Функционал библиотек языка Python по применению метода Манна-Уитни для оценивания различий групп обучающихся

Векслер В.А.
Vitalv74@mail.ru
Саратовский национальный исследовательский государственный университет имени Н.Г. Чернышевского, Саратов, Россия

Аннотация. В статье рассматриваются особенности использования непараметрического теста Манна-Уитни для определения различий в несвязанных группах учащихся при проведении педагогических экспериментов. Особенности вычисления характеристик и примеры использования представлены на базе библиотек scipy.stats и pingouin языка программирования Python.

Ключевые слова: Педагогические измерения, программирование, python, scipy, pingouin, метод Манна-Уитни

В педагогическом исследовании зачастую изучаются не результаты отдельных обучающихся, а обобщенные данные, по которым требуется сделать выводы: отличие групп друг от друга по ряду признаков, значимость воздействие одной методики обучения на результаты в одной группе в отличии от другой методики в другой группе. Экспериментальное педагогическое исследование всегда связано с проверкой некоторой исходной гипотезы. В рамках исследования предполагаю две гипотезы: нулевую и альтернативную.

Нулевая гипотеза – это выдвинутое предположение об отсутствии каких-либо значимых изменений измеряемого параметра. Это вариант утверждения, которое исследователь хочет опровергнуть, обозначают ее как правило H0.

Альтернативная гипотеза – противоположное суждение, выдвинутое автором исследования экспериментальное предположение о значимости различий измеряемого параметра между сравниваемыми группами. Это то, что исследователь желает доказать в своем педагогическом эксперименте, ее стандартное обозначение H1.

Например, рассмотрим следующую задачу: В двух независимых (несвязанных друг с другом) группах обучающихся проводилось итоговое тестирование по химии с максимально возможным числом баллов 10. Результаты (набранные обучающимися баллы) представлены в списках. Можно ли с определенной достоверностью утверждать, что уровни освоения химии в этих группах различаются?

Списки с данными об оценивании (порядок данных не важен внутри групп):

```
group1=[5, 5, 6, 7, 7, 3, 5, 7, 8, 3, 10, 10, 8, 8, 9, 8],
group2 = [5, 6, 7, 3, 1, 3, 4, 4, 8, 3, 4, 7, 6, 7, 6].
```

Выдвинем две гипотезы:

– H0: Различие в уровне усвоения химии между обеими группами отсутствует.

– H1: Учащиеся группы 1 имеют более высокий уровень усвоения химии.

Гипотезы всегда носят строго статистический характер, это означает, что их истинность не может быть доказана с абсолютной достоверностью.

Из-за действия набора независимых от исследователя факторов вполне может оказаться, например, что гипотеза H1 будет отвергнута, хотя на самом деле изменения имеются.

Выделяют разновидности ошибок выбора статистической гипотезы по итогам исследования:

– ошибка 1-го рода – когда была отклонена нулевая гипотезаH0, хотя она на самом деле оказалась верна;

– ошибка 2-го рода – была принята альтернативная гипотеза H0, хотя она на самом деле неверна.

Надежность любой гипотезы должна быть установлена за счет величины, которая называется уровнем статистической значимости – это вероятность того, что была в рамках исследования допущена некорректность, проявляющаяся в виде ошибки 1-го рода.

В педагогических исследованиях как правило используются два уровня значимости: $p < 0,05$ и $p < 0,01$. Эти уровни численно означаю следующее, что надежность принятия альтернативной гипотезы H1 будет не менее 95% и 99%, соответственно. Чем будет ниже значение p, тем более неожиданными являются приведенные доказательства, тем более нелепой становится наша нулевая гипотеза. Если p-значение примет значения ниже заданного уровня значимости, тогда мы отвергаем нулевую гипотезу.

Одним из способов определения различий является критерий Манна-Уитни U – это непараметрический тест нулевой гипотезы о том, что распределение, лежащее в основе выборки первой группы, совпадает с распределением, лежащим в основе выборки второй группы.

Критерий предназначен для проверки в рамках исследования статистической достоверности различий между двумя независимыми наборами данных (результаты в группах обучающихся) по уровню признака, измеренного по шкале порядка (оценки, показатели, баллы и пр.).

Алгоритм расчета базируется на следующем:

– значениям выявленного признака приписываются ранги, при этом, ранжирование проводится одновременно по обеим наборам данных;

– потом по вычисленным рангам рассчитывается экспериментальное значение U-критерия, который должен отразить степень перекрытия интервалов значений рангов в двух наборах данных;

– чем меньше Uэксп (экспериментально выявленное статистическое значение), тем меньше факт перекрытия интервалов признаков рассматриваемых групп и, следовательно, существует большая вероятность того, что различие между исследуемыми наборами достоверно.

Для проверки гипотез Uэксп должно сопоставится с табличным критическим значением (выбираемым в зависимости от объемов наборов данных и статистической значимости): при Uэксп > Uкр принимается H0 как статистически достоверное, в противоположном случае – H1.

Выделен ряд ограничений для применения U-критерия:

1) объем данных по группам должен быть не менее трех, при этом возможно существование всего двух значений в одной из групп, но при этом во второй их должно быть не менее пяти;

2) объем данных в каждой из групп не должен превышать 60 (это связано с определённой ограниченностью рассчитанных таблиц критических значений).

Для выполнения вычислительных операций можно использовать следующий функционал внешних библиотек:

1. В модуле scipy.stats определена функция mannwhitneyu():

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

При использовании (без установки дополнительных опций) можно следовать следующим правилам:

В случае принятия альтернативной гипотезы должна подойти статистика (экспериментальное значение) - быть меньше табличного и pvalue (p значение) так же должна быть меньше 0.05.

Возможные случаи:

1. Если статистика большая и pvalue больше 0.05 - выбираем нулевую гипотезу

2. Если статистика подходит и pvalue меньше 0.05 - выбираем альтернативную гипотезу

3. Если статистика подходит, но pvalue больше 0.05 - нулевая гипотеза не выполняется, но и предложенная альтернативная тоже не выполняется (меняем группы местами - которые сравниваем).

4. Если статистика большая, но pvalue меньше 0.05 - значит нулевая не работает, но нужно поменять местами группы в функции и перепроверить.

Применим функцию для примера, рассмотренного выше:

```
from scipy.stats import mannwhitneyu
mannwhitneyu(group2, group1)
```

На второе место в перечне групп ставим данные, которые могут быть потенциально больше распределены.

Результат:

```
MannwhitneyuResult(statistic=62.5, pvalue=0.022834934525754674)
```

В группах 16 и 15 учащихся. Согласно табличным данным, критическое значение равно 70 при уровне значимости 0.05. Согласно результату, принимаем гипотезу H1 (альтернативную).

В случае если бы мы подали данные наоборот – получим следующие результаты:

```
mannwhitneyu(group1, group2)
```

Результат:

```
MannwhitneyuResult(statistic=177.5, pvalue=0.022834934525754674)
```

Полученные данные демонстрируют большое экспериментальное значение, намного больше критического табличного, но маленькое p значение подсказывает о том, что нулевую гипотезу H0 принять нельзя – нужно попробовать изменить последовательность групп.

2. В модуле pingouin так же определена функция mwu:
https://pingouin-stats.org/build/html/generated/pingouin.mwu.html

Кроме экспериментального и p значения функция возвращает RBC (rank-biserial correlation): ранговая бисериальная корреляция, разница между долей благоприятных доказательств за вычетом доли неблагоприятных доказательств. Значения варьируются от -1 до 1, причем отрицательные значения указывают на то, что вторая группа больше первой , а положительные – на то, что первая группа больше второй.

Так же приводится CLES (common language effect size): размер эффекта общего языка, доля пар, где первая группа больше, чем вторая.

Для нашего примера в зависимости от расстановки групп получим следующие результаты:

```
import pingouin as pg
group1=[5, 5, 6, 7, 7, 3, 5, 7, 8, 3, 10, 10, 8, 8, 9, 8]
group2 = [5, 6, 7, 3, 1, 3, 4, 4, 8, 3, 4, 7, 6, 7, 6]
pg.mwu(group1, group2)
```

Результат (см. таблицу 1)

Таблица 1. Результаты расчета.

	U-val	alternative	p-val	RBC	CLES
MWU	177.5	two-sided	0.022835	0.479167	0.739583

В варианте другой расстановки групп (результат в таблице 2):

Таблица 2. Результаты расчета.

	U-val	alternative	p-val	RBC	CLES
MWU	62.5	two-sided	0.022835	-0.479167	0.260417

Во второй таблице видим приемлемые результаты, благодаря которым можем принять альтернативную гипотезу.

Студентам педагогических специальностей может быть предложены следующие упражнения в рамках педагогических исследований:

Педагогическое исследование (анализ групп учащихся)

Представим, что исследователь провел измерения какого-либо критериального показателя у учащихся разных групп. Возникает вопрос: существует ли значимое различие между двумя наборами значений, перекрывающее разброс в пределах каждого из наборов?

Предварительно установите библиотеку scipy: pip install scipy

Упражнения:

1. Результаты тестирования по 30-бальной шкале для группы X и группы Y представлены в таблице 2. Сравнить эффективность двух методов обучения студентов в двух группах

Таблица 3. Результаты тестирования

X	18	10	7	15	14	11	13				
Y	15	20	10	8	16	10	19	7	15	14	29

Решение:

Нулевая гипотеза – различий нет, альтернативная гипотеза – вторая группа показала лучшие результаты (значит методика, применяемая в этой группе более эффективна).

```
from scipy.stats import mannwhitneyu
group1 = [18, 10, 7, 15, 14, 11, 13]
group2 = [15, 20, 10, 8, 16, 10, 19, 7, 15, 14, 29]
mannwhitneyu(group2, group1)
Результат:
MannwhitneyuResult(statistic=30.0, pvalue=0.4664292213967257)
```

Вывод:

В первой группе 7 человек, во второй 11. Согласно таблице критических значений Манна-Уитни – на пересечении 7 и 11 стоит – 16. Вычисленное значение 30, оно больше табличного – значит принимаем нулевую гипотезу

2. В двух группах у обучающихся измерили показатели уровня сформированности определенного умения по следующей шкале: «высокий», «достаточный», «недостаточный». Результаты приведены в таблицах 3 и 4. Можно ли утверждать, что в целом обучающихся группы 1 данные умения были сформированы лучше по сравнению с обучающимися группы 2?

H0: Различие в уровнях сформированности умений у испытуемых обеих групп отсутствует.

H1: У испытуемых группы 1 уровень сформированности умения выше.

Таблица 4. Данные группы 1.

Группа 1	
Фамилия	***Уровень умений***
Иванов	Высокий
Петров	Достаточный
Сидоров	Высокий
Семенов	Высокий
Кузнецова	Достаточный
Котова	Высокий
Лебедева	Достаточный
Шилов	Высокий
Мягкова	Высокий

Таблица 5. Данные группы 2.

Группа 1	
Фамилия	***Уровень умений***
Локтев	Недостаточный
Антонов	Достаточный
Колобов	Высокий
Шитова	Недостаточный
Махит	Достаточный
Розанер	Достаточный
Литвин	Недостаточный
Осин	Достаточный

Всем качественным градациям признака присвоить номера в порядке их расположения на шкале: «недостаточный» – 1, «достаточный» – 2, «высокий» – 3.

Решение:

```
group1 = [3,2,3,3,2,3,2, 3, 3]

group2 = [1,2,3,1,2, 2, 1,2]

mannwhitneyu(group2, group1)
```

Результат:

```
MannwhitneyuResult(statistic=12.0, pvalue=0.014627455199509639)
```

Вывод: В первой группе 9 человек, во второй 8. Согласно таблице критических значений Манна-Уитни на пресечении 6 и 5 стоит - 15. Наше вычисленное значение 12 - оно меньше табличного, значит принимаем альтернативную гипотезу.

Таким образом, рассмотренный критерий Манна-Уитни прост к использованию и может найти достаточно широкое применение в разнообразных педагогических исследованиях. Его целевое назначение - оценка различий между двумя независимыми выборками по уровню какого-либо признака, количественно измеренного (между малыми выборками).

Список литературы:

[1]. Васильева, Л.А. Статистические методы в биологии, медицине и сельском хозяйстве: учеб. пособие / Л.А. Васильева. – Новосибирск.: Институт цитологии и генетики СО РАН, 2007.-124 с.

[2]. Красильников, В.В. Высшая математика. Вероятность. Статистика. Исследование операций: учеб. пособие / В.В. Красильников. - Набережные Челны.: Печатный двор, 1996. - 225 с.

[3]. Стариченко Б.Е. Обработка и представление данных педагогических исследований с помощью компьютера/ Урал. гос. пед. ун-т. Екатеринбург, 2004. – 218 с.