МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ

Кафедра функционального анализа и аналитической экономики

Критерий Манна-Уитни

Курсовая работа

Мартиневского Д. А.
студента 2 курса,
специальность «Математика и
информационные технологии»
Научный руководитель:
доцент, кандидат физ.-мат. наук
Сташулёнок С. П.

Минск 2020

Содержание

1. Статистический критерий **3**
2. Гипотеза однородности и её альтернативы **4**
2. Критерии Уилкоксона и Манна-Уитни **5**
3. Эквивалентность критериев Уилкоксона и Манна-Уитни **6**
4. Пример решения задачи **8**
5. Примеры использования критерия Манна-Уитни **9**
Литература **11**

1. Статистический критерий

Правило, позволяющее принять или отвергнуть гипотезу H на основе реализации выборки $x_1, \dots, x_n$, называется *статистическим критерием*. Обычно критерий задается при помощи *статистики критерия* $T(x_1, \dots, x_n)$ такой, что для нее типично принимать умеренные значения в случае, когда гипотеза H верна, и большие (малые) значения, когда H не выполняется.

При проверке гипотез с помощью критериев всегда присутствует возможность ошибочно отвергнуть гипотезу H, когда на самом деле она верна. В общем случае задается малое число α — вероятность, с которой мы можем позволить себе отвергнуть верную гипотезу H (например, $\alpha = 0.05$). Это число называют *уровнем значимости*. Исходя из предположения, что гипотеза H верна, определяется наименьшее значение $x_{1-\alpha}$, удовлетворяющее условию $P(T(x_1, \dots, x_n) \geq x_{1-\alpha}) \leq \alpha$. Такое $x_{1-\alpha}$ называют *критическим значением*: гипотеза H отвергается, если $t_0 = T(x_1, \dots, x_n) \geq x_{1-\alpha}$ (произошло маловероятное событие), и принимается — в противном случае.

При этом величина $\alpha_0 = P(T(x_1, \dots, x_n) \geq t_0)$ задает *фактический уровень значимости*. Он равен вероятности того, что статистика T (измеряющая степень отклонения полученной реализации от наиболее типичной) за счет случайности примет значение t_0 или даже больше. Фактический уровень значимости — наименьший уровень, на котором проверяемая гипотеза H принимается.

Пример. Предположим, что кто-то подбросил 10 раз монетку, и в 8 случаях она упала гербом вверх. Можно ли считать эту монетку симметричной?

Используем для описания эксперимента схему Бернулли, т.е. будем считать данные эксперимента реализацией выборки $X = (x_1, \dots, x_{10})$, где $x_i = 1$ (выпадет герб) с вероятностью θ и $x_i = 0$ (выпадет решка) с вероятностью $1 - \theta$. Проверим гипотезу $H: \theta = \frac{1}{2}$ на уровне значимости $\alpha = 0.05$ и вычислим α_0.

В качестве статистики T можно взять сумму $x_1 + \dots + x_n$. Тогда гипотезе $H: \theta = \frac{1}{2}$ противоречат значения, которые близки к 0 или n. Известно, что сумма $T = x_1 + \dots + x_n$ имеет биномиальное распределение:

$$P(T \geq k) = \sum_{i=k}^{n} C_n^i \theta^i (1-\theta)^{n-i}$$

Для $\theta = \frac{1}{2}$ правая часть этого выражения при $k = 8$ равна $\frac{45+10+1}{1024} \approx 0.055$ и при $k = 9$ равна $\frac{10+1}{1024} \approx 0.011$. Поэтому для $\alpha = 0.05$ наименьшим $x_{1-\alpha}$ удовлетворяющим условию $P(T(x_1, \dots, x_n) \geq x_{1-\alpha}) \leq \alpha$, будет 9. Поскольку полученное в эксперименте значение $t_0 = T(x_1, \dots, x_n) = 8 < 9$, на заданном уровне значимости гипотеза $H: \theta = \frac{1}{2}$ принимается.

С другой стороны, фактический уровень значимости $\alpha_0 = P(T \geq 8) \approx 0.055$, что всего на 0.005 превосходит заданный уровень: уже при $\alpha = 0.06$ гипотезу H следует отклонить.

На основе данных эксперимента нельзя уверенно принять или отвергнуть гипотезу H (хотя последнее представляется более правдоподобным). Следовало бы еще несколько раз подбросить монетку, чтобы прийти к более взвешенному заключению.

Вычисление фактического уровня значимости нередко позволяет избегнуть категоричных (и при этом — ошибочных) выводов, сделанных лишь на основе сравнения t_0 с критическим значением $x_{1-\alpha}$ найденным для формально заданного α.

2. Гипотеза однородности и её альтернативы

Пусть есть два набора наблюдений $x_1, \dots, x_n$ и $y_1, \dots, y_m$, будем рассматривать их как реализовавшиеся значения случайных величин $X_1, \dots, X_n$ и $Y_1, \dots, Y_m$ таких, что:

1. Случайные величины $X_1, \dots, X_n$ независимы и имеют общую функцию распределения $F(x)$.
2. Случайные величины $Y_1, \dots, Y_m$ независимы и имеют общую функцию распределения $G(x)$.
3. Обе функции F и G неизвестны, но принадлежат множеству всех непрерывных функций распределения.

Гипотеза однородности. $H_0: \forall\, x : G(x) = F(x)$.

В качестве гипотез, конкурирующих с H_0, выделим следующие **альтернативы:**

а) *Неоднородности* $H_1: \exists\, x : G(x) \neq F(x)$;
б) *Доминирования* $H_2: \forall\, x : G(x) \leq F(x),\ \exists\, x : G(x) < F(x)$;

в) *Правого сдвига* $H_3: \forall\, x : G(x) = F(x - \theta)$, где параметр $\theta > 0$ (эта альтернатива — частный случай предыдущей);

г) *Масштаба* $H_4: \forall\, x : G(x) = F(\frac{x}{\theta})$, где параметр $\theta > 0, \theta \neq 1$;

Причины, по которым следует рассматривать конкурирующие гипотезы, отличные от H_1, таковы:

1. С практической точки зрения бывает важно уловить отклонения от H_0 только определенного вида (например, наличие систематического прироста у y_i по сравнению с x_i);
2. За счет сужения (по сравнению с H_1) множества пар распределений (F, G), составляющих альтернативное подмножество, обычно удается построить более эффективные (чувствительные) критерии, настроенные на обнаружение отклонений от H_0 конкретного вида;

2. Критерии Уилкоксона и Манна-Уитни

Данные критерии применяются для проверки гипотезы однородности H_0 против *альтернативы доминирования* H_2, в частности, — против *альтернативы правого сдвига* H_3.

Пусть заданы две выборки $X = (x_1, \dots, x_n)$ и $Y = (y_1, \dots, y_m)$, такие что $\forall\, x \in X, y \in Y: x \neq y$, рассмотрим объединенную выборку $(x_1, \dots, x_n, y_1, \dots, y_m)$ и построим ее вариационный ряд, где под вариационным рядом понимается неубывающая последовательность из значений выборки.

Для элемента x_i рассмотрим значение R_i, которое будет равно номеру места (рангу) элемента x_i в построенном нами объединенном вариационном ряду.

Пусть $T = R_1 + \cdots + R_n$. Т.е. T – сумма номеров мест (рангов), которые занимают в объединенном вариационном ряду элементы выборки X. Критерий, основанный на ранговой статистике T, и есть критерий Уилкоксона.

Пусть $U = \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij}$, где $f_{ij} = \begin{cases} 1, \text{если } x_i < y_j \\ 0, \text{в противном случае} \end{cases}$. Т.е. U – общее число тех случаев, когда элементы выборки Y в объединенном вариационном ряду идут после элементов выборки X. Критерий, основанный на ранговой статистике U, и есть критерий Манна-Уитни.

Критерий, основанный на статистике Т, был предложен Ф. Уилкоксоном в 1945 г. для выборок одинакового размера и распространен на случай $n \neq m$ Х. Манном и Д. Уитни в 1947 г.

После того, как мы посчитали значение статистики U, необходимо определить критическое значение критерия для данных n и m. Если полученное значение U ***меньше либо равно*** табличного, то принимается ***альтернативная гипотеза***. Если же полученное значение U ***больше*** табличного, принимается ***нулевая гипотеза***.

3. Эквивалентность критериев Уилкоксона и Манна-Уитни

Покажем, что критерии Уилкоксона и Манна-Уитни эквивалентны, для этого покажем, что одно можно выразить из другого. Докажем, что верно равенство

$$\boldsymbol{T + U = nm + \frac{n(n+1)}{2}}.$$

Для начала заметим, что $\frac{n(n+1)}{2} = 1 + 2 + \cdots + n$. Преобразуем исходное равенство в равенство $(T - \frac{n(n+1)}{2}) + U = nm$. Рассмотрим его часть $T - \frac{n(n+1)}{2} = T - (1 + 2 + \cdots + n) = (R_1 + \cdots + R_n) - (1 + 2 + \cdots + n)$. Теперь заменим слагаемое $(R_1 + \cdots + R_n)$ на $\left(R_{(1)} + \cdots + R_{(n)}\right)$, где $\left(R_{(1)}, \dots, R_{(n)}\right)$ - вариационный ряд $(R_1, \dots, R_n)$, т.к. сумма не изменилась мы имеем право сделать такую замену.

$$T - \frac{n(n+1)}{2} = \left(R_{(1)} + \cdots + R_{(n)}\right) - (1 + 2 + \cdots + n) =$$
$$\left(R_{(1)} - 1\right) + \left(R_{(2)} - 2\right) + \cdots + (R_{(n)} - n).$$

Заметим, что каждое его слагаемое равно количеству элементов выборки Y идущих перед i-ым по значению элементом выборки X в объединенном вариационном ряду.

Действительно, если $R_{(1)}$ – номер места минимального элемента выборки X в построенном нами объединенном вариационном ряду, то $R_{(1)} - 1$ будет равно количеству элементов в объединенном вариационном ряду идущих перед ним, а т.к. мы выбрали минимальный элемент выборки X, то перед ним могут быть только либо только элементы выборки Y, либо ничего.

Если $R_{(2)}$ – номер места второго по значению элемента выборки X в построенном нами объединенном вариационном ряду, то $R_{(2)} - 1$ будет равно количеству элементов в объединенном вариационном ряду идущих перед ним, т.к. мы выбрали второй по значению элемент выборки X, следовательно перед ним есть ровно один элемент выборки X и чтобы получить количество элементов выборки Y надо отнять единицу, т.е. количество элементов выборки Y будет равно $R_{(2)} - 2$.

Если $R_{(n)}$ – номер места максимального элемента выборки X в построенном нами объединенном вариационном ряду, то $R_{(n)} - 1$ будет равно количеству элементов в объединенном вариационном ряду идущих перед ним, т.к. мы выбрали максимальный элемент выборки X, следовательно перед ним есть ровно $n - 1$ элементов выборки X и чтобы получить количество элементов выборки Y надо отнять их количество, т.е. количество элементов выборки Y будет равно $R_{(n)} - 1 - (n-1) = R_{(n)} - n$.

Тогда выражение $T - \frac{n(n+1)}{2}$ будет равно общему число тех случаев, когда элементы выборки Y в объединенном вариационном ряду идут перед элементом выборки X. Другими словами $T - \frac{n(n+1)}{2} = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}^*$, где $f_{ij}^* = \begin{cases} 1, \text{если } x_i > y_j \\ 0, \text{в противном случае} \end{cases} = \begin{cases} 0, \text{если } x_i < y_j \\ 1, \text{в противном случае} \end{cases}$ (т.к. $\forall\, x \in X, y \in Y: x \neq y$, мы можем упустить знак $=$ в неравенстве $x_i < y_j$).

Тогда,

$$\left(T - \frac{n(n+1)}{2}\right) + U = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}^* + \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} (f_{ij}^* + f_{ij}) = \sum_{i=1}^{n}\sum_{j=1}^{m} 1 = nm.$$

4. Пример решения задачи

Пусть даны две нормально распределенные ($\mu = 0, \sigma = 1$) выборки $X = (-1.75, -0.29, -0.93, -0.45, 0.51)$ и $Y = (-0.33, 0.09, 0.13, -0.24, -0.88)$. Требуется проверить гипотезу однородности для этих выборок.

Покажем, как найти статистику U двумя способами:

1. Будем использовать формулу
   $U = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}$, где $f_{ij} = \begin{cases} 1, \text{если } x_i < y_j \\ 0, \text{в противном случае} \end{cases}$
   Т.е. будем фиксировать элемент $x \in X$ и считать, сколько элементов $y \in Y$ строго больше него.
   Количество элементов из Y строго больших -1.75: 5
   Количество элементов из Y строго больших -0.29: 3
   Количество элементов из Y строго больших -0.93: 5
   Количество элементов из Y строго больших -0.45: 4
   Количество элементов из Y строго больших 0.51: 0
   $U = 5 + 3 + 5 + 4 + 0 = 17$

2. Будем использовать формулу
   $$T + U = nm + \frac{n(n+1)}{2} \Leftrightarrow U = nm + \frac{n(n+1)}{2} - T$$
   Найдем значение T, для этого построим объединенный вариационный ряд выборок X и Y (черным цветом обозначим элементы выборки X, красным – элементы выборки Y) и расставим ранги:

$$\begin{pmatrix} -1.75 & -0.93 & -0.88 & -0.45 & -0.33 & -0.29 & -0.24 & 0.09 & 0.13 & 0.51 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{pmatrix}$$

$T = 1 + 2 + 4 + 6 + 10 = 23$

$$U = nm + \frac{n(n+1)}{2} - T = 5 * 5 + \frac{5 * (5+1)}{2} - 23 = 25 + 15 - 23 = 17$$

Т.к. критическое значение для $n = 5$ и $m = 5$ равно 2 и $U = 17 > 2$, следовательно принимается нулевая гипотеза, т.е. выборки однородны.

5. Примеры использования критерия Манна-Уитни

Для получения результатов критерия Манна-Уитни будем использовать функцию из библиотеки SciPy для языка программирования Python. Будем работать с выборками из 30-ти элементов. Сгенерируем выборки 10 раз. Уровень значимости примем равным $\alpha = 0.05$.

Пример 1. Рассмотрим две нормально распределенные выборки с параметрами $(\mu = 0, \sigma = 1)$ и $(\mu = 0, \sigma = 2)$.

Номер	Значение статистики	p-значение	Результат
1	343	0.05768	Однородны
2	392	0.19763	Однородны
3	438	0.43249	Однородны
4	416	0.31020	Однородны
5	347	0.06483	Однородны
6	356	0.08343	Однородны
7	336	0.04667	Неоднородны
8	236	0.00079	Неоднородны
9	433	0.40363	Однородны
10	372	0.12594	Однородны

Пример 2. Рассмотрим две выборки, первая будет нормально распределенной выборкой с параметрами $(\mu = 0, \sigma = 1)$, вторая будет равномерно распределенной выборкой с параметрами $(a = -1, b = 1)$.

Номер	Значение статистики	p-значение	Результат
1	411	0.28461	Однородны
2	382	0.15915	Однородны
3	362	0.09789	Однородны
4	409	0.27466	Однородны
5	401	0.23667	Однородны
6	312	0.02103	Неоднородны
7	431	0.39222	Однородны
8	405	0.25529	Однородны
9	364	0.10310	Однородны
10	401	0.23667	Однородны

Пример 3. Рассмотрим две равномерно распределенные выборки с параметрами $(a = -2, b = 2)$ и $(a = -2.5, b = 2.5)$.

Номер	Значение статистики	p-значение	Результат
1	434	0.40937	Однородны
2	371	0.12290	Однородны
3	402	0.24125	Однородны
4	446	0.47936	Однородны
5	279	0.00585	Неоднородны
6	365	0.10578	Однородны
7	441	0.44999	Однородны
8	301	0.01406	Неоднородны
9	448	0.49115	Однородны
10	433	0.40363	Однородны

Литература

1. Лагутин М. Б., Наглядная математическая статистика
2. Ивченко Г. И, Медведев Ю. И., Математическая статистика
3. http://www.saburchill.com/IBbiology/downloads/002.pdf
4. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html