Занятие шестое. Критерии однородности или как заметить отличие

Старые знакомые

Мы будем рассматривать две задачи: проверку гипотезы однородности — по нескольким выборкам $X_{1,1}, ..., X_{1,n_1}, X_{2,1}, ..., X_{2,n_2}, ..., X_{k,1}, ..., X_{k,n_k}$, $n_1 + ... + n_k = n$, проверить гипотезу о том, что выборки из одного распределения.

При этом будет изучать две альтернативы — общую H_1 : выборки из разных распределений или доминирования $H_2 : F_1(x) \geq F_2(x) ... \geq F_k(x)$ 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 $k = 2$ и модификации известных нам критериев.

1) Критерий Смирнова.

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

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

Оказывается, что $\sqrt{n_1 n_2} n_1 + n_2 D_n \to K \sim K(x)$, где $K(x)$ — ф.р. Колмогорова, $n_1, n_2 \to \infty$, откуда получаем критическое множество

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

В случае альтернативы доминирования H_2 рассматривают

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

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

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

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

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

где $\hat{H}_{n_1,n_2}(x) = (n_1 \hat{F}_{1,n_1}(x) + n_2 \hat{F}_{2,n_2}(x))/(n_1 + n_2)$. 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.

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

3) Критерий хи-квадрат.

Однородность нескольких выборок можно проверять с помощью критерия хи-квадрат. Для этого положим Δ_1,...,Δ_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

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

Для бернуллиевских выборок он реализован функцией prop.test. В этом случае может быть также ис-

пользован явный критерий Фишера fisher.test.

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

1) Критерий перестановок (Permutation test)

Этот критерий олицетворяет общую идею, которая, в частности, может использоваться для проверки однородности. Рассмотрим какую-то асимметричную статистику $T(x_1, ..., x_n)$, которая, как мы считаем, должна замечать 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

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

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

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

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

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

где σ_i — выбранные перестановки.

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

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

Вопрос 2. Проверить однородность выборок $(X_1, X_2, Y_1) = (1, 9, 3)$.

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

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

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

Многократные повторные выборки

В ряде случаев мы сталкиваемся с зависимостью следующего характера — в выборках

$X_{1,1},...,X_{1,n},...,X_{k,1},...,X_{k,n}$ величины $X_{i,j}$ зависимы между собой при разных i. Типичным примером этого может быть исследования 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 $X_{i,j}$, $i \le k$, $j \le n$, т.ч. векторы $(X_{1,j},...,X_{k,j})$ 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 справедливы 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 $n \to \infty$, а можно в явном виде.

2) Критерий знаковых рангов Уилкоксона.

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

$$T = R_1 U_1 + ...R_n U_n,$$

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

В R критерий Уилкоксона задан все той же функцией wilcox.test, но с указание 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 а) критерия Фишера (F-критерия) для проверки равенства дисперсий, основанном на том, что статистика

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

при равенстве дисперсий имеет распределение Фишера; б) критерия Стьюдента или t-критерия, основывающемся на статистике

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

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

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

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 обобщает пункт 1) на случай k выборок. В предположении равенства

дисперсий, положим $\overline{X}_{\cdot,j}$ — среднее по выборке j, $\overline{X}_{\cdot,\cdot}$ — среднее по всем выборкам

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

При этом $SSE \sim \chi^2_{n_1+...+n_k-k}$ из леммы Фишера, при выполнении гипотезы $SSTR \sim \chi^2_{k-1}$ 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

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

В R получить такую таблицу можно следующий 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 $anova(lm(X \sim I))$ 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) $a_1 = ... = a_5$, 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), можно использовать и ранговые критерии. Если $X_{i,j} = \mu_i + \varepsilon_{i,j}$, $i \le k$, $j \le n_i$, где $\varepsilon_{i,j}$ имеют одинаковое непрерывное (но уже необязательно нормальное) распределение, то для проверки гипотезы $\mu_1 = ...\mu_k$ можно использовать критерий Краскелла-Уоллеса:

1) Критерий Краскелла-Уоллеса.

Положим $R_{i,j}$ — ранг $X_{i,j}$ в общем вариационном ряду. Тогда пусть $S_i = \sum_{j=1}^{n_i} R_{i,j}$, $R_{i,\cdot} = S_i/n_i$, $R_{\cdot,\cdot} = \sum R_{i,j}/N$, $N = n_1 + ... + n_k$. 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^2_{k-1} при $n_i \to \infty$.

2) Для альтернативы $\mu_1 \le \mu_2.. \le \mu_k$ используют критерий Джонкхиера-Терпстры, 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-Уоллиса функцией kruskal.test из пакета stats.