборочный коэффициент корреляции Пирсона $\widehat{\rho}_{XY}$ применяется в задачах корреляционного

анализа, когда данные эксперимента можно считать случайными и выбранными из совокупности, распределенной по многомерному нормальному закону.

Для ответа на вопрос, является ли линейная корреляционная зависимость закономерностью при $\widehat{\rho}_{XY} \neq 0$, проверяется гипотеза *о значимости коэффициента корреляции*. При отклонении основной гипотезы $H_0 : \rho(X, Y) = 0$ в пользу альтернативной $H_1 : \rho(X, Y) \neq 0$ говорят о *значимости* коэффициента корреляции на заданном уровне значимости α.

Тестовая статистика вычисляется на основе выборочного коэффициента корреляции Пирсона:

$$T = \frac{\widehat{\rho}_{XY}\sqrt{n-2}}{1-\widehat{\rho}^2_{XY}}$$

и имеет распределение Стьюдента с $n-2$ степенями свободы. Критическая область является двусторонней и удовлетворяет соотношению

$$\mathbb{P}(|T| > t_{1-\alpha,n-2} | H_0 \text{ верна }) = \alpha,$$

где $t_{1-\alpha,n-2}$ – квантиль распределения Стьюдента.

В случаях, когда нет уверенности, что данные распределены нормально, либо при наличии в выборке так называемых «выбросов» (англ. outlier) для проверки гипотезы о независимости используются ранговые критерии, основанные на *коэффициенте ранговой корреляции*. Рассмотрим два наиболее известных ранговых критерия Спирмена и Кендалла.

2.3.2 Критерий Спирмена

Будем называть *рангом* порядковый номер наблюдения в вариационном ряду. Пусть аналогично (2.8) имеются пары наблюдений $(x_i, y_i),\ i \in [1, n]$ и соответствующие им ранги (R_i, S_i). Коэффициент корреляции Спирмена (статистика Спирмена) является коэффициентом Пирсона для пар рангов и считается по формуле

$$\widehat{\rho}_{XY} = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}},$$

где $\overline{R},\ \overline{S}$ – арифметические средние ранги x и y соответственно. Поскольку

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2},$$

то

$$\sum_{i=1}^{n}(R_i - \overline{R})^2 = \sum_{i=1}^{n} i^2 - n\big(\frac{n+1}{2}\big)^2 = \frac{n(n+1)(2n+1)}{6} - n\big(\frac{n+1}{2}\big)^2.$$

Повторив аналогичные вычисления для рангов S_i, получим статистику Спирмена:

$$\begin{aligned}\widehat{\rho}_{XY} &= \frac{12}{n(n^2-1)}\sum_{i=1}^{n}\big(R_i - \frac{n+1}{2}\big)\big(S_i - \frac{n+1}{2}\big)\\ &= 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - S_i)^2. \end{aligned} \tag{2.9}$$

Иногда для проведения вычислений удобнее упорядочить пары рангов в порядке возрастания первой компоненты R_i, таким образом, получив набор пар

$$(1, T_1), \ldots, (n, T_n). \tag{2.10}$$

Тогда статистика Спирмена считается по формуле

$$\begin{aligned}\widehat{\rho}_{XY} &= \frac{12}{n(n^2-1)}\sum_{i=1}^{n}\big(i - \frac{n+1}{2}\big)\big(T_i - \frac{n+1}{2}\big)\\ &= 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(T_i - i)^2. \end{aligned} \tag{2.11}$$

При больших n справедливо соотношение

$$\mathbb{P}(\sqrt{n}|\widehat{\rho}_{XY}| > t_{\alpha/2} | H_0 \text{ верна }) \approx 2\Phi(-t_{\alpha/2}) = \alpha,$$

тогда критическая область задается в виде

$$W_1 = \{|\widehat{\rho}_{XY}| > t_{\alpha/2}/\sqrt{n}\},$$

где $t_{\alpha/2}$ – квантиль стандартного нормального распределения.

2.3.3 Критерий Кендалла

Данный критерий, предложенный М. Кендаллом, является еще одним примером рангового критерия независимости. Ранги T_i вычисляются аналогично (2.10) при расчете статистики Спирмена. Функция sign возвращает знак и равна 1 и -1 для положительных и отрицательных аргументов, соответственно.

Критерий Кендалла основан на статистике

$$\tau = \frac{2}{n(n-1)} S_\tau, \quad S_\tau = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sign}(T_j - T_i). \tag{2.12}$$

Фактически, упорядочиваем выборку по значению признака X, ранжируем значения показателя Y, это означает, что для каждого элемента y_i ставим столько «+», сколько элементов, стоящих правее y_i в вариационном ряду, будут больше, чем y_i; и столько «-», сколько следующих за ним элементов в вариационном ряду будут меньше y_i. Далее из количества плюсов вычитается количество минусов и получается ранг элемента, тогда S_τ есть сумма рангов.

Отметим, что если гипотеза независимости H_0 верна, то

$$\mathbb{M}\tau = 0, \quad \mathbb{D}\tau = \frac{2(2n+5)}{9n(n-1)}.$$

На практике статистику Кендалла рекомедуется вычислять по формуле

$$\tau = \frac{4}{n(n-1)} - 1, \tag{2.13}$$

где N – число тех индексов $(i,\, j)$, $i > j$, для которых $T_i < T_j$. Можно показать

$$S_\tau = 2N - n(n-1)/2,$$

что влечет эквивалентность формул (2.12) и (2.13).

Критическая точка: k_α – квантиль нормального распределения с параметрами $\mathbb{N}\big(0,\, \sqrt{\frac{2(2n+5)}{9n(n-1)}}\big)$ порядка $1 - \frac{\alpha}{2}$.

$$H_0 \text{ отвергается } \iff \tau > k_\alpha.$$

Список литературы

[1] Ивченко Г. И. Введение в математическую статистику : учебник / Г. И. Ивченко, Ю. И. Медведев. — Москва : ЛКИ, — 2010. — 600 с.

[2] Айвазян С. А., Прикладная статистика. Основы эконометрики : учебник для вузов. В 2 т. / С. А. Айвазян, В. С. Мхитарян. — Москва : Юнити-Дана, — 2001. — 1000 с.

[3] Кремер Н. Ш. Теория вероятностей и математическая статистика. — 2-е изд. / Н. Ш. Кремер. — Москва : Юнити-Дана, — 2012. — 538 с.

[4] Лемешко Б. Ю. Критерии проверки гипотез о случайности и отсутствии тренда. Руководство по применению / Б. Ю. Лемешко, И. В. Веретельникова. — НИЦ ИНФРА-М, — 2021. — 221 с.

[5] Data Science and Machine Learning: Mathematical and Statistical Methods / D. P. Kroese, Z. I. Botev, T. Taimre, R. Vaisman // Chapman and Hall:CRC, Boca Raton. — 2019. — 531 p.

[6] Ефимов А. В. Сборник задач по математике для втузов : учебное пособие. В 4 ч. Ч 4 / А. В. Ефимов, А. С. Поспелов. — Москва : Изд-во физ.-мат. лит., — 2003. — 432с.

[7] Савельев В. Статистика и котики / В. Савельев. —Москва : Времена, — 2020. — 192 с.

[8] Гмурман В. Е. Теория вероятностей и математическая статистика : учебное пособие для вузов / В. Е. Гмурман. — 9-е изд., — Москва: Высшая школа, — 2003. — 479 с.

Приложение

Основные распределения и критерии

1) $\mathbb{N}(a,\, \sigma^2)$: нормальное распределение с параметрами $a,\, \sigma^2$.

$$\begin{aligned} \xi &\sim \mathbb{N}(a,\, \sigma^2), \\ \mathsf{P}(\xi < x) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(y-a)^2}{2\sigma}}\, dy, \\ \mathbb{M}\xi = a, &\qquad \mathbb{D}\xi = \sigma^2. \end{aligned}$$

   $\mathbb{N}(0,\, 1)$ – стандартизованное (стандартное) нормальное распределение.

2) $\chi^2(n)$: распределение хи-квадрат с n степенями свободы.

   Пусть $\xi_1, \ldots, \xi_n$ – независимые стандартизованные нормальные случайные величины, т. е. $\xi_i \sim \mathbb{N}(0,\, 1)$, $i = 1, \ldots, n$, тогда

$$\xi_1^2 + \cdots + \xi_n^2 \sim \chi^2(n).$$

3) $\mathbb{B}(m,\, p)$: биномиальное распределение – число успехов в серии из m независимых испытаний, если вероятность успеха в одном испытании равна p.

$$\mathsf{P}\big(\mathbb{B}(m,\, p) = k\big) = C_m^k p^k (1-p)^{m-k}, \quad k = 0, \ldots,\, m,$$

   биномиальный коэффициент: $C_m^k = \frac{m!}{k!(m-k)!}$.

Таблица 1 Статистические критерии

Название	Статистика, критич. точка	H_0 отверг
Критерии согласия. $H_0:\ F_\xi(x)=F(x)$, $H_1:\ F_\xi(x)\neq F(x)$, $F(x)$ известно		
Колмогорова	$\rho=\sqrt{n}\max_{1\le i\le n}\left[\left\lvert F_n(x_i)-\frac{2i-1}{2n}\right\rvert+\frac{1}{2n}\right]$, где x_i – элемент вариационного ряда. $k_\alpha=K^{-1}(1-\alpha)$, где $K(z)$ – ф-ция Колмогорова	$\rho>k_\alpha$
Пирсона χ^2	$\chi^2_n=\sum_{j=1}^{N}\frac{(\nu_j-n\cdot p_j)^2}{n\cdot p_j}$, где ν_j – кол-во элементов в j-м интервале. $k_\alpha=\chi^2_{1-\alpha,\,N-1}$ – квантиль распр-ия хи-квадрат с $N-1$ степенями свободы	$\chi^2_n>k_\alpha$
Критерии однородности (2 выборки). $H_0:\ F_\xi(x)=F_\eta(x)$, $H_1:\ F_\xi(x)\neq F_\eta(x)$		
Смирнова (Колмогорова – Смирнова)	$\rho=\max(D^-_{n,\,m},\,D^+_{n,\,m})\sqrt{\frac{n\,m}{n+m}}$, $D^-_{n,\,m}=\max_{1\le i\le n}\left\lvert\frac{i}{n}-\overline{F}_m(x_i)\right\rvert$, $D^+_{n,\,m}=\max_{1\le j\le m}\left\lvert\frac{j}{m}-\overline{F}_n(y_j)\right\rvert$, $k_\alpha=K^{-1}(1-\alpha)\sqrt{1/n+1/m}$	$\rho>k_\alpha$
Знаков	$H_0:\mathsf{P}(\xi<\eta)=0,5$ и $H_1:\mathsf{P}(\xi<\eta)\neq 0,5$, $b=\frac{1}{2^n}\sum_{i=1}^{u}C^i_n$, где $u=\sum_{i=1}^{n}\mathbb{I}(x_i>y_i)$	$b<\alpha/2$, $b>1-\alpha/2$
Манна – Уитни	$R_1=\sum_{k=1}^{n}r(x_k)$; $U_1=nm+\frac{n(n+1)}{2}-R_1$, $R_2=\sum_{k=1}^{m}r(y_k)$; $U_2=nm+\frac{m(m+1)}{2}-R_2$, $U=\min(U_1,\,U_2)$, $t_\alpha(n,\,m)$ – квантиль $\mathbb{N}(\frac{mn}{2},\,\frac{nm}{12}(m+n+1))$	$H_1:F_\xi>F_\eta$, $U<t_\alpha(n,\,m)$
Серий	$Z=\frac{N-\frac{2n_1n_2}{n_1+n_2}-1}{\sqrt{\frac{2n_1n_2(2n_1n_2-(n_1+n_2))}{(n_1+n_2)^2(n_1+n_2-1)}}}$, $n_1,\ n_2$– кол-во эл-ов в сериях; N – число серий, $k_\alpha=F^{-1}_{\mathbb{N}(0,\,1)}\left(\frac{1-\alpha}{2}\right)$	$\lvert Z\rvert<k_\alpha$
Краскелла – Уоллиса	$K=\sum_{j=1}^{k}\frac{R_j^2}{n_j}-\frac{n(n+1)^2}{4}$, R – ср. ранг объед. выборки, R_j – ср. ранг выборки j; $k_\alpha=\chi^2_{1-\alpha,\,k-1}$	$\frac{12}{n(n+1)}K\ge k_\alpha$
Медианный	$M=\sum_j\frac{(n_j^-)^2}{\hat{n}_j}+\sum_j\frac{(n_j^+)^2}{\hat{n}_j}-n$, $n_j^-,\ n_j^+$ – число эл-в < и > медианы, $\hat{n}_j$– ожидаемые кол-ва эл-в < (>) медианы	$M>\chi^2_{1-\alpha,\,k-1}$

Учебное издание

Бородина Александра Валентиновна
Некрасова Руслана Сергеевна

СТАТИСТИЧЕСКИЕ КРИТЕРИИ В АНАЛИЗЕ ДАННЫХ

Учебное пособие
для обучающихся по направлениям подготовки бакалавриата
«Математика», «Прикладная математика и информатика»,
«Программная инженерия», «Информационные системы и технологии»

Редактор *А. Б. Соболева*
Компьютерная верстка *А. В. Бородиной и Р. С. Некрасовой*

Подписано в печать 26.12.2022. Формат 60x84 1/16
Бумага офсетная. Усл. печ. л. 2.67. Тираж 50 экз. Изд. № 122

Федеральное государственное бюджетное
образовательное учреждение высшего образования
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Отпечатано в типографии Издательства ПетрГУ
185910, г. Петрозаводск, пр. Ленина, 33