Статистика: ранговые критерии

В этой лекции мы расширим аппарат статистических критериев, добавив ранговые критерии, не основанные на центральной предельной теореме.

10.1 Зачем нужны ранговые критерии

Существует достаточно много приложений, где данных мало. Это может быть связано с самыми разными факторами.

Во-первых, данных может быть мало в силу специфики задачи. К примеру, при тестировании лекарств есть ограниченное число испытаний, часто речь идёт о десятках людей, чего не хватает для уверенного использования асимптотических критериев вроде z-тестов. Данные могут быть дорогими в сборе: представьте себе

- клинические испытания лекарства от редкой болезни, где нужно подобрать очень специфических кандидатов;
- анкету из 100 вопросов и специального человека на улице, который уговаривает прохожих её заполнить;
- дорогостоящий физический эксперимент, который и несколько раз запустить крайне сложно, из-за чего приходиться по максимуму собирать данные на ходу;
- или фокус-группы, где надо платить людям за участие.

Во-вторых, распределение выборки или вычисляемых по ней величин не обязательно гауссовское, чего требует, например, t-тест (критерии Стьюдента) или F-тест (критерии Фишера). Проверить гауссовость можно, но критерии согласия требуют большой выборки, чтобы набрать мощность. Использование этих критериев в условиях невыполненных предположений может вести к некорректным выводам. Единственная хорошая новость: критерии Стьюдента и Фишера применимы в асимптотическом режиме (при большой выборке), но, как мы говорили выше, с большой выборкой тоже проблемы.

Возникает вопрос: можно ли построить статистический критерий такой, чтобы он

1. Не был бы асимптотическим, то есть, математически работал бы при (почти) любом размере выборки;

2. Не зависел бы от конкретного распределения данных (например, не требовал бы гауссовости) и допускались бы только общие и не очень ограничивающие предположения.

Оказывается, такие критерии можно построить, используя достаточно простые идеи и мысленно вернувшись обратно на этап, когда только что закончился курс теории вероятности и только начался курс статистики.

10.2 Проверка гипотез о сдвиге

Вспомним, что несколько лекций назад мы обсуждали гипотезу об однородности. Один из вариантов постановки такой задачи – гипотеза о сдвиге распределения. В такой постановке нам даны две выборки $X_1, .., X_n \sim F$ и $Y_1, .., Y_m \sim G$ из независимых величин и мы рассматриваем гипотезу об однородности

$$H_0 : \forall x \; F_X(x) = F_Y(x)$$

против альтернативы о сдвиге

$$H_A : \exists \mu \neq 0 \; \forall x \; F_X(x) = F_Y(x - \mu).$$

10.2.1 Критерий знаков

Раньше мы с помощью z-теста проверяли гипотезу о матожидании

$$H_0 : \mathbb{E}[X] = \mu_0, \quad H_A : \mathbb{E}[X] \neq \mu_0$$

и она нам позволила сделать z-тест для сравнения среднего в двух выборках. Давайте попробуем вместо этого рассмотреть гипотезу о медиане:

$$H_0 : med(X) = m, \quad H_A : med(X) \neq m.$$

Если медиана отличается, то мы предполагаем, что сдвиг есть. Определяющее свойство медианы – это то, что она делит распределение на две области с вероятностью $1/2$; на основе этого предложим статистику

$$T = \sum_{i=1}^{n} \mathbb{1}(X_i \geq m),$$

которая, оказывается, при верной гипотезе имеет распределение $Bin(n, 1/2)$, у которого можно несложно вычислить квантили и критическую область разместить по краям – в пользу альтернативы говорят либо слишком большие T (много наблюдений справа), либо слишком маленькие (много наблюдений слева). Критерий готов.

Такой критерий называется *критерием знаков*. Его можно применить для проверки однородности выборок, которые называют связанными. Две таких выборки описывают одни и те же субъекты (например, людей), которые сравниваются в разных обстоятельствах. Посмотрим на примере.

Пример 10.1. *Школа Насти организует курсы по подготовке к ЕГЭ. Для того, чтобы понять, насколько курсы эффективны, независимо выбирается* 100 *учеников из разных школ и измеряются результаты пробного ЕГЭ в начале 11-го года (в сентябре) и ближе к концу (в апреле). Как проверить, есть ли положительный эффект от курсов Насти?*

Наши данные представлены результатами ЕГЭ в сентябре $X_1, .., X_{100}$ *и в апреле* $Y_1, .., Y_{100}$. *Определим изменение как* $Z_i = Y_i - X_i$ *(так можно, потому что выборки связаны:* X_i *и* Y_i *– оценки одного и того же ученика). Тогда для новой выборки мы можем применить критерий знаков, задав гипотезу и альтернативу как*

$$H_0 : med(Z) = 0, \quad H_A : med(Z) > 0,$$

а критическую область поместить справа от квантили $b_{1-\alpha}$.

У этого критерия всё хорошо, но у него есть проблема: он очень слабый в смысле мощности. Статистика принимает целые значения, а в критической область попадает всего несколько чисел при малых выборках.

10.2.2 Критерий Манна-Уитни

Другая полезная идея для построения критерия состоит в анализе специально упорядоченных выборок. Таким критерием является, например, критерий Манна-Уитни, который поддерживает любые размеры выборок; кроме того, выборки могут быть несвязанными. Получив две выборки $X_1, .., X_n \sim F_X$ и $Y_1, .., Y_m \sim F_Y$, проверяем гипотезу

$$H_0 : \forall x \;\; F_X(x) = F_Y(x)$$

против альтернативы

$$H_A : \exists \mu \neq 0 \;\; F_X(x) = F_Y(x - \mu)$$

Алгоритм следующий:

1. Скинуть все наблюдения в одну выборку $Z_1, .., Z_{n+m}$, остортировать её по возрастанию.

2. Каждому наблюдению X_i назначить ранг R_i, а наблюдению Y_j – ранг S_j, равный порядковому номеру элемента в смешанном ряду, если несколько элементов равны, назначить каждому ранг, равный среднему арифметическому их рангов.

3. Посчитать сумму рангов R_X, R_Y наблюдений из выборок X, Y и отклонения
$$U_X = nm + \frac{n(n+1)}{2} - R_X,$$
$$U_Y = nm + \frac{m(m+1)}{2} - R_Y;$$
4. Вычислить статистику Манна-Уитни $U = min(U_X, U_Y)$ и принять решение с использованием квантилей.

Квантили статистики Манна-Уитни – комбинаторная задача и они явно выписаны для различных n, m, причём при размере выборок порядка нескольких десятков эта статистика уже распределена почти по нормальному закону и можно использовать z-квантили.

В пользу альтернативы говорят маленькие значения статистики U. Если присмотреться, то числа в левой части $nm+n(n+1)/2$ и $nm+m(m+1)/2$ соответствуют максимально возможным суммам рангов для выборки X и Y. Таким образом, малое отклонение соответствует случаю, когда в упорядоченной выборке наблюдения из одной выборки в целом расположены ближе к концу. Альтернатива про $\mu \neq 0$, поэтому в качестве U берётся минимум из U_X, U_Y, что обрабатывает случай, что какая-то из выборок смещена ближе к концу.

10.3 Проверка гипотез о масштабе

Ранее мы задавались вопросом, как можно для выборок $X_1, .., X_n \sim F_X$ и $Y_1, .., Y_m \sim F_Y$ проверить гипотезу об однородности
$$H_0 : \forall x \;\; F_X(x) = F_Y(x)$$
против альтернативы
$$H_A : \exists\, \alpha > 1 \; \forall x \;\; F_X(x) = F_Y(\alpha x).$$
Ответ на вопрос в случае гауссовских (или очень больших) выборок мы получали путём переформулирования задачи в виде гипотезы о равенстве дисперсий:
$$H_0 : \frac{\sigma_X}{\sigma_Y} = 1, \quad H_A : \frac{\sigma_X}{\sigma_Y} > 1,$$
– и получали критерий Фишера. Оказывается, что есть ранговый критерий, который решает похожую задачу, но без требования гауссовости и очень большой выборки.

10.3.1 Критерий Ансари-Брэдли

Проверяется гипотеза
$$H_0 : \; \forall x \;\; F_X(x) = F_Y(x), \quad H_A : \; \exists \alpha > 1 \; F_X(\alpha x) = F_Y(x).$$

Алгоритм критерия:

1. Вычесть выборочную медиану из каждой выборки (!);

2. Скинуть все наблюдения в одну выборку $Z_1, .., Z_{n+m}$ и отсортировать её по возрастанию.

3. Каждому элементу назначить ранг, равный порядковому номеру в ряду. В случае, если есть повторения, можно использовать средний ранг (как выше), но есть тонкости и лучше использовать немного другую концепцию mid-rank.

4. Посчитать статистику теста: полагая R_i рангами наблюдений X_i, вычислить

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

5. Используя посчитанные квантили или аппроксимацию нормальным распределением принять решение.

Заметьте, что критерий очень похож на критерий Манна-Уитни, так как использует ранги, но адресует вопрос их разброса. Для данного критерия есть предпосчитанные квантили, а для выборок порядка десятков уже работает процедура нормальной аппроксимации [11, Гл. 9.3], позволяющая использовать z-квантили.

10.4 Проверка гипотез о независимости

Затронем ещё два критерия, которые имеют прямой аналог в области анализа корреляций в линейной регрессии. Рассмотрим две связанные выборки $X_1, .., X_n$ и $Y_1, .., Y_n$. Все мы видели выборочный коэффициент корреляции Пирсона

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}.$$

На его основе построен корреляционный анализ и что-то похожее естественно возникает в задаче одномерной линейной регрессии как оценка одного из коэффициентов. Коррелированность (близость коэффициента к 1 или -1) позволяет задуматься о том, что, возможно, X_i и Y_i линейно связаны. Можно проверить гипотезу о корреляции (возможны односторонние альтернативы)

$$H_0 : \ \rho_{XY} = 0, \quad H_A : \ \rho_{XY} \neq 0$$

с помощью z- или t-теста из аппарата линейной регрессии.

Но у подобного подхода есть несколько недостатков.

- Нужно выполнение предположений линейной регрессии...

- Не поддерживает случай, где X_i и Y_i замерены в порядковой шкале без чёткого понимания расстояния между делениями.

Другой путь состоит в том, чтобы использовать статистику

$$T = \frac{\hat{\rho}_{XY}}{\sigma(\hat{\rho}_{XY})} = \hat{\rho}_{XY}\sqrt{\frac{n-2}{1-\hat{\rho}_{XY}^2}},$$

имеющую асимптотическое нормальное распределение или t-распределение с $n-2$ степенями свободы при гауссовских выборках, но здесь приходится опираться либо на нормальность, либо на асимптотику. И мы всё ещё не учли ограничения порядковой шкалы.

Если первые проблемы решаются дополнительными данными, то ограничение порядковой шкалы так не решить, хотя на практике подобные данные, даже не категориальные – очень частый случай.

Пример 10.2. *Действительно, если, скажем, в некотором университете оценки ставятся от* 0 *до* 10*, то, вероятно, расстояние между* 3*(незачёт) и* 4*(удовлетворительно) кажется несоизмеримо большим, чем между* 7*(хорошо) и* 8*(отлично).*

Пример 10.3. *Если мы сравниваем показатели хоккеистов, например, количество забитых шайб, и игровой опыт (в годах), то мы быстро поймём, что между 15 и 17 лет опыта разница незаметна, а между 2 и 4 года существенно (спортсмен активно развивается). С другой стороны, 1 или 2 забитых шайбы за матч – хороший результат и 2 чуть лучше, но добиться 3 уже гораздо сложнее.*

Пример 10.4. *Как понять, как хорошо ваша генеративная модель генерирует текст, речь или картинки? Объективные метрики либо не всегда отражают то, что надо, либо вообще недоступны, как например в задаче text-to-speech. В этих случаях предлагается выкатить семплы инференса своей модели (сгенерированную речь или сгенерированное по тексту изображение) на Толоку и спросить реальных людей: модель сгенерировала качественный/реалистичный/осмысленный семпл? Подобные метрики в статьях так и обозначаются: Mean Opinion Score (MOS).*

Однако такое мероприятие требует бюджета, требует хорошо продуманной анкеты с чёткими критериями для оценщика и времени, чтобы собрать экспертные оценки. Исследователям хочется вместо таких активностей тратить время на конструирование и усовершенствование моделей. Поэтому стали появляться автоматизированные метрики, которые моделируют экспертное мнение и могут выдать оценку очень быстро и дёшево. Некоторые примеры:

- *Картинки: Perceptual Quality Metric (2020), Inception Score (2016), SuperGlue(2020).*
- *Дополненная реальность: ARIQUA (2022).*

- *Звук: DNSMOS(2021) в задаче денойзинга Deep Noise Suppression Challenge, Wav2Vec MOS (2020), NORESQA speech quality (2021).*
- *Тексты: GLUE(2018), RussianSuperGlue(2020), BERTScore (2020).*

Но как всем доказать, что ваша метрика, сколь бы она ни была обоснована интуитивно и физически, действительно отражает экспертное мнение? Распределения непонятные, зависимости между MOS и вашей новой метрикой необязательно линейные, но хотелось бы чтобы рост MOS сопровождался и ростом вашей кастомной метрики.

Всё это говорит о том, что нам нужен некоторый аналог коэффициента корреляции, который бы использовал только порядковую информацию в выборке и помогал бы понять, есть ли зависимость между двумя показателями, возможно, нелинейная.

10.4.1 Коэффициент корреляции Кендалла

Оказывается, можно попробовать сравнивать направление изменения каждой из шкал. Это нас приводит к идее коэффициента Кендалла. С аналогичной критерию знаков идеей мы вводим *параметр согласованности* двух пар случайных величин $(X_1, Y_1), (X_2, Y_2)$

$$\tau_{XY} = 1 - 2\mathbb{P}\left((X_2 - X_1)(Y_2 - Y_1) < 0\right),$$

который так же, как и коэффициент корреляции, принимает значения в $[-1, 1]$, а края достигаются, если зависимость $Y = \phi(X)$ монотонная; попробуйте подставить и применить свойство монотонности. Так, параметр согласованности, в отличие от коэффициента корреляции, способен улавливать более общие зависимости. С другой стороны, в случае независимости $\tau_{XY} = 0$, и всё ещё есть примеры, где коэффициент нулевой, а зависимость есть.

Пример 10.5. *Попробуйте взять X с чётной относительно нуля плотностью и задать $Y = X^2$.*

Мы можем состоятельно и несмещённо оценить параметр согласованности по данной выборке:

$$\hat{\tau}_{XY} = 1 - \frac{4K}{n(n-1)},$$

получим (выборочный) *коэффициент Кендалла*, где K – количество *несогласованных пар*, то есть, таких, где $(X_2 - X_1)(Y_2 - Y_1) < 0$; всего возможных пар $n(n-1)/2$.

Интересно, что этот коэффициент можно также записать используя ранги R_i по X и S_i по Y отдельно:

$$\hat{\tau}_{XY} = \frac{2(Q-K)}{n(n-1)} = \frac{2\sum_{1\le i<j\le n} sign\left((R_i - R_j)(S_i - S_j)\right)}{n(n-1)}.$$

А для случая, когда есть повторения, предусмотрены поправки [11, Гл. 9.3].

Мы можем использовать этот коэффициент для проверки гипотезы о независимости против альтернативы о монотонной зависимости. Для этого введём гипотезу и альтернативу как

$$H_0 : F_{XY}(x, y) = F_X(x)F_Y(y) \quad , \; H_A : \tau_{XY} \neq 0;$$

односторонние альтернативы тоже возможны. Статистика критерия Кендалла задаётся как

$$T = \frac{\hat{\tau}_{XY}}{\sqrt{\frac{4n+10}{9n(n-1)}}},$$

где в знаменателе стоит точно вычисленная дисперсия выборочного коэффициента Кендалла. Эта статистика асимптотически нормальна – z-квантили используются уже при десятках наблюдений; для меньших выборок квантили вычислены точно и даются таблицами.

10.4.2 Коэффициент корреляции Спирмена

Если критерий Кендалла мы изначально вводили с мыслью о знаках, то критерий Спирмена сразу пытается зайти через ранги. На самом деле, этот коэффициент – прямая аналогия обычного коэффициента корреляции в случае рангов.

Пусть нам дана выборка пар $(X_1, Y_1), .., (X_n, Y_n)$, а также заданы ранги R_i, S_i для элементов X_i и Y_i в отсортированных по возрастанию выборках X, Y. *Коэффициент Спирмена* определяется как

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

Если в рангах есть повторения, то вводят дополнительную поправку [11, Гл. 9.3].

Коэффициент Спирмена уже можно использовать как статистику в проверке гипотезы

$$H_0 : F_{XY}(x, y) = F_X(x)F_Y(y) \quad , \; H_A : \rho_{XY} \neq 0,$$

квантили для выборок порядка десятков вычислены и записаны таблично, а статистика

$$T = \sqrt{n-1}\hat{\rho}_{XY}$$

при больших n имеет нормальное распределение.

Коэффициента Спирмена и Кендалла очень похожи по мощности и охватываемым альтернативам. Оказывается даже, что коэффициенты Кендалла и Спирмена сильно коррелированы и более того коэффициент корреляции равен

$$\frac{2(n+1)}{\sqrt{2n(2n+5)}} \to 1, \quad n \to \infty.$$

Литература

[1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit"Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.

[2] Peter C Austin, Muhammad M Mamdani, David N Juurlink, and Janet E Hux. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, 59(9):964–969, July 2006.

[3] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, Sep 2011.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[6] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[7] S.J. Levitt, S.D. Dubner. *Freakonomics*. NY: Harper Trophy, 2006.

[8] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[9] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[10] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[11] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных*. Изд. дом Высшей школы экономики, 2012.

[12] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования*. М.: МЦНМО, 2013.

[13] А.Н. Ширяев. *Основы стохастической финансовой математики*. МЦНМО, 2016.