

Занятие шестое. Критерии однородности или как заметить отличие

Старые знакомые

Мы будем рассматривать две задачи: проверку гипотезы однородности — по нескольким выборкам $X_{1,1}, \dots, X_{1,n_1}, X_{2,1}, \dots, X_{2,n_2}, \dots, X_{k,1}, \dots, X_{k,n_k}$, $n_1 + \dots + n_k = n$, проверить гипотезу о том, что выборки из одного распределения. В англоязычной литературе такого рода критерии называют k -sample test.

При этом будет изучать две альтернативы — общую H_1 : выборки из разных распределений или доминирования H_2 : $F_1(x) \geq F_2(x) \dots \geq F_k(x)$ при всех x , т.е. каждое следующее распределение в целом выдает более большие наблюдения.

Вопрос 1. Показать, что $F_X(x) \geq F_Y(x)$ при всех x равносильно тому, что на некотором вероятностном пространстве можно задать $X_1 \stackrel{d}{=} X$, $Y_1 \stackrel{d}{=} Y$, т.ч. $Y_1 \geq X_1$ п.н.

Начнем с модификации известных нам критериев.

1) Критерий Смирнова.

Для $k = 2$ и непрерывных распределений F_X, F_Y можно модернизировать критерий Колмогорова.

Пусть \hat{F}_{i,n_i} — ЭФР, распределения F_i непрерывны, $i = 1, 2$. Тогда положим

$$D_n = \sup |\hat{F}_{1,n_1} - \hat{F}_{2,n_2}|.$$

Мы понимаем, что если гипотеза верна и функции распределения выборок одинаковы, то их эмпирические функции распределения близки. Более точно, оказывается, что при выполнении гипотезы $\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_n \rightarrow K \sim K(x)$, где $K(x)$ — ф.р. Колмогорова, $n_1, n_2 \rightarrow \infty$, откуда получаем критическое множество

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_n > k_{1-\alpha}.$$

В случае альтернативы доминирования H_2 рассматривают статистику

$$D_n^+ = \sup(\hat{F}_{2,n_2} - \hat{F}_{1,n_1}).$$

При выполнении основной гипотезы $\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_n^+ \rightarrow 1 - e^{-2x^2}$ при $n_1, n_2 \rightarrow \infty$ откуда получаем критическое множество

$$\frac{\sqrt{n_1 n_2}}{\sqrt{n_1 + n_2}} D_n^+ > \sqrt{-\ln \sqrt{\alpha}}.$$

В R критерий Колмогорова доступен в той же функции `ks.test`, параметр `alternative` может быть выбран "less" или "greater" для одностороннего критерия.

2) Критерий Розенблатта (two sample Lehman test).

Этот критерий является модификацией критерия Крамера-Мизеса. Тестовая статистика критерия

$$\int_{\mathbb{R}} (\hat{F}_{1,n_1}(x) - \hat{F}_{2,n_2}(x))^2 d\hat{H}_2(x),$$

где $\hat{H}_2(x) = (n_1 \hat{F}_{1,n_1}(x) + n_2 \hat{F}_{2,n_2}(x)) / (n_1 + n_2)$ — совместная эмпирическая функция распределения двух выборок. При выполнении гипотезы однородности она будет сходиться к тому же распределению, что и статистика Крамера-Мизеса.

Другой формой той же статистики является

$$\frac{1}{n_1 n_2} \left(\frac{1}{6} + \frac{1}{n_2} \sum_{i=1}^{n_1} (R_i - i)^2 + \frac{1}{n_1} \sum_{i=1}^{n_2} (S_i - i)^2 \right) - \frac{2}{3},$$

где R_i — ранги элементов $X_{1,i}$, S_i — ранги элементов $X_{2,i}$ в общем вариационном ряду.

Этот критерий является модификацией критерий Крамера-Мизеса.

Упражнение 1. Проверить критерии на выборках из $\mathcal{N}(0, 1)$ и Лапласа размера а) 30 б) 50 в) 100.

3) Критерий Андерсона-Дарлингга.

Для нескольких выборок можно модернизировать и критерий Андерсона-Дарлингга. Такой критерий

был предложен F.W. Scholz, M.A. Stephens в 1987 году. Они предложили рассматривать статистику

$$A_k^2 = \sum_{i=1}^k n_i \int_B \frac{(\hat{F}_{i,n_i}(x) - H_k(x))^2}{H_k(x)(1 - H_k(x))} dH_k(x),$$

где H_k — совместная ЭФР всех выборок, $B = \{x \in \mathbb{R} : H_k(x) < 1\}$. Оказывается, что для непрерывных величин A_k^2 с ростом длин выборок сходится по распределению к $\sum_{j=1}^{\infty} \frac{1}{j(j+1)} Y_j$, $Y_j \sim \chi_{j-1}^2$ независимы. Критерий можно применять и в не непрерывном случае. В R этот критерий представлен функцией AndersonDarling пакета MissMech со слегка неудобным интерфейсом — для ее использования выборки нужно склеить в одну и второй переменной указать вектор из длин выборок.

4) Критерии хи-квадрат и отношения правдоподобия.

Однородность нескольких выборок можно проверять с помощью критерия хи-квадрат. Для этого положим $\Delta_1, \dots, \Delta_m$ — разбиение области значений наших величин на множества, найдем $\nu_{i,j}$ — количество $X_{i,l}$, попавших в Δ_j , $\nu_{\cdot,j} = \sum_i \nu_{i,j}$, $\nu_{i,\cdot} = \sum_j \nu_{i,j}$. Тогда при выполнении основной гипотезы

$$\sum_{i=1}^k \sum_{j=1}^m \frac{(\nu_{i,j} - \frac{\nu_{i,\cdot} \nu_{\cdot,j}}{n})^2}{\nu_{\cdot,j} \nu_{i,\cdot} / n} \xrightarrow{d} Y \sim \chi_{mk-k-m+1}^2.$$

Аналогично

$$2 \sum_{i=1}^k \sum_{j=1}^m \ln \left(\frac{\nu_{i,j} n}{\nu_{i,\cdot} \nu_{\cdot,j}} \right) \xrightarrow{d} Y \sim \chi_{mk-k-m+1}^2.$$

Вопрос 2. Как вывести этот результат из общего критерия отношения правдоподобий?

Критерий хи-квадрат для проверки однородности двух выборок реализован в `chisq.test`. Первым аргументом ему задается так называемая таблица сопряженности, то есть матрица из $\nu_{i,j}$. Для дискретной выборки получить ее строки можно с помощью команды `table()`.

В случае двух бернуллиевских выборок может быть также использован явный критерий Фишера `fisher.test`, которые не апеллируют к асимптотическим свойствам.

Замечание. Приведенные критерии асимптотические, при небольших выборках p -value в указанных функциях рассчитывается не из предельного распределения, а исходя из метода Монте-Карло.

Упражнение 2. Перед вами результаты медицинских исследований. Из 1500 мужчин, испытывающих лекарство, выздоровели 700, из 210 не принимавших выздоровели 80. Из 220 принимавших женщин — 150, из 680 не принимавших — 400. Проверить, влияет ли лекарство на мужчин? На женщин? На людей обоих полов?

Непараметрические тесты

К непараметрическим критериям 1) и 2) из прошлого раздела добавим еще несколько:

1) Критерий перестановок (Permutation test).

Этот критерий олицетворяет общую идею, которая, в частности, может использоваться для проверки однородности. Рассмотрим какую-то асимметричную статистику $T(x_1, \dots, x_n)$, которая, как мы считаем, должна быть большой при альтернативе и небольшой при гипотезе. Тогда мы можем рассмотреть критерий вида $\{T > c\}$, где c — какая-то константа, но не знаем, как найти фактический уровень значимости такого критерий.

Критерий предлагает нам представить, что среди возможных выборок есть только перестановки нашей выборки. Они равновероятны, поэтому вероятность $T > c$ есть

$$\frac{1}{n!} \sum_{\sigma \in S_n} I_{T(X_{\sigma(1)}, \dots, X_{\sigma(n)}) > c},$$

где S_n — множество всех перестановок. Значит, фактический уровень значимости на реализации x_1, \dots, x_n есть

$$\frac{1}{n!} \sum_{\sigma \in S_n} I_{T(x_{\sigma(1)}, \dots, x_{\sigma(n)}) > T(x_1, \dots, x_n)}.$$

Поскольку рассматривать все перестановки слишком трудоемко, достаточно выбрать некоторое количество N случайных перестановок и оценить фактический уровень значимости величиной

$$\frac{1}{N} \sum_{i=1}^N I_{T(x_{\sigma_i(1)}, \dots, x_{\sigma_i(n)}) > T(x_1, \dots, x_n)},$$

где σ_i — выбранные перестановки.

Для случая проверки однородности X_1, \dots, X_n и Y_1, \dots, Y_m в качестве тестовой статистики можно использовать, например, $|\bar{X} - \bar{Y}|$ или как в t-критерии

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}}}.$$

Вопрос 3. Что можно сказать про однородность выборок $(X_1, X_2) = (1, 9), Y_1 = 3$.

Метод достаточно прост и может быть легко запрограммирован, а может быть взят, например, `permTS` из пакета `perm`.

Упражнение 3. Массив `Salary` из пакета `wPerm` содержит информацию о зарплатах сотрудников частных или публичных университетов. Проверить гипотезу о том, что эти зарплаты одинаковы.

2) Ранговый критерий Манна-Уитни-Уилкоксона.

Этот критерий используется для проверки гипотезы однородности против альтернативы доминирования H_2 .

Для величин $X_1, \dots, X_n, Y_1, \dots, Y_m$ найдем ранги R_1, \dots, R_m величин $Y_{(1)}, \dots, Y_{(m)}$ в общем вариационном ряду. На их основе подсчитаем $V = R_1 + \dots + R_m$. При выполнении гипотезы эта статистика распределена также как $Z_1 + \dots + Z_m$, где Z_i — выборка без возвращения из чисел $\{1, \dots, n + m\}$. При больших n, m верна сходимость

$$\frac{V - (n + m + 1)m/2}{\sqrt{nm(n + m + 1)/12}} \xrightarrow{d} Z \sim \mathcal{N}(0, 1),$$

при малых параметрах оценить фактический уровень значимости можно методом Монте-Карло. В R этот критерий задается `wilcox.test`.

Вопрос 4. Убедиться, что при выполнении гипотезы $EV = (n + m + 1)m/2, DV = nm(n + m + 1)/12$.

Многократные повторные выборки

В ряде случаев мы сталкиваемся с зависимостью следующего характера — в выборках $X_{1,1}, \dots, X_{1,n}, \dots, X_{k,1}, \dots, X_{k,n}$ величины $X_{i,j}$ зависимы между собой при разных i . Типичным примером этого может быть исследование каких-либо объектов в изменяющихся условиях. Например, мы изучали параметры цветов на грядке до применения удобрений и после. При этом наблюдения до и после зависимы, поскольку растения одни и те же.

Итак, многократными повторными выборками мы будем называть $X_{i,j}, i \leq k, j \leq n$, т.ч. векторы $(X_{1,j}, \dots, X_{k,j})$ независимы и одинаково распределены.

Рассмотрим случай парных повторных выборок ($k = 2$). Положим $Z_j = X_{2,j} - X_{1,j}$ и предположим, что Z_j имеют непрерывные распределения, причем $Z_j = \theta + \varepsilon_j$, где $P(\varepsilon_j < 0) = P(\varepsilon_j > 0) = 1/2$. Гипотеза H_0 будет заключаться в том, что $\theta = 0$, альтернатива $H_1 : \theta \neq 0$. Тогда используют следующие критерии:

1) Критерий знаков.

Рассмотрим S — число $Z_i > 0$. Тогда при гипотезе S будет иметь биномиальное $n, 0.5$ распределение, а при альтернативе биномиальное с другим параметром p . Эту гипотезу можно проверить асимптотически при $n \rightarrow \infty$, а можно в явном виде.

2) Критерий знаковых рангов Уилкоксона.

При дополнительном предположении симметричности распределений Z_j , справедлив критерий Уилкоксона, основывающийся на статистике

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

где $U_i = I_{Z_i > 0}, R_i$ — ранг $|Z|_i$ в ряду $|Z|_1, \dots, |Z|_n$. При выполнении гипотезы $(T - ET)/\sqrt{DT}$ сходится к

нормальному распределению.

В R критерий Уилкоксона задан все той же функцией `wilcox.test`, но с указанием `paired = TRUE`.

Упражнение 4. Проверить гипотезу для Z_i а) $Z_i \sim \mathcal{N}(0, 1)$, б) $Z_i \sim \mathcal{N}(0.5, 3)$, в) $Z_i + 1 \sim \exp(1)$.

Нормальные выборки

1) Критерии Фишера и Стьюдента.

Классический подход предлагал сравнивать однородность нормальных выборок путем применения

а) критерия Фишера (F -критерия) для проверки равенства дисперсий, основанном на том, что статистика

$$\frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 / (n_1 - 1)}{\sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 / (n_2 - 1)}$$

при равенстве дисперсий имеет распределение Фишера;

б) критерия Стьюдента или t -критерия, основывающемся на статистике

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2}{n_1 + n_2 - 2}}},$$

которая при равенстве средних и дисперсий имеет распределение Стьюдента. Существенным недостатком классического подхода заключается в том, что критерий Стьюдента значимо опирается на равенство дисперсий, в котором мы не можем быть уверенным наверняка.

В R критерии Стьюдента и Фишера заданы функциями `t.test` и `var.test`.

Можно убрать шаг проверки равенства дисперсий, используя более сложные подходы, в частности, подход Уэлча (Welch). Этот критерий рассматривает статистику

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\sum_{i=1}^{n_1} (X_i - \bar{X})^2 / n_1 + \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 / n_2}}$$

и приближенно находит ее распределение. По умолчанию `t.test` использует именно критерий Уэлча для случая различных дисперсий.

Упражнение 5. Испытать `t.test` для сравнения средних а) $\mathcal{N}(0, 1)$ и $\mathcal{N}(0, 3)$, б) $\mathcal{N}(0, 1)$ и $T(0, 1, 2, 3, 0.01)$, в) $\mathcal{N}(0, 1)$ и $T(0, 1, 2, 3, 0.05)$, г) $T(0, 1, 2, 3, 0.05)$ и $T(0, 1, 2, 3, 0.05)$ д) t_7 и t_7 , где t — критерий Стьюдента, на выборках размера 100, 500, 1000.

Здесь $T(a_1, \sigma_1, a_2, \sigma_2, p)$ — смесь $\mathcal{N}(a_1, \sigma_1^2)$ и $\mathcal{N}(a_2, \sigma_2^2)$, где первое распределение берется с вероятностью $1 - p$, а второе — p .

2) Дисперсионный анализ ANOVA.

Этот метод обобщает пункт 1) на случай k выборок. В предположении равенства дисперсий, положим $\bar{X}_{\cdot, j}$ — среднее по выборке j , $\bar{X}_{\cdot, \cdot}$ — среднее по всем выборкам

$$SSTR = \sum_{j=1}^k n_j (\bar{X}_{\cdot, j} - \bar{X}_{\cdot, \cdot})^2, \quad SSE = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{i, j} - \bar{X}_{\cdot, j})^2.$$

При этом $SSE \sim \chi_{n_1 + \dots + n_k - k}^2$ из леммы Фишера, при выполнении гипотезы $SSTR \sim \chi_{k-1}^2$ и эти величины независимы. Значит если $MSTR = SSTR / (k - 1)$, $MSE = SSE / (n - k)$, то $F = MSTR / MSE$ будет иметь распределение Фишера с $k - 1, n - k$ степенями свободы при выполнении гипотезы и будет принимать большие значения иначе. Так называемые ANOVA table, иллюстрирующие данные анализа ANOVA, имеют вид

$$\begin{array}{cccc} SSTR & MSTR & F & p\text{-value} \\ SSE & MSE & & \end{array}.$$

В R получить такую таблицу можно следующим образом. Запишем все выборки в один вектор X подряд. В другой вектор I на место i запишем номер выборки, к которой принадлежит X_i . Тогда `anova(lm(X ~ I))` построит таблицу ANOVA.

Упражнение 6. Сгенерировать 5 выборок размера 100 из $\mathcal{N}(a_i, 1)$ или $R[a_i - 1/2, a_i + 1/2]$ при а)

$a_1 = \dots = a_5$, б) $a_1 = a_2 = a_3 = a_4 = 0$, $a_5 = 0.5$ в) $a_i = i/10$ и исследовать их с помощью ANOVA. Повторить эксперимент в случае, если у последней выборки дисперсия 2.

Полезно также использовать метод Тьюки, позволяющий оценить, какие из переменных значимо отличаются. Интерфейс этого метода в R $TukeyHSD(aov(X \sim I))$. Здесь I как и прежде вектор, идентифицирующий принадлежность элементов вектора X к различным выборкам. Для метода Тьюки он должен быть в формате factor, это можно сделать функцией factor(I).

Упражнение 7. Применить метод Тьюки к массиву данных warpbreaks для того, чтобы понять, какие уровни натяжения (tension) значимо отличаются с точки зрения количества разрывов волокна (breaks).

И еще немного рангов

В однофакторной модели, рассмотренной нами в части 2) предыдущего раздела, можно использовать и ранговые критерии. Если $X_{i,j} = \mu_i + \varepsilon_{i,j}$, $i \leq k$, $j \leq n_i$, где $\varepsilon_{i,j}$ имеют одинаковое непрерывное (но уже необязательно нормальное) распределение, то для проверки гипотезы $\mu_1 = \dots = \mu_k$ можно использовать критерий Краскелла-Уоллеса:

1) Критерий Краскелла-Уоллеса. Положим $R_{i,j}$ — ранг $X_{i,j}$ в общем вариационном ряду. Тогда пусть $S_i = \sum_{j=1}^{n_i} R_{i,j}$, $R_{i,\cdot} = S_i/n_i$, $R_{\cdot,\cdot} = \sum R_{i,j}/N$, $N = n_1 + \dots + n_k$. Критерий Краскелла-Уоллеса основан на статистике

$$W = \frac{12}{N(N+1)} \sum_{i=1}^k n_i (R_{i,\cdot} - R_{\cdot,\cdot})^2,$$

сходящейся к χ_{k-1}^2 при $n_i \rightarrow \infty$.

2) Критерий Джонкхиера-Терпстры.

Для альтернативы $\mu_1 \leq \mu_2 \leq \dots \leq \mu_k$ используют критерий, задаваемый статистикой

$$\sum_{r < s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} I_{X_{i,r} < X_{j,s}}.$$

В R этот критерий задан JonckheereTerpstraTest из пакета DescTools, а критерий Краскелла-Уоллеса функцией kruskal.test из пакета stats.

Ответы на вопросы

1. Если $F_X(x) \geq F_Y(x)$, то $F_X^{-1}(x) \leq F_Y^{-1}(x)$. Тогда $X_1 = F_X^{-1}(R) \leq F_Y^{-1}(R) = Y_1$, где $R \sim R[0, 1]$, что в силу метода обратной функции и дает требуемое.

2. Правдоподобие для нашей задачи имеет вид

$$L = \prod_{i=1}^k \prod_{j=1}^m p_{i,j}^{\nu_{i,j}},$$

где $\nu_{i,j}$ — количество наблюдений вида j в i -й выборке, $p_{i,j}$ — соответствующая вероятность, $\sum_j p_{i,j} = 1$. Наше параметрическое семейство в общем случае имеет размерность $k(m-1)$, поскольку последняя вероятность $p_{i,m}$ определяется остальными при каждом i . ОМП, как мы убедились во втором семинаре, имеет вид

$$\hat{p}_{i,j} = \nu_{i,j} / \nu_{i,\cdot}, \quad \nu_{i,\cdot} = \sum_{j=1}^m \nu_{i,j}.$$

При гипотезе параметров $m-1$, правдоподобие имеет вид

$$L = \prod_{j=1}^m p_{i,j}^{\nu_{i,j}}, \quad \nu_{i,j} = \sum_{i=1}^k \nu_{i,j},$$

а ОМП $\hat{p}_{i,j} = \nu_{i,j}/n$. Остается поделить полученные правдоподобия и применить критерий отношения правдоподобий.

3 Величина $|\bar{X} - \bar{Y}|$ при данной выборке (и перестановке, меняющей местами 1 и 9) принимает значение 1.5, а при остальных перестановках 5, 5, 7 и 7. Следовательно, фактический уровень значимости 1/3.

4 В силу представления $V = Z_1 + \dots + Z_m$, имеем $\mathbf{E}V = \mathbf{E}Z_1 + \dots + \mathbf{E}Z_m = m(n+m+1)/2$. Математическое ожидание каждой из Z_i очевидно равно $(n+m+1)/2$, поскольку Z_i равновероятно принимает значения от 1 до $n+m$.

С дисперсией ситуация обстоит чуть сложнее, поскольку $\mathbf{D}V = \mathbf{D}Z_1 + \dots + \mathbf{D}Z_m + 2 \sum_{i>j} \text{cov}(Z_i, Z_j) = m\mathbf{D}Z_1 + m(m-1)\text{cov}(Z_1, Z_2)$. При этом

$$\mathbf{E}Z_1^2 = \frac{1}{n+m} \sum_{i=1}^{n+m} i^2 = \frac{(n+m)(n+m+1)(2n+2m+1)}{6(n+m)} = \frac{(n+m+1)(2n+2m+1)}{6},$$

$$\begin{aligned} \mathbf{E}Z_1Z_2 &= \frac{1}{n+m} \frac{1}{n+m-1} \sum_{i \neq j} ij = \frac{1}{(n+m)(n+m-1)} \left(\sum_{i=1}^{n+m} i \sum_{j=1}^{n+m} j - \sum_{i=1}^{n+m} i^2 \right) = \\ &= \frac{((n+m)(n+m+1))^2/4 - (n+m)(n+m+1)(2n+2m+1)/6}{(n+m)(n+m-1)}. \end{aligned}$$

Итого

$$\begin{aligned} &\frac{m(n+m+1)(2n+2m+1)}{6} + \frac{m(m-1)(n+m+1)(3(n+m)(n+m+1) - 2(2n+2m+1))}{12(n+m-1)} - \\ &\frac{m^2(n+m+1)^2}{4} = \frac{mn(n+m+1)(2n+2m+1)}{6(n+m-1)} + \frac{m(n+m+1)^2((n+m)(m-1) - (n+m-1)m)}{4(n+m-1)} = \\ &\qquad\qquad\qquad \frac{nm(n+m+1)}{12}. \end{aligned}$$