3 Проверка однородности

3.1 Теория

Задача проверки однородности двух выборок состоит в проверке гипотезе $H_0 : F = G$, где выборка $(X_1, \dots, X_n)$ имеет структуру $(Y_1, \dots, Y_{n_1}, Z_1, \dots, Z_{n_2})$, где $Y_i \sim F$, $Z_i \sim G$. Мы не касаемся так называемых парных повторных наблюдений, где Y_i и Z_i могут быть зависимыми.

Рассмотрим несколько подходов

1. Критерий хи-квадрат (для дискретных данных реализован в scipy). Данные дискретизируются, подсчитываются количества $\nu_{i,j}$ попадания i-й выборки в j-й интервал, вводится статистика

$$\sum_{i=1}^{2}\sum_{j=1}^{k}\frac{\left(\nu_{i,j}-\frac{n_i\nu_{\cdot,j}}{n}\right)^2}{\frac{n_i\nu_{\cdot,j}}{n}} > y_{1-\alpha}, \quad \nu_{\cdot,j} = \nu_{1,j}+\nu_{2,j},$$

где y – квантиль χ^2_{k-1}.

2. Критерий Смирнова (реализован в scipy) имеет вид

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\sup|\widehat{F}_{n_1}-\widehat{G}_{n_2}| > k_{1-\alpha},$$

где k – квантиль распределения Колмогорова.

3. Критерий Стивенса-Шольца (он же k-выборочный критерий Андерсона-Дарлинга) реализован в scipy) имеет вид

$$\frac{n_1 n_2}{n_1+n_2}\int\frac{(\widehat{F}_{n_1}-\widehat{G}_{n_2})^2}{\widehat{H}_{n_1,n_2}(x)(1-\widehat{H}_{n_1,n_2}(x))} > A_{1-\alpha},$$

где A – квантиль распределения Андерсона-Дарлинга, $\widehat{H}$ – ЭФР объединенной выборки.

4. Критерий Баумгартнера-Вейсса-Шиндлера предлагает рассматривать статистику

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

где R_i – ранги (упорядоченные по возрастанию) первой выборки, S_i – ранги (упорядоченные по возрастанию) второй выборки в общем вариационном ряду. Далее предлагают сравнивать ее с квантилями распределения Андерсона-Дарлинга, которые, опять же, можно определять методом Монте-Карло.

5. t-критерий основан на статистике

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{\overline{Y}-\overline{Z}}{\sqrt{\frac{n_1 S_Y^2+n_2 S_Z^2}{n_1+n_2-2}}}$$

При верной гипотезе она сходится к $\mathcal{N}(0,1)$ распределению. Эту статистику используют для критерия однородности с альтернативой доминирования $F \le G$, то есть $F(x) \le G(x)$ при всех x, причем $F(x_0) < G(x_0)$ для некоторого x_0. В Python он есть здесь.

6. Критерий Манна-Уитни-Уилкоксона базируется на статистике

$$\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\left(I_{Y_i\ge Z_j}-\frac{1}{2}\right).$$

При выполнении гипотезы статистика при нормировке $\sqrt{(n_1+n_2)n_1n_2/12}$ сходится к величине с $\mathcal{N}(0,1)$ распределением, $n_1, n_2 \to \infty$. Отсюда получается соответствующий критерий для проверки гипотезы однородности с альтернативой доминирования.

Опишем ряд подходов, пригодных для m выборок.

1. Критерий хи-квадрат для m выборок имеет вид

$$\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\left(\nu_{i,j}-\frac{n_i\nu_{\cdot,j}}{n}\right)^2}{\frac{n_i\nu_{\cdot,j}}{n}} > y_{1-\alpha}, \quad \nu_{\cdot,j}=\nu_{1,j}+\nu_{2,j},$$

где y – квантиль χ^2_{k-1}.

2. Критерий Стивенса-Шольца:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{n_i}{n}\int\frac{(\widehat{F}_{n_i}-\widehat{H})^2}{\widehat{H}_n(x)(1-\widehat{H}_n(x))}d\widehat{H}_n > A_{1-\alpha},$$

где $\widehat{H}$ – ЭФР объединенной выборки. Критерий реализован в Python здесь.

3. Критерий Краскелла-Уоллиса (реализован в Python здесь):

$$\sum_{i=1}^{m} n_i\left(\overline{R}_i-\frac{n+1}{2}\right)^2,$$

где $\overline{R}_i$ – среднее арифметическое рангов i-й выборки. При выполнении гипотезы статистика сходится к величине с распределением хи-квадрат с $m-1$ степенями. Отсюда получается соответствующий критерий для проверки гипотезы однородности с альтернативой, что хоть для одной пары выборок выполнена альтернатива доминирования.

3.2 Задачи

1. Рассмотрим t-критерий и критерий Манна-Уитни: применим их для сравнения однородности двух выборок из распределения из а) $\mathcal{N}(0,1)$ распределения и $\mathcal{N}(\mu,1)$ распределения, б) распределения $Laplace(0,1)$ и $Laplace(\mu,2)$ распределения, взяв размеры выборок равные а) 10, б) 50, в) 100. Используйте разные виды критериев, меняя настройки: для t-критерия equal_var и permutations, для Манна-Уитни – exact и asymptotic method. Построить график мощности всех версий критериев в зависимости от μ.

2. Реализовать критерий Баумгартнера-Вейсса-Шиндлера в перестановочной версии. Проверить его работу на а) $R[0,1]$ и $R[0,1]$ выборках б) $R[0,1]$ и $R[0.1,1.1]$ выборках. Сравнить с встроенным критерием Стивенса-Шольца (scipy.stats.anderson_ksamp).

3. Сравнить (построив ЭФР p-value) критерии Манна - Уитни, критерий Смирнова (scipy.stats.ks_2samp), критерий Стивенса-Шольца (scipy.stats.anderson_ksamp) и BWS на примере следующих модельных данных:

   (a) $X_i, Y_j \sim N(0,1)$,
   (b) $X_i \sim N(0,1), Y_j \sim N(0.3,1)$,
   (c) $X_i \sim N(0,1), Y_j \sim N(0,3/2)$,
   (d) $X_i \sim N(0,1), Y_j \sim t_k$, где t_k – распределение Стьюдента с k степенями свободы,
   (e) $X_i \sim N(0,1), Y_j$ – центрированная нормированная сумма k независимых с.в. из равномерного распределения $R[-1,1]$.

Размер выборки в каждом случае выбирать так, чтобы он был поменьше среди тех, когда часть критериев замечает разницу.

4. * Проверить на однородность k выборок, используя критерии Стивенса-Шольца, хи-квадрат и Краскелла-Уоллиса:

   (a) $X_{i,j} \sim N(0, 2+i)$, $i \leq 4$, $j \leq n$;

   (b) $X_{i,j} \sim N(i/4, 1)$, $i \leq 4$, $j \leq n$;

   (c) $X_{i,j} \sim t_{3+i}$, $i \leq 4$, $j \leq n$, где t – распределение Стьюдента.