Оглавление

Глава 1

Введение

1.1 Критерии однородности и независимости: общее введение

1.1.1 Актуальность проблемы

Какие критерии всплывают в памяти при вопросе "Какие критерии однородности вы знаете?". Пользователи AB-тестов, вероятно, назовут t-критерий, любители непараметрической статистики наверняка вспомнят критерий Манна-Уитни. Многие, возможно, вспомнят критерии хи-квадрат. Посещавшие курсы дополнительных глав, вполне вероятно, знакомы с критериями Смирнова и возможно даже Андерсона-Дарлинга в соответствующей модификации.

На аналогичный вопрос о критериях для проверки независимости названы будут критерии Пирсона, Спирмена и Кендалла. И, конечно же, критерий хи-квадрат.

Однако, из всех названных критериев только критерий хи-квадрат применим к многомерным данным (и тот имеет достаточно спорные качества), хотя в настоящее время задача анализа многомерных данных, пожалуй, более актуальна чем одномерных. Более того, большая часть названных критериев имеет существенные ограничения. Все это требует новых подходов к тем же задачам. Такой прорыв случился в начале 2000х годов с развитием вычислительной техники: за период 2005–2020 года появился целый ряд популярных ныне критериев: Energy test и Covariance Distance Секея и Риццо, Maximal Mean Discrepancy и Hilbert-Schmidt Independence Criterion Греттона и соавторов, критерий HHG авторства Хеллера, Хеллера и Горфина, Maximal Information Coefficient Решефа с соавторами, Multiscale Graph Correlation и другие подходы. При этом задача по-

строения критерия независимости по существу свелась к построению критерия однородности.

Цель этого курса – описать более современную базу критериев, исследовать их эмпирически, а также сформировать общие взгляды на указанные подходы.

1.1.2 Базовая терминология

Критерий и уровень значимости

Напомню, основные термины. Пусть есть выборка $X_1, \ldots, X_n$ (вообще говоря, элементы могут быть разнораспределенными и зависимыми, в общей постановке это некритично). При этом $(X_1, \ldots, X_n) \sim F_n$, где F_n – некоторая ф.р.

Определение 1. *Гипотезой* мы будем называть утверждение $H_0 : F_n \in \mathcal{F}_0$, где $\mathcal{F}_0$ – некоторое семейство функций распределения.

Определение 2. *Альтернативой* мы будем называть утверждение $H_1 : F_n \in \mathcal{F}_1$, где $\mathcal{F}_1$ – некоторое семейство функций распределения, не пересекающееся с $\mathcal{F}_0$.

Определение 3. *Критерием* мы будем называть функцию ψ, которая выборке сопоставляет значение из $\{0, 1\}$.

Физически критерий по выборке сообщает принимать ли гипотезу или отвергать.

При этом функция $\psi(X_1, X_2, \ldots, X_n)$ имеет вид $I_{\vec{X} \in D}$, I – индикатор, D – некоторое множество.

Определение 4. Множество D называется *критическим*.

Определение 5. Набор величин

$$\mathbf{E}_{F_n} \psi(\vec{X}) = \mathbf{P}_{F_n}(\vec{X} \in D)$$

при $F_n \in \mathcal{F}_0$ называют *вероятностями ошибки первого рода*, а при $F_n \in \mathcal{F}_1$ – *мощностью критерия*.

Мы бы хотели, чтобы вероятность ошибки первого рода не превосходила α при всех $F_n \in \mathcal{F}_0$. Здесь α – заданное число, называемое *уровнем значимости*.

В основном мы будем рассматривать *асимптотические критерии*, то есть те, для которых

$$\lim_{n \to \infty} \mathbf{P}_{F_n}(\vec{X} \in D_n) \leq \alpha$$

при всех рассматриваемых F_n.

Как правило D имеет вид $\{(x_1, \ldots, x_n) : T(x_1, \ldots, x_n) > c\}$, где T – некоторая измеримая функция выборки (статистика), которая называется *статистикой критерия*.

Эмпирическая мера

Пусть $X_i \sim F$, где F – функция распределения (ф.р.) соответствующая мере P, где $X_i \in \mathbb{R}^k$. Здесь предполагается, что X_i независимы и одинаково распределены (н.о.р). Тогда для оценки ф.р. F естественной оценкой является *эмпирическая ф.р.* (ЭФР)

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \le x},$$

а для меры $\mathbf{P}$ – *эмпирическая мера* (ЭМ)

$$\widehat{P}_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \in A}, \quad A \in \mathcal{B}(\mathbb{R}^k).$$

Эти оценки хорошо себя ведут при фиксированном x или A. ЭФР достаточно удачная оценка и в смысле равномерной метрики в $\mathbb{R}$ (про это мы поговорим позднее), а вот эмпирическая мера, вообще говоря, не слишком близка к исходной. Например, для любой непрерывной меры P

$$\widehat{P}_n(A_n) = 1, \ P(A_n) = 0,$$

где $A_n = \{x_1, \ldots, x_n\}$.

1.2 Задача однородности

1.2.1 Общий подход

Задача проверки гипотезы однородности для двух выборок формулируется так. Имеется две выборки из некоторых распределений $Y_i \sim F$, $i \le n_1$, $Z_i \sim G$, $i \le n_2$, $n_1 + n_2 = n$. При этом все элементы выборок независимы, а выборки независимы между собой. Соответственно, выборка в этом случае есть $(Y_1, \ldots, Y_{n_1}, Z_1, \ldots, Z_{n_2})$, где

$$F_n(x_1, \ldots, x_n) = \prod_{i=1}^{n_1} F(x_i) \prod_{i=n_1+1}^{n} G(x_i).$$

Гипотеза H_0 заключается в том, что $F = G$, а альтернатива H_1 – что $F \neq G$. В некоторых случаях мы будем рассматривать более частные альтернативы.

Общий подход к задаче однородности таков – вводится некоторое отображение $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}^+$, где $\mathcal{P}$ – множество интересующих нас вероятностных мер. При этом $d(P, P) = 0$ при любом $P \in \mathcal{P}$. Можно рассматривать взамен $d'(F, G)$, если нам удобнее выражать ответ в терминах ф.р.

Отметим, что $d(P, Q)$ или $d'(F, G)$ не обязаны быть метриками (неравенство треугольника нам совершенно не требуется) и даже полуметрикой (то есть $d(P, Q) = 0$ может выполняться и в случае, когда $P \neq Q$). Полуметрики для нас будут более ценными, однако, это необязательно.

Соответственно, $d(P, Q)$ мы будем называть *характеристикой критерия* (это не общепринятое название, но надо же нам это как-то называть), где P соответствует ф.р. F, а Q – ф.р. G. В качестве статистики критерия при этом используется естественная оценка $d(\widehat{P}_{n_1}, \widehat{Q}_{n_2})$, где $\widehat{P}_{n_1}$ – ЭМ, построенная по Y_i, а $\widehat{Q}_{n_2}$ – по Z_i.

Вообще говоря, оценка $d(\widehat{P}_{n_1}, \widehat{Q}_{n_2})$ не обязана сходиться к $d(P, Q)$ (пример будет ниже), однако, мы будем рассматривать ситуации, когда такая сходимость будет иметься. Это даст нам возможность проверять гипотезу: при $P = Q$ $d(P, Q) = 0$, а при тех P, Q, которые мы рассматриваем в альтернативе, нужно чтобы выполнялось неравенство $d(P, Q) > 0$. Отметим, что если d – полуметрика, то это неравенство верно при любой альтернативе.

Соответственно, критерий будет иметь вид $d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) > C$, где C – некоторая константа.

Лучше всего, если при этом распределение $d(\widehat{P}_{n_1}, \widehat{Q}_{n_2})$ будет при $P = Q$ некоторым фиксированным распределением R. Тогда условие на уровень значимости имеет вид

$$\sup_{F=G} \mathbf{P}_{F,G}(d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) > C) = \alpha$$

сведется к

$$R(C) = 1 - \alpha,$$

то есть $C = R^{-1}(1-\alpha) =: y_{1-\alpha}$ – квантиль уровня $1-\alpha$ распределения R. В случае асимптотических критериев мы можем требовать инвариантность предельного распределения

$$d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) \to U \sim R$$

при $n_1 \to \infty$, $n_2 \to \infty$. Увы, как мы увидим, не так часто удается достичь такого в случае многомерных критериев.

Кроме того, мы бы хотели чтобы критерий был состоятелен:

$$\lim_{n_1 \to \infty, n_2 \to \infty} \mathbf{P}(d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) > C) = 1.$$

Это гарантирует, что при больших выборках критерий имеет мощность, близкую к 1.

Зачастую критерии будут использовать некоторую вспомогательную ф.р. $J(x)$ (или соответствующую меру $S(A)$), как правило, имеющую вид $aF(x) + bG(x)$, где $a, b > 0$, $a + b = 1$.

1.2.2 Примеры критериев однородности

Итак, нужно выбрать характеристику d, которая будет различать наши распределения. Приведем некоторые примеры.

- $d(P, Q) = \sup_A |P(A) - Q(A)|$ – расстояние по вариации. Это плохая характеристика, поскольку для непрерывных распределений

  $$d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) = 1$$

  Попробуйте самостоятельно доказать почему это так.

- Критерии ω^2, Смирнова и другие родственные критерии.

  - $d'(F, G) = \sup_x |F(x) - G(x)|$.

    Это естественное расстояние между функциями распределения. Соответствующий критерий (он называется критерием Смирнова) предлагает рассматривать

    $$T = d'(\widehat{F}_{n_1}, \widehat{G}_{n_2})$$

    и строить отсюда критерий. На следующих лекциях мы обсудим почему эта величина при верной гипотезе имеет некоторое фиксированное распределение, не зависящее от вида F, если она непрерывна (но зависящее от n_1, n_2), а также обсудим ее предельное поведение. В том числе покажем состоятельность критерия.

  - $d'(F, G) = \int_{\mathbb{R}} (F(x) - G(x))^2 dJ(x)$, где $J = aF + bG$ – некоторая ф.р., $a, b \in (0, 1)$, $a + b = 1$. Это также разумное расстояние – расстояние в $L^2(H)$.

    При этом соответствующий критерий (его принято называть критерием Розенблатта) имеет вид

    $$\int_{\mathbb{R}} (\widehat{F}_{n_1}(x) - \widehat{G}_{n_2}(x))^2 d\widehat{J}_{n_1,n_2}(x), \quad \widehat{J}_{n_1,n_2}(x) = \frac{n_1}{n_1 + n_2}\widehat{F}_{n_1}(x) + \frac{n_2}{n_1 + n_2}\widehat{G}_{n_2}(x).$$

Такой выбор коэффициентов a, b обусловен тем, что при верной гипотезе $\widehat{J}$ образует ЭФР объединенной выборки $(\vec{Y}, \vec{Z})$.

При этом выполнены те же свойства – распределение статистики критерия при гипотезе не зависит от конкретного вида F (если та непрерывна), можно найти предельное распределение статистики критерия гипотезы и показать, что критерий состоятелен.

– Третий критерий из того семейства предлагает брать

$$d'(F,G) = \int_{\mathbb{R}} \frac{1}{J(x)(1-J(x))}(F(x)-G(x))^2 dJ(x).$$

Весовой коэффициент позволяет с большим весом учитывать "хвосты" – т.е. точки x, для которых $J(x)$ очень маленькое или очень большое.

Такой критерий продолжает известный тест Андерсона-Дарлинга для проверки гипотезы о виде распределения в область критериев однородности. Эту версию критерия предложили Стивенс и Шольц (Scholz и Stephens, 1987).

Опять же, статистика критерия не зависит от распределения, если гипотеза верна и распределение непрерывно, опять же можно найти предельное распределение статистики критерия и показать, что критерий состоятелен.

– Четвертый родственный критерий – критерий Баумгартнера-Вейсса-Шиндлера (Baumgartner и др., 1998). Он предлагает взять

$$d'(F,G) = c_1 \int_{\mathbb{R}} \frac{(J(x)-F(x))^2}{F(x)(1-F(x))} dF(x) + c_2 \int_{\mathbb{R}} \frac{(J(x)-G(x))^2}{G(x)(1-G(x))} dG(x),$$

где c_1, c_2 – некоторые параметры. В критерии берется статистика

$$d'(\widehat{F},\widehat{G}) = \frac{n_1}{n_2} \int_{\mathbb{R}} \frac{(\widehat{J}_{n_1,n_2}(x)-\widehat{F}_{n_1}(x))^2}{\widehat{F}_{n_1}(x)(1-\widehat{F}_{n_1}(x))} d\widehat{F}_{n_1}(x) + \frac{n_2}{n_1} \int_{\mathbb{R}} \frac{(\widehat{J}_{n_1,n_2}(x)-\widehat{G}_{n_2}(x))^2}{\widehat{G}_{n_2}(x)(1-\widehat{G}_{n_2}(x))} d\widehat{G}_{n_2}(x).$$

В действительности, в оригинальной работе рассматривается слегка другая статистика, асимптотически эквивалентная нашей.

Свойства критерия достаточно близки к двум предыдущим.

Несмотря на то, что два критерия выше выглядят менее естественно, на практике они оказываются довольно эффективными.

- Можно было вместо вычитания ф.р. вычесть другие характеристики, описывающие распределение. Например, взять

$$d(\mathbf{P},\mathbf{Q}) = \sup_t \left| \int_{\mathbb{R}} e^{itx}\mathbf{P}(dx) - \int_{\mathbb{R}} e^{itx}\mathbf{Q}(dx) \right|$$

  или

$$d(\mathbf{P},\mathbf{Q}) = \int_{\mathbb{R}} \left| \int_{\mathbb{R}} e^{itx}\mathbf{P}(dx) - \int_{\mathbb{R}} e^{itx}\mathbf{Q}(dx) \right|^2 \mu(dt),$$

  где μ – некоторая конечная мера на $\mathbb{R}$. Подобные подходы менее популярны, но что-то похожее можно найти, например, в работе Epps и Singleton, 1986 или Fernandez и др., 2008.

- Критерий хи-квадрат предлагает рассмотреть некоторое разбиение $\{A_1, \ldots, A_l\}$ прямой и в качестве d взять

$$d(P,Q) = \frac{1}{a^{-1}+b^{-1}} \sum_{i=1}^{l} \frac{(P(A_i)-Q(A_i))^2}{S(A_i)}.$$

  Здесь $S(A_i) = aP(A_i) + bQ(B_i)$. Нормировка перед расстоянием выглядит неожиданно, однако, параметры подобраны так, чтобы получить предельное распределение χ^2 (про это мы подробно поговорим позднее). При этом, как и прежде, при подсчете $d(P_{n_1}, Q_{n_2})$ в качестве a, b берутся $n_1/(n_1+n_2)$ и $n_2/(n_1+n_2)$.

  Этот критерий не состоятелен (из-за выбора A_i, если меры совпадают на A_i, то критерий их не различит),

- Можно рассматривать расстояние в форме

$$d(P,Q) = \sup_{g\in\mathcal{G}} |\mathbf{E}_P g(X) - \mathbf{E}_Q g(X)|,$$

  где $\mathcal{G}$ – некоторое семейство функций $g : \mathbb{R}^k \to \mathbb{R}$.

  Простой частный случай $\mathcal{G} = \{I_{x\le y}\}$ даст нам критерии Смирнова, а $\mathcal{G} = \{I_{x\in A}\}$ даст пример из первого пункта.

  Для хорошего критерия $\mathcal{G}$ должно состоять из функций, которые не могут слишком быстро изменяться (например, монотонными, липшицевыми или что-то такое). Позднее мы рассмотрим два таких критерия: критерий в котором $\mathcal{G}$ – липшецевы функции с константой Липшица 1 (это один из случаев критерия на основе расстояния Вассерштейна, описанного в Ramdas и

др., 2017), критерий в котором $\mathcal{G}$ это единичный шар в некотором гильбертовом пространстве определенной структуры, которое мы позднее назовем Reproducing Kernel Hilbert Space (RKHS, **Gretton-a**).

Задача 1. Запишите явный вид статистик полученных критериев.

1.2.3 Критерии с частной альтернативой

Зачастую, по существу нам интересна не общая альтернатива. Например, если мы испытываем работу лекарства, то гипотеза однородности будет отвергнута всегда, когда лекарство как-то изменяет наблюдаемые показатели. Однако, если мы захотим узнать увеличивает ли лекарство, скажем, продолжительность жизни, то нам придется проверять гипотезу однородности с альтернативой "доминирования", когда $F(x) \leq G(x)$, но $F(x) \neq G(x)$. Если гипотеза будет отвергнута, то мы уверенно сможем сказать, что лекарство изменило что-то в лучшую сторону.

Поэтому зачастую мы по сути меняем альтернативу на альтернативу доминирования

$$H_1' : \ F(x) \leq G(x), \exists x_0 : F(x_0) < G(x_0).$$

Эту альтернативу можно проверять с помощью значительно большего спектра функций d.

- Первый вариант предлагает рассмотреть

$$d(P, Q) = \int_{\mathbb{R}} xQ(dx) - \int_{\mathbb{R}} xP(dx).$$

Если распределения совпадают, то и математические ожидания равны (в предположении, что они существуют), а если есть доминирование, то и математические ожидания разные. При этом

$$d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) = \overline{Y} - \overline{X}$$

имеет распределение, зависящее от меры $\mathbf{P}$. Для инвариантности предельного распределения d нужно отнормировать разность на дисперсию, например, рассматривая

$$d(P, Q) = \frac{\int_{\mathbb{R}} xQ(dx) - \int_{\mathbb{R}} xP(dx)}{\sqrt{a\left(\int_{\mathbb{R}} x^2 P(dx) - \left(\int_{\mathbb{R}} xP(dx)\right)^2\right) + b\left(\int_{\mathbb{R}} x^2 Q(dx) - \left(\int_{\mathbb{R}} xQ(dx)\right)^2\right)}}.$$

Полагая $a = n_1/(n_1+n_2)$, $b = n_2/(n_1+n_2)$, получаем статистику критерия

$$d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) = \frac{\overline{Z} - \overline{Y}}{\sqrt{(n_1 S_Y^2 + n_2 S_Z^2)/(n_1+n_2)}}.$$

Тогда

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}} d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

откуда и получаем асимптотический критерий. Этот критерий называют t-критерием (как правило слегка в другой форме, о чем мы поговорим позднее).

- Можно использовать, например, медиану, однако, здесь поиск нормировки окажется заметно сложнее, поскольку асимптотическая дисперсия выборочной медианы имеет вид $1/(4p^2(x_{1/2}))$, где $x_{1/2}$ – медиана, p – плотность. Оценка плотности здесь достаточно трудна, хотя можно построить критерий, используя подходы, о которых мы поговорим чуть позже.

- Можно взять

$$d'(F,G) = \int_{\mathbb{R}} F(x) dG(x) - \frac{1}{2} = \mathbf{P}(Y < Z) - \frac{1}{2},$$

где $Y \sim F$, $Z \sim G$ независимы. При гипотезе однородности $d'(F,G) = 0$, а при альтернативе доминирования $d'(F,G) > 0$.

При этом статистика будет иметь вид

$$d'(\widehat{F}_{n_1}, \widehat{G}_{n_2}) = \int_{\mathbb{R}} \widehat{F}_{n_1}(x) d\widehat{G}_{n_2}(x) - \frac{1}{2} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(I_{y_i < z_j} - \frac{1}{2} \right).$$

Можно найти предельное распределение d', откуда получится критерий, называемый критерием Манна-Уитни-Уилкоксона. Как правило, опять же, рассматривают асимптотически эквивалентный критерий со слегка другими параметрами.

Задача 2. Запишите явный вид статистик полученных критериев

1.3 Несколько подходов к построению критериев

Обсудим варианты как строить критерии со статистиками, чье распределение не зависит от конкретной точки гипотезы (т.е. при $F = G$ не зависят от конкретного F).

Мы предлагаем три варианта:

1. построение предельной теоремы для статистики критерия, где предельное распределение не зависит от наблюдений;

2. переход к рангам;

3. перестановочный тест.

Первый подход мы уже рассматривали выше, однако, зачастую он затрудняется тем, что на многомерный случай он не переносится. Например, критерии Смирнова или Стивенса-Шольца в многомерных версиях уже не будут работать напрямую – теперь предельное распределение статистик будет зависеть от совместного распределения координат наших векторов.

1.3.1 Перестановочный подход

Значительно более универсальным (но крайне вычислительно требовательным) является перестановочный подход. Представим себе, что $T(x_1, \dots, x_n)$ – статистика интересующего нас критерия (здесь $x_1, \dots, x_n$ – объединеннная выборка). Мы предполагаем, что при верной гипотезе T принимает маленькие значения, а при верной альтернативе, как правило, более большие. Тогда возьмем и перемешаем нашу исходную выборку, игнорируя разделения на группы. Если гипотеза была верна, то любая перестановка оставит наши данные реализацией н.о.р. величин. Если же верна альтернатива, то для перемешанных данных, скорее всего, по существу верна гипотеза (поскольку мы смешали наши точки вместе). Значит, если я найду значения $T_1, \dots, T_N$, полученные из N случайно выбранных перестановок наших исходных данных, то при верной гипотезе T равновероятно займет любой из $N + 1$ места в нашей выборке, а при альтернативе окажется где-то справа. Значит отвергая гипотезу, если T оказалась среди $[(N + 1)\alpha]$ самых больших значений, я получу критерий с вероятностью ошибки, близкой к α.

Этот подход позволяет использовать любую статистику в качестве статистики критерия (отчасти это и вызвало бум новых критериев, когда появились

быстро работающие компьютеры) – нам не нужно уметь доказывать предельные теоремы или искать распределения. Но платить за это придется N-кратным подсчетом нашей статистики, где N – велико (иначе, будет низкая точность критерия). А ведь иногда наши статистики будут также вычислительно сложными. Скажем, если статистика считается за $O(n^2)$ операций (а это даже очень неплохо), то при $N = n = 1000$ понадобится миллиард операций. Что же будет при выборках в десятки или сотни тысяч наблюдений?

Главная проблема здесь именно в мультипликативном увеличении вычислительной сложности критерия. Тем не менее, этот подход дает ответ в любой ситуации и зачастую другого ответа у нас не будет.

1.3.2 Метод Монте-Карло

Впрочем, ситуацию можно заметно улучшить, если мы знаем, что распределение (предельное распределение) статистики критерия при гипотезе не зависит от конкретной точки гипотезы (например, потому что статистика ранговая), но не умеем его в явном виде оценивать. Тогда можно сгенерировать M выборок из того же распределения, посчитать для каждой из них значение статистики T_i и достаточно сравнить наше значение статистики T с T_i – если наше значение оказалось среди $[\alpha(M+1)]$ самых больших, то гипотезу можно отвергать.

Если у вас возникло дежа вю, то это совершенно напрасно. Подход в чем-то похож на перестановочный, но имеет существенные отличие. Перестановочный подход не требует независимости от распределения, но зато требует проделывать манипуляции с генерированием новых данных при каждом запуске. А вот наш метод (назовем его методом Монте-Карло) может быть запущен разово при создании критерия. Пусть даже потратив большое время, мы можем составить список T_i, а потом сохранить его и использовать как эталон при каждом запуске. В итоге первый запуск будет достаточно длительным, а следующие – значительно быстрее.

Проблема в том, что нужно иметь набор эталонных статистик T_i при всевозможных наборах n_1, n_2, ведь значения статистик T_i равноправны только если размер выборки один и тот же. Как правило, делают так – при маленьких n_1, n_2 составляют такие таблицы в явном виде, а при больших доказывают какую-то предельную теорему (возможно с довольно неприятным предельным распределением) и оценивают уже предельное распределение.

1.4 Краткое резюме

Итак, мы с вами ввели понятия критерия однородности, поняли, что для задания критерия нужно выбрать хорошую характеристику "расстояния" между распределениями $d(P, Q)$.

Если мы хотим построить критерий с общей альтернативой, то лучше чтобы $d(P, Q)$ была полуметрикой, то есть обнулялась только при $P = Q$. Если же нам интересны более узкие альтернативы (зачем это нужно – пример был выше), то выполнение указанного условия полуметрики нужно только на допустимых альтернативах.

Критерий будет иметь вид $\{\vec{x} : d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) > C\}$. Хорошо, чтобы при этом $d(\widehat{P}_{n_1}, \widehat{Q}_{n_2}) \to d(P, Q)$, это обеспечит нашему критерию состоятельность.

Остается выбрать C. Это можно сделать:

- с помощью перестановочного подхода;
- с помощью поиска предельного или точного распределения при верной гипотезе (если оно не зависит от конкретной точки гипотезы);
- с помощью метода Монте-Карло, если известно, что распределение не зависит от конкретной точки гипотезы.

По существу остается лишь две проблемы: а) выбрать хорошую характеристику d, подходящую для ваших целей;
б) если вам важна скорость, то показать независимость распределения (предельного распределения) статистики от гипотезы и либо найти это распределение, либо оценить его методом Монте-Карло.

Глава 2

Критерии однородности и независимости. Введение, часть вторая

2.1 Критерии однородности в многомерном случае

2.1.1 Ранговые критерии и проблема многомерных рангов

Если распределение непрерывно и одномерно, то мы можем перейти от X_i к $\widehat{F}_n(X_i)$. Величина $n\widehat{F}_n(X_i)$ представляет собой порядковый номер X_i среди $X_1, \ldots, X_n$ и называется *рангом*.

При этом набор

$$\left(n\widehat{F}_n(X_1), \ldots, n\widehat{F}_n(X_n)\right)$$

представляет собой перестановку чисел от 1 до n, причем все перестановки равновероятны. Таким образом, при верной гипотезе однородности мы можем перейти от $X_{i,j}$, $i \leq n_j$, $j \leq k$ к $n\widehat{H}_n(X_{i,j})$, $n = n_1 + \cdots + n_k$. Это будут ранги наблюдений в общей выборке, при верной гипотезе однородности они представляют собой случайную перестановку чисел от 1 до n. Именно по этому принципу построен критерий Манна-Уитни.

Проблема в том, что в многомерном случае этот подход не работает (по-

крайней мере в наивной форме). В этом случае

$$\widehat{F}_n(\vec{x}) = \frac{1}{n}\sum_{j=1}^{n} I_{X_{j,i}\le x_i, i=1,\dots,d},$$

где d – размерность пространства.Величины $n\widehat{F}_n(X_i)$ в многомерном случае могут быть, например, все равны 1 даже в случае непрерывного распределения. Например, если $X_1 = (1,2)$, $X_2 = (2,1)$, то

$$n\widehat{F}_n(X_1) = I_{1\le 2,2\le 1} + I_{2\le 2,1\le 1} = 1, \quad n\widehat{F}_n(X_2) = I_{1\le 1,2\le 2} + I_{1\le 2,2\le 1} = 1.$$

Та же проблема возникает, если мы перейдем к набору рангов по каждой из координат. По каждой из координат мы получим набор $(1,2,\dots,n)$, то система сочетания этих наборов будет различной и будет зависеть от распределения. Например, для независимых (X,Y) мы получим n случайных пар из набора $\{(i,j), i\le n, j\le n\}$, а для $X = Y$ мы получим n пар $\{(1,1),\dots,(n,n)\}$.

Проблема оказалась столь серьезной, что решить ее сумели только "дорогими" с вычислительной точки зрения методами в конце 20–начале 21 века. Один из подходов, предложенный в работе Chernozhukov, 2017 в 2017 году мы рассмотрим ближе к концу курса.

2.1.2 Копулы

Посмотрим более детально на те же проблемы с многомерными данными с точки зрения теории копул.

Итак, мы знаем, что если F – непрерывная ф.р., то $F(X) \sim R[0,1]$. А что же будет если мы рассмотрим

$$(F_1(X_1), F_2(X_2)),$$

где $X_1 \sim F_1$, $X_2 \sim F_2$, вообще говоря, зависимы. Это будет случайный вектор, у которого координаты равномерны.

Определение 6. Распределение, чьи маргинальные распределения равномерны, называют копулой:

$$C(x_1,\dots,x_m) = \mathbf{P}(U_1 \le x_1, U_2 \le x_2, \dots, U_m \le x_m),$$

где U_i – зависимые равномерные, т.е. $C(x_1,1,\dots,1) = x_1$, $C(1,x_2,1,\dots,1) = x_2$,..., $C(1,1,\dots,1,x_n) = x_n$.

Таких функций много, например,

- $C(x, y) = xy$ (независимые величины),
- $C(x, y) = \min(x, y)$ (совпадающие величины $U_1 = U_2$),
- $C(x, y) = \max(x + y - 1, 0)$ ($U_1 = 1 - U_2$),
- $C(x, y) = \Psi^{-1}(\Psi(x) + \Psi(y))$, где Ψ – строго монотонно убывающщая строго выпуклая функция, стремящаяся к бесконечности при $x \to 0$ и равная нулю в единице.

Проблема как обобщения критериев типа Смирнова, так и обобщения рангов на многомерный случай в том, что распределение вектора

$$(F_1(X_1), \ldots, F_n(X_n))$$

образует некоторую копулу, которую мы не знаем. Это в одномерном случае копула была всего одна (какая?), а вот в многомерном случае их много и в зависимости от ее структуры распределение статистики от $F_i(X_i)$ будет разным. Существуют подходы, которые основаны на оценивании соответствующей копулы, однако, это заметно более сложный путь.

2.1.3 Работа построенных критериев для $\mathbb{R}^k$

Рассмотрим еще раз построенные выше критерии и посмотрим, какие из них допускают адаптации на случай многомерных данных.

- Три подхода (MMD, Energy test и поход на основе расстояния Вассерштейна), которые базировались на статитике вида

  $$d(P, Q) = \sup_{g \in \mathcal{G}} |\mathbf{E}_P g(X) - \mathbf{E}_Q g(X)|,$$

  где $\mathcal{G}$ – некоторое семейство функций $g : \mathbb{R}^k \to \mathbb{R}$, могут быть без каких-либо сложностей адаптированы на другие метрические (или даже не метрические) пространства. Соответственно, изначально эти критерии строились для многомерного случая или даже ситуации, когда наблюдения имеют более общую структуру.

- Критерии однородности хи-квадрат и обобщенного отношения правдоподобий не привязаны к геометрии и размерности модели и без изменений применимы к любой вероятностной модели.

- Критерии Смирнова и ω^2 могут быть перенесены на многомерный случай, если рассмотреть те же статистики для многомерных функций распределения. Например, критерий Смирнова может быть рассмотрен в форме

$$T = \sup_{\vec{x}} |\widehat{F}_n(\vec{x}) - \widehat{G}_m(\vec{x})| > C.$$

  Однако, если в случае непрерывных ф.р. T в одномерном случае имела некоторое фиксированное распределение, то в многомерном случае это не так. Об этом подробно поговорим в один из следующих раз, пока лишь отметим, что это порождает множество различных попыток обходить эту проблему, например, столь изысканный как в Friedman и Rafsky, 1979. Может показаться удачным подход, предложенный в (Arboretti и др., 2020) – можно взять характеристику вида

$$\sum_{i=1}^{k} \sup_{x} |F_i(x) - G_i(x)|$$

  для критерия Смирнова или, как было предложено в названной выше работе, складывать маргинальные характеристику критерия Андерсона-Дарлинга. Однако, во-первых, это уже не полуметрика (из равенства маргинальных распределений не следует равенство распределений), во-вторых, распределение при гипотезе в этом случае не будет инвариантным.

- Критерий на основе многомерной характеристической функцией можно построить так же как и на основе одномерной, однако, здесь также будет проблема неинвариантности предельного распределения при верной гипотезе.

- Критерий Стьюдента в одномерном случае основывался на характеристике

$$|\mathbf{E}_F X - \mathbf{E}_G Y|,$$

  которую для инвариантности предельного распределения переводят в форму

$$\frac{|\mathbf{E}_F X - \mathbf{E}_G Y|}{\sqrt{a\mathbf{D}X + b\mathbf{D}Y}}.$$

  В многомерном случае естественно рассматривать взамен

$$(\mathbf{E}_F \vec{X} - \mathbf{E}_G \vec{Y})^T (a\Sigma_X + b\Sigma_Y)^{-1} (\mathbf{E}_F \vec{X} - \mathbf{E}_G \vec{Y}),$$

где Σ – матрицы ковариаций нашего распределения. Соответствующий критерий для нормальных выборок известен как критерий Хотеллинга (Hotelling, 1992).

- Критерий Манна-Уитни рассматривал характеристику $\mathbf{P}(X < Y)$. Опять же на что заменить неравенство в многомерном случае – большое вопрос. Ранговую структуру критерия при этом мы все равно потеряем. Таким образом, модернизация критерия в многомерном случае требует некоторого нового взгляда на подход. Опять же, ряд таких попыток предпринимались, но, насколько я могу судить, ни одна не стала особенно популярной.

2.1.4 Критерий однородности нескольких выборок

Пусть $X_{i,j}$, $i \le n_i$, $j \le k$, набор выборок,

$$X_{i,j} \sim F_i.$$

Гипотеза однородности k выборок звучит в форме $H_0 : F_1 = F_2 = \ldots = F_k$, а альтернатива, как правило, рассматривается общая.

Для множества выборок, как правило, используют величину

$$d(P_1, \ldots, P_k) = d(P_1, S) + \cdots + d(P_k, S), \quad S(A) = a_1 P_1(A) + \cdots + a_k P_k(A).$$

Здесь как всегда $\sum a_i = 1$, $a_i > 0$. Как правило, строя статистику критерия, мы будем брать $a_i = n_i/(n_1 + \cdots + n_k)$

- Критерий Стивенса и Шольца (Андерсона-Дарлинга для проверки однородности) в такой версии базируется на статистике

$$\sum_{i=1}^{k} \frac{n_i}{n_1 + \cdots + n_k} \int_{\mathbb{R}} \frac{1}{\widehat{J}_n(x)(1 - \widehat{J}_n(x))} (\widehat{F}_{i,n_i}(x) - \widehat{J}_n(x))^2 d\widehat{J}_n(x).$$

- Критерий BWS может быть обобщен, если взять статистику

$$\sum_{i=1}^{k} \frac{n_i}{n_1 + \cdots + n_k} \int_{\mathbb{R}} \frac{1}{\widehat{F}_{i,n_i}(x)(1 - \widehat{F}_{i,n_i}(x))} (\widehat{F}_{i,n_i}(x) - \widehat{J}_n(x))^2 d\widehat{J}_n(x).$$

  Достаточно похоже на предыдущий критерий, но слегка другой знаменатель.

- Критерий хи-квадрат превращается в критерий на основе функции

$$d(P,H)=\sum_{i=1}^{l}\frac{(P(A_i)-J(A_i))^2}{J(A_i)}.$$

Таким образом, получаем вид статистики критерия

$$\sum_{j=1}^{k}\sum_{i=1}^{l}\frac{(\widehat{P}_{j,n_j}(A_i)-\widehat{S}_n(A_i))^2}{\widehat{S}_n(A_i)},$$

где $\widehat{S}_n$ – ЭМ, построенная по всем выборкам вместе.

- Рассматривая аналогичную статистике Манна-Уитни величину, получаем

$$d(F,J)=\left(\int_{\mathbb{R}}J(x)dF(x)-\frac{1}{2}\right)^2$$

и статистику критерия

$$\sum_{i=1}^{k}n_i\left(\int_{\mathbb{R}}\widehat{J}_n(x)d\widehat{F}_{n_i}(x)-\frac{1}{2}\right)^2.$$

Этот критерий (его называют критерием Краскелла-Уоллиса), вообще говоря, не состоятелен, но если хотя бы одно распределение доминирует над каким-либо другим, то $\sum_{i=1}^{k}d(F_i,H)\neq 0$.

Задача 3. Запишите явный вид статистик полученных критериев

2.2 Критерии независимости

2.2.1 Обший подход

Гипотеза независимости проверяется на основе данных $(X_i,Y_i)\sim H$, где $\vec{X}\in\mathbb{R}^p$ имеет ф.р. F, $\vec{Y}\in\mathbb{R}^q$ имеет ф.р. G. Гипотеза H_0 говорит о том, что $H(x)=F(x)G(x)$ при всех x, а альтернатива H_1 – что $H(x)\neq F(x)G(x)$ при некотором x.

В действительности, можно заметить, что это все та же гипотеза однородности – мы проверяем верно ли, что набор (X_i,Y_j), $i,j\leq n$ имеет то же распределение, что и (X_i,Y_i), $i\leq n$. Напрямую применить критерий при этом не

получится, поскольку данные в двух выборках зависимы, но подход мы можем использовать тот же, рассматривая характеристику $d'(F \times G, H)$ и статистику критерия $d'(\widehat{F} \times \widehat{G}, \widehat{H})$. или $d(P \times Q, R)$, где P, Q, R – меры, соответствующие F, G, H. Каждый многомерный (!) критерий однородности, который мы построим, будет порождать соответствующий критерий независимости. При $p = q = 1$ не возникнет и описанной выше проблемы с многомерностью и копулами – мы знаем структуру зависимости координат $\mathbf{P} \times \mathbf{Q}$ (зависимости нет), а при верной гипотезе и $\mathbf{R}$ (зависимости также нет). При $\max(p, q) > 1$ проблемы с многомерными данными сохранятся.

2.2.2 Некоторые частные случаи

- Критерий хи-квадрат для проверки независимости использует характеристику
$$d(P \times Q, R) = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{(P(A_i)Q(B_j) - R(A_i \times B_j))^2}{P(A_i)Q(B_j)}$$
и статистику
$$d(\widehat{P}_n \times \widehat{Q}_n, \widehat{R}_n) = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{(\widehat{P}_n(A_i)\widehat{Q}_n(B_j) - \widehat{R}_n(A_i \times B_j))^2}{\widehat{P}_n(A_i)\widehat{Q}_n(B_j)},$$
где $\{A_i, i \leq l\}$ – разбиение $\mathbb{R}^{d_1}$ и $\{B_j, j \leq m\}$ – разбиение $\mathbb{R}^{d_2}$. Как мы увидим позже, соответствующий критерий совпадает с критерием однородности.

- Довольно естественно в векторном случае рассматривать характеристики
$$d(\mathbf{R}, \mathbf{P} \times \mathbf{Q}) = \int_{\mathbb{R}^p \times \mathbb{R}^q} (H(x, y) - F(x)G(y))^2 J(x, y) dx dy,$$
где J – некоторая весовая функция (возможно, зависящая от $\mathbf{P}, \mathbf{Q}, \mathbf{R}$. Такой критерий предлагался Хефдингом (Hoeffding, 1994), однако, насколько я могу судить, не приобрел большой популярности.

- Аналоги критерия Смирнова и других критериев могут быть построены в рассматриваемом случае, причем при верной гипотезе в случае двумерных данных ($p = q = 1$) можно показать инвариантность требуемого распределения. Отсюда получаются критерии на основе характеристик,
$$\sup_{x,y} |H(x, y) - F(x)G(y)|, \int_{\mathbb{R}^2} (H(x, y) - F(x)G(y))^2 dH(x, y),$$

ряд таких критериев описан в Blum и др., 1961

- Аналогично вполне естественно брать

$$d(\mathbf{R}, \mathbf{P} \times \mathbf{Q}) = \int_{\mathbb{R}^p \times \mathbb{R}^q} (\psi_{X,Y}(s,t) - \psi_X(s)\psi_Y(t))^2 J(x,y)dxdy,$$

где ψ – х.ф. соответствующих векторов. Такой критерий предложен Секеем и Риццо в Székely и Rizzo, 2009, хотя мы дадим множество различных интерпретаций их результату. Этот критерий стал очень популярен на практике.

- Можно рассмотреть

$$\mathbf{E}_R f(X,Y) - \mathbf{E}_{P\times Q} f(X,Y),$$

где f какая-то функция или семейство функций (второй случае мы рассмотрим во второй половине семестра, изучая подход RKHS, что описан выше).

  - Для случая $d_1 = d_2 = 1$ можно рассмотреть $f(x,y) = xy$, тогда

$$d(P \times Q, R) = \left| \int_{\mathbb{R}^{\not\varkappa}} xyP(dx)Q(dy) - \int_{\mathbb{R}^2} xyR(dx \times dy) \right|.$$

Это классическая ковариация. Соответствующая статистика имеет вид

$$d(\widehat{P}_n \times \widehat{Q}_n, \widehat{R}_n) = \left| \int_{\mathbb{R}^2} xy\widehat{P}_n(dx)\widehat{Q}_n(dy) - \int_{\mathbb{R}^2} xy\widehat{R}_n(dx,dy) \right| = |\overline{XY} - \overline{X}\,\overline{Y}|$$

Однако, данная величина при гипотезе имеет распределение, зависящее от распределения координат (что, впрочем, не мешает использовать перестановочный подход). Поэтому как правило переходят к корреляции

$$d(\mathbf{P} \times \mathbf{Q}, \mathbf{R}) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\mathbf{D}X\mathbf{D}Y}}$$

и выборочной корреляции

$$\widehat{\rho}_P = d(\widehat{P}_n \times \widehat{Q}_n, \widehat{R}_n) = \frac{|\overline{XY} - \overline{X}\,\overline{Y}|}{S_X S_Y},$$

где

$$S_X^2 = \int_{\mathbb{R}} x^2 d\widehat{F}_{n_1}(x) - \left(\int_{\mathbb{R}} x d\widehat{F}_{n_1}(x)\right)^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \overline{x})^2,$$

и S_Y^2 аналогичным образом определяется по y_i.

Задача 4. Какой характеристике соответствует выборочная корреляция?

Величина $\widehat{\rho}_P$ стремится к некоторому предельному нормальному распределению, чья дисперсия зависит от третьего и четвертого моментов распределения (позднее мы это предельное распределение получим), а в случае двумерного нормального распределения величин как правило используют величину $\sqrt{n-2}\rho_P/\sqrt{1-\rho_P^2}$, имеющую распределение Стьюдента.

Критерий не является состоятельным, поскольку его характеристика не является полуметрикой: для зависимых величин коэффициент корреляции также бывает нулевым. Однако, это и плюс коэффициент – он сможет выловить то, что с увеличением X величина Y также склонна увеличиваться/уменьшаться.

– Можно избавиться от зависимости от распределения, взяв $f(x,y) = F(x)G(y)$:

$$d'(F\times G, H) = \left|\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} F(x)G(y)dH(x,y) - \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} F(x)G(y)dF(x)dG(y)\right| = \left|\int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} F(x)G(y)dH(x,y) - \frac{1}{4}\right|.$$

Если среди наблюдений нет повторений, то статистика критерия имеет вид

$$\left|\frac{1}{n}\sum_{i=1}^{n}\widehat{F}_n(X_i)\widehat{G}_n(Y_i) - \frac{1}{4}\right| = \left|\frac{1}{n}\sum_{i=1}^{n}U_iV_i - \frac{1}{2}\right|,$$

где U_i, V_i – ранги X_i среди X_j и Y_i среди Y_j соответственно.

Как правило, здесь также рассматривают корреляцию

$$\widehat{\rho}_S = \frac{n}{\sum_{i=1}^{n}(U_i-\overline{U})^2}\left(\frac{1}{n}\sum_{i=1}^{n}U_iV_i - \frac{1}{2}\right) = \frac{12}{n^2-1}\left(\frac{1}{n}\sum_{i=1}^{n}U_iV_i - \frac{1}{2}\right),$$

которую называют *коэффициентом корреляции Спирмена*. Его модуль и используют как статистику. Его распределение при верной гипотезе не зависит от функции распределения (если та непрерывна). Позже мы покажем, что ее предельное распределение нормально, откуда можно построить критерий для выборок большого размера.

Это опять же не состоятельный критерий – для зависимых величин $\operatorname{corr}(F(X), G(Y))$ совсем не обязана быть ненулевой. Вектор $(F(X), G(Y))$ имеет в качестве ф.р. копулу, но копулы с зависимыми координатами с нулевой корреляцией вполне себе встречаются.

Задача 5. Постройте пример такой копулы.

- Родственным критерию Спирмена является критерий Кенделла. Он использует

$$d(P\times Q, R) = \left| \int_{\mathbb{R}} \operatorname{sgn}((F(x_1) - F(x_2))(G(y_1) - G(y_2)))H(dx_1, dy_1)H(dx_2, dy_2) \right|.$$

Соответствующая статистика критерия имеет вид

$$\frac{1}{n^2} \left| \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{sgn}(U_i - U_j) \operatorname{sgn}(V_i - V_j) \right|,$$

где U, V как и прежде ранги X и Y соответственно. Это разность количества согласованных и несогласованных пар наблюдений, нормированная n^2. Эту величину называют *коэффициентом корреляции Кендалла* (точнее, как правило, используют асимптотически эквивалентную величину с другой нормировкой).

Задача 6. Убедитесь в этом соотношении.

Опять же распределение статистики при гипотезе не зависит от распределения наблюдений (если оно непрерывно), предельное распределение статистики (после перенормировки) нормальное.

Опять же соответствующий критерий не состоятелен, поскольку соотношение

$$\mathbf{E} \operatorname{sgn}((X_1 - X_2)(Y_1 - Y_2)) = 0$$

для н.о.р. пар (X_i, Y_i), $i \leq 2$, возможно даже если компоненты зависимы.

Критерии Спирмена и Кендалла, как нередко говорят, отслеживают "монотонную зависимость". Что это значит, сформулировать не так просто, но как минимум, а) при неслучайной монотонно возрастающей зависимости коэффициенты будут 1, а при убывающей будет -1), б) при монотонно возрастающих преобразованиях каждой из координат коэффициенты не меняются.

Глава 3

t-критерий, критерий Манна-Уитни и Краскелла-Уоллиса

3.1 Критерий Стьюдента и его модификации

Мы рассматриваем две независимых выборки $Y_1, \ldots, Y_{n_1}$ с ф.р. F и $Z_1, \ldots, Z_{n_2}$ с ф.р. G, проверяя для них гипотезу $H_0 : F = G$.

3.1.1 Общая философия

Начнем с нескольких вспомогательных утверждений, касающихся сходимости по распределению.

Лемма 1. *Пусть* $\vec{X}_n \xrightarrow{d} \vec{X}$, $f : \mathbb{R}^k \to \mathbb{R}$ – *непрерывна, тогда* $f(\vec{X}_n) \to f(\vec{X})$.

Лемма 2. *1) Пусть* $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$, *где* $(X_n, n \geq 1, X)$ *не зависит от* $(Y_n, n \geq 1, Y)$. *Тогда* (X_n, Y_n) *сходится к* (X, Y).
2) Пусть $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$, *где* c – *некоторая константа. Тогда* (X_n, Y_n) *сходится к* $(X,)$.

Задача 7. Докажите первую часть леммы, используя х.ф.

Задача 8. Докажите вторую часть леммы, используя а) то, что для доказательства слабой сходимости вектора достаточно доказать слабую б) лемму Слуцкого: если $X_n + Y_n \xrightarrow{d} X + c$.

Итак, начнем мы с достаточно хорошо известного критерия Стьюдента. Представим себе, что я хочу сравнивать два распределения с помощью характеристики

$$d(P,Q) = \left| \int_{\mathbb{R}} xP(dx) - \int_{\mathbb{R}} xQ(dx) \right|.$$

Распределения с равными математическими ожиданиями я считаю одинаковыми, а с разными – разными. Это не очень помогает в общей альтернативе гипотезы однородности, но вполне эффективно в более частных случаях, например, альтернативе доминирования.

Статистикой критерия, таким образом, я должен назвать при этом $\overline{Y} - \overline{Z}$. Это состоятельная оценка нашей характеристики в силу ЗБЧ. Распределение статистики (и точное, и асимптотическое) при верной гипотезе зависит от распределения выборок, хотя я могу использовать соответствующий перестановочный критерий.

Задача 9. Реализовать в Python, R, C или другом удобном вам языке перестановочный критерий, основанный на данной характеристике.

Какое же предельное распределение у этой величины при верной гипотезе? Пусть среднее нашего распределение μ, а дисперсия $\sigma^2 > 0$. Тогда

$$\sqrt{n_1}(\overline{Y} - \mu) \xrightarrow{d} U \sim \mathcal{N}(0, \sigma^2),\ \sqrt{n_2}(\overline{Z} - \mu) \xrightarrow{d} V \sim \mathcal{N}(0, \sigma^2).$$

Предположим, что n_1, n_2 стремятся к бесконечности так, что $n_1/(n_1+n_2) \to p \in [0,1]$. Тогда

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}(\overline{Y}-\overline{Z}) = \sqrt{\frac{n_2}{n_1+n_2}}\sqrt{n_1}(\overline{Y}-\mu) - \sqrt{\frac{n_1}{n_1+n_2}}\sqrt{n_2}(\overline{Z}-\mu) \xrightarrow{d} \sqrt{1-p}U + \sqrt{p}V \sim \mathcal{N}(0,\sigma^2),$$

где U, V н.о.р. $\mathcal{N}(0,\sigma^2)$.

Остается избавиться от неизвестной дисперсии. Это нетрудно сделать, используя любую состоятельную оценку $\widehat{\theta}$ величины σ, ведь

$$\left(\sqrt{\frac{n_1 n_2}{n_1+n_2}}(\overline{Y}-\overline{Z}), \widehat{\theta}\right) \xrightarrow{d} (U, \sigma)$$

в силу леммы 2. В силу леммы 1

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{(\overline{Y}-\overline{Z})}{\widehat{\theta}} \overset{d}{U} \sim \mathcal{N}(0,1).$$

Например, в качестве $\widehat{\theta}$ можно взять

$$S_Y = \sqrt{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(Y_i - \overline{Y})^2}$$

или S_Z или же

$$S = \sqrt{\frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{n_1}(Y_i - \overline{Y})^2 + \sum_{i=1}^{n_2}(Z_i - \overline{Z})^2\right)},$$

где X_i – объединенная выборка. Как правило, используют последний вариант, поскольку эта оценка использует обе выборки.

Итого получаем критерий

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\frac{|\overline{Y} - \overline{Z}|}{S} > z_{1-\alpha/2},$$

где z – квантили стандартного нормального распределения. Этот критерий можно назвать критерием Стьюдента проверки однородности.

В случае, если выборки имеют распределение $F((x - \theta_1)/\theta_2)$, где θ_1, θ_2 неизвестны, а F известна, то величина

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\frac{(\overline{Y} - \overline{Z})}{S}$$

будет иметь распределение, не зависящее от параметров. В случае нормального распределения данная статистика будет иметь распределение Стьюдента с $n-1$ степенями свободы. В этом случае мы можем использовать точный критерий, который также называют критерием Стьюдента.

3.1.2 О критерии равенства средних

Отметим, что мы проверяем именно гипотезу $H_0 : F = G$ и явно используем то, что при верной гипотезе равны распределения (по-крайней мере дисперсии). Проверка гипотезы $H_0 : \mathbf{E}_F Y = \mathbf{E}_G Z$ будет осложнена тем, что распределения Y и Z, вообще говоря, отличаются. Как следствие, при построении асимптотического критерия

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\overline{Y} - \overline{Z}) \xrightarrow{d} U \sim \mathcal{N}(0, \sigma_1^2(1-p) + \sigma_2^2 p).$$

Соответственно, оценивать дисперсию придется уже другим путем (впрочем, результат вновь назовут критерием Стьюдента):

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{\overline{Y}-\overline{Z}}{\sqrt{\frac{n_2}{n_1+n_2}S_1^2+\frac{n_1}{n_1+n_2}S_2^2}} \xrightarrow{d} U \sim \mathcal{N}(0,1).$$

С точным же критерием все еще больше усложнится. Упомянутый выше подход для нормальных выборок с одинаковыми дисперсиями будет работать, а для нормальных выборок с, вообще говоря, разными дисперсиями потребует модификации. Эта проблема называется проблемой Беренса-Фишера и наиболее популярный на сегодня вариант ее решения – критерий Уэлча, который использует для той же статистики приближение распределением Стьюдента с числом свободы, зависящим от выборки.

3.1.3 ANOVA

Напомним, что для проверки гипотезы на основе k выборок предлагается рассмотреть характеристику $\sum_{i=1}^{k} d(F_i, H)$, где H – общая ф.р.

В нашем случае это привело бы к рассмотрению статистики

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{\overline{X}_i-\overline{X}}{\sqrt{\frac{n}{n_1+n}S_1^2+\frac{n}{n_1+n}S_2^2}},$$

где $\overline{X}$ – среднее объединенной выборки, а $\overline{X}_i$ – отдельных. Эта статистика неудобна по ряду причин. Давайте взамен этого вспомним, что основная интересующая нас характеристика это $d(\mathbf{P},\mathbf{Q}) = |\mathbf{E}_P X - \mathbf{E}_Q X|$. Соответственно и суммировать мы также будем их, но возведенные в квадрат (и умноженные на n_i), что удобнее чисто технически, а двухвыборочный критерий задает тот же, что и раньше. То есть мы рассмотрим величину

$$S_{out}^2 = \sum_{i=1}^{k} n_i(\overline{X}_i - \overline{X})^2,$$

которую называют межклассовой дисперсией.

Эта величина в случае однородных нормальных выборок с одинаковыми дисперсиями будет иметь χ^2_{k-1} распределение (с точностью до множителя σ^2). Давайте покажем, что то же асимптотическое распределение в условиях $n_i/n \to p_i$ она будет иметь и в случае выборок с произвольным распределением с конечной дисперсией.

Теорема 1. *Пусть* $X_{i,j} \sim F$, $\mathbf{E}X_{i,j} = \mu$, $\mathbf{D}X_{i,j} = \sigma^2 > 0$. *Тогда*

$$\frac{1}{2}S^2_{out} \xrightarrow{d} Y \sim \chi^2_{k-1}.$$

Доказательство. Без ограничения общности, будем считать, что $\mathbf{E}X_{i,j} = 0$, поскольку от вычитания константы из всех наблюдений S^2_{out} не изменится. Заметим, что тогда

$$W_n = (\sqrt{n_1}\overline{X}_1, \ldots, \sqrt{n_k}\overline{X}_k) \xrightarrow{d} U \sim \mathcal{N}(0, \sigma^2 E).$$

Здесь мы воспользовались ЦПТ для каждого из наборов $X_{i,\cdot}$ и первой частью леммы 2.

При этом

$$\overline{X} = \sum_{i=1}^{k} \frac{n_i}{n}\overline{X}_i,$$

откуда

$$(\sqrt{n_1}(\overline{X}_1 - \overline{X}), \ldots, \sqrt{n_k}(\overline{X}_k - \overline{X})) = C_n W_n,$$

где

$$C_n = \begin{pmatrix} 1 - n_1/n & -\sqrt{n_1 n_2}/n & \ldots & -\sqrt{n_1 n_k}/n \\ -\sqrt{n_2 n_1}/n & 1 - n_2/n & \ldots & -\sqrt{n_2 n_k}/n \\ & & \ldots & \\ -\sqrt{n_k n_1} & -\sqrt{n_k n_2} & \ldots & 1 - n_k/n \end{pmatrix} \to E - vv^T$$

при $n \to \infty$, где $v = (\sqrt{p_1}, \ldots, \sqrt{p_k})$. Следовательно,

$$S^2_{out} = W_n^T C_n^2 W_n \xrightarrow{d} ||(E - vv^T)U||^2$$

при $n \to \infty$. При этом вектор $(E - vv^T)U$ имеет матрицу ковариации

$$\sigma^2(E - vv^T)(E - vv^T) = \sigma^2(E - 2vv^T + vv^T vv^T) = \sigma^2(E - vv^T),$$

т.к. $v^T v = 1$. При этом с.з. матрицы $E - vv^T$ легко находятся, т.к. все векторы $u \perp v$ являются собственными с с.з. 1, а v – с с.з. 1.

Задача 10. Проверьте это!

Остается воспользоваться леммой:

Лемма 3. *Если вектор W имеет ковариационную функцию Σ, то*

$$||W||^2 \stackrel{d}{=} \sum_{i=1}^{k} \lambda_i Z_i^2,$$

где $U_i \sim \mathcal{N}(0,1)$ н.о.р., λ_i – с.з. Σ.

Доказательство. Заметим, что $\Sigma = D^T \Lambda D$, где D – ортогональная, а $\Lambda = diag(\lambda_i)$. Но тогда

$$||W||^2 = ||DW||^2 = \sum_{i=1}^{k} \lambda_i Z_i^2,$$

поскольку $DW \sim \mathcal{N}(0, \Lambda)$, а значит его координаты независимые $\mathcal{N}(0, \lambda_i)$. □

Следовательно,

$$\frac{1}{\sigma^2}||(E - vv^T)U||^2 = \sum_{i=1}^{k-1} Z_i^2 \sim \chi^2_{k-1},$$

поскольку $k-1$ с.з. нашей матрицы равны 1, а одно – нулю. □

Увы, σ^2 мы не знаем и ее надо оценить. Для нормировки используют

$$S^2_{int} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i,j} - \overline{Y}_i)^2,$$

которая для однородных нормальных выборок имеет (с точностью до множителя σ^2) χ^2_{n-k} распределение и не зависит от S^2_{out}, а для общих нормальных при делении на $n-k$ сходится к σ^2. Статистикой критерия однофакторного дисперсионного анализа (ANOVA) является величина

$$\frac{S^2_{out}/(k-1)}{S^2_{int}/(n-k)},$$

имеющая распределение Фишера с $k-1, n-k$ степенями свободы для конечных нормальных выборок и χ^2_{k-1} в предельном случае для произвольных выборок.

3.2 Критерии Манна-Уитни-Уилкоксона и Краскелла-Уоллиса

3.2.1 Общие соображения

Идея в том, чтобы рассмотреть характеристику $|\mathbf{P}(Y \le Z) - 0.5|$. При гипотезе однородности в случае непрерывного распределения это $1/2$, при некоторых альтернативах (например, доминирования) – не $1/2$. Опять же характеристика не универсальная и для общей альтернативы не годится.

Соответствующая статистика имеет вид

$$\left| \int_{\mathbb{R}} \widehat{F}_{n_1}(x) d\widehat{G}_{n_2}(x) - \frac{1}{2} \right| = \frac{1}{n_1 n_2} \left| \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left(I_{Y_i \le Z_j} - \frac{1}{2} \right) \right|.$$

прежде всего заменим X_i на $U_i = F(X_i)$, а Y_i на $V_i = G(Y_i)$ и заметим, что

$$\widehat{F}_{n_1}(F^{-1}(x)) = \frac{1}{n_1} \sum_{i=1}^{n_1} I_{X_i \le F^{-1}(x)} = \frac{1}{n_1} \sum_{i=1}^{n_1} I_{U_i \le x}.$$

Следовательно,

$$T = \int_{\mathbb{R}} \widehat{F}_{n_1}(x) d\widehat{G}_{n_2}(x) = \int_0^1 \widehat{F}^*_{n_1}(F(G^{-1}(x))) d\widehat{G}^*_{n_2}(x),$$

где $\widehat{F}^*$, $\widehat{G}^*$ – ЭФР для U_i и V_i.

Отсюда, во-первых, следует, что наша статистика – состоятельная оценка соответствующей характеристики, поскольку функция под знаком интеграла равномерно сходится к $F(G^{-1}(x)$, а функция под дифференциалом – к x.

Во-вторых, это позволяет при гипотезе рассматривать только случай когда все наблюдения имеют $R[0,1]$ распределения.

Разложим величину

$$\int_0^1 \left(\widehat{F}_{n_1}(x) - \frac{1}{2} \right) d\widehat{G}_{n_2}(x) = \int_0^1 \left(\widehat{F}_{n_1}(x) - x \right) dx +$$
$$\int_0^1 \left(x - \frac{1}{2} \right) d(\widehat{G}_{n_2}(x) - x) + \int_{\mathbb{R}} \left(\widehat{F}_{n_1}(x) - x \right) d(\widehat{G}_{n_2}(x) - x).$$

При этом

$$\sqrt{n_1}\int_{\mathbb{R}}\left(\widehat{F}_{n_1}(x)-x\right)dx=\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\left(1-X_i-\frac{1}{2}\right)\xrightarrow{d}U,$$

$$\sqrt{n_2}\int_{\mathbb{R}}xd\left(\widehat{G}_{n_2}(x)-x\right)=\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2}\left(Y_i-\frac{1}{2}\right)\xrightarrow{d}V,$$

$$\sqrt{n}\int_{\mathbb{R}}\left(\widehat{F}_{n_1}(x)-x\right)d(\widehat{G}_{n_2}(x)-x)\to 0,$$

где

$$U\sim\mathcal{N}(0,\mathbf{D}X),\quad V\sim\mathcal{N}(0,\mathbf{D}Y).$$

Первые два соотношения вытекают из ЦПТ, причем в силу независимости рассматриваемых величин можно сказать, что и соответствующая векторная последовательность сходится к паре (U,V), где U,V независимы. Третье соотношение докажем вручную, рассмотрев второй момент указанной величины:

$$\mathbf{E}\left(\sqrt{n}\int_{\mathbb{R}}\left(\widehat{F}_{n_1}(x)-x\right)d(\widehat{G}_{n_2}(x)-x)\right)^2=\frac{n}{n_1^2n_2^2}\times$$
$$\mathbf{E}\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\int_{\mathbb{R}}(I_{X_i\le x}-x)d(I_{Y_j\le x}-x)\right)^2.$$

Раскрывая второй момент суммы в предыдущем выражении, получаем,

$$\frac{n}{n_1^2n_2^2}\sum_{i_1=1}^{n_1}\sum_{j_1=1}^{n_2}\sum_{i_2=1}^{n_1}\sum_{j_2=1}^{n_2}\mathbf{E}\left(\int_{\mathbb{R}}(I_{X_{i_1}\le x}-x)d(I_{Y_{j_1}\le x}-x)\ \int_{\mathbb{R}}(I_{X_{i_2}\le x}-x)d(I_{Y_{j_2}\le x}-x)\right).$$

Заметим, что если $i_1\neq i_2$, то

$$\mathbf{E}\left(\int_{\mathbb{R}}(I_{X_{i_1}\le x}-x)d(I_{Y_{j_1}\le x}-x)\ \int_{\mathbb{R}}(I_{X_{i_2}\le x}-x)d(I_{Y_{j_2}\le x}-x)\middle|Y_{j_1},Y_{j_2}\right)$$

имеет нулевое значение. Аналогичная ситуация получится в случае $j_1\neq j_2$. Таким образом, число ненулевых слагаемых не превышает n_1n_2, а каждое из них единицы. Значит, второй момент рассмотренной величины стремится к 0, то есть величина сходится к 0 в L^2, а значит и по распределению. При этом $\mathbf{D}X=\mathbf{D}Y=1/12$ в силу равномерности рассматриваемых величин. Тем самым, при $n_1,n_2\to\infty$, $n_1/(n_1+n_2)\to p\in[0,1]$ выполнено соотношение

$$\sqrt{\frac{n_1n_2}{n_1+n_2}}T\xrightarrow{d}W\sim\mathcal{N}\left(0,\frac{1}{12}\right)$$

Итак, критерий приобретает вид

$$\left|\sqrt{\frac{n_1 n_2}{n_1+n_2}}T\right| > z_{1-\alpha/2}.$$

Отметим, что критерий по существу проверяет гипотезу однородности против альтернативы $H_1 : \mathbf{P}(X \le Y) \ne 0.5$. Однако, для проверки гипотезы $H_0 : \mathbf{P}(X \le Y) = 0.5$ с той же альтернативой он не годится.

Задача 11. Покажите это.

3.2.2 Критерий Краскелла-Уоллиса

Опять же применим наш способ работы с критериями для k выборок – возьмем

$$d(F, J) = \left(\int_{\mathbb{R}} J(x)dF(x) - \frac{1}{2}\right)^2$$

и рассмотрим

$$\sum_{i=1}^{k} d(F_i, J)$$

в качестве характеристики. При этом статистика критерия имеет вид

$$T = \sum_{i=1}^{k}\left(\int_{\mathbb{R}}\left(\widehat{J}(x) - \frac{1}{2}\right)d\widehat{F}_i(x)\right)^2.$$

По тем же причинам, что и ранее, статистика состоятельна для своей характеристики.

Найдем ее предельное распределение при $n_i \to \infty$, $n_i/(n_1 + \cdots + n_k) \to p_i \in (0, 1)$. Будем использовать тот же прием, сперва перейдем к равномерным распределениям и будем считать, что F_i – равномерны. Далее, воспользуемся разложением

$$\int_{\mathbb{R}}\left(\widehat{J}(x) - \frac{1}{2}\right)d\widehat{F}_i(x) = \int_{\mathbb{R}}\left(\widehat{J}(x) - x\right)dx +$$
$$\int_{\mathbb{R}} xd(\widehat{F}_i(x) - x) + \int_{\mathbb{R}}\left(\widehat{J}(x) - x\right)d(\widehat{F}_i(x) - x)$$

и получим

$$T = \sum_{i=1}^{k}\left(\int_{\mathbb{R}}\left(\widehat{J}(x) - x\right)dx + \int_{\mathbb{R}} xd(\widehat{F}_i(x) - x)\right)^2 + r_{n_1,\ldots,n_k},$$

где $nr_{n_1,\dots,n_k} \xrightarrow{d} 0$, $n_i \to \infty$, из тех же соображений, что и для критерия Манна-Уитни. При этом

$$\sqrt{n_i}\int_{\mathbb{R}} xd(\widehat{F}_i(x)-x) = \frac{1}{\sqrt{n_i}}\sum_{j=1}^{n_i}(Y_{i,j}-1/2) \xrightarrow{d} U_i,$$

$$\sqrt{n}\int_{\mathbb{R}}\left(\widehat{J}(x)-x\right)dx = \sum_{i=1}^{k}\frac{\sqrt{n_i}}{\sqrt{n}}\frac{1}{\sqrt{n_i}}\sum_{j=1}^{n_i}(1/2-Y_{i,j}) \xrightarrow{d} -(\sqrt{p_1}U_1+\dots+\sqrt{p_k}U_k) =: U$$

причем соответствующий вектор сходится к соответствующему вектору, где $U_i \sim \mathcal{N}(0,1/12)$ независимы.

Таким образом, nT сходится к

$$\sum_{i=1}^{k}(U_i - p_i^{1/2}(\sqrt{p_1}U_1+\dots+\sqrt{p_k}U_k))^2.$$

Но правую величину мы уже исследовали ранее, она имеет χ^2_{k-1} распределение. Отсюда имеем критерий

$$V > y_{1-\alpha},$$

где y – квантиль χ^2_{k-1} распределения. Этот критерий называется критерием Краскелла-Уоллиса. Впрочем, как правило, его статистику записывают в другой форме, асимптотически не отличающейся от данной.

Глава 4

Сходимость в функциональных пространствах

Мы с вами использовали выражение для статистик критерия в терминах эмпирической функции (ЭФР) или эмпирической меры (ЭМ). Неплохо бы разобраться с тем как ведет себя ЭФР или ЭМ с ростом выборки. Конечно, что-то на этот счет мы знаем, однако, нам потребуются более фундаментальные знания.

Для понимания происходящего нам будет удобно использовать язык так называемых "эмпирических процессов". Эта лекция посвящена обзору теории сходимости случайных процессов.

Чтобы придать чуть больше мотивации дальнейшим построения, заметим, что нас интересует предельное поведение

4.1 Слабая сходимость

4.1.1 Определение случайного процесса

Представим себе измеримое пространство $(S, \mathcal{S})$.

Определение 7. *Случайным элементом* на нем называется отображение X из $(\Omega, \mathcal{F}, \mathbf{P})$ в $(S, \mathcal{S})$, обладающее свойством измеримости, то есть $X^{-1}(B) \in \mathcal{F}$ при любом $B \in \mathcal{S}$.

Мы будем рассматривать S – подмножество пространства L_∞ ограниченных функций из области $T \subseteq \mathbb{R}^k$ в $\mathbb{R}$. Тогда случайный элемент будем называть *случайным процессом.*

Здесь возникает вопрос что такое $\mathcal{S}$. Будем использовать стандартный подход и построим *цилиндрическую сигма-алгебру*. Цилиндрическая сигма-алгебра, что вполне естественно, порождается цилиндрами

$$I_{t_1,\dots,t_k;B} = \{f : T \to \mathbb{R} : (f(t_1), f(t_2), \dots, f(t_k)) \in B\},$$

где $k \in \mathbb{N}$, $t_i \in T$, $i \le k$, $B \in \mathcal{B}(\mathbb{R}^k)$. Итак, $\mathcal{S}$ – минимальная сигма-алгебра, содержащая все такие цилиндры.

Увы, $\mathcal{S}$ весьма бедна и можно описать все множества в ней: $B \in \mathcal{S}$ тогда и только тогда, когда найдется некоторая последовательность $\{t_i\}_{i=1}^{\infty}$ и множество $C \in \mathcal{B}(\mathbb{R}^\infty)$, что

$$B = \{f : (f(t_1), f(t_2), \dots, f(t_k), ...) \in \mathcal{B}(\mathbb{R}^\infty)\}.$$

Это очень скромная система множеств, в нее не попадают даже множество ограниченных функций или непрерывных функций.

Задача 12. Покажите, что указанные два множества не имеют такого вида.

Зато легко проверять измеримость (оказывается, что случайный процесс – любой набор случайных величин на одном пространстве, индексированный множеством T), легко задавать вероятностную меру на $\mathcal{S}$ – достаточно задать конечномерные распределения

$$\mathbf{P}(X \in I_{t_1,\dots,t_k;B}) = \mathbf{P}(\omega : (X(t_1,\omega), \dots, X(t_k,\omega)) \in B).$$

Здесь $B \in \mathcal{B}(\mathbb{R}^k)$, а $X(t_i,\omega) = (X(\omega))(t_i)$ – значение функции $X(\omega)$ в точке t_i.

Если S – топологическое (или метрическое) пространство, то в некоторых случаях цилиндрическая сигма-алгебра совпадает с шаровой, то есть порожденными всеми открытыми шарами. Например, в пространстве непрерывных функций $C[0,1]$ или пространстве непрерывных справа и имеющих предел слева $D[0,1]$ функций с равномерной нормой любой замкнутый шар $\overline{U_g(\delta)} = \{f : ||f-g|| \le \delta\}$ представим в виде

$$\{f : |f(q_i) - g(q_i)| \le \delta\},$$

где q_i – упорядоченные рациональные числа, а открытый шар $U_g(\delta)$ это объединение $\overline{U_g(\delta - 1/n)}$ по всем n.

Более удачной для была бы *борелевская сигма-алгебра* $\mathcal{B}(S)$.

Определение 8. *Борелевской* называется минимальная сигма-алгебра, которая содержит все открытые подмножества S.

К сожалению, непонятно как проверять измеримость и как задавать меры на этой сигма-алгебре – у нее не видно простой системы порождающих. В сепарабельных пространствах она совпадает с шаровой, а вот в общих – нет.

Пример 1. Пусть $\rho(x, y) = I_{x \neq y}$. Тогда все шары – это все отдельные точки или вся прямая, а открытые множества – все множества. Но минимальная сигма-алгебра, содержащая все отдельные точки – это множество всех не более чем счетных множеств и всех их дополнений.

Если сузить пространство функций, то можно добиться того, что $\mathcal{B}(S) = \mathcal{S}$, например, это так для пространства непрерывных функций (непрерывных справа функций) с равномерной нормой. Такой подход используется, например, в популярной книге Биллингсли, 1977.

В нашем случае $\widehat{F}_n$ не является непрерывной, поэтому нам придется рассматривать пространство непрерывных справа функций. Для того, чтобы это пространство было полным и сепарабельным (в частности, чтобы борелевская сигма-алгебра совпала с цилиндрической) нельзя использовать равномерную норму, а нужно вводить другую топологию, так называемую топологию Скорохода. Нам это не очень удобно, поэтому мы будем рассматривать равномерную норму и пользоваться тем, что для наших целей (сходимости) если предельный процесс непрерывен п.н., то сходимость к нему можно рассматривать и по равномерной норме. Можно было бы воспользовать понятием C-сходимости (**Bor**), но мы будем придерживаться линии дополнения ко второму изданию Биллингсли Billingsley, 1999 и использовать $\circ$-сходимость.

Существует и более продвинутая версия этого подхода, когда мы отказываемся от сепарабельности и полноты, рассчитывая лишь на положительные свойства предельной меры. Этот подход был предложен Хоффманом-Йоргенсоном (к сожалению, не могу указать явной ссылки, но основы концепции подробно, хотя и весьма тяжеловесно излагаются в Wellner, 2013) и требует постоянной работы с внешней мерой (поскольку большинство событий не измеримо). Зато он применим к крайне общим пространствам (не требуется рассматривать пространство функций из $T \subset \mathbb{R}^k$, а можно рассматривать более общие пространства). В рамках нашего курса это слишком трудозатратно, но это наиболее популярный на сегодня подход при анализе асимптотического распределения статистик критериев.

4.1.2 Гауссовские процессы, броуновское движение, броуновский мост

В качестве пределов в классической предельной теории зачастую возникают нормальные распределения. А в соответствующей теории для процессов возникают гауссовские процессы. По существу это те же нормальные векторы, только несчетной размерности.

Определение 9. Процесс X_t называется гауссовским, если его распределения являются многомерными нормальными.

При этом чтобы задать распределение гауссовского вектора $(X_{t_1}, \dots, X_{t_n})$, нужно задать его вектор средних и матрицу ковариации. Естественно, каждый элемент такого вектора $m(t_i)$ и каждый элемент такой матрицы $\Sigma(t_i, t_j)$ должен быть одним и тем же, в какой бы компании мы не взяли индексы t_i, t_j. Таким образом, должны существовать некоторые функции $m : T \to \mathbb{R}$ и $K : T \times T \to \mathbb{R}$, которые и задают средние и ковариации. С другой стороны, любая такая функция m и любая неотрицательно определенная функция K зададут свой гауссовский процесс. Это можно непосредственно проверить с помощью теоремы о согласованности.

Важным примером для нас является винеровский процесс: гауссовский процесс с $K(t, s) = \min(t, s)$, $m(t) = 0$ и непрерывными траекториями. Как правило, его определяют иначе – как процесс с независимыми $\mathcal{N}(0, t - s)$ приращениями.

Для нас ключевым будет другой гауссовский процесс – броуновский мост, то есть $W_t^0 = W_t - tW_1$. Этот процесс гауссовский, поскольку $(W_{t_1}^0, \dots, W_{t_k}^0)$ представляется в виде $A(W_{t_1}, \dots, W_{t_k}, W_1)$, где A – некоторая матрица, а значит, это линейное преобразование гауссовского вектора, то есть гауссовский вектор. При этом

$$\mathbf{E}W_t^0 = 0,\ \operatorname{cov}(W_t^0, W_s^0) = \operatorname{cov}(W_t, W_s) - s\operatorname{cov}(W_t, W_1) - t\operatorname{cov}(W_1, W_s) + ts\operatorname{cov}(W_1, W_1) = \min(t, s) - ts.$$

Это можно взять за определение броуновского моста – гауссовский процесс с непрерывными траекториями, нулевым средним и ковариационной функцией $\min(t, s) - ts$.

4.1.3 Слабая сходимость в функциональных пространствах

Пусть $X_n : \Omega_n \to S$ – случайные элементы, где $(S, \mathcal{S})$ – некоторое пространство с сигма-алгеброй на нем, снабженное метрикой ρ.

Классическое определение слабой сходимости X_n к X подразумевает следующее:

$$\mathbf{E}f(X_n) \to \mathbf{E}f(X),$$

при всех непрерывных ограниченных отображениях f из S в $\mathbb{R}$.

Осталось разобраться с несколькими вопросами.

Во-первых, что такое непрерывный? На этот вопрос у нас ответ есть, поскольку есть метрика – f непрерывно, если $f(g_n) \to f(g)$ при $\rho(g_n, g) \to 0$.

Во-вторых, что такое $\mathbf{E}f(X_n)$? Вопрос может показаться нелепым, но на самом деле он не так прост. Дело в том, что в привычном нам случае если X_n случайная величина, то и $f(X_n)$ – тоже, поскольку для любого открытого U

$$\{f(X_n) \in U\} = \{X_n \in f^{-1}(U)\} \in \mathcal{F}.$$

В последнем соотношении мы заметили, что $f^{-1}(U)$ открыто как прообраз открытого при непрерывном отображении. Но если X_n – произвольный случайный элемент, то почему $X_n \in f^{-1}(U)$ открыто? Ведь $\mathcal{S}$ – это, вообще говоря, не содержит открытых множеств, если это не борелевская или включающая борелевскую сигма-алгебра.

Мы будем работать со случаем, когда $\mathcal{S} = \mathcal{B}_0$ – минимальная сигма-алгебра, содержащая все открытые шары (вообще-то этого можно достичь в достаточно большом спектре задач, но в нашем частном случае это есть и так). Однако, эта сигма-алгебра, вообще говоря, не совпадает с борелевской, если пространство не сепарабельно (а, увы, такое будет случаться).

Поэтому прямое определение слабой сходимости для наших случайных процессов сгодится только, а) когда цилиндрическая сигма-алгебра совпадает с борелевской, б) когда мы поменяем определение слабой сходимости.

Мы используем второй подход – назовем $\circ$-сходимостью X_n к X (где X имеет непрерывные траектории) такую сходимость, когда

$$\mathbf{E}f(X_n) \to \mathbf{E}f(X)$$

для каждого функционала f, являющегося измеримым относительно $(S, \mathcal{B}_0)$. То есть мы изменили условие на f, однако, для нас это не столь критично.

Отсюда следует ключевое для нас замечание: Пусть

$$X_n \xrightarrow{\circ} X,$$

а $f : D[0,1] \to \mathbb{R}^k$ – а) измеримый относительно сигма-алгебры $\mathcal{B}_0$ функционал б) непрерывный функционал. Тогда

$$f(X_n) \xrightarrow{d} f(X),$$

где сходимость рассматривается уже в обычном смысле слабой сходимости векторов.

Доказательство. Для доказательства слабой сходимости векторов достаточно показать, что для любой непрерывной ограниченной функции $g : R^k \to \mathbb{R}$ выполнено соотношение

$$\mathbf{E}g(f(X_n)) \to \mathbf{E}g(f(X)).$$

Но $g(f(\cdot))$ ограничен (т.к. g ограничена), непрерывен (т.к. оба отображения непрерывны) и измерим относительно шаровой сигма-алгебры $\mathcal{B}_0$ как композиция измеримых функционалов. Значит нужное соотношение вытекает из определения $\circ$-сходимости. □

Более того, достаточно рассматривать только функционалы, чьи точки разрыва образуют множество A, для которого $\mathbf{P}(X \in A) = 0$. У нас, как правило, X будет иметь непрерывные траектории, поэтому мы сможем существенно расширить класс функционалов.

4.1.4 Слабая сходимость в $D[0,1]$ с равномерной нормой

Пусть $D[0,1]$ – пространство функций, непрерывных справа и имеющих предел слева (cadlag = continue à droite, limite à gauche), на котором мы вводим норму

$$||f(x)|| = \sup |f(x)|.$$

К сожалению, это пространство с такой нормой не сепарабельно – функции $f_y(x) = I_{x<y}$ находятся на попарном расстоянии 1, хотя таких функций континуум.

Однако, нас будет интересовать сходимость к процессу с непрерывными траекториями, поэтому мы будем использовать понятие $\circ$-сходимости.

Основная требующаяся нам теорема выглядит следующим образом

Теорема 2. *Пусть X_i – н.о.р. $R[0,1]$ величины, $\widehat{F}_n(x)$ – ЭФР. Тогда*

$$\sqrt{n}(\widehat{F}_n(x) - x) \overset{\circ}{\to} W_x^0$$

при $n \to \infty$.

Эта теорема более известна с обычной слабой сходимостью в топологии Скорохода (Billingsley, 1999, Section 14), однако, она верна и в рассматриваемом смысле (Billingsley, 1999, Section 15). Нам последнее будет удобнее, поскольку позволяет работать с привычной нам равномерной нормой.

Глава 5

Применение теоремы о сходимости эмпирических процессов

Основная теорема позволяет нам исследовать поведение ряда интересующих нас статистик.

5.1 Одномерный случай

5.1.1 Критерий Смирнова

Пусть $Y_i \sim F$, $Z_i \sim G$. Рассмотрим

$$\widehat{D}(x) = \widehat{F}_n(x) - \widehat{G}_m(x) = \widehat{F}_n(x)(x) - F(x) - \widehat{G}_m(x) + G(x) + F(x) - G(x).$$

При этом, если гипотеза однородности $H_0 : F = G$ неверна, то рассматриваемая величина $\widehat{D}(x)$ при $n, m \to \infty$ п.н. стремится к

$$F(x) - G(x)$$

в смысле равномерной сходимости. Это следствие теоремы Гливенко-Кантелли о сходимости разности ЭФР и ф.р. к нулю. А вот если гипотеза верна, то этот предел 0. Этого результата достаточно, чтобы понять, что $\sup_x |\widehat{D}|$ – хорошая статистика для различения гипотезы и альтернативы (стремится к 0 при гипотезе и к положительной константе иначе), но для построения критерия хорошо бы узнать еще и предельное распределение при верной гипотезе. Для этого заметим, что

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x} = \frac{1}{n}\sum_{i=1}^{n} I_{F^{-1}(R_i) \le x} = \frac{1}{n}\sum_{i=1}^{n} I_{R_i \le F(x)} = \widehat{R}_n(F(x)),$$

где R_i – н.о.р. равномерные величины, а $\widehat{R}_n$ – их ЭФР.

Тем самым

$$\sup_x |\widehat{F}_n(x) - \widehat{G}_m(x)| = \sup_x |\widehat{R}_{n,1}(F(x)) - \widehat{R}_{m,2}(F(x))| = \sup_{y \in ImF} |\widehat{R}_{n,1}(y) - \widehat{R}_{m,2}(y)|,$$

где $\widehat{R}_{n,1}(y)$, $\widehat{R}_{m,2}(y)$ – две независимых ЭФР, построенных по равномерной выборке. При этом

$$\sqrt{\frac{nm}{n+m}}(\widehat{R}_{n,1}(x) - \widehat{R}_{m,2}(x)) = \sqrt{\frac{m}{n+m}}\sqrt{n}(\widehat{R}_{n,1}(x) - x) - \sqrt{\frac{n}{n+m}}\sqrt{m}(\widehat{R}_{m,2}(x) - x) \to \sqrt{1-p}W^0_{t,1} - \sqrt{p}W^0_{t,2} = Y_t$$

при $n/(n+m) \to p$, где $W^0_{t,1}$, $W^0_{t,2}$ – независимые броуновские мосты. При этом процесс Y_t гауссовский и

$$\operatorname{cov}(Y_t, Y_s) = (1-p)\operatorname{cov}(W^0_{t,1}, W^0_{s,1}) + p\operatorname{cov}(W^0_{t,2}, W^0_{s,2}) = \min(t,s) - ts,$$

то есть вновь броуновский мост. Следовательно,

$$\sqrt{\frac{nm}{n+m}}\widehat{D} \xrightarrow{d} \sup_{ImF} |W^0_t|.$$

Здесь мы пользуемся тем, что $\sup_A |f(x)|$ – а) непрерывный функционал по равномерной норме, б) измерим относительно шаровой сигма-алгебры в $D[0,1]$.

Задача 13. Докажите это.

Как правило, критерий используют в случае непрерывных F, когда $ImF = [0,1]$. Можно найти распределение модуля супремума броуновского моста (см. Billingsley, 1999), это будет известное вам распределение Колмогорова.

Отсюда получаем следующий критерий проверки однородности:

$$\sqrt{\frac{nm}{n+m}}\widehat{D} > k_{1-\alpha},$$

где k – квантиль распределения Колмогорова. Этот критерий состоятельный, асимптотического уровня α при непрерывной ф.р. F, асимптотического уровня не больше α при разрывной ф.р. F, т.к. п.н.

$$\sup_{ImF} |W^0_t| \le \sup_{[0,1]} |W^0_t|.$$

Для альтернативы доминирования ($F(x) \geq G(x)$ при всех x), мы могли бы использовать характеристику

$$\sup_x F(x) - G(x),$$

которая состоятельно оценивается величиной

$$\sup_x(\widehat{F}_n(x) - \widehat{G}_m(x)),$$

сходящей при той же нормировке к $\sup_{\mathrm{Im}\, F} W_x^0$. Опять же используется критерий

$$\sup_x(\widehat{F}_n(x) - \widehat{G}_m(x)) > k^+_{1-\alpha},$$

где k^+ – квантиль распределения $\sup W_x^0$. Это распределение также известно и имеет ф.р. $1 - e^{-x^2/2}$.

5.1.2 Критерий Розенблатта

Напомним, что этот критерий базировался на статистике

$$T = \frac{nm}{n+m}\int_{\mathbb{R}}(\widehat{F}_n(x) - \widehat{G}_m(x))^2 d\widehat{H}_{n,m}(x).$$

При этом

$$\int_{\mathbb{R}}(\widehat{F}_n(x) - \widehat{G}_m(x))^2 d\widehat{H}_{n,m}(x) \to \int_{\mathbb{R}}(F(x) - G(x))^2 d(pF(x) + (1-p)G(x))$$

в силу все той же теоремы Гливенко-Кантелли. То есть естественная оценка является состоятельной оценкой характеристики.

При этом при верной гипотезе

$$\int_{\mathbb{R}}(\widehat{F}_n(x) - \widehat{G}_m(x))^2 d\widehat{H}_{n,m}(x) = \int_{\mathbb{R}}(\widehat{R}_{n,1}(F(x)) - \widehat{R}_{m,2}(F(x)))^2 d\widehat{R}_{n,m,3}(F(x)) =$$
$$= \int_{[0,1]}(\widehat{R}_{n,1}(y) - \widehat{R}_{m,2}(y)^2 d\widehat{R}_{n,m,3}(y).$$

При этом функционал

$$f(D, H) = \int_0^1 D^2(y) dH(y)$$

на множестве $D[0,1] \times \mathcal{F}$, где $\mathcal{F}$ – множество функций распределения, непрерывен по паре аргументов в точке $(D(x), x)$ для любой непрерывной g:

$$|f(D_n, H_n) - f(g,x)| \leq \left| \int_0^1 (D_n^2(x) - g^2(x)) dH_n(x) + \int_0^1 D^2(x) d(x - H_n(x)) \right| \leq$$
$$||D_n^2 - g^2|| + \left| \int_0^1 D^2(x) dH_n(x) - \int_0^1 D^2(x) dx \right|,$$

где мы воспользовались тем, что из $||H_n(x) - x|| \to 0$ следует слабая сходимость, а из нее сходимость интегралов от непрерывных ограниченных функций.

Более того, этот функционал измерим относительно шаровой сигма-алгебры, поскольку

$$\{g, h : f(g,h) \leq u\} = \bigcap_n \left\{ \sum_{i=1}^{k} g^2 \left(\frac{1}{n}\right) \left(h\left(\frac{k+1}{n}\right) - h\left(\frac{k}{n}\right)\right) \leq u \right\},$$

а множества в правой части при любом n есть цилиндры.

Следовательно, величина

$$\frac{nm}{n+m} \int_{\mathbb{R}} (\widehat{F}_n(x) - \widehat{G}_m(x))^2 d\widehat{H}_{n,m}(x)$$

сходится по распределению к величине

$$\int_0^1 (\sqrt{p} W_{t,1}^0 - \sqrt{1-p} W_{t,2}^0)^2 dt$$

Эта величина, в силу тех же соображений, что и прежде, совпадает по распределению с

$$\int_0^1 (W_t^0)^2 dt.$$

Распределение этой величины можно описать явно (см., например, Shorack и Wellner, 2009, 3.8), но мы этого делать не будем.

5.1.3 Критерий Баумгартнера-Вейсса-Шиндлера

Критерий BWS (Baumgartner, Weiss, Schindler, Baumgartner и др., 1998) предлагает рассмотреть характеристику

$$d'(F,G) = c_1 \int_{\mathbb{R}} \frac{(H(x) - F(x))^2}{F(x)(1-F(x))} dF(x) + c_2 \int_{\mathbb{R}} \frac{(H(x) - G(x))^2}{G(x)(1-G(x))} dG(x).$$

По существу, это подход "введи расстояние между F и общей ф.р., а затем сложи по всем имеющимся распределениям", который мы использовали для построения многомерных критериев. Соответственно, рассматривают статистику

$$d'(\widehat{F},\widehat{G}) = \frac{n_1}{n_2}\int_{\mathbb{R}} \frac{(\widehat{H}_{n_1,n_2}(x)-\widehat{F}_{n_1}(x))^2}{\widehat{F}_{n_1}(x)(1-\widehat{F}_{n_1}(x))} d\widehat{F}_{n_1}(x) + \frac{n_2}{n_1}\int_{\mathbb{R}} \frac{(\widehat{H}_{n_1,n_2}(x)-\widehat{G}_{n_2}(x))^2}{\widehat{G}_{n_2}(x)(1-\widehat{G}_{n_2}(x))} d\widehat{G}_{n_2}(x),$$

При этом справедливо представление

$$\int_{\mathbb{R}} \frac{(\widehat{H}_{n_1,n_2}(x)-\widehat{F}_{n_1}(x))^2}{\widehat{F}_{n_1}(x)(1-\widehat{F}_{n_1}(x))} d\widehat{F}_{n_1}(x) = \int_0^1 \frac{(\widehat{R}_{n_1,n_2,3}(x)-\widehat{R}_{n_1,1}(x))^2}{\widehat{R}_{n_1,1}(x)(1-\widehat{R}_{n_1,1}(x))} d\widehat{R}_{n_1,1}(x),$$

где $\widehat{R}_{n_1,1}(x)$ – ЭФР для равномерной выборки, $\widehat{R}_{n_2,2}(x)$ – для независящей от нее равномерной выборки, а

$$\widehat{R}_{n_1,n_2,3} = \frac{n_1}{n}\widehat{R}_{n_1,1}(x) + \frac{n_2}{n}\widehat{R}_{n_2,2}(x)$$

их совместная ЭФР. Правая часть преобразуется к

$$\frac{n_2^2}{n^2}\int_0^1 \frac{(\widehat{R}_{n_1,1}(x)-\widehat{R}_{n_2,2}(x))^2}{\widehat{R}_{n_1,1}(x)(1-\widehat{R}_{n_1,1}(x))} d\widehat{R}_{n_1,1}(x),$$

а вся статистика к виду

$$\frac{n_1 n_2}{n^2}\int_0^1 (\widehat{R}_{n_1,1}(x)-\widehat{R}_{n_2,2}(x))^2 \left(\frac{d\widehat{R}_{n_1}(x)}{\widehat{R}_{n_1}(x)(1-\widehat{R}_{n_1}(x))} + \frac{d\widehat{R}_{n_2}(x)}{\widehat{R}_{n_2}(x)(1-\widehat{R}_{n_2}(x))}\right)$$

Здесь бы нам перейти к броуновскому мосту, наивно получив

$$\frac{n}{2} d'(\widehat{F},\widehat{G}) \to \int_0^1 (W_t^0)^2 \frac{dt}{t(1-t)}.$$

Увы, аккуратный анализ такого рода величины требует более тонких оценок. В частности, потребуется оценка поведения процесса $(\widehat{R}_i(x)-x)$ в окрестности точек 0 и 1, которую можно найти в **Shorack**, 3.7. Оценки, приведенные в Baumgartner и др., 1998, на мой взгляд, проведены небрежно и недостаточны, чтобы считаться доказательством. Однако, более технически хлопотные оценки показывают, что указанное выше утверждение все-таки верно, откуда и выводится критерий BWS.

Родственный критерий Стивенса-Шольца предлагает рассматривать

$$d(F,G) = \int_{\mathbb{R}} \frac{1}{H(x)(1-H(x))}(F(x)-G(x))^2 dH(x).$$

и, тем самым, статистику

$$d'(\widehat{F},\widehat{G}) = \int_{\mathbb{R}} \frac{1}{\widehat{H}_{n_1,n_2}(x)(1-\widehat{H}_{n_1,n_2}(x))}(\widehat{F}_{n_1}(x)-\widehat{G}_{n_2}(x))^2 dx$$

По тем же причинам, мы можем ожидать, что статистика при домножении на $n_1 n_2/n$ сойдется к тому же интегралу, однако, это вновь требует более тонких оценок.

5.2 Многомерный случай

5.2.1 Теорема о сходимости эмпирических процессов

Пусть $X \in \mathbb{R}^d$. Сможем ли мы сформулировать основную теорему о сходимости в этом случае?

В этом случае мы рассматриваем

$$\widehat{F}_n(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} I_{X_{i,1}\le x_1,\dots,X_{i,d}\le x_d},$$

пространство непрерывных справа (со всех сторон) функций из $[0,1]^d$ в $[0,1]$, имеющих в каждой точке предел "односторонний" предел. Скажем, в двумерном случае я имею ввиду, что есть предел слева-слева, слева-справа, справа-слева и справа-справа по паре координат, сами пределы, вообще говоря, разные, но последний из них совпадает со значением функции в нашей точке.

Теорема 3. *Пусть X_i – н.о.р. векторы с ф.р. C, где C – некоторая копула (это замена равномерным величинам в многомерном случае), $\widehat{F}_n(x)$ – ЭФР (многомерная). Тогда*

$$\sqrt{n}(\widehat{F}_n(x) - F(x)) \overset{\circ}{\to} Y_x$$

при $n \to \infty$, где $\{Y_x,\ x \in [0,1]^d\}$ – гауссовский процесс (то есть набор таких гауссовских векторов, что любой собранный из них гауссовский вектор является гауссовским) с ковариационной функцией

$$K(\vec{t},\vec{s}) = C(\min(t_1,s_1),\dots,\min(t_{d_1},s_{d_1})) - C(t_1,\dots,t_{d_1})C(s_1,\dots,s_{d_1}).$$

Первые шаги в построении критерия, скажем, Смирнова, остаются теми же:

$$\sup_{\vec{x}\in\mathbb{R}^d} |\widehat{F}_n(\vec{x}) - \widehat{G}_m(\vec{x})| = \sup_{\vec{x}\in\mathbb{R}^d} |\widehat{R}_{n,1}(F_1(x_1),\dots,F_d(x_d)) - \widehat{R}_{m,2}(G_1(x_1),\dots,G_d(x_d))|.$$

Здесь $F_i(x_i)$ – маргинальные ф.р. для F, $\widehat{R}_{n,i}$ – ЭФР, построенные по независимым векторам, каждая из компонент которых является $R[0,1]$.

При верной гипотезе и непрерывных F_i эта величина переписывается в виде

$$\sup_{\vec{y}\in[0,1]^d} |\widehat{R}_{n,1}(\vec{y}) - \widehat{R}_{m,2}(\vec{y}))|.$$

При этом в этом случае ЭФР $\widehat{R}$ – ЭФР, построенным по двум выборкам из некоторой копулы C.

Опять же функционал $\sup_{\vec{y}\in[0,1]^d} |f|$ является непрерывным и измеримым относительно шаровой сигма-алгебры, поэтому

$$\sup_{\vec{y}\in[0,1]^d} |\widehat{R}_{n,1}(\vec{y}) - \widehat{R}_{m,2}(\vec{y}))| \to \sup_{\vec{y}\in[0,1]^d} |Y_y|,$$

где Y_y тот самый гауссовский процесс, описанный выше. Только теперь он зависит от неизвестной копулы C и критерий строить значительно сложнее – либо нужно оценить C (такой подход предложен в Scaillet, 2005), либо использовать перестановочный подход. Поэтому критерии такого типа, насколько я могу судить, не приобрели большой популярности.

Глава 6

Критерии Пирсона, Спирмена и Кендалла

6.1 Корреляция

6.1.1 Общий подход

Напомним наш общий к проверке независимости. Мы рассматриваем выборку $(Y_i, Z_i), i \leq n$, для проверки гипотезы H_0 о независимости Y и Z внутри вектора (сегодня мы считаем их скалярными) мы строим величину

$$d(P \times Q, H),$$

где P, Q – маргинальные распределения, а H совместное, оцениваем ее с помощью

$$d(\widehat{P}_n \times \widehat{Q}_n, \widehat{H}_n),$$

которую и будем использовать как статистику критерия.

6.1.2 Критерий Пирсона

Предлагается взять в качестве характеристики

$$|cov(Y, Z)| = d(P \times Q, H) = \left| \int_{\mathbb{R}^2} yzH(dy, dz) - \int_{\mathbb{R}^2} yzP(dy)Q(dz), \right|$$

а соответствующая статистика будет иметь вид модуля выборочной ковариации

$$|Cov(Y, Z)| = |\overline{YZ} - \overline{ZY}|.$$

Величина $d(\widehat{P}_n \times \widehat{Q}_n, \widehat{H}_n)$ – состоятельная оценка $d(P \times Q, H)$ (если ковариация вообще определена) в силу ЗБЧ. Поэтому наш будущий критерий будет отличать гипотезу от альтернативы $cov(Z, Y) \neq 0$.

При этом мы можем использовать данную статистику в перестановочном формате. В противном случае нам придется решать проблему с нормировку соответствующей статистики. Без ограничения общности будем считать, что $\mathbf{E}Z = \mathbf{E}Y = 1$. Рассмотрим предельное распределение $Cov(Z, Y) = \overline{ZY} - \overline{Z}\,\overline{Y}$ при верной гипотезе независимости. Для этого представим ее в виде $g(\overline{ZY}, \overline{Z}, \overline{Y})$, где $g(x, y, z) = x - yz$. Воспользуемся центральной предельной теоремой для векторов:

$$\sqrt{n}(\overline{ZY}, \overline{Z}, \overline{Y}) \xrightarrow{d} \vec{U} \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{pmatrix} \mathbf{D}ZY & 0 & 0 \\ 0 & \mathbf{D}Z & 0 \\ 0 & 0 & \mathbf{D}Y \end{pmatrix},$$

где мы воспользовались тем, что $\mathrm{cov}(ZY, Z) = \mathrm{cov}(ZY, Y) = \mathrm{cov}(Z, Y) = 0$ в силу независимости и равенства средних нулю.

В нашем случае

$$\mathbf{D}ZY = \mathbf{E}Z^2Y^2 = \mathbf{E}Z^2\mathbf{E}Y^2 = \mathbf{D}Z\mathbf{D}Y.$$

Нам понадобится следующая лемма об асимптотической нормальности.

Лемма 4. *Пусть*

$$\sqrt{n}(\widehat{\theta}(X_1, \ldots, X_n) - \theta) \xrightarrow{d} \vec{U} \sim \mathcal{N}(\vec{0}, \Sigma),$$

g – дифференцируемая функция из $\mathbb{R}^d \to \mathbb{R}$. Тогда

$$\sqrt{n}(g(\widehat{\theta})(X_1, \ldots, X_n) - g(\theta)) \xrightarrow{d} \vec{U} \sim \mathcal{N}(\vec{0}, J\Sigma J^T), \quad J = \left(\frac{\partial}{\partial \theta_i} g(\theta), i > 0\right).$$

В нашем случае

$$\frac{\partial g}{\partial x} = 1, \ \frac{\partial g}{\partial y} = -z, \ \frac{\partial g}{\partial z} = -y, \ J = (1, 0, 0).$$

То есть

$$\sqrt{n}Cov(X, Y) \xrightarrow{d} Z \sim \mathcal{N}(0, \mathbf{D}X\mathbf{D}Y).$$

Значит,

$$|Corr(Z, Y)| = \left|\frac{\sqrt{n}Cov(Z, Y)}{S_Z S_Y}\right| > z_{1-\alpha/2},$$

где S_Z, S_Y – выборочные стандартные отклонения (впрочем, подошли бы любые состоятельные оценки $\sqrt{\mathbf{D}Z}$, $\sqrt{\mathbf{D}Y}$) будет задавать асимптотический критерий уровня α. Это асимптотический критерий Пирсона проверки гипотезы независимости.

6.1.3 Точное распределение коэффициента при нормальных данных

В случае, когда (Y, Z) – нормальный вектор, то распределение $Corr(Z, Y)$ при верной гипотезе может быть исследовано напрямую. Для этого без ограничения общности будем рассматривать случай $(Y, Z) \sim \mathcal{N}(\vec{0}, E)$, поскольку выборочный коэффициент корреляции не меняется при добавлении к наблюдениям константы или умножении их на константу. Положим для краткости $\widehat{\rho} = Corr(Y, Z)$ и представим величину $\widehat{\rho}$ в виде

$$\widehat{\rho} = \frac{\sum_{i=1}^n U_i(Z_i - \overline{Z})}{\sqrt{\sum_{i=1}^n (Z_i - \overline{Z})^2}}, \quad U_i = \frac{Y_i - \overline{Y}}{\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

Рассмотрим распределение $\widehat{\rho}$ при фиксированном векторе $U_i = u_i$. Введем ортонормированный базис $(y_1, \ldots, y_n)$, первые два элемента которого

$$y_1 = (n^{-1/2}, \ldots, n^{-1/2}),\ y_2 = (u_1, \ldots, u_n).$$

При этом мы используем то, что $||y_1||^2 = ||y_2||^2 = 1$, $\langle y_1, y_2 \rangle = 0$ по определению u_i. Перейдем от $Z_1, \ldots, Z_n$ к

$$V_i = \sum_{j=1}^n y_{i,j} Z_j.$$

Поскольку замена ортогональная, то $\vec{Z}_i$ останутся стандартными нормальными. При этом наша статистика имеет вид

$$\widehat{\rho} = \frac{V_2}{\sqrt{V_1^2 + \cdots + V_n^2 - V_1^2}} = \frac{V_2}{\sqrt{V_2^2 + \cdots + V_n^2}},$$

где мы воспользовались тем, что замена ортогональная, а значит

$$Z_1^2 + \cdots + Z_n^2 = V_1^2 + \cdots + V_n^2, \quad Z_1^2 + \cdots + Z_n^2 - n\overline{Z}^2 = V_2^2 + \cdots + V_n^2.$$

Отсюда плотность $\widehat{\rho}$ при условии $(U_1, \ldots, U_n)$ имеет некоторый фиксированный вид – отношение стандартной нормальной величины V_2 к длине вектора $(V_2, \ldots, V_n)$ со стандартным нормальным распределением. Но тогда

$$\mathbf{P}(\widehat{\rho} \in A) = \mathbf{E}(\mathbf{E}(I_{\widehat{\rho} \in A} | U_1, \ldots, U_n)) = \mathbf{E}\mathbf{P}\left(\frac{V_2}{\sqrt{V_2^2 + \cdots + V_n^2}} \in A\right) = \mathbf{P}\left(\frac{V_2}{\sqrt{V_2^2 + \cdots + V_n^2}} \in A\right).$$

Можно получить распределение $\widehat{\rho}$ непосредственно отсюда, однако, удобнее использовать статистику

$$T = \frac{\widehat{\rho}}{\sqrt{1-\widehat{\rho}^2}} \stackrel{d}{=} \frac{V_2^2}{\sqrt{V_3^2+\cdots+V_n^2}}.$$

Следовательно, $\sqrt{n-2}T$ имеет t_{n-2} распределение, откуда получаем точный критерий

$$\sqrt{n-2}\,|T| > y_{1-\alpha/2},$$

где y – квантиль t_{n-2}.

Итак, мы получаем точный критерий Пирсона для нормальной выборки.

6.2 Ранговые критерии

6.2.1 Критерий Спирмена

Для критерия Спирмена мы рассматриваем характеристику $|cov(F(Y), G(Z))|$, предполаг которой соответствует статистика

$$\left|\frac{1}{n}\sum_{i=1}^n \widehat{F}_n(X_i)\widehat{G}_n(X_i) - \frac{1}{n}\sum_{i=1}^n \widehat{F}_n(Y_i)\ \frac{1}{n}\sum_{i=1}^n \widehat{G}_n(Z_i)\right| = \left|\frac{1}{n}\sum_{i=1}^n \widehat{F}_n(X_i)\widehat{G}_n(X_i) - \frac{1}{4}\right|.$$

Наша статистика состоятельна для характеристики, поэтому критерий будет работать против любой альтернативы, для которой $cov(F(X), G(Y)) \neq 0$.

Найдем распределение статистики при верной гипотезе. Запишем величину

$$\widehat{\rho}_S = \int_{\mathbb{R}^2} \widehat{F}_n(x)\widehat{G}_n(x)d\widehat{H}_n(x,y) = \int_{[0,1]^2} \widehat{F}_n(F^{-1}(x))\widehat{G}_n(G^{-1}(y))d\widehat{H}_n(F^{-1}(x), G^{-1}(y)).$$

Как мы уже видели, это позволяет перейти от (X_i, Y_i) к $(F(X_i), G(Y_i))$, имеющим равномерное распределение. Тем самым, можно рассматривать только случай, когда X_i, Y_i – н.о.р. $R[0,1]$ величины. При этом

$$T = \int_{[0,1]^2} (\widehat{F}_n(x) - x)ydxdy + \int_{[0,1]^2} (\widehat{G}_n(y) - y)xdxdy + \int_{[0,1]^2} xyd(\widehat{H}(x,y) - xy) + R_n,$$

где R_n состоит из четырех слагаемых $R_{n,i}$, $i \leq 4$:

$$\int_{[0,1]^2} (\widehat{F}_n(x) - x)(\widehat{G}_n(y) - y)dxdy + \int_{[0,1]^2} (\widehat{G}_n(y) - y)xd(\widehat{H}_n(x,y) - xy) +$$

$$\int_{[0,1]^2} (\widehat{F}_n(x) - x)yd(\widehat{H}(x,y) - xy) + \int_{[0,1]^2} (\widehat{F}_n(x) - x)(\widehat{G}_n(y) - y)d(\widehat{H}_n(x,y) - xy).$$

Покажем, что $\sqrt{n}R_n(x)$ сходится к нулю. Для этого заметим, что $R_{n,1}$ при умножении на $\sqrt{n}$ стремится к нулю (поскольку $\sqrt{n}(\widehat{F}_n(x)-x)(\widehat{G}_n(y)-y)$ равномерно сходится к 0 при $n \to \infty$), а любое из $R_{n,i}$, $i \in \{2,3,4\}$, можно представить в форме

$$\frac{1}{n^3} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} \int_{[0,1]}^{2} F_{i_1}(x) G_{i_2}(y) d(I_{X_{i_3} \leq x, Y_{i_3} \leq y} - xy),$$

где $F_{i_1}(x)$ имеет либо вид $I_{X_{i_1} \leq x} - x$, либо x, а $G_{i_2}(y)$ либо вид $I_{X_{i_2} \leq y} - y$, либо y (причем хотя бы в одной из пар будет первое слагаемое, имеющее нулевое среднее). Следовательно, второй момент $\sqrt{n}R_{n,i}$, $i \in 2,3,4$, имеет вид

$$\frac{1}{n^5} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} \sum_{i_4=1}^{n} \sum_{i_5=1}^{n} \sum_{i_6=1}^{n} \mathbf{E} h(i_1, i_2, i_3, i_4, i_5, i_6),$$

$$h(\vec{i}) = \int_{[0,1]^2} F_{i_1}(x) G_{i_2}(y) d(I_{X_{i_3} \leq x, Y_{i_3} \leq y} - xy) \int_{[0,1]^2} F_{i_4}(x) G_{i_5}(y) d(I_{X_{i_6} \leq x, Y_{i_6} \leq y} - xy).$$

Обратите внимание, что это случайная величина. Предположим, что $i_3 \notin \{i_l, l \neq 3\}$, тогда

$$\mathbf{E}\left(h(i_1, i_2, i_3, i_4, i_5, i_6) | X_{i_l}, l \in \{2,3\}, Y_{i_l}, l \leq 3\right) = 0,$$

откуда нулевым будет и искомое математическое ожидание. Значит, ненулевые слагаемые встречаются только если i_3 и i_6 повторяются. Аналогичным образом обнулятся те слагаемые, где все 4 индекса i_1, i_2, i_4, i_5 уникальны. Значит, все ненулевые слагаемые попадают в объединение множеств

$$\bigcup_{k \in \{1,2,4,5\}} \{i_k = i_6 = i_3\} \bigcup \bigcup_{k \neq l \in \{1,2,4,5\}} \{i_k = i_3, i_l = i_3\}.$$

Однако, каждое из этих множеств имеет мощность $O(n^4)$, $n \to \infty$, откуда

$$n\mathbf{E}R_n^2 = o(1),\ n \to \infty,$$

откуда вытекает сходимость к 0 в L^2, а значит и по распределению.

Остается заметить, что

$$\sqrt{n}\int_{[0,1]^2}(\widehat{F}_n(x)-x)ydxdy=\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{2}-X_i\right),$$
$$\sqrt{n}\int_{[0,1]^2}(\widehat{G}_n(y)-y)xdxdy=\frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{2}-Y_i\right)$$
$$\sqrt{n}\int_{[0,1]^2}xyd(\widehat{H}(x,y)-xy)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_iY_i-\frac{1}{4}\right).$$

Таким образом, в силу ЦПТ

$$\sqrt{n}\widehat{\rho}_S\xrightarrow{d}\mathcal{N}(0,\mathbf{D}Z),\ Z=\left(\frac{1}{4}-\frac{(X_1+Y_1)}{2}+X_1Y_1\right)=\left(\frac{1}{2}-X_1\right)\left(\frac{1}{2}-Y_1\right).$$

Остается заметить, что

$$\mathbf{D}Z=\mathbf{E}\left(\frac{1}{2}-X_1\right)^2\mathbf{E}\left(\frac{1}{2}-Y_1\right)^2=\frac{1}{144}.$$

Отсюда асимптотический критерий Спирмена имеет вид $|12\sqrt{n}\widehat{\rho}_S|>z_{1-\alpha/2}$, где z – квантиль нормального распределения.

6.2.2 Критерий Кенделла

Другим ранговым критерием (достаточно похожим на критерий Спирмена) является критерий Кенделла, чья характеристика имеет вид

$$\mathbf{E}\operatorname{sgn}(X_1-X_2)\operatorname{sgn}(Y_1-Y_2),$$

а соответствующая статистика критерия

$$\widehat{\theta}=\int_{\mathbb{R}^4}\operatorname{sgn}(x_1-x_2)\operatorname{sgn}(y_1-y_2)d\widehat{F}_n(x_1,y_1)d\widehat{F}_n(x_2,y_2)=\frac{1}{n^2}\sum_{i\neq j}\operatorname{sgn}((X_i-X_j)(Y_i-Y_j)).$$

Другой формой представления статистики является $(N^+-N^-)/n^2$, где N^+ – число согласованных пар (пар наблюдений, в которых один из векторов поокоординатно больше другого), а N^- – число не согласованных пар. Чаще вместо n^4 рассматривают $n^2(n-1)^2$, но для нас это не столь критично, поскольку мы изучаем асимптотические свойства критерия.

Опять статистика состоятельна (что прямо следует из ее вида), покажем ее асимптотическую нормальность при верной гипотезе. Для этого вновь перейдем к равномерному распределению, представив статистику (пользуясь монотонностью F, G) в виде

$$\widehat{\rho}_K = \int_{\mathbb{R}^4} \operatorname{sgn}(F(x_1) - F(x_2)) \operatorname{sgn}(G(y_1) - G(y_2)) d\widehat{F}_n(x_1, y_1) d\widehat{F}_n(x_2, y_2)$$

и перейдя к $F(x_i), G(y_i)$ получим, что достаточно рассматривать только равномерный случай. Используем тот же трюк (его можно назвать "методом проекций")

$$\widehat{\rho}_K = 2 \int_{[0,1]^4} \operatorname{sgn}(x_1 - x_2) \operatorname{sgn}(y_1 - y_2) dF(x_1, y_1) d(\widehat{F}_n(x_2, y_2) - F(x_2, y_2)) + \int_{[0,1]^4} \operatorname{sgn}(x_1 - x_2) \operatorname{sgn}(y_1 - y_2) d(\widehat{F}_n(x_1, y_1) - x_1 y_1) d(\widehat{F}_n(x_2, y_2) - x_2 y_2).$$

При этом нам нужно показать, что второе слагаемое R_n при умножении на $\sqrt{n}$ стремится к нулю по распределению. Опять же разложим соответствующий второй момент в виде суммы и получим

$$\frac{1}{n^4} \sum_{i_j, j \le 4} \mathbf{E} \left(\int_{[0,1]^8} \operatorname{sgn}(x_1 - x_2) \operatorname{sgn}(y_1 - y_2) \operatorname{sgn}(x_3 - x_4) \operatorname{sgn}(y_3 - y_4) \prod_{l=1}^{4} d(I_{X_{i_l} \le x_l, Y_{i_l} \le y_l} - x_l y_l) \right).$$

Возьмем интеграл по первым четырем переменным, получим

$$I_{i_1, i_2} = \operatorname{sgn}((X_{i_1} - X_{i_2})(Y_{i_1} - Y_{i_2})) - \int_{[0,1]^2} \operatorname{sgn}((X_{i_1} - x_2)(Y_{i_1} - y_2)) dx_2 dy_2 - \int_{[0,1]^2} \operatorname{sgn}((x_1 - X_{i_2})(y_1 - Y_{i_2})) dx_1 dy_1.$$

Заметим, что если индекс i_1 не лежит в $\{i_2, i_3, i_4\}$, то

$$\mathbf{E}(I_{i_1, i_2} I_{i_3, i_4} | X_{i_2}, X_{i_3}, X_{i_4}) = I_{i_3, i_4} = 0,$$

поскольку среднее второго слагаемоего нулевого, а первое слагаемое после усреднения совпадает с третьим. Следовательно, искомые второй момент включает в себя лишь те наборы, где $i_1 = i_3, i_2 = i_4$ или $i_1 = i_4, i_2 = i_3$ (набор $i_1 = i_2$ или $i_3 = i_4$ дают нулевое значение I). Таким образом, существует менее $2n^2$

слагаемых, каждое из которых оценивается сверху по модулю единицей. Следовательно,

$$n\mathbf{E}R_n^2 = o(1), n \to \infty,$$

откуда $\sqrt{n}R_n \xrightarrow{d} 0$.

В свою очередь, первое слагаемое имеет вид

$$\int_{[0,1]^4} \operatorname{sgn}(x_1 - x_2)\operatorname{sgn}(y_1 - y_2)dx_1y_1d(\widehat{F}_n(x_2, y_2) - x_2y_2) = \tag{6.1}$$

$$\frac{1}{n}\sum_{i=1}^{n}\int_{[0,1]^2} \operatorname{sgn}(x_1 - X_i)\operatorname{sgn}(y_1 - Y_i)dx_1dy_1 = \tag{6.2}$$

$$\frac{1}{n}\sum_{i=1}^{n}((1 - X_i)(1 - Y_i) + X_iY_i - (1 - X_i)Y_i - (1 - Y_i)X_i). \tag{6.3}$$

Упростим величину под знаком суммы до формы $1 - 2X_i - 2Y_i + 4X_iY_i = (2X_i - 1)(2Y_i - 1)$. У данной величины дисперсия имеет вид

$$\mathbf{E}(2X_1 - 1)^2\mathbf{E}(2Y_1 - 1)^2 = \frac{16}{12^2}$$

В силу ЦПТ при умножении на $\sqrt{n}$ полученная сумма сойдется к

$$Z \sim \mathcal{N}\left(0, \frac{4}{9}\right).$$

Получаем критерий

$$12\sqrt{n}|\widehat{\rho}_K| > z_{1-\alpha/2},$$

где z – квантиль стандартного нормального распределения.

Отметим, что в силу наших рассуждений обе статистики $\widehat{\rho}_S$, $\widehat{\rho}_P$ аппроксимируются величинами

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2} - F(X_1)\right)\left(\frac{1}{2} - G(Y_1)\right)$$

с точностью до постоянного множителя. Из-за этого, полученные статистики сильно кореллированы при верной гипотезе.

Глава 7

Критерии, основанные на разности функционалов

Из середины прошлого века мы перемещаемся ближе к нашим дням. Следующие критерии однородности, которые мы будем рассматривать, имеют характеристику

$$d(\mathbf{P}, \mathbf{Q}) = \sup_{f \in \mathcal{F}} \mathbf{E}_P f(Y) - \mathbf{E}_Q f(Z).$$

Здесь $\mathcal{F}$ – некоторое семейство отображений из $\mathcal{X} \to \mathbb{R}$.

Здесь мы не требуем, чтобы элементы выборки лежали в $\mathbb{R}$, метрическое пространство $\mathcal{X}$ может быть достаточно общего вида.

7.1 Критерий, основанный на расстоянии Канторовича-Вассерштейна

7.1.1 Общий обзор

Удобное для нас свойство $\mathcal{F}$ – если это определяющий распределение класс, то есть из $\mathbf{E}_P f(X) = \mathbf{E}_Q f(X)$ при всех $f \in \mathcal{F}$ следует условие $P(A) = Q(A)$ при всех A.

Тогда d будет полуметрикой и в случае, если $\widehat{d}$ состоятельна, мы получим состоятельный критерий.

Два примера такого подхода мы уже рассматривали.

Если $\mathcal{X} = \mathbb{R}$, а

$$\mathcal{F} = \{I_{X \le x}, x \in \mathbb{R}\} \bigcup \{I_{X > x}, x \in \mathbb{R}\},$$

то получим критерий Смирнова с характеристикой

$$d(\mathbf{P}, \mathbf{Q}) = |\mathbf{P}(Y \le x) - \mathbf{Q}(Z \le x)|.$$

Задача 14. Убедитесь в этом.

Если $\mathcal{X} = \mathbb{R}$, а

$$\mathcal{F} = \{I_{X \in A}, A \in \mathcal{B}(\mathbb{R})\},$$

то мы получим непригодную для построения критерия характеристику, что уже обсуждалось на первой лекции.

Таким образом, мы оказываемся между двух противоречащих друг другу требований: с одной стороны хотелось бы, чтобы класс $\mathcal{F}$ был как можно шире (тогда критерий будет более чувствителен к различным альтернативам), а с другой стороны – чтобы класс $\mathcal{F}$ был достаточно узким, чтобы статистика не реагировала на сам факт отличия значений элементов выборок.

Для $\mathcal{X} = \mathbb{R}^d$ естественным выбором класса $\mathcal{F}$ является $\mathcal{F}_{Lip}$ – класс всех липшицевых функции с параметром 1:

$$|f(x) - f(y)| \le \rho(x, y),$$

где $\rho(x, y) = \sum_{i=1}^{n} |x_i - y_i|$. С одной стороны, это не даст функциям слишком быстро колебаться, с другой стороны, класс содержит достаточно много функций.

Величина

$$\rho_K(\mathbf{P}, \mathbf{Q}) = \sup_{F \in \mathcal{F}_{Lip}} \mathbf{E}_P f(X) - \mathbf{E}_Q f(X)$$

называется расстоянием Канторовича-Вассерштейна между мерами $\mathbf{P}$, $\mathbf{Q}$.

7.1.2 Другое определение расстояния Канторовича-Вассерштейна

Лемма 5. *Справедливо представление*

$$\rho_K(\mathbf{P}, \mathbf{Q}) = \inf_{Y \sim \mathbf{P}, Z \sim \mathbf{Q}} \mathbf{E}\rho(Y, Z).$$

Иными словами, расстояние Канторовича-Вассерштейна между мерами – это среднее расстояние между величинами, у которых совместное распределение подобрано оптимальным образом по заданным маргинальным.

Доказательство. Будем называть второе выражение 1-расстоянием Вассерштейна и в данном доказательстве обозначать его ρ_W. Расстояния Канторовича и Вассерштейна совпадают, однако, исторически они возникли разными путями.

Тогда ρ_K очевидно не превосходит ρ_W, поскольку

$$\sup_{f\in\mathcal{F}_{Lip}} E_{\mathbf{P}}f(X) - \mathbf{E}_{\mathbf{Q}}f(X) \leq \mathbf{E}||X-Y||$$

при любых $X\sim\mathbf{P}, Y\sim\mathbf{Q}$, откуда правая часть не превосходит и ρ_W. Обратное соотношение заметно сложнее и мы проговорим его поверхностно, не погружаясь в детали оптимизации. Главная идея здесь в том, что для поиска ρ_W нужно составить Лагранжиан

$$L(\mathbf{R},f,g) = \left(\int \rho(y,z)\mathbf{R}(dy,dz) - \int g(y)\left(\int \mathbf{R}(dy,dz) - \mathbf{P}(dy)\right) - \right.$$
$$\left.\int h(z)\left(\int \mathbf{R}(dy,dz) - \mathbf{Q}(dz)\right)\right) = \int(\rho(y,z)-g(y)-h(z))\mathbf{R}(dy,dz) +$$
$$\int g(y)\mathbf{P}(dy) + \int h(z)\mathbf{Q}(dz).$$

где $\mathbf{R}$ – произвольная мера, g,h – ограниченные измеримые функции. Здесь подставленные интегралы соответствуют условиям на то, что маргинальные распределения $\mathbf{R}$ – это $\mathbf{P}$ и $\mathbf{Q}$. Тем самым,

$$\rho_W(\mathbf{P},\mathbf{Q}) = \inf_{\mathbf{R}}\sup_{g,h} L(\mathbf{R},g,h)$$

Здесь мы без каких-либо обоснований воспользуемся так называемой сильной двойственностью и скажем, что он представим в виде

$$\sup_{g,h}\inf_{\mathbf{R}} L(\mathbf{R},g,h).$$

Если при этом $\rho(y,z)-g(y)-h(z)\leq 0$ при каком-то y,z, то сосредоточив меру $\mathbf{R}$, равную n, в этой точке, мы сможем довести наш инфимум до минус бесконечности. Тем самым, можно рассматривать $g(y)+h(z)\leq\rho(y,z)$. Но на этом множестве первое слагаемое неположительно и выгоднее всего положить $\mathbf{R}$ нулевой и получить

$$\rho_W(\mathbf{P},\mathbf{Q}) = \sup_{g,h:g(y)+h(z)\leq\rho(y,z)} (\mathbf{E}_P g(Y) + \mathbf{E}_Q h(Z)).$$

Рассмотрим при фиксированной g

$$\varkappa(y) = \inf_z(\rho(y,z) - h(z)),$$

так как h – ограниченная, то инфимум конечен. При этом при всех y

$$\varkappa(y) \le \rho(y,z) - h(z) \le \rho(y',z) - h(y) + \rho(y,y'),$$

откуда

$$\varkappa(y) \le \varkappa(y') + \rho(y,y').$$

Значит, $\varkappa(\cdot)$ – липшецева функция с параметром 1. При этом, если $g(y)+h(z) \le \rho(x,y)$, то $g(z) \le \varkappa(z)$, $h(z) \le -\varkappa(z)$ откуда

$$\rho_W(\mathbf{P},\mathbf{Q}) \le \mathbf{E}_P\varkappa(Z) - \mathbf{E}_Q\varkappa(Z) \le \rho_K(\mathbf{P},\mathbf{Q}),$$

что и требовалось доказать. □

Отсюда мы получаем следующий критерий – если $\rho_W(\widehat{\mathbf{P}}_n, \widehat{\mathbf{Q}}_m) > C$, то отвергаем гипотезу, а иначе принимаем. Остаются два ключевых вопроса – как найти статистику критерия и какое у нее распределение при верной гипотезе.

7.1.3 Расстояние Вассерштейна между дискретными мерами

Пусть $Y \sim \mathbf{P}$, $Z \sim \mathbf{Q}$ дискретные случайные величины, $(Y,Z) \sim H$. Тогда

$$\rho_W(\mathbf{P},\mathbf{Q}) = \sum_{i=1}^{n}\sum_{j=1}^{m} H(y_i,z_j)\rho(y_i,z_j), \tag{7.1}$$

откуда поиск ρ_W сводится к поиску матрицы $H_{i,j} = H(y_i,z_j)$, минимизирующей (7.1) с условиями $\sum_{i=1}^{n} H_{i,j} = \mathbf{Q}(j)$, $\sum_{j=1}^{m} H_{i,j} = \mathbf{P}(i)$.

При построении статистики критерия, соответственно, в роли $\mathbf{P}$ выступает $\widehat{P}_n$, в роли $\mathbf{Q}$ – $\widehat{Q}_m$, откуда

$$\sum_{i=1}^{n} H_{i,j} = \frac{1}{m}, \quad \sum_{j=1}^{m} H_{i,j} = \frac{1}{n}.$$

Эта задача представляет собой задачу линейного программирования, решение которой реализовано в linprog пакета scipy.optimize в Python и подробно описано в следующем мануале.

Таким образом, мы можем вычислить статистику критерия за полиномиальное (от размера выборок) время.

Уровень значимости при этом может быть определен с помощью перестановочного подхода. Отметим, что скорость сходимости эмпирического расстояния к теоретическому не слишком высока (см. Ramdas и др., 2017), поэтому при сколько-то больших размерностях (больше 2) выгоднее рассматривать другие подходы этого типа.

7.1.4 Одномерный случай

В одномерном случае расстояние приобретает простой явный вид.

Лемма 6. *Пусть* $\mathbf{P}$, $\mathbf{Q}$ – *меры на* $\mathbb{R}$, $\rho(x,y) = |x-y|$. *Тогда справедливо соотношение*

$$\rho_W(\mathbf{P},\mathbf{Q}) = \int_0^1 |F^{-1}(x) - G^{-1}(x)|dx.$$

В указанном построении H – это распределение с маргинальными распределениями F и G, значит, $H(x,y) = C(F^{-1}(x), G^{-1}(y))$, где C – некоторая копула. Однако, глядя на ответ, мы видим, что экстремум достигается на копуле $C(x,y) = \min(x,y)$. Иными словами, оптимальное совместное распределение $H(x)$ соответствует паре $(F^{-1}(R), G^{-1}(R))$, где R – равномерная величина.

Итак, в одномерном случае мы оцениваем характеристику

$$\int_0^1 |F^{-1}(x) - G^{-1}(x)|dx$$

$$\int_0^1 |\widehat{F}_n^{-1}(x) - \widehat{G}_m^{-1}(x)|dx.$$

Полученный критерий оказывается достаточно неплохим, хотя и зависит от распределения выборки.

7.2 Гильбертово пространство, воспроизводящее ядро

Другой подход к выбору $\mathcal{F}$ работает и для более общих пространств. Основная идея заключается в том, чтобы найти такое гильбертово пространство $\mathcal{H}$, что

$$\mathbf{E}_P f(X) = \langle f, g_P \rangle_{\mathcal{H}},\ f \in \mathcal{F},$$

где g_P – некоторый элемент $\mathcal{H}$, соответствующий мере $\mathbf{P}$, а $\mathcal{F}$ – единичный шар в $\mathcal{H}$. Тогда

$$\mathbf{E}_P f(X) - \mathbf{E}_Q f(X) = \langle f, g_P - g_Q \rangle.$$

Теперь задача приобрела прозрачный вид – найти элемент единичного шара, который имеет максимальную проекцию на $g_P - g_Q$. Разумеется, это $f = (g_P - g_Q)/||g_P - g_Q||$.

Глава 8

О построении ядра

8.0.1 Конструкция RKHS

Остается построить указанное пространство $\mathcal{H}$. Фиксируем функцию $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, которую мы назовем ядром. Рассмотрим подпространство множества функций из $\mathcal{X} \to \mathbb{R}$ вида

$$\mathcal{H}_0 = \left\{ f(x) = \sum_{i=1}^{k} a_i k(x, x_i),\ x_i \in \mathcal{X},\ a_i \in \mathbb{R} \right\}.$$

На этом пространстве введем скалярное пространство

$$\langle f, g \rangle = \sum_{i=1}^{m} \sum_{j=1}^{l} a_i b_j k(x_i, x_j),$$

где

$$f(x) = \sum_{i=1}^{m} a_i k(x, x_i), \quad g(x) = \sum_{j=1}^{l} b_j k(x, x_j)$$

Задача 15. Показать, что если k – симметричная неотрицательно определенная функция, то это действительно скалярное произведение.

Теперь пополним пространство $\mathcal{H}_0$ и получим пространство $\mathcal{H}$ с соответствующим скалярным произведением.

Данное пространство обладает важным свойством, которое принято называть *воспроизводящим свойством ядра*. Полагая $k_y(x) = k(x, y)$, получаем при $f \in \mathcal{H}_0$

$$\langle f, k_y \rangle = \left\langle \sum_{i=1}^{k} a_i k(x, x_i), k_y(x) \right\rangle = \sum_{i=1}^{k} a_i k(x_i, y) = f(y).$$

При пополнении данное свойство сохраняется, откуда верно общее свойство:

$$\langle f, k_y \rangle = f(y).$$

Это удобное свойство, позволяющее восстанавливать значение функции в точках, оперируя с ней как с элементом $\mathcal{H}$.

Отметим также, что

$$\langle \sum_{i=1}^{k} a_i k(x, x_i), \sum_{i=1}^{m} b_i k(x, y_i) \rangle = \sum_{i=1}^{k} \sum_{j=1}^{m} a_i b_j k(x_i, y_j) = \vec{a}^T K \vec{b},$$

где $K_{i,j} = (k(x_i, y_j))$. В частности,

$$||k_x||^2 = k(x, x).$$

Следующий наш шаг – построение элемента g_P, соответствующего мере $\mathbf{P}$ на $\mathcal{X}$, как это описано выше. Для этого рассмотрим функционал $h : \mathcal{H} \to \mathbb{R}$ вида

$$h_P(f) = \int_{\mathcal{X}} f(x) \mathbf{P}(dx).$$

Это линейный функционал на $\mathcal{H}$. Оценим его норму, используя соотношение

$$\left|\left| \int_{\mathcal{X}} f(x) \mathbf{P}(dx) \right|\right| \leq \int_{\mathcal{X}} |f(x)| \mathbf{P}(dx) = \int_{\mathcal{X}} |\langle f, k_x \rangle| \mathbf{P}(dx) \leq ||f||_{\mathcal{H}} \int_{\mathcal{X}} ||J_x||_{\mathcal{H}} \mathbf{P}(dx),$$

откуда в силу воспроизводящего свойства ядра

$$||h_P|| \leq \int_{\mathcal{X}} \sqrt{k(x, x)} \mathbf{P}(dx).$$

Предположим, что $\sqrt{k(x, x)} \in L^1(\mathbf{P})$, тогда наш функционал линеен и ограничен. Значит, в силу теоремы Риса о представлении

$$h_P(f) = \langle g_P, f \rangle$$

для некоторого элемента $g_P \in \mathcal{H}$.

Для него справедливо соотношение

$$\sup_{f \in \mathcal{F}} \mathbf{E}_P f(X) - \mathbf{E}_Q f(X) = ||g_P - g_Q||_{\mathcal{H}}.$$

При этом в силу воспроизводящего свойства ядра и определения элементов g_P, g_Q:

$$\langle g_P, g_Q\rangle_{\mathcal{H}} = \int_{\mathcal{X}} g_Q(x)\mathbf{P}(dx) = \int_{\mathcal{X}} \langle g_Q, k_x\rangle \mathbf{P}(dx) = \int_{\mathcal{X}^2} k(x,y)\mathbf{P}(dx)\mathbf{Q}(dy),$$

откуда

$$||g_P - g_Q||^2_{\mathcal{H}} = \int_{\mathcal{X}^2} k(x,y)(\mathbf{P}(dx) - \mathbf{Q}(dx))(\mathbf{P}(dy) - \mathbf{Q}(dy)),$$

где под указанным интегралом мы понимаем

$$\mathbf{E}_{P\times P} k(X_1, X_2) + \mathbf{E}_{Q\times Q} k(Y_1, Y_2) - 2\mathbf{E}_{P\times Q} k(X_1, Y_1).$$

8.0.2 Характеристика и статистика критерия MMD

Резюмируем проделанные нами построения (по существу мы следуем работе Gretton, 2012). Мы рассматриваем неотрицательно определенную симметричную функцию $k : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$, называемую ядром. Мы вводим характеристику

$$d(\mathbf{P}, \mathbf{Q}) = \sup_{f\in\mathcal{F}} \mathbf{E}_P f(X) - \mathbf{E}_Q f(X) = \sqrt{\int_{\mathcal{X}^2} k(x,y)(\mathbf{P}(dx) - \mathbf{Q}(dx))(\mathbf{P}(dy) - \mathbf{Q}(dy))},$$

которую оцениваем статистикой

$$MMD_b = \sqrt{\int_{\mathcal{X}^2} k(x,y)(\widehat{\mathbf{P}}_n(dx) - \widehat{\mathbf{Q}}_m(dx))(\widehat{\mathbf{P}}_n(dy) - \widehat{\mathbf{Q}}_m(dy))} =$$

$$\sqrt{\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n K(Y_i, Y_j) + \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(Z_i, Z_j) - \frac{2}{nm}\sum_{i=1}^m\sum_{j=1}^n k(Y_i, Z_j)}.$$

Можно взамен использовать несмещенную оценку квадрата характеристики

$$MMD_u^2 = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j=1}^n k(Y_i, Y_j) + \frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j=1}^m k(Z_i, Z_j) - \frac{2}{nm}\sum_{i=1}^m\sum_{j=1}^n k(Y_i, Z_j).$$

Асимптотически статистики близкие, но вторая зачастую удобнее.

Данный критерий уже можно применять, используя перестановочный подход, однако, мы проведем более глубокое исследование свойств нашей статистики.

Задача 16. Изменить коэффициенты перед суммами в оценке MMD так, чтобы MMD^2 стало несмещенной оценкой $d(\mathbf{P}, \mathbf{Q})$.

8.0.3 Свойства характеристики критерия

Напомним следующее определение.

Определение 10. Метрическое пространство $\mathcal{X}$ хаусдорфово, если для любых $x \neq y$ из $\mathcal{X}$ найдутся такие положительные δ_1, δ_2, что $U_{\delta_1}(x) \cap U_{\delta_2}(y) = \emptyset$.

Лемма 7. *Пусть $\mathcal{X}$ – хаусдорфово компактное метрическое пространство, на котором рассматривается $\mathcal{F}$ – минимальная сигма-алгебра, содержащая все компактные множества, функция K непрерывна, а $\mathcal{H}$ плотно в $C(\mathcal{X})$ по равномерной норме. Тогда $d(\mathbf{P}, \mathbf{Q}) = 0$ тогда и только тогда, когда $\mathbf{P} = \mathbf{Q}$.*

Доказательство. В одну сторону утверждение очевидно, докажем в другую.

Пусть $f \in C(\mathcal{X})$, тогда (в силу указанной плотности) при любом положительном ε найдется $f_\varepsilon \in \mathcal{H}$:

$$\sup_{x \in \mathcal{X}} |f(x) - f_\varepsilon(x)| < \varepsilon.$$

Тогда

$$|\mathbf{E}_P f(Y) - \mathbf{E}_Q f(Z)| \leq 2\varepsilon + |\mathbf{E}_P f_\varepsilon(Y) - \mathbf{E}_Q f_\varepsilon(Z)|.$$

При этом последнее слагаемое имеет вид

$$|\langle g_P - g_Q, f_\varepsilon \rangle|$$

и в силу условия равно нулю. Остается заметить, что если $\mathbf{E}_P f(Y) = \mathbf{E}_Q f(Z)$ при всех $f \in C(\mathcal{X})$, то $\mathbf{P} = \mathbf{Q}$, поскольку в компактном метрическом хаусдорфоровом пространстве непрерывные функции плотны в L^1 (см., например, рассуждения по ссылке), а значит

$$\mathbf{E}_P I_A = \mathbf{E}_Q I_A$$

при любом компактном A, а значит и всех A из $\mathcal{F}$. □

Тем самым, в компактных метрических пространствах (или пространствах, представимых в виде счетного объединения расширяющихся компактных) d – полуметрика.

8.0.4 Свойства статистики критерия

Лемма 8. *При выполнении условий $\mathbf{D}_P k(X_1, X_1) < +\infty$, $\mathbf{D}_Q k(X_1, X_1) < +\infty$ статистика MMD является состоятельной оценкой $d(\mathbf{P}, \mathbf{Q})$ при $n, m \to \infty$.*

Доказательство. Достаточно доказать, что MMD^2 сходится к $d^2(\mathbf{P}, \mathbf{Q})$, поскольку непрерывные функции сохраняют сходимость по вероятности. При этом

$$\mathbf{D}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(X_i, X_j) = \frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_{p=1}^{n}\operatorname{cov}(k(X_i, X_j), k(X_l, X_p)).$$

Все слагаемые в которых i, j, l, p – четыре различных индекса, равны 0 в силу независимости рассматриваемых величин, а остальных слагаемых не больше $6n^3$, причем каждое из них ограничено величиной

$$\max(\mathbf{D}_P k(X_1, X_2), \mathbf{D}_P k(X_1, X_2))$$

в силу неравенства Коши-Буняковского. При этом в силу неотрицательной определенности ядра $2|k(x, y)| \leq k(x, x) + k(y, y)$. Аналогичные выкладки действуют для двух остальных слагаемых в представлении MMD. Значит,

$$\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(Y_i, Y_j) - \mathbf{E}_P k(Y_1, Y_2)\right) \overset{L^2}{\to} 0$$

при $n \to \infty$, откуда следует сходимость по вероятности. □

Глава 9

Сходимость статистики MMD^2 и ее применения

9.1 Предельное распределение для MMD^2

9.1.1 Основная теорема

В следующей теореме будем рассматривать несмещенную оценку

$$MMD_u^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} k(Y_i, Y_j) + \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i} k(Z_i, Z_j) - \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} k(Y_i, Z_j).$$

Теорема 4. *Пусть ядро k интегрируемо в $L^2(X\times X, \mathbf{P}\times\mathbf{P})$, X – сепарабельное метрическое пространство. Тогда при верной гипотезе и $n/(n+m)\to p\in(0,1)$ выполнено соотношение*

$$(n+m)MMD_u^2 \xrightarrow{d} W,$$

где W – случайная величина, распределение которой задается соотношением (10.1).

Доказательство леммы разобьем на несколько частей.

9.1.2 Альтернативное представление статистики критерия

Без ограничения общности можно считать, что $\mathbf{E}_P k(x, Y) = 0$ при любом x, поскольку при переходе от $k(x, y)$ к

$$k(x,y) - \mathbf{E}_P k(x,Y) - \mathbf{E}_P k(y,Z) + \mathbf{E}_{P\times P} k(Y,Z)$$

статистика MMD_u^2 не меняет своей формы.

Задача 17. Убедитесь в этом.

Однако, стоит отметить, что полученное выражение для W выражается в терминах именно такого, центрированного, ядра.

Формула для статистики MMD_u^2 неудобна – она использует суммы попарных произведений наших величин, с чем мы плохо умеем работать. Куда удобнее было бы оставить только сумму или ее квадрат. Это позволяет сделать следующая лемма.

Лемма 9. *В предположениях основной теоремы выполнено соотношение*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}k(Y_i,Y_j)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{\infty}\lambda_k\psi_k(Y_i)\psi_k(Y_j),$$

где λ_i – некоторые константы, ψ – некоторые функции.

Доказательство. Рассмотрим оператор $A : L^2(\mathbf{P}) \to L^2(\mathbf{P})$, заданный соотношением

$$Af(x)=\int_{\mathcal{X}}k(x,y)f(y)\mathbf{P}(dy).$$

Во-первых, убедимся, что результат действия оператора лежит в $L^2(\mathbf{P})$. Это вытекает из того, что

$$||Af||^2_{L^2(\mathbf{P})}=\int_{\mathcal{X}}\left(\int_{\mathcal{X}}k(x,y)f(y)\mathbf{P}(dy)\right)^2\mathbf{P}(dx)\leq\int_{\mathcal{X}}||k(x,\cdot)||^2_{L^2(P)}||f||^2_{L^2(P)}\mathbf{P}(dx)\leq$$
$$\int_{\mathcal{X}\times\mathcal{X}}k(x,y)^2\mathbf{P}(dx)\mathbf{P}(dy)||f||^2_{L^2(P)}=||k||^2_{L^2(\mathbf{P}\times\mathbf{P})}||f||^2_{L^2(\mathbf{P})}.$$

Во-вторых, мы видим, что оператор ограниченный и его норма не превосходит $||k||_{L^2(\mathbf{P}\times\mathbf{P})}$. В-третьих, утверждается, что оператор A компактный самосопряженный оператор.

Этот факт можно найти в Reed и Simon, 1980, Theorem VI.23. Приведу здесь линию доказательства, однако, не буду требовать от вас досконального знания указанного материала. Рассмотрим в $L^2(\mathbf{P})$ ортонормированный базис $\{\phi_k\}_{k=1}^{\infty}$ (он существует, поскольку это сепарабельное гильбертово пространство, сепарабельность обоснована в Reed и Simon, 1980, IV, задача 43). Тогда $\{\phi_k(x)\phi_l(y)\}_{k,l=1}^{\infty}$

образуют базис в $L^2(\mathbf{P}\times\mathbf{P})$ (Reed и Simon, 1980, II.4 Proposition 2 и последующее замечание для данного случая). Значит,

$$k(x,y)=\sum_{k,l=1}^{\infty}a_{k,l}\phi_k(x)\overline{\phi_l(y)}$$

при некоторых коэффициентах $a_{k,l}$. Рассмотрим

$$k_N(x,y)=\sum_{k,l=1}^{N}a_{k,l}\phi_k(x)\overline{\phi_l(y)}.$$

Тогда

$$A_Ng=\int_{\mathcal{X}}k_N(x,y)g(y)\mathbf{P}(dy)$$

является конечномерным оператором (т.е. образ данного оператора является конечномерным пространством). При этом

$$||A-A_N||\leq||k-k_N||_{L^2(\mathbf{P}\times\mathbf{P})}\to 0,$$

где первая норма рассматривается в операторном смысле. Следовательно, A является пределом (по норме) конечномерных операторов, а значит является компактным (Reed и Simon, 1980, Theorem VI.12).

В четвертых,

$$Af=\sum_{i=1}^{\infty}\lambda_i\langle f,\psi_i\rangle_{L^2(P)}\psi_i,$$

где ψ_i – собственные функции оператора A, образующие ортонормированный базис в $L^2(\mathbf{P})$, а λ_i – соответствующие собственные значения. Сам факт, что собственные функции компактного самосопряженного оператора A образуют ортогональный базис, известен как теорема Гильберта-Шмидта (Reed и Simon, 1980, VI.16). Указанное представление есть каноническое представление компактного самосопряженного оператора (Reed и Simon, 1980, VI.17).

В-пятых, ряд $\sum_{i=1}^{\infty}\lambda_i^2$ сходится к $||k||_{L^2(\mathbf{P}\times\mathbf{P})}$ в силу равенства Парсеваля (Reed и Simon, 1980, теорема VI.22).

Поскольку $\{\psi_i\}$ – ортонормированный базис в $L^2(\mathbf{P})$, то

$$\int_{\mathcal{X}}k(x,y)\psi_k(y)\mathbf{P}(dy)=\sum_{i=1}^{\infty}\lambda_i\langle\psi_k,\psi_i\rangle\psi_i(x)=\lambda_k\psi_k(x).$$

Эта формула похожа на ту, которую мы получали пока доказывали компактность оператора, только базис ψ это и еще и собственный для A базис (к тому же вещественный), а ϕ был произвольным.

Значит, $k(x,y)$ как функция y имеет коэффициенты $\lambda_k\psi_k(x)$ разложения по базису $\psi_k(y)$, откуда

$$k(x,y) = \sum_{k=1}^{\infty} \lambda_k \psi_k(x)\psi_k(y).$$

Следовательно,

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} k(Y_i,Y_j) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{\infty} \lambda_k\psi_k(Y_i)\psi_k(Y_j).$$

□

9.1.3 Структура $\psi_k(Y_i)$.

Следующая лемма описывает структуру величин $\psi_k(Y_i)$.

Лемма 10. *Величины $\psi_k(Y_1)$ имеют нулевое среднее и единичную дисперсию при каждом k и некоррелированы при различных k.*

Доказательство. Исследуем случайные величины $\psi_k(Y_i)$. Во-первых,

$$\lambda_k \mathbf{E}_P \psi_k(Y_i) = \lambda_k \int_{\mathcal{X}} \psi_k(x)\mathbf{P}(dx) = \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)\psi_k(y)\mathbf{P}(dy)\mathbf{P}(dx) = 0,$$

в силу условия $\mathbf{E}_P k(Y,x) = 0$ при всех x (интегралы можно поменять в силу теоремы Фубини). Более того,

$$\mathbf{E}_P \psi_k^2(Y_i) = \int_{\mathcal{X}} \psi_k(x)^2 \mathbf{P}(dx) = 1$$

в силу ортонормированности. Кроме того,

$$\mathbf{E}_P \psi_k(Y_i)\psi_l(Y_i) = \int_{\mathcal{X}} \psi_k(x)\psi_l(x)\mathbf{P}(dx) = 0$$

в силу ортогональности. □

9.1.4 О предельном распределении ряда

Если бы сумма по k в нашем представлении была бы конечной, то мы без труда нашли бы ее совместное распределение:

$$T_{N,n} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{N}\lambda_k\psi_k(Y_i)\psi_k(Y_j) = \frac{1}{n}\sum_{k=1}^{N}\lambda_k\left(\left(\sum_{i=1}^{n}\psi_k(Y_i)\right)^2 - \sum_{i=1}^{n}\psi_k(Y_i)^2\right),$$

где вектор

$$\left(\sum_{i=1}^{n}\psi_1(Y_i),\ldots,\sum_{i=1}^{n}\psi_N(Y_i),\sum_{i=1}^{n}(\psi_1^2(Y_i)-1),\ldots\sum_{i=1}^{n}(\psi_N^2(Y_i)-1)\right)$$

при домножении на $\sqrt{n}$ сходится к нормальному вектору в силу ЦПТ. Значит, мы без проблем найдем предельное распределение. Увы, у нас ряд бесконечный.

Лемма 11.

$$T_{N,n} \xrightarrow{d} \sum_{k=1}^{\infty}\lambda_k(U_k^2-1),$$

где U_k – н.о.р. $\mathcal{N}(0,1)$ величины.

Доказательство. При этом п.н. при $i \neq j$

$$\sum_{k=1}^{N}\lambda_k\psi_k(Y_i)\psi_k(Y_j) \to \sum_{k=1}^{\infty}\lambda_k\psi_k(Y_i)\psi_k(Y_j),\ N\to\infty,$$

поскольку последовательность в левой части образует мартингал, причем

$$\mathbf{E}\left|\sum_{k=1}^{N}\lambda_k\psi_k(Y_i)\psi_k(Y_j)\right| \leq \sqrt{\mathbf{E}\left(\sum_{k=1}^{N}\lambda_k\psi_k(Y_i)\psi_k(Y_j)\right)^2} = \sqrt{\sum_{k=1}^{N}\lambda_k^2(\mathbf{D}\psi_k(Y_1))^2} \leq \sqrt{\sum_{k=1}^{\infty}\lambda_k^2}$$

В силу теоремы о сходимости мартингалов частичные суммы имеют предел п.н. Следовательно, п.н. верно соотношение

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{\infty}\lambda_k\psi_k(Y_i)\psi_k(Y_j) = \frac{1}{n}\sum_{k=1}^{\infty}\lambda_k\left(\left(\sum_{i=1}^{n}\psi_k(Y_i)\right)^2 - \sum_{i=1}^{n}\psi_k(Y_i)^2\right).$$

При этом

$$\mathbf{E}_P \left(\frac{1}{n} \sum_{k=N}^{M} \lambda_k \sum_{i \neq j} \psi_k(Y_i)\psi_k(Y_j) \right)^2 = \tag{9.1}$$

$$\frac{1}{n^2} \sum_{k,k' \in [N,M]} \sum_{i \neq j} \sum_{i' \neq j'} \lambda_k \lambda'_k \mathbf{E}_P \psi_k(Y_i)\psi_k(Y_j)\psi_{k'}(Y_{i'})\psi_{k'}(Y_{j'}).$$

При $k \neq k'$ слагаемые в рассматриваемой сумме равны нулю, при $k = k'$ и наличии уникального индекса в $\{i, i', j, j'\}$ также равны нулю. Значит, останется лишь два случай с ненулевыми слагаемыми: $i = i', j = j'$ и $i = j', j = i'$, откуда правая часть (9.1) есть

$$\frac{2(n-1)}{n} \sum_{k=N}^{M} \lambda_k^2.$$

Положим

$$T_{N,n} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{N} \lambda_k \psi_k(Y_i)\psi_k(Y_j), \quad T_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{\infty} \lambda_k \psi_k(Y_i)\psi_k(Y_j),$$

тогда

$$\mathbf{E}(T_n - T_{N,n})^2 \leq 2 \sum_{k=N+1}^{\infty} \lambda_k^2$$

Тогда в силу неравенств

$$\left| e^{itx} - e^{ity} \right| \leq \left| e^{it(x-y)} - 1 \right| \leq |t||x-y|$$

имеем

$$|\mathbf{E} \exp(itT_n) - \mathbf{E} \exp(itT_{N,n})| \leq |t| \mathbf{E}|T_n - T_{N,n}| \leq |t| \sqrt{\mathbf{E}(T_n - T_{N,n})^2} \leq |t| \sqrt{2 \sum_{k=N+1}^{\infty} \lambda_k^2} < \varepsilon$$

при любом $\varepsilon > 0$ и достаточно большом N. При этом

$$T_N = \frac{1}{n} \sum_{k=1}^{N} \lambda_k \left(\sum_{i=1}^{n} \psi_k(Y_i) \right)^2 - \frac{1}{n} \sum_{k=1}^{N} \lambda_k \sum_{i=1}^{n} \psi_k(Y_i)^2.$$

В силу ЦПТ

$$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_1(Y_i), \ldots, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_N(Y_i) \right) \xrightarrow{d} (U_1, \ldots, U_N),$$

где U_i – н.о.р. $\mathcal{N}(0,1)$ величины (здесь мы воспользовались тем, что $\psi_i(Y_1)$, $\psi_j(Y_1)$ не коррелируют при $i \neq j$). Следовательно,

$$\frac{1}{n}\sum_{k=1}^{N}\lambda_k\left(\sum_{i=1}^{n}\psi_k(Y_i)\right)^2 \xrightarrow{d} \sum_{k=1}^{N}\lambda_k U_k^2, \quad \frac{1}{n}\sum_{k=1}^{N}\lambda_k\sum_{i=1}^{n}\psi_k(Y_i)^2 \xrightarrow{d} \sum_{k=1}^{N}\lambda_k,$$

где во втором случае мы воспользовались ЗБЧ. Следовательно,

$$\lim_{n\to\infty}\mathbf{E}\exp(itT_{N,n}) \to \mathbf{E}\exp\left(it\sum_{k=1}^{N}\lambda_k(U_k^2-1)\right).$$

Наконец, при всех достаточно больших N

$$\left|\mathbf{E}\exp\left(it\sum_{k=1}^{N}\lambda_k(U_k^2-1)\right) - \mathbf{E}\exp\left(it\sum_{k=1}^{\infty}\lambda_k(U_k^2-1)\right)\right| \leq \varepsilon.$$

Здесь важно, что ряд

$$T = \sum_{k=1}^{\infty}\lambda_k(U_k^2-1)$$

сходится п.н. в силу все тех же рассуждений на основе теоремы о сходимости мартингалов.

Задача 18. Докажите это.

Таким образом,

$$|\mathbf{E}\exp(itT_n) - \mathbf{E}\exp(itT)| \leq 3\varepsilon$$

при всех достаточно больших n, откуда и вытекает требуемая сходимость. □

Окончание доказательства основной теоремы проведем на следующей лекции.

Глава 10

Завершение доказательства и некоторые замечания

10.1 Теорема о предельном распределении MMD_u^2

10.1.1 Завершение доказательства основной теоремы

Доказательство. Аналогичным образом,

$$\frac{1}{m}\sum_{i=1}^{m}\sum_{j\neq i}\sum_{k=1}^{\infty}\lambda_k\psi_k(Z_i)\psi_k(Z_j) \xrightarrow{d} \sum_{k=1}^{\infty}\lambda_k(V_j^2-1).$$

Наконец,

$$\frac{1}{\sqrt{nm}}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{\infty}\lambda_k\psi_k(Y_i)\psi_k(Z_j) \xrightarrow{d} \sum_{k=1}^{\infty}\lambda_k U_i V_j,$$

где U_i, V_j – н.о.р. стандартные нормальные величины. Однако, это еще не означает, что

$$\frac{nm}{n+m}MMD_u^2 \xrightarrow{d} p(1-p)\sum_{k=1}^{\infty}\lambda_k\left(\frac{U_k^2-1}{p}+\frac{V_k^2-1}{1-p}-\frac{2U_kV_k}{\sqrt{p(1-p)}}\right) =: W, \quad (10.1)$$

где U_i, V_i – н.о.р. стандартные нормальные величины. Проблема в том, что из сходимости по распределению слагаемых не следует сходимость по распределению их суммы.

Задача 19. Докажите требуемое утверждение самостоятельно, используя то, что совместная характеристическая функция наших трех слагаемых сходится к

совместной х.ф. трех слагаемых в представлении W. Используйте тот же трюк, с ограничением сразу всех трех сумм по k до конечных сумм.

Отметим, что можно представить W в виде

$$\sum_{k=1}^{\infty} \lambda_k \left((\sqrt{1-p}U_k - \sqrt{p}V_k)^2 - 1 \right) = \sum_{k=1}^{\infty} \lambda_k T_k^2,$$

где $T_k \sim \mathcal{N}(0,1)$ независимы.

В случае, если сходится ряд $\sum \lambda_i$, аналогичное доказательство приводит к

$$\frac{nm}{n+m} MMD_b^2 \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k T_k^2.$$

Однако, условие сходимости ряда более сильное, чем ряда из квадратов. □

С одной стороны, мы доказали вожделенную предельную теорему, с другой стороны, не вполне понятно как ее использовать – предельное распределение зависит от λ_k, являющихся характеристикой оператора $\mathbf{E}_P k(x, X)$, зависящего от неизвестного распределения $\mathbf{P}$. Приведем три возможных применения данной теоремы для построения критерия.

1. Ускоренный перестановочный критерий.
   Проблема перестановочного критерия в том, что для проведения K перестановок нам потребуется $O(mnK)$ операций. Скажем, при $K = 1000$, $m = 1000$, $n = 1000$ это потребует 10^9 операций. Однако, имея доказанную выше теорему, можно несколько схитрить. Возьмем при каждом i из нашей исходной выборки m', n' наблюдений, где $m'/n' = m/n$ (или, покрайней мере, величины близки). Тогда перемешаем наши наблюдения и посчитаем статистику $(m'+n')(MMD_u^2)_i$ по нашим наблюдениям, $i \leq N$. Тогда величины $(m'+n')(MMD_u^2)_i$ будут при верной гипотезе распределены практически также как и $(m+n)MMD_u^2$. Исходя из этого, мы можем ориентироваться на долю $(m_i+n_i)(MMD_u^2)_i$, расположенных правее $(m+n)MMD_u^2$ и использовать ее в качестве p-value. При этом нужно уже $O(m'n'N+mn)$ операций, что может быть значительно быстрее. Конечно, при этом m', n' все равно должны быть достаточно большими, чтобы мы были близки к предельному распределению.

2. Метод, использующий оценку с.з. оператора A_K.

Второй естественный вариант - оценить с.з. оператора A_K. Оказывается, что при верной гипотезе справедлив следующий тезис:

$$\sum_{k=1}^{n+m} \widehat{\lambda}_k U_k^2 \xrightarrow{d} \sum_{k=1}^{\infty} \lambda_k U_k^2,\ n+m \to \infty,$$

где $\widehat{\lambda}_k$ – с.з. матрицы $M = LKL$, где

$$K = (k(x_i, x_j), i, j \leq n+m), \quad L = E - \frac{1}{n+m}(1,1,\ldots,1)^T(1,1\ldots,1),$$

где $\{x_i,\ i \leq n+m\}$ – объединенная выборка.

К сожалению, это требует конечности ряда $\sum_k \sqrt{\lambda_k}$, что не очень понятно как проверять (впрочем, работает как минимум для конечномерных операторов A).

Таким образом, мы можем найти с.з. матрицы M, эмпирически оценить (сгенерировав достаточное число величин) квантили распределения

$$\sum_{i=1}^{n+m} \widehat{\lambda}_k U_k^2,$$

а затем сравнить наше значение статистики с квантилью уровня $1-\alpha$.

3. Метод кривых Пирсона.

   Другим популярным методом является использование так называемых кривых Пирсона. Дело в том, что существует так называемое семейство Pearson IV распределений, задающихся 4 параметрами, в котором распределение полностью определяется первыми 4 моментами. При этом распределения конечных взвешенных сумм хи-квадрат распределения аппроксимируются бесконечными (Solomon и Stephens, 1977). У нас, увы, сумма бесконечная, однако, она аппроксимируется конечной. Вот и получается, что мы можем вычислить первые 4 момента и аппроксимировать соответствующей пирсоновской кривой. При этом первые два момента влияют на распределение очевидным образом: $X = \sigma Y + a$, где $\mathbf{E}X = a$, $\mathbf{E}Y = 0$, $\mathbf{D}X = \sigma^2$, $\mathbf{D}Y = 1$, причем Y также имеет распределение Пирсона. Поэтому используют специальные таблицы квантилей наших распределений, соответствующих заданным асимметрии $\alpha_3 = \mathbf{E}Y^3$ и эксцессу $\alpha_4 = \mathbf{E}Y^4$.

При этом

$$\mathbf{E}MMD_u^2 = 0, \quad \mathbf{E}(MMD_u^2)^2 = \frac{2(m+n-2)(m+n-1)}{m(m-1)n(n-1)}\mathbf{E}_{P\times P}k(X_1,X_2)^2,$$

$$\mathbf{E}(MMD_u^2)^3 = \frac{8(m+n)^3}{n^3m^3}\mathbf{E}_{P\times P\times P}k(X_1,X_2)k(X_2,X_3)k(X_3,X_1) + O\left(\frac{1}{m^3}+\frac{1}{n^3}\right).$$

Задача 20. Докажите это.

Мы можем оценить

$$\mathbf{E}_{P\times P}k(X_1,X_2)^2 \approx \overline{k(X_1,X_2)^2},$$
$$\mathbf{E}_{P\times P\times P}k(X_1,X_2)k(X_2,X_3)k(X_3,X_1) \approx \overline{k(X_1,X_2)k(X_2,X_3)k(X_3,X_1)},$$

где под соответствующими средними мы подразумеваем средние по всем парам. К сожалению, данная формула использует ядро

$$\widetilde{k} = k(x,y) - \mathbf{E}_P k(x,X) - \mathbf{E}_P k(y,X) + \mathbf{E}_{P\times P}k(X,Y),$$

которое зависит от неизвестной нам $\mathbf{P}$. Вместо этого используют

$$\widehat{k}(x_i,x_j) = k(x_i,x_j) - \frac{1}{n+m}\sum_{l=1}^{n+m}k(x_i,x_l) - \frac{1}{n+m}\sum_{l=1}^{n+m}k(x_j,x_l) + \frac{1}{(n+m)^2}\sum_{l,p}k(x_l,x_p) = LKL = M,$$

где матрица M была введена выше. Можно показать, что соответствующая оценка состоятельна.

Вместо эксцесса, как правило, используют нижнюю оценку $\alpha_4 \geq (\alpha_3)^2 + 1$, поскольку его оценивание требует $O(m^4+n^4)$ операций.

10.1.2 Виды ядер

Для ядра сгодится а) любая симметричная неотрицательно определенная функцию б) любая функция, представимая в виде $k(x,y) = \langle \phi_x, \phi_y \rangle_{\mathcal{H}}$. Это эквивалентные условия, но удобными для использования бывают оба.

1. Если k – ядро, то αk тоже ядро, $\alpha > 0$.

   Действительно, симметричность очевидна, а неотрицательная определенность следует из

   $$\sum_{i,j}\alpha k(x_i,x_j)t_it_j = \alpha\sum_{i,j}\alpha k(x_i,x_j)t_it_j \geq 0.$$

2. Если k_1, k_2 ядра, то $k_1 + k_2$ также ядро.
3. Если $f : Y \to X$, а k – ядро на $X \times X$, то $\widetilde{k}(x, y) = k(f(x), f(y))$.
4. Если k_1 ядро на $X \times X$, k_2 на $Y \times Y$, то $k((x, y), (x', y')) = k_1(x, x')k_2(y, y')$ – также ядро.
5. Если k_1 ядро на $X \times X$, k_2 на $X \times X$, то $k(x, x') = k_1(x, x')k_2(x, x')$ – также ядро.
6. Пусть $\phi(x) = \sum a_n x^n$, $a_n \geq 0$. Показать, что $\phi(k(x, y))$ также ядро.

Задача 21. Докажите свойства 2)-3) с помощью неотрицательной определенности, в 4), 5) докажите неотрицательную определенность с помощью представления через ϕ, используя базисное представление, 6) используя прошлые свойства.

Популярными видами ядер в $\mathbb{R}^k$ являются

1. Полиномиальное $(c + \langle x, y\rangle)^d$, $d \in \mathbb{N}$, $c \geq 0$.
2. Гауссовское $\exp(-\sigma^{-2}||x - y||^2)$.
3. Лапласово (Абелево) ядро $\exp(-\alpha||x - y||)$.

О другом способе строить ядра поговорим позднее.

Параметры σ, α, d приходится выбирать исходя из успешности работы критерия (с этим стоит быть осторожным и, например, использовать обучающую и тестовую выборку или кросс-валидацию) или же их из эвристических соображений оценивают по выборке. Например, σ в одномерном гауссовском ядре часто берут равным медиане разниц x_i и x_j.

Глава 11

HSIC, DCov и Energy Test

11.1 HSIC – критерий независимости

11.1.1 Подход к проверке независимости на основе RKHS

Как мы описывали прежде, для проверки независимости компонент выборки (Y, Z), $Y \in \mathcal{Y}$, $Z \in \mathcal{Z}$, мы можем приспособить наш критерий путем использования характеристики

$$d(\mathbf{R}, \mathbf{P} \times \mathbf{Q}) = \mathcal{V}^2 =$$

$$\int_{(\mathcal{Y}\times\mathcal{Z})^2} k((y,z),(y',z'))(\mathbf{R}(dy,dz) - \mathbf{P}(dy)\mathbf{Q}(dz))(\mathbf{R}(dy',dz') - \mathbf{P}(dy')\mathbf{Q}(dz')),$$

где k – ядро на $(Y \times Z) \times (Y \times Z)$. Как правило, используют $k((y,z),(y',z')) = k_Y(y,y')k_Z(z,z')$, где k_Y, k_Z – ядра на $\mathcal{Y}$, $\mathcal{Z}$. Соответствующий критерий называют HSIC – Hilbert-Schmidt Independence Criterion.

При этом статистика приобретает вид

$$HSIC^2 = \frac{1}{n^2}\sum_{i,j} k_Y(Y_i,Y_j)k_Z(Z_i,Z_j) - \frac{2}{n^3}\sum_{i,j,l} k_Y(Y_i,Y_j)k_Z(Z_i,Z_l) + \frac{1}{n^4}\sum_{i,j,l,p} k_Y(Y_i,Y_j)k_Z(Z_l,Z_p).$$

Как правило, ее записывают в матричной форме $HSIC^2 = tr(K_Y H K_Z H)/n^2$, где K_Y, K_Z – матрицы из $k_Y(y_i,y_j)$ и $k_Z(z_i,z_j)$ соответственно, а $H = E - \frac{1}{n}(1,\dots,1)^T(1,\dots,1)$.

Задача 22. Убедитесь в обоих представлениях $HSIC^2$.

Для нашей статистики справедливы леммы, аналогичные леммам для MMD^2.

Лемма 12. *Пусть* $\mathbf{E}k_Y(Y_1,Y_1)^2k_Z(Z_1,Z_1)^2 < +\infty$. *Тогда при* $n \to \infty$ *справедливо соотношение*

$$HSIC \xrightarrow{P} \mathcal{V}.$$

Доказательство. Принаждлежность HSIC к $L^2(\mathbf{P})$ вытекает из условия. Будем доказывать сходимость в L^2 для каждого из трех слагаемых в $HSIC^2$ к соответствующему множителю:

$$\mathbf{E}\left(\frac{1}{n^2}\sum_{i,j=1}^{n}(k_Y(Y_i,Y_j)k_Z(Z_i,Z_j) - \mathbf{E}k_Y(Y_1,Y_2)k_Z(Z_1,Z_2))\right)^2 = \frac{1}{n^4}\sum_{i,j,p,q}\mathbf{E}(k_Y(Y_i,Y_j)k_Z(Z_i,Z_j) - \mathbf{E}k_Y(Y_1,Y_2)k_Z(Z_1,Z_2)) (k_Y(Y_p,Y_q)k_Z(Z_p,Z_q) - \mathbf{E}k_Y(Y_1,Y_2)k_Z(Z_1,Z_2)).$$

При этом если i,j,p,q – четыре различных индекса, то математическое ожидание 0 по определению, а все случаи, когда какие-то два индекса совпадают, дают $O(n^3)$ слагаемых. Следовательно,

$$\frac{1}{n^2}\sum_{i,j=1}^{n}k_Y(Y_i,Y_j)k_Z(Z_i,Z_j) \xrightarrow{P} \mathbf{E}k_Y(Y_1,Y_2)k_Z(Z_1,Z_2).$$

Аналогичным путем

$$\mathbf{E}\left(\frac{2}{n^3}\sum_{i,j,l}(k_Y(Y_i,Y_j)k_Z(Z_i,Z_l) - \mathbf{E}k_Y(Y_1,Y_2)k_Z(Z_1,Z_3))\right)^2 = \frac{4}{n^6}\sum_{i,j,l,p,q,r}\mathbf{E}(k_Y(Y_i,Y_j)k_Z(Z_i,Z_l) - \mathbf{E}k_Y(Y_1,Y_2)k_Z(Z_1,Z_3)) \times (k_Y(Y_p,Y_q)k_Z(Z_p,Z_r) - \mathbf{E}k_Y(Y_1,Y_2)k_Z(Z_1,Z_3)).$$

В случаях, если все индексы i,j,l,p,q,r различны, то слагаемые в полученной сумме равны 0, а остальных случаев $O(n^5)$. Аналогичным образом рассматривается третье слагаемое. Значит,

$$HSIC^2 - \mathcal{V}^2 \xrightarrow{P} 0,\ n \to \infty,$$

что и требовалось доказать. □

Лемма 13. *Пусть k_Y, k_Z интегрируемы в квадрате по мерам* $\mathbf{P}$, $\mathbf{Q}$ *соответственно. Пусть (X, Y) независимы, тогда*

$$nHSIC^2 \xrightarrow{d} W,$$

где W определена в (11.1).

Доказательство. Как и прежде, можно считать, что

$$\mathbf{E}(k_Z(Z_1, Z_2)|Z_2, \ldots, Z_n) = 0, \quad \mathbf{E}(k_Y(Y_1, Y_2)|Y_2, \ldots, Y_n) = 0.$$

Действительно, запишем представление

$$HSIC^2 = \sum_{i,j} k_Y(Y_i, Y_j) \left(k_Z(Z_i, Z_j) - \frac{1}{n}\sum_l k_Z(Z_i, Z_l) - \frac{1}{n}\sum_l k_Z(Z_j, Z_l) + \frac{1}{n^2}\sum_{l,p} k_Z(Z_p, Z_l) \right).$$

Нетрудно заметить, что если перейти от k_Z к

$$\widetilde{k}_Z(z, z') = k_Z(z, z') - \mathbf{E}_P k_Z(z, Z) - \mathbf{E}_P k_Z(z', Z) + \mathbf{E}_{P\times P} k_Z(Z, Z'),$$

то выражение выше не изменится, однако, теперь мы получим ядро с указанными выше свойствами. После этого можно раскрыть формулу обратно и свернуть ее аналогичным образом относительно k_Y.

Рассмотрим операторы

$$A_{k_Y}(f)(y) = \int_{\mathcal{Y}} k_Y(y, y') f(y') \mathbf{P}(dy'), \quad A_{k_Z}(f)(z) = \int_{\mathcal{Z}} k_Y(z, z') f(z') \mathbf{Q}(dz').$$

Это по тем же причинам, что и прежде, компактные операторы и

$$k_Y(y, y') = \sum_{k=1}^{\infty} \lambda_k \psi_k(y) \psi_k(y'), \;\; k_Z(z, z') = \sum_{k=1}^{\infty} \mu_k \phi_k(z) \phi_k(z'),$$

где $\{\phi_k\}$ – собственные функции A_{k_Z}, $\{\psi_k\}$ – A_{k_Y}, $\{\lambda_k\}$, $\{\mu_k\}$ – соответствующие наборы собственных значений. Отметим, что интегируемость ядра влечет сходимость рядов

$$\sum_{k=1}^{\infty} \lambda_k, \quad \sum_{k=1}^{\infty} \mu_k.$$

Представим $nHSIC^2$ в виде $I_1 - 2I_2 + I_3$, где

$$I_{n,1} = \frac{1}{n}\sum_{i,j=1}^{n}\sum_{q,r=1}^{\infty}\lambda_q\mu_r\psi_q(Y_i)\psi_q(Y_j)\phi_r(Z_i)\phi_r(Z_j),$$

$$I_{n,2} = \frac{1}{n^2}\sum_{i,j,l=1}^{n}\sum_{q,r=1}^{\infty}\lambda_q\mu_r\psi_q(Y_i)\psi_q(Y_j)\phi_r(Z_i)\phi_r(Z_l),$$

$$I_{n,3} = \frac{1}{n^3}\sum_{i,j,l,p=1}^{n}\sum_{q,r=1}^{\infty}\lambda_q\mu_r\psi_q(Y_i)\psi_q(Y_j)\phi_r(Z_l)\phi_r(Z_p).$$

Дальнейшая работа достаточно близка к той, которая была проделана при работе с MMD. Для этого заметим, что $(\psi_q(Y_j), q \geq 1)$, $(\phi_r(Z_i), r \geq 1)$ – набор независимых последовательностей, причем в каждой из последовательностей все члены некореллированы, средние у каждой из величины нулевые, а дисперсии единичные. Также как и прежде мы можем утверждать, что ряды в I_1, I_2, I_3 сходятся п.н., откуда можно поменять порядок суммирования. Также как и прежде, исследуем

$$I_{N,n,1} = \frac{1}{n}\sum_{q,r=1}^{N}\lambda_q\mu_r\sum_{i,j=1}^{n}\psi_q(Y_i)\psi_q(Y_j)\phi_r(Z_i)\phi_r(Z_j) = \sum_{q,r=1}^{N}\lambda_q\mu_r\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_q(Y_i)\psi_r(Z_i)\right)^2,$$

$$I_{N,n,2} = \frac{1}{n^2}\sum_{q,r=1}^{N}\lambda_q\mu_r\sum_{i,j,l=1}^{n}\psi_q(Y_i)\psi_q(Y_j)\phi_r(Z_i)\phi_r(Z_l) =$$

$$\frac{1}{\sqrt{n}}\sum_{q,r=1}^{N}\lambda_q\mu_r\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_q(Y_i)\phi_r(Z_i)\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_q(Y_i)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_r(Z_i),$$

$$I_{N,n,3} = \frac{1}{n^3}\sum_{q,r=1}^{N}\lambda_q\mu_r\sum_{i,j,l,p=1}^{n}\psi_q(Y_i)\psi_q(Y_j)\phi_r(Z_l)\phi_r(Z_p) =$$

$$\frac{1}{n}\sum_{q,r=1}^{N}\lambda_q\mu_r\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_q(Y_i)\right)^2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_r(Z_i)\right)^2.$$

При этом

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\psi_q(Y_i)\phi_r(Z_i), q, r \in \{1,\ldots,N\},\right) \xrightarrow{d} \vec{U} \sim \mathcal{N}(\vec{0}, E),$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_q(Y_i) \xrightarrow{d} V \sim \mathcal{N}(0,1),\ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_r(Z_i) \xrightarrow{d} V \sim \mathcal{N}(0,1).$$

Задача 23. Докажите первое соотношение.

Тем самым, $I_{N,n,2} \xrightarrow{P} 0$, $I_{N,n,3} \xrightarrow{P} 0$ при $n \to \infty$ и любом N, поскольку сумма из них имеет предел, а множитель перед суммой стремится к нулю.

Следовательно,

$$I_{N,n,1} - 2I_{N,n,2} + I_{N,n,3} \xrightarrow{d} \sum_{q,r=1}^{N} \lambda_q \mu_r U_{q,r}^2,$$

где $U_{q,r}$ – н.о.р. $\mathcal{N}(0,1)$ величины. Следовательно, при любом $\varepsilon > 0$, $N \in \mathbb{N}$ и всех достаточно больших n выполнено выражение

$$\left| \mathbf{E} \exp\left(it(I_{N,n,1} - 2I_{N,n,2} + I_{N,n,3})\right) - \mathbf{E} \exp\left(it \sum_{q,r=1}^{N} \lambda_q \mu_r U_{q,r}^2 \right) \right| < \varepsilon.$$

При этом те же рассуждения, что и в аналогичной лемме об MMD, приводят нас к тому, что при достаточно больших N и всех n выполнены соотношения

$$|\mathbf{E} \exp\left(it(I_{n,1} - 2I_{n,2} + I_{n,3})\right) - \mathbf{E} \exp\left(it(I_{N,n,1} - 2I_{N,n,2} + I_{N,n,3})\right)| < \varepsilon,$$
$$\left| \mathbf{E} \exp\left(it \sum_{q,r=1}^{\infty} \lambda_q \mu_r U_{q,r}^2 \right) - \mathbf{E} \exp\left(it \sum_{q,r=1}^{N} \lambda_q \mu_r U_{q,r}^2 \right) \right| < \varepsilon.$$

Отсюда

$$nHSIC^2 \xrightarrow{d} \sum_{q,r=1}^{\infty} \lambda_q \mu_r U_{q,r}^2 =: W. \tag{11.1}$$

□

Критерий отсюда можно построить тем же путем, что и для MMD.

11.1.2 Связь ядра и полуметрики

Определение 11. Полуметрикой назовем симметричную функцию $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, для которой при всех $x, x' \in \mathcal{X}$ выполнено $\rho(x, x') = 0$ тогда и только тогда, когда $x = x'$.

Определение 12. Полуметрика ρ имеет отрицательный тип, если

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(x_i, x_j) \leq 0$$

при любых α_i, в сумме дающих 0, и любых $x_i \in \mathcal{X}$.

Заметим, что справедливы следующие соотношения, доказательства которых мы приводить не будем:

Лемма 14. *1) Если ρ – полуметрика отрицательного типа, то ρ^q также полуметрика отрицательного типа при $q \in (0,1)$.*
2) ρ – полуметрика отрицательного типа тогда и только тогда, когда существует такие гильбертово пространство $\mathcal{H}$ и отображение $\phi : \mathcal{X} \to \mathcal{H}$, что

$$\rho(x,x') = ||\phi(x) - \phi(x')||^2_{\mathcal{H}}.$$

В частности, евклидова метрика в $\mathbb{R}^k$ является полуметрикой отрицательного типа.

Зачем нам полуметрики отрицательного типа? Потому что они задают ядра. Пусть x_0 – некоторая фиксированная точка.

Определение 13. Ядро

$$k(x,x') = (\rho(x,x_0) + \rho(x',x_0) - \rho(x,x'))/2 \tag{11.2}$$

называется индуцированным полуметрикой ρ.

Определение 14. Ядро называется невырожденным, если функции $k(x,x')$ и $k(x,x'')$ различны при $x' \neq x''$.

Лемма 15. *Если ρ – полуметрика отрицательного типа, то k – неотрицательно определенное невырожденное ядро.*

Доказательство. Рассмотрим $t_i \in \mathbb{R}$, $x_i \in \mathcal{X}$:

$$\sum_{i,j=1}^{n} k(x_i,x_j)t_it_j = \sum_{i,j=1}^{n} \rho(x_i,x_0)t_i \sum t_j + \sum_{i,j=1}^{n} \rho(x_0,x_j)t_j \sum t_i - \sum_{i,j=1}^{n} \rho(x_i,x_j)t_it_j = \\ - \sum_{i=0}^{n}\sum_{j=0}^{n} \rho(x_i,x_j)t_it_j,$$

где $t_0 = -\sum_{i=1}^{n} t_i$. В силу отрицательного типа полуметрики получаем неотрицательную определенность ядра.

Предположим, что ядро вырождено, то есть $k(x,x') = k(x,x'')$ при всех x. Тогда

$$\rho(x',x) - \rho(x'',x)$$

не зависит от x. Однако, при $x = x'$ мы получаем отрицательную величину, а при $x = x''$ положительную. □

Заметим, что ядер таких несколько (из-за различного z_0), а полуметрику можно восстановить

$$\rho(x, x') = k(x, x) + k(x', x') - 2k(x, x').$$

Более того, любое невырожденное ядро по той же формуле определяет полуметрика отрицательного типа.

Итак, полуметрики позволяют строить ядра, а ядра полуметрики. Если быть точным, то невырожденные ядра образуют классы эквивалентности, каждому из которых соответствует своя полуметрика.

11.2 Подход Секея и Риццо

11.2.1 Energy Test

Секей и Риццо в работе Székely и M., 2005 предложили рассматривать для проверки однородности полуметрику

$$d(\mathbf{P}, \mathbf{Q}) = 2\int_{\mathcal{X}\times\mathcal{X}} \rho(x, x')\mathbf{P}(dx)\mathbf{Q}(dx') - \int_{\mathcal{X}\times\mathcal{X}} \rho(x, x')\mathbf{P}(dx)\mathbf{P}(dx') -$$

$$\int_{\mathcal{X}\times\mathcal{X}} \rho(x, x')\mathbf{Q}(dx)\mathbf{Q}(dx') = -\int_{\mathcal{X}\times\mathcal{X}} \rho(x, x')(\mathbf{P}(dx) - \mathbf{Q}(dx))(\mathbf{P}(dx') - \mathbf{Q}(dx')),$$

где ρ – некоторая полуметрика (исходно рассматривалось евклидово расстояние).

Соответствующая статистика приобретает вид

$$T = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \rho(Y_i, Z_j) - \frac{1}{n^2}\sum_{i,j=1}^{n} \rho(Y_i, Y_j) - \frac{1}{m^2}\sum_{i,j=1}^{m} \rho(Z_i, Z_j).$$

Если от полуметрики ρ перейти к ядру k, то

$$d(\mathbf{P}, \mathbf{Q}) = \int_{\mathcal{X}\times\mathcal{X}} k(x, x')(\mathbf{P}(dx) - \mathbf{Q}(dx))(\mathbf{P}(dx') - \mathbf{Q}(dx')),$$

в чем нетрудно убедиться, пользуясь соотношением между полуметрикой и индуцированным ей ядром.

Таким образом, energy distance является частным случаем MMD (а в указанной постановке фактически равносильна ему).

11.2.2 Distance covariance

Секей и Риццо в работе **Szekely-DCov** предложили рассматривать в качестве расстояния для проверки независимости

$$d(\mathbf{R}, \mathbf{P}\times\mathbf{Q}) = \int_{\mathcal{X}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \rho_X(x,x')\rho_Y(y,y')\mathbf{R}(dx,dy)\mathbf{R}(dx',dy') - 2\int_{\mathcal{X}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \rho_X(x,x')\rho_Y(y,y')\mathbf{R}(dx,dy)\mathbf{P}(dx')\mathbf{P}(dy') + \int_{\mathcal{X}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}} \rho_X(x,x')\rho_Y(y,y')\mathbf{P}(dx)\mathbf{Q}(dy)\mathbf{P}(dx')\mathbf{Q}(dy').$$

Это соответствует тому же Energy Distance, но в случае независимости называется Distance covariance. Соответственно, мы можем опять таки записать это в виде HSIC для того же ядра, соответствующего полуметрике.

Отметим интересную интерпретацию HSIC с точки зрения проверки независимости. Пусть ядра $k_Y(y,y')$ и $k_Z(z,z')$ на $\mathbb{R}^p \times \mathbb{R}^p$ и $\mathbb{R}^q \times \mathbb{R}^q$ непрерывны и зависят от разности, то есть $k_Y(y,y') = k_1(y-y')$, $k_Z(z,z') = k_2(z-z')$.

Справедлива теорема Бохнера-Хинчина:

$$k_1(y) = \int_{\mathbb{R}^p} \exp(i\langle s, y\rangle)\mu_1(ds), \quad k_2(z) = \int_{\mathbb{R}^q} \exp(i\langle t, z\rangle)\mu_2(dt).$$

Значит,

$$d(\mathbf{R}, \mathbf{P}\times\mathbf{Q}) = \int_{\mathbb{R}^p\times\mathbb{R}^p\times\mathbb{R}^q\times\mathbb{R}^q} \int_{\mathbb{R}^p\times\mathbb{R}^q} e^{i\langle y-y',t\rangle + i\langle z-z',s\rangle}\mu_1(ds)\mu_2(dt) \times (\mathbf{R}(dx,dy) - \mathbf{P}(dx)\mathbf{Q}(dy))(\mathbf{R}(dx',dy') - \mathbf{P}(dx')\mathbf{Q}(dy')).$$

Меняя порядок интегрирования, что можно сделать коль скоро меры μ_1, μ_2 конечны, получаем

$$d(\mathbf{R}, \mathbf{P}\times\mathbf{Q}) = \int_{\mathbb{R}^p}\int_{\mathbb{R}^q} \left(|\psi_{Y,Z}(t,s)|^2 - Re(\psi_{Y,Z}(t,s)\psi_Y(t)\psi_Z(s)) + |\psi_Y(t)|^2|\psi_Z(s)|^2\right)\mu_1(ds)\mu_2(dt),$$

где $\psi_{Y,Z}$ – совместная х.ф. наших величин, а ψ_Y, ψ_Z – маргинальные. Данную величину можно представить в форме

$$\int_{\mathbb{R}^p}\int_{\mathbb{R}^q} |\psi_{Y,Z}(t,s) - \psi_Y(t)\psi_Z(s)|^2\mu_1(ds)\mu_2(dt).$$

Это дает удачную интерпретацию критерия Distance covariance в случае евклидова пространства – мы основываемся на близости совместной характеристической функции к произведению маргинальных. Этот подход вполне естественен и мы рассматривали его на первой лекции.

11.2.3 Реализация и применение

Отсюда мы получаем критерии Секея-Риццо для проверки гипотез однородности (energy test) и независимости (dcov). Оба критерия могут примениться в указанных нами вариантах (перестановочном, основанном на оценке с.з. оператора или с помощью кривых Пирсона). Как правило, в качестве полуметрики для векторных величин используют еклидову норму в квадрате.

Оба названных критерия, как и общие версии MMD и HSIC, реализованы в пакете hyppo в Python, в R соответственные критерии реализованы в пакетах Energy (energy + dcov), Ecume (mmd), kpcalg (hsic). Учтите, что реализованные версии достаточно урезанные, используют лишь один из описанных нами подходов и по умолчанию используют простые настройки ядра/полуметрики.

Глава 12

Модификации критериев обобщенного отношения правдоподобий и хи-квадрат

Забравшись на вершину ядерных подходов, спустим к самым началам статистики – к критериям обобщенного отношения правдоподобий и хи-квадрат.

12.1 Критерий обобщенного отношения правдоподобий

Начнем с критерий обобщенного отношения правдоподобий (likelihood ratio test). К данному критерию можно прийти множеством путей, в частности, стартуя от классической теоремы Уилкса. Об этом вы можете прочитать в Roussas, 1997, 13.7, 13.8. Мы же будем верны подходу характеристика–статистика и поэтому оставим в сторону классический взгляд на этот вопрос.

12.1.1 Критерий однородности LLR

Итак, пусть $\mathbf{P}$, $\mathbf{Q}$ – дискретные меры с общим множеством атомов $\{u_1, \ldots, u_k\}$, которое мы предполагаем конечным. Тогда расстоянием (или дивергенцией) Кульбака-Лейблера называют

$$\rho_{KL}(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{k} p_i \ln \frac{p_i}{q_i},$$

где $p_i = \mathbf{P}(u_i)$, $q_i = \mathbf{Q}(u_i)$. Это не метрика и даже не полуметрика (она не симметрична), зато неотрицательная величина, обнуляющаяся лишь в случе $\mathbf{P} = \mathbf{Q}$. Докажем это:

$$\rho(\mathbf{P}, \mathbf{Q}) = \mathbf{E}_\mathbf{P} \ln \frac{p_X}{q_X} = -\mathbf{E}_\mathbf{P} \ln \frac{q_X}{p_X} \geq -\ln \mathbf{E}_P \frac{q_X}{p_X} = -\ln \sum_{i=1}^{n} q_i = 0,$$

где неравенство появилось из неравенства Иенсена. Функция $\ln x$ строго выпукла и равенство в неравенство Иенсена возможно лишь в случае вырожденной величины q_X/p_X, что бывает лишь в случае $\mathbf{Q} = \mathbf{P}$.

Построим из расстояния Кульбака-Лейблера желаннную характеристику по принципу

$$d(\mathbf{P}, \mathbf{Q}) = \alpha\rho(\mathbf{P}, \mathbf{R}) + (1 - \alpha)\rho(\mathbf{Q}, \mathbf{R}),$$

где $\mathbf{R} = \alpha\mathbf{P} + (1 - \alpha)\mathbf{Q}$, $\alpha \in (0, 1)$. На практике мы, как и прежде, возьмем $\alpha = n/(n + m)$.

Подставляя выражение расстояние Кульбака-Лейблера, получаем

$$d(\mathbf{P}, \mathbf{Q}) = \alpha \sum_{i=1}^{k} p_i \ln \frac{p_i}{\alpha p_i + (1 - \alpha)q_i} + (1 - \alpha) \sum_{i=1}^{k} q_i \ln \frac{q_i}{\alpha p_i + (1 - \alpha)q_i}.$$

Будем считать, что выборка при этом представляет набор частот $(\nu_{1,1}, \ldots, \nu_{k,1})$ ($\nu_{j,1}$ – число x_j среди первой выборки) и $(\nu_{1,2}, \ldots, \nu_{k,2})$ ($\nu_{j,2}$ – число x_j среди второй выборки). При этом общий объем первой выборки равен n, то есть $\nu_{1,1} + \cdots + \nu_{k,1} = n$, а второй – m, то есть $\nu_{1,2} + \cdots + \nu_{k,2} = m$.

Статистика критерия при этом принимает вид

$$\widehat{\lambda} = \sum_{i=1}^{k} \frac{\nu_{i,1}}{n + m} \ln \frac{\nu_{i,1}/n}{(\nu_{i,1} + \nu_{i,2})/(n + m)} + \sum_{i=1}^{k} \frac{\nu_{i,2}}{m + n} \ln \frac{\nu_{i,2}/m}{(\nu_{i,1} + \nu_{i,2})/(n + m)}.$$

Теорема 5. *Пусть гипотеза однородности верна. Тогда* $2\lambda \xrightarrow{d} Y \sim \chi^2_{k-1}$ *при* $n, m \to \infty$, $n/(n + m) \to a \in (0, 1)$.

Доказательство. Прежде всего заметим, что вектор

$$\sqrt{n + m}(\nu_{i,1}/n_i, \nu_{i,2}/m - p_i, i \leq k)$$

сходится к некоторому нормальному вектору $(\vec{U}, \vec{V})$ в силу ЦПТ. Пользуясь теоремой Скорохода, выберем новое пространство $(\Omega', \mathcal{F}', \mathbf{P}')$ и $(\nu'_{i,1}, \nu'_{i,2}, i \leq k)$ с тем же распределением, что

$$\sqrt{n + m}(\nu_{i,1}/n_i - p_i, i \leq k, \nu_{i,2}/m - p_i, i \leq k) \to (U', V')$$

п.н. Фиксируем $\omega' \in \Omega'$ при котором сходимость выше верна. Положим $\widehat{p}'_{i,1} = \nu'_{i,1}/n$, $\widehat{p}'_{i,2} = \nu'_{i,2}/m$, $\widehat{p}'_{i,3} = (\nu'_{i,1} + \nu'_{i,2})/(n+m)$. Тогда

$$\widehat{p}'_{i,1} - p_i = O\left(\frac{1}{\sqrt{n}}\right), \quad \widehat{p}'_{i,2} - p_i = O\left(\frac{1}{\sqrt{n}}\right), \quad \widehat{p}'_{i,3} - p_i = O\left(\frac{1}{\sqrt{n}}\right)$$

при $n \to \infty$. Следовательно,

$$\ln\left(1 + \frac{\widehat{p}'_{i,3} - \widehat{p}'_{i,1}}{\widehat{p}'_{i,1}}\right) = \frac{\widehat{p}'_{i,3} - \widehat{p}'_{i,1}}{\widehat{p}'_{i,1}} - \frac{1}{2}\frac{\left(\widehat{p}'_{i,3} - \widehat{p}'_{i,1}\right)^2}{(\widehat{p}'_{i,1})^2} +$$
$$+o\left(\frac{1}{n}\right) = \frac{\widehat{p}'_{i,3} - \widehat{p}'_{i,1}}{\widehat{p}'_{i,1}} - \frac{\left(\widehat{p}'_{i,3} - \widehat{p}'_{i,1}\right)^2}{2p_i\widehat{p}'_{i,1}} + o\left(\frac{1}{n}\right), \ n \to \infty.$$

Обратите внимание, что все рассматриваемые величины не случайные, поскольку мы фиксировали ω'. Мы пользуемся обычной теорией сходимости для числовых последовательностей.

Отсюда

$$\frac{\nu'_{i,1}}{n} \ln \frac{\nu'_{i,1}/n}{(\nu'_{i,1} + \nu'_{i,2})/(n+m)} = -(\widehat{p}'_{i,3} - \widehat{p}'_{i,1}) + \frac{\left(\widehat{p}'_{i,3} - \widehat{p}'_{i,1}\right)^2}{2p_i} + o\left(\frac{1}{n}\right).$$

Заметим, что при суммировании по $i \in \{1, \ldots, k\}$ первое слагаемое обнулится. Таким образом,

$$2\lambda' = \frac{n}{n+m}\sum_{i=1}^{k}\frac{\left(\widehat{p}'_{i,3} - \widehat{p}'_{i,1}\right)^2}{p_i} + \frac{m}{n+m}\sum_{i=1}^{k}\frac{\left(\widehat{p}'_{i,3} - \widehat{p}'_{i,2}\right)^2}{p_i} + o\left(\frac{1}{n}\right) =$$
$$\sum_{i=1}^{k}\frac{(a(1-a)^2 + (1-a)a^2)(\widehat{p}'_{i,1} - \widehat{p}'_{i,2})^2}{p_i} + o\left(\frac{1}{n}\right),$$

где мы воспользовались тем, что

$$\widehat{p}'_{i,3} = \frac{n\widehat{p}'_{i,1}}{n+m} + \frac{m\widehat{p}'_{i,2}}{n+m},$$

а также тем, что $n/(n+m) \to a$, $n, m \to \infty$. Следовательно,

$$2(n+m)\lambda' \to a(1-a)\sum_{i=1}^{k}\frac{(U'_i - V'_i)^2}{p_i},$$

откуда $2(n+m)\widehat{\lambda}$ сходится к той же величине по распределению. Остается понять какое распределение имеет величина в правой части. Заметим, что векторы (U_i') и (V_i') независимы, причем оба имеют нулевые средние и матрицы ковариаций Σ/a и $\Sigma/(1-a)$, где

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_k \\ -p_1p_2 & p_2(1-p_2) & \dots & -p_2p_k \\ & & \dots & \\ -p_1p_k & -p_2p_k & \dots & p_k(1-p_k). \end{pmatrix}.$$

Следовательно, вектор $(\sqrt{a(1-a)}(U_i - V_i)/\sqrt{p_i}, i \le k)$ нормальный с матрицей ковариаций

$$E - uu^T, \quad \vec{u} = (\sqrt{p_1}, \dots, \sqrt{p_k})^T.$$

Искомая статистика есть квадрат длины этого вектора, а значит распределена как

$$\sum_{i=1}^{k} \lambda_i Z_i^2,$$

где Z_i н.о.р. $\mathcal{N}(0,1)$, λ_i – с.з. матрицы ковариции. Однако, собственный базис матрицы легко угадывается:

$$(E - \vec{u}\vec{u}^T)\vec{u} = \vec{u} - \vec{u}\langle\vec{u},\vec{u}\rangle = 0, \quad (E - \vec{u}\vec{u}^T)\vec{v} = \vec{v} - \vec{v}\langle\vec{u},\vec{u}\rangle = \vec{v},$$

где $\vec{v}$ – любой ортогональный $\vec{u}$ вектор. Значит, с.з. 1 имеет кратность k-1, а 0 – единичную кратность, то есть

$$\sum_{i=1}^{k} \lambda_i Z_i^2 = Z_1^2 + \dots + Z_{k-1}^2 \sim \chi_{k-1}^2,$$

что и требовалось доказать. □

Итак, получаем критерий $2\widehat{\lambda} > y_{1-\alpha}$, где y – квантиль χ^2_{k-1}-распределения.

12.1.2 Критерий однородности для l выборок

Для l выборок мы используем характеристику

$$\sum_{i=1}^{l} \alpha_i \rho_{KL}(\mathbf{P}_i, \mathbf{R}),\ \mathbf{R}(A) = \sum_{i=1}^{k} \alpha_i \mathbf{P}_i(A).$$

Соответствующая статистика критерия есть

$$\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\nu_{i,j}}{n}\ln\left(\frac{\nu_{i,j}/n_j}{\sum_{j=1}^{k}\nu_{i,j}/n}\right),$$

где $\nu_{i,j}$ – число x_i в j-й выборке, n_i – число элементов i-й выборки, $n = n_1 + \cdots + n_k$ – число элементов всех выборок.

Аналогичные предыдущим рассуждения приводят к тезису о том, что при верной гипотезе

$$\sum_{i=1}^{l}\sum_{j=1}^{k}\frac{\nu_{i,j}}{n}\ln\left(\frac{\nu_{i,j}/n_j}{\sum_{j=1}^{k}\nu_{i,j}/n}\right) \xrightarrow{d} Y \sim \chi^2_{(k-1)(l-1)},$$

откуда получаем критерий.

12.1.3 Критерий независимости

Проверка независимости с помощью того же подхода соответствует дискретной выборки (Y_i, Z_i), где $Y_i \in \mathcal{Y}$, $|\mathcal{Y}| = k$, $Z_i \in \mathcal{Z}$, $|\mathcal{Z}| = l$. Будем использовать $\nu_{i,j}$ – число наблюдений с i-м значением первой координаты и одновременно j-м значением второй.

Наш подход естественно приводит к характеристике

$$\rho_{KL}(\mathbf{P}\times\mathbf{Q},\mathbf{R}) = \sum_{i=1}^{k}\sum_{j=1}^{l} p_i q_j \ln\frac{p_i q_j}{r_{i,j}},$$

где $r_{i,j}$ – совместное распределение, p_i, q_j – маргинальные.

Соответствующая статистика имеет вид

$$\lambda = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{\nu_{i,\cdot}}{n}\frac{\nu_{\cdot,j}}{n}\ln\frac{\nu_{i,\cdot}\nu_{\cdot,j}}{n\nu_{i,j}},$$

где

$$\nu_{\cdot,j} = \sum_{i=1}^{k}\nu_{i,j}, \quad \nu_{i,\cdot} = \sum_{j=1}^{l}\nu_{i,j}.$$

Представим себе таблицу (такие таблицы называют таблицами сопряженности) следующего вида (где отчеркнутые линии отделяют суммы по столбцам или

строкам):

$\nu_{1,1}$	$\nu_{1,2}$	$\dots$	$\nu_{1,l}$	$\nu_{1,\cdot}$
$\nu_{2,1}$	$\nu_{2,2}$	$\dots$	$\nu_{2,l}$	$\nu_{2,\cdot}$
		$\dots$		$\dots$
$\nu_{k,1}$	$\nu_{k,2}$	$\dots$	$\nu_{k,l}$	$\nu_{k,\cdot}$
$\nu_{\cdot,1}$	$\nu_{\cdot,2}$	$\dots$	$\nu_{\cdot,l}$	n

Заметим, что если построить такую же таблицу для критерия однородности, то статистика будет иметь в точности тот же вид (только $\nu_{i,\cdot}$ у нас будет представлять n_i). Поэтому два наших критерия дают в точности одно и то же.

Почему же задачи проверки независимости и однородности оказываются одним и тем же? Потому что выборки дискретные и независимость сводится к проверке того, что векторы вероятностей $\mathbf{P}(Y = y_i|Z = z_j)$, $i \le k$, совпадают при всех $j \le l$. Это суть та же однородность.

Правда, вероятностная модель несколько меняет вид – раньше фиксированы были количества наблюдений в каждой строке, а теперь количества наблюдений в каждой строке случайно, а фиксировано только общее количество n. Тем не менее, теорема остается той же:

$$2 \ln \widehat{\lambda} \xrightarrow{d} Y \sim \chi^2_{(k-1)(l-1)}.$$

Отсюда получаем критерий независимости.

В Python его можно найти в scipy.stats.chi2_contingency, а в R chisq.test.

12.1.4 Критерий хи-квадрат

Критерий LLR строился на основе расстояния Кульбака-Лейблера. Можно рассмотреть взамен величину

$$\rho^2_{\chi^2}(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{k} \frac{(p_i - q_i)^2}{q_i}.$$

Это опять же несимметричная характеристика, которая локально соответствует расстоянию Кульбака-Лейблера (главный член у них при $p_i \approx q_i$ одинаковый).

Соответственно, для критерия однородности двух выборок предлагается рассматривать характеристику

$$d_{\chi^2}(\mathbf{P}, \mathbf{Q}) = \alpha\rho_{\chi^2}(\mathbf{P}, \mathbf{R}) + (1-\alpha)\rho_{\chi^2}(\mathbf{Q}, \mathbf{R}) = \alpha(1-\alpha)\sum_{i=1}^{k}(p_i - q_i)^2 a p_i + (1-p)q_i.$$

Соответственно, статистика имеет вид

$$\widehat{\chi}^2 = \frac{n^2}{(n+m)^2} \sum_{i=1}^{k} \frac{(\nu_{i,1}/n - \nu_{i,2}/m)^2}{(\nu_{i,1}+\nu_{i,2})/(n+m)}.$$

Лемма 16. *При верной гипотезе однородности*

$$(n+m)\widehat{\chi}^2 \xrightarrow{d} Y \sim \chi^2_{k-1}.$$

Доказательство. Проделаем тот же трюк, что и для LLR критерия: перейдем к другому пространству $(\Omega', \mathcal{F}', \mathbf{P}')$ и на нем получим тождество

$$\sqrt{n+m}(\nu'_{i,1}/n - p_i, i \le k, \nu'_{i,2}/m - p_i, i \le k) \to (U', V'),$$

все величины при этом определены также как и ранее. Тогда

$$(n+m)\sum_{i=1}^{k} \frac{(\widehat{p}'_{i,1} - \widehat{p}'_{i,2}/m)^2}{\widehat{p}_{i,3}} = (n+m)\sum_{i=1}^{k} \frac{(\nu'_{i,1}/n - \nu'_{i,2}/m)^2}{p_i} + o(1) \to \sum_{i=1}^{k} \frac{(U'_i - V'_i)^2}{p_i},$$

откуда

$$\sqrt{n+m}\widehat{\chi}^2 \xrightarrow{d} a(1-a)\sum_{i=1}^{k} \frac{(U_i - V_i)^2}{p_i},$$

а распределение величины в правой части уже исследовалось нами прежде. □

Для критерия однородности l выборок предлагается рассматривать характеристику

$$d_{\chi^2}(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{l} \alpha_i \rho_{\chi^2}(\mathbf{P}_i, \mathbf{R}),$$

а статистика имеет вид

$$\widehat{\chi}^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(\nu_{i,j}/n_j - \nu_{i,\cdot}/n)^2}{\nu_{i,\cdot}/n_j},$$

где n_j – размер j-й выборки, $\nu_{i,\cdot}$ – число i-х наблюдений. Предельное распределение величины $(n+m)\widehat{\chi}^2$ при этом, как и прежде, будет $\chi^2_{(k-1)(l-1)}$.

Как и для LLR критерий независимости по существу совпадает с критерием однородности.

12.2 Адаптации критериев для общего случая

12.2.1 Критерии хи-квадрат и к.о.о.п. для недискретного случая

Если Y_i, Z_i произвольные величины со значениями в произвольном множестве S, то мы можем произвести дискретизацию, рассмотрев разбиение $\bigcup \Delta_i = S$ и заменив исходные величины на

$$\nu_{i,1} = \#\{j : Y_j \in \Delta_i\}, \quad \nu_{i,2} = \#\{j : Z_j \in \Delta_i\}.$$

Однако, здесь становится существенным вопрос "Что такое Δ_i?" Мощность критерия напрямую будет зависеть от удачности выбора распределения.

Может возникнуть идея перебрать всевозможные разбиения на Δ_i какого-то вида. Например, на прямой можно рассмотреть множества $\Delta_1 = (-\infty, x]$ и $\Delta_2 = (x, +\infty)$, откуда статистика приобретет вид

$$\sup_x \left| \frac{n}{n+m}\widehat{F}_n(x) \ln \frac{\widehat{F}_n(x)}{\widehat{H}_{n,m}(x)} + \frac{n}{n+m}(1-\widehat{F}_n(x)) \ln \frac{(1-\widehat{F}_n(x))}{(1-\widehat{H}_{n,m}(x))} + \right.$$
$$\left. \frac{m}{n+m}\widehat{G}_m(x) \ln \frac{\widehat{G}_m(x)}{\widehat{H}_{n,m}(x)} + \frac{m}{n+m}(1-\widehat{G}_m(x)) \ln \frac{(1-\widehat{G}_m(x))}{(1-\widehat{H}_{n,m}(x))} \right|.$$

Можно вместо супремума по x проинтегрировать данную величину по x по множеству $\widehat{H}_{n,m}(x)$.

В более общей постановке аналогичная статистика была предложена Хеллером, Хеллером и Ко (Heller и др., 2016) в одномерном случае. Они предложили зафиксировать параметр k и всевозможные целые положительные решения уравнения $i_1 + i_2 + \ldots + i_k = n + m$. При этом мы хотим сформировать Δ_j так, что первый из них содержит i_1 левых наблюдений, следующий i_2 следующих и так далее. При этом посчитаем для каждого такого разбиения $\nu_{i,j}$ – количество элементов j-й выборки в i-м элементе разбиения. При этом считается статистика критерия отношения правдоподобий

$$2 \ln \widehat{\lambda}_{i_1,\ldots,i_k},$$

после чего полученные статистики максимизируются (точка зрения подхода Колмогорова или, иными словами, взгляд подхода пересечений-объединений) или суммируются (точка зрения подхода Крамера-фон Мизеса, этакий байесовский взгляд). Данная величина и называется статистикой T_{HHG} критерия Хеллера, Хеллера, Горфина и Ко.

Для одномерных данных эта статистика, очевидно, зависит только от рангов Y_i в вариационном ряду X_i, а значит не зависит от распределения выборок при верной гипотезе, если только указанное распределение непрерывное. Тем самым, мы получаем критерий второго типа, который можно использовать на основе метода Монте-Карло.

Аналогичным образом строится статистика на основе хи-квадрат. Более того, для k выборок построения также аналогичны.

Наконец, для гипотезы независимости мы также можем построить аналогичную статистику, рассматривая всевозможные $i_1, \ldots, i_k$ по горизонтали и $j_1, \ldots, j_k$ по вертикали, а затем строя таблицу сопряженности и считая статистику хи-квадрат или отношения правдоподобий.

12.2.2 Критерий Хеллера-Хеллера-Горфина

Итак, введем для гипотезы однородности двух одномерных выборок статистику Хеллера-Хеллера-Горфина следующим образом. Пусть $\mathcal{I}_k$ – всевозможные разбиения индексов $1, \dots, n$ на k непересекающихся отрезков $I_1, \dots, I_k$, введем для каждого разбиения статистику хи-квадрат в форме

$$\widehat{\chi}^2_{I_1,\dots,I_k} = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{\left(\nu_i(j) - \frac{n_i \#\{I_j\}}{n}\right)^2}{\frac{n_i \#\{I_j\}}{n}},$$

где $\nu_i(j)$ – количество наблюдений в i-й выборке, ранги которых находятся в диапазоне I_j. Аналогично определяется статистика для однородности m выборок.

При этом положим

$$S_k = \sum_{I\in\mathcal{I}_k} \widehat{\chi}^2_{I_1,\dots,I_k}, \quad M_k = \max_{I\in\mathcal{I}_k} \widehat{\chi}^2_{I_1,\dots,I_k}.$$

Это статистика, а какую характеристику она оценивает? Для первой статистики такой факт получен в Heller и др., 2016 (Theorem 3, переведенная на язык критериев однородности).

Теорема 6. *Пусть $n_i/n \to c_i \in (0,1)$, $i \le k$, i-я выборка имеет плотность f_i. Тогда при $k = k_n \to \infty$, $k_n/n \to 0$, $n \to \infty$ выполнено соотношение*

$$\frac{S_k}{nC_{n-1}^{k_n-1}} \xrightarrow{P} \sum_{i=1}^{m} c_i \mathbf{E}_{P_i} \ln \frac{f_i(X_i)}{\sum_{i=1}^m c_i f_i(X_i)},$$

то есть взвешенной сумме расстояния Кульбака-Лейблера между каждой из плотностей и усредненной плотностью.

Таким образом, получается состоятельная оценка такой, вполне разумной, характеристики критерия. Тот же факт верен для статистики критерия отношения правдоподобий, рассматриваемой вместо χ^2.

Аналогичная ситуация с критерием независимости. Если (Y_i, Z_i), $i \le n$, наши наблюдения, то положим

$$\widehat{\chi}^2_{I_1,\dots,I_k,J_1,\dots,J_k} = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\left(\nu_{i,j} - \frac{\#\{J_i\}\#\{I_j\}}{n}\right)^2}{\frac{\#\{J_i\}\#\{I_j\}}{n}},$$

где разбиения $I = (I_j, j \le k)$, $J = (J_i, i \le k)$ берутся по переменным Y и Z, $nu_{i,j}$ – число наблюдений (Y_l, Z_l) в $I_j \times J_i$. Аналогичным образом верна следующая теорема.

Теорема 7. *Пусть выборка (Y, Z) имеет непрерывную плотность и совместная информация*

$$I(x, y) = \int f_{Y,Z}(y, z) \ln \frac{f_{Y,Z}(y, z)}{f_Y(y) f_Z(z)} dydz < +\infty.$$

При $k_n/\sqrt{n} \to 0$, $k = k_n \to \infty$, $n \to \infty$ выполнено соотношение

$$\frac{S_k}{n \left(C_{n-1}^{k_n - 1}\right)^2} \xrightarrow{P} I.$$

Итак, характеристика разумная, критерий также получается достаточно неплохой, к тому же второго типа, что позволяет эффективно рассчитывать предельное распределение статистики. Вот только сам подсчет статистики достаточно длителен, что приводит к весьма неспешной работе критерия.

Мы советуем просмотреть статью Heller и др., 2016, в которой проводится весьма информативное сравнение критерия с другими.

12.3 Многомерные ранги

12.3.1 Многомерные теоретические ранги и квантили

Как мы отмечали ранее, при переходе в большую чем единица размерность построения критерия I или II типа достаточно затруднительно.

В одномерном случае в случае непрерывного распределения $X \sim F$ выполнялись соотношения $F(X) \sim R[0, 1]$, $F^{-1}(R) \sim F$, где R – равномерная случайная величина.

Мы бы хотели определить понятия, аналогичные "квантили" и "функции распределения", которые бы позволяли вектор с заданным распределением переводить в другое заданное распределение. Для определенности будем вместо $R[0, 1]$ использовать распределение U вектора $\vec{Y}$, у которого полярный радиус $||\vec{Y}||$ равномерного распределен на $[0, 1]$, а вектор единичной длины $\vec{Y}/||Y||$, сонаправленный с нашим, равномерно распределен на единичной сфере.

Предположим, что $\vec{X}$ имеет выпуклый носитель в $\mathbb{R}^d$. Найдем для нашего вектора $\vec{X}$ вектор $\vec{Y} \sim U$, дающий минимум в задаче минимизации расстояния

Канторовича-Вассерштейна, которая возникала у нас в лекции 7:

$$\sup_{\vec{X}\sim F,\vec{Y}} \mathbf{E}\left|\vec{Y}-\vec{X}\right|^2 \to \min.$$

При этом минимум достигается при $\vec{X} = T(\vec{Y})$ п.н., где T – некоторое отображение, которое называют *оптимальным транспортным планом*. Соответственно, $T(\vec{Y})$ называют квантилью (в роли уровня теперь выступает не число, а вектор внутри единичного шара), $R(x) = T^{-1}(x)$ называют векторным рангом точки x, $r(x) = ||R(x)||$ – скалярным рангом, $s(x) = R(x)/||R(x)||$ – векторным знаком.

При этом $T(\vec{Y}) \sim F$, если $\vec{Y} \sim U$, а если $\vec{X} \sim F$, то $R(\vec{X}) \sim U$, $R(x) \sim R[0,1]$, $s(X) \sim R\{S^d\}$, где S^d – единичная сфера.

В случае, если F имеет плотность относительно меры Лебега с компактным носителем, которая на этом носителе является непрерывной и гельдеровой с параметром $\alpha \in (0,1)$ (назовем это свойством регулярности плотности), отображение T является гомеоморфизмом, а R – градиентом некоторой выпуклой функции (Villani, 2021, 4.2.2 Theorem 4.14).

Любопытной характеристикой являются также "глубины" Монжа-Канторовича (depth). Глубиной называют отображение D из $\mathbb{R}^d \to \mathbb{R}$, являющееся непрерывным. Будем говорить, что глубина индуцирует отношение порядка, говоря, что $x \geq_D y$ тогда и только тогда, когда $D(x) \geq D(y)$. Глубина задает линии уровня, то есть кривые вида $D(x) = const$.

Соответственно, глубиной Монжа-Канторовича назовем образы сфер с центром в нуле радиуса u при действии отображения T. Учитывая, что распределение U имеет постоянную плотность на каждой из сфер, это некоторые линии уровня, характеризующие распределение. Линии уровня при указанных выше условиях регулярности плотности будут гладкими. Ранг будет согласован с глубиной – чем больше глубина (больше u), тем больше скалярный ранг.

12.3.2 Многомерные выборочные глубина, ранги и квантили

Мы будем строить аналогичные характеристики на основе выборки.

Рассмотрим выборку $X_1, \ldots, X_n \sim F$ и $U_1, \ldots, U_n$, лежащие в единичном шаре, для которых эмпирическое распределение с ростом n сходится к U. Рассмотрим оптимальный транспорт из эмпирической меры на $X_1, \ldots, X_n$ в эмпирическую меру на $U_1, \ldots, U_n$. Построение такого отображения обсуждалось в разделе 7. Оно задает выборочную квантиль $\widehat{T}_n$, выборочный ранг $\widehat{R}_n$.

Теорема 8 (Chernozhukov и др., 2017, Theorem 3.1)**.** *Пусть F имеет компактный носитель и абсолютно-неперерывна. Тогда*

$$\sup_x ||\widehat{R}_n(x) - R(x)|| \xrightarrow{P} 0, \quad \sup_x ||\widehat{T}_n(x) - T(x)|| \xrightarrow{P} 0, \ n \to \infty.$$

Иными словами, выборочные ранги и квантили равномерно состоятельны как оценки рангов и квантилей.

12.3.3 Выбор Y_n

Два основных варианта выбора $U_1, \dots, U_n$ – это выбор фиксированных точек в шаре, например, сетки S вида (R_i, ϕ_j), где ϕ_j – равномерное разбиение $[0, 2\pi]$, R_i – равномерное разбиение на $[0, 1]$. Обычно добавляют несколько точек в нуле, чтобы исправить потенциальное наличие остатка.

Это соответствует обычным рангам, только трансформированным в $[-n, n]$. При этом н.о.р. выборка $X_1, \dots, X_n$ перейдет в случайную перестановку на множестве S (то есть станет зависимой).

Другой вариант – выбор в качестве Y_i случайных точек в шаре с распределением U. Тогда образы X_i будут н.о.р. величинами с распределение U.

Таким образом, мы можем взять исходную выборку, осуществить оптимальную транспортировку их в случайные точки в шаре U_i, называть U_i соответствующую X_i выборочным рангом R_i и использовать соответствующий подход для работы с рангами.

12.3.4 Ранговые критерии однородности

Пусть $(X_1, \dots, X_n)$ – объединенная выборка из $Y_1, \dots, Y_{n_1}$, $Z_1, \dots, Z_{n_2}$ – выборки из $\mathbb{R}^d$ для которых мы хотим проверить однородность. Тогда построим оптимальный транспортный план по перевозу в $(X_i, i \le n)$ в $(U_i, i \le n)$.

При этом Y_i будут соответствовать эмпирические ранги $R_{i,1}$, а Z_i – эмпирические ранги $R_{i,2}$. При верной гипотезе однородности набор рангов R_i – случайная перестановка на множестве $U_1, \dots, U_n$ объединенных значений рангов, поскольку все перестановки X_i "равновероятны".

Таким образом, в случае, если сеть S для рангов берется заранее фиксированной, можно использовать любой критерий однородности, модифицировав, впрочем, соответствующие предельные теоремы для данных, представляющих случайную перестановку на S, а не н.о.р. наблюдения.

Если же сеть берется случайным размещением точек с распределением U, то модификация не требуется – ранги при верной гипотезе будут н.о.р. с распределением U, а значит можно без изменений применять ту же технику, что и ранее для этого случая.

Благодарности

Разработка данного курса поддержана грантом «Спецкурс – Механико-математический факультет» Фонда развития теоретической физики и математики «БАЗИС».

Список литературы

Scholz, F. & Stephens, M. (1987). K-sample Anderson–Darling tests. *Journal of the American Statistical Association*, *82*(399), 918—924.

Baumgartner, W., Weiss, P. & Schindler, H. (1998). A nonparametric test for the general two-sample problem. *Biometrics*, 1129—1135.

Epps, T. & Singleton, K. (1986). An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, *26*(3-4), 177—203.

Fernandez, A., Gamero, J. & Garcia, M. (2008). A test for the two-sample problem based on empirical characteristic functions. *Computational statistics & data analysis*, *52*(7), 3730—3748.

Ramdas, A., García, T. & Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, *19*(47), 1—15.

Chernozhukov, V. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, *45*(1), 223—256.

Friedman, J. & Rafsky, L. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 697—717.

Arboretti, R., Bathke, A., Carrozzo, E., Pesarin, F. & Salmaso, L. (2020). Multivariate permutation tests for two sample testing in presence of nondetects with application to microarray data. *Statistical Methods in Medical Research*, *29*(1), 258—271.

Hotelling, H. (1992). The generalization of Student's ratio. *Breakthroughs in statistics* (с. 54—65). Springer.

Hoeffding, W. (1994). A non-parametric test of independence, 214—226.

Blum, J., Kiefer, J. & Rosenblatt, M. (1961). *Distribution free tests of independence based on the sample distribution function*. Sandia Corporation.

Székely, G. & Rizzo, M. (2009). Brownian distance covariance. *The annals of applied statistics*, *3*(4), 1236—1265.

Биллингсли, П. (1977). *Сходимость вероятностных мер.* (Т. 351). Наука.

Billingsley, P. (1999). *Convergence of probability measures*. John Wiley & Sons.

Wellner, J. (2013). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Shorack, G. & Wellner, J. (2009). *Empirical processes with applications to statistics*. SIAM.

Scaillet, O. (2005). A Kolmogorov-Smirnov type test for positive quadrant dependence. *Canadian Journal of Statistics*, *33*(3), 415—427.

Gretton, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, *13*(1), 723—773.

Reed, M. & Simon, B. (1980). *Functional analysis. Revised and Enlarged Edition*. Academic Press.

Solomon, H. & Stephens, M. (1977). Distribution of a sum of weighted chi-square variables. *Journal of the American Statistical Association*, *72*(360), 881—885.

Székely, G. & M., R. (2005). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, *8*(143), 1249—1272.

Roussas, G. (1997). *A Course in mathematical statistics*. Academic Press.

Heller, R., Heller, Y., Kaufman, S., Brill, B. & Gorfine, M. (2016). Consistent distribution-free k-sample and independence tests for univariate random variables. *The Journal of Machine Learning Research*, *17*(1), 978—1031.

Villani, C. (2021). *Topics in optimal transportation* (T. 58). American Mathematical Soc.

Chernozhukov, V., Galichon, A., Hallin, M. & Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, *45*(1), 223—256.