Прикладная статистика

П. С. Рузанкин

Глава 1

Ранговые критерии

1.1 Критерий Манна–Уитни (Mann-Whitney U-test, Wilcoxon rank sum test)

Рассматриваются две независимые выборки: $X_1, ..., X_m$ из распределения F и $Y_1, ..., Y_m$ из распределения G. Для простоты мы будет считать распределения F и G непрерывными.

Нулевая гипотеза H_0: $F = G$.
Альтернативная гипотеза: $F \neq G$.

Если в "смешанной" выборке $X_1, ..., X_m,\ Y_1, ..., Y_n$ упорядочить все наблюдения по возрастанию и пронумеровать их от 1 до $m + n$, то эти номера и будут рангами наблюдений. Обозначим ранги наблюдений $X_1, ..., X_m$ через $R_1, ..., R_m$, а ранги наблюдений $Y_1, ...Y_n$ через $S_1, ..., S_n$, соответственно.

Для критерия используются следующие статистики:

$$U = \sum_{i=1}^{m} R_i - \frac{m(m+1)}{2},$$

$$V = \sum_{j=1}^{n} S_j - \frac{n(n+1)}{2}.$$

Можно представить эти статистики и в другом виде. Поскульку

$$R_i = \#\{j : Y_j < X_i\} + \#\{k : X_k < X_i\} + 1,$$

$$\begin{aligned}\sum_i R_i &= \sum_i (\#\{j : Y_j < X_i\} + \#\{k : X_k < X_i\} + 1) \\ &= \sum_i \sum_j I(Y_j < X_i) + \sum_i \sum_k I(X_k < X_i) + m \\ &= \sum_i \sum_j I(Y_j < X_i) + \frac{m(m-1)}{2} + m \\ &= \sum_i \sum_j I(Y_j < X_i) + \frac{m(m+1)}{2},\end{aligned}$$

где I обозначает индикаторную функцию. Поэтому

$$U = \sum_i \sum_j I(Y_j < X_i)$$

и аналогично

$$V = \sum_i \sum_j I(X_i < Y_j).$$

Заметим также, что так как

$$\sum_i R_i + \sum_j S_j = \sum_i \sum_j \big(I(Y_j < X_i) + I(X_i < Y_j)\big) = \sum_i \sum_j 1 = \frac{(m+n)(m+n+1)}{2},$$

то $U + V = mn$.

Статистики U, V могут принимать все целые значения на отрезке $[0, mn]$.

Как вычисляется этот критерий. Задана вероятность ошибки первого рода α (обычно $\alpha = 0,05$). Находим интервал C_α такой, что $\mathbf{P}(U \in C_\alpha) \geq 1 - \alpha$ при выполнении H_0. Мы принимаем H_0 тогда и только тогда, когда $U \in C_\alpha$.

1.1.1 Асимптотическая нормальность статистик

Прямыми вычислениями можно проверить, что при выполнении H_0

$$\mathbf{E}U = \mathbf{E}V = \frac{mn}{2},$$

$$\mathbf{D}U = \mathbf{D}V = \frac{mn(m+n+1)}{12}.$$

Имеет место асимптотическая нормальность статистик при выполнении H_0: распределения величин $\frac{U-\mathbf{E}U}{\sqrt{\mathbf{D}U}}$ и $\frac{V-\mathbf{E}V}{\sqrt{\mathbf{D}V}}$ слабо сходятся к стандартному нормальному распределению. Строгое доказательство этого факта возможно методом моментов либо с помощью проекций Хаджека, применяемых в доказательствах асимптотической нормальности U-статистик (это отдельный термин, который мы не будем здесь рассматривать).

1.1.2 Оценка параметра сдвига (параметра разности положений)

В этом разделе мы будем рассматривать модель, в которой наблюдения $Y_1, ..., Y_n$ берутся из распределения F (распределения наблюдений $X_1, ..., X_m$), сдвинутого на θ, то есть $G(t) = F(t - \theta)$.

Оценка Ходжеса-Лемана (Hodges–Lehmann estimator) для θ:

$$\theta^* = \text{median}\{Y_j - X_i : i = 1, ..., m;\ j = 1, ..., n\}.$$

Доверительный интервал для θ (Bauer). Для числа Δ определим статистику $V(\Delta)$ по наблюдениям $X_1, ..., X_m,\ Y_1-\Delta, ..., Y_n-\Delta$ аналогично тому, как определяется статистика V по наблюдениям $X_1, ..., X_m,\ Y_1, ..., Y_n$. Затем вычислим интервал C_α такой, что при нулевой гипотезе $\theta = 0$ имеет место $\mathbf{P}(V \in C_\alpha) \geq 1 - \alpha$. Тогда доверительный интервал для θ есть $\{\Delta : V(\Delta) \in C_\alpha\}$.

1.2 Критерий Вилкоксона (Wilcoxon signed rank test)

Выборка $Z_1, ..., Z_n$ берется из распределения F. Для простоты мы будет считать это распределение непрерывным.

Нулевая гипотеза H_0: распределение F симметрично относительно 0.

Критерий может применяться для парных наблюдений $(X_1, Y_1), ..., (X_n, Y_n)$. Вэтом случае мы полагаем $Z_i = Y_i - X_i$.

R_i — ранг наблюдения Z_i в последовательности $|Z_1|, ..., |Z_n|$.

Статистика

$$T^+ = \sum_{i=1}^{n} R_i I(Z_i > 0).$$

T^+ принимает целые значения от 0 до $\frac{n(n+1)}{2}$.

Выберем интервал C_α такой, что $\mathbf{P}(T^+ \in C_\alpha) \geq 1-\alpha$ при выполнении H_0. Мы принимаем H_0 тогда и только тогда, когда $T^+ \in C_\alpha$.

1.2.1 Асимптотическая нормальность статистики

Можно вычислить, что при выполнении H_0

$$\mathbf{E}T^+ = \frac{n(n+1)}{4},$$

$$\mathbf{D}T^+ = \frac{n(n+1)(2n+1)}{24}.$$

При выполнении H_0 имеет место асимптотическая нормальность: распределение $\frac{T^+ - \mathbf{E}T^+}{\sqrt{\mathbf{D}T^+}}$ слабо сходится к стандартному нормальному распределению.