Statistical Criterion: Mann-Whitney U-Test

Alexandr Yagin

E-mail: alexandr.an.yagin@gmail.com

Accepted for Publication: 2023
Published Date: October 2023

Abstract

The purpose of this research was to identify the need for the Mann-Whitney U-test in society, as such tests are primarily applied in statistics only. We make assumptions every day based on our calculations and often neglect to verify the hypotheses we put forward. This study aimed to establish the applicability of the Mann-Whitney test in everyday life by relying on class data.

The research consisted of several stages. The first stage involved identifying general areas of human activity suitable for the Mann-Whitney test. The second stage comprised searching for the necessary information to understand the operation of statistical tests, compiling a sample, and formulating hypotheses. The third stage involved selecting real-life examples to confirm these hypotheses.

Throughout the project, we aimed to conduct accurate analyses of student work, collect reliable information, select comparable periods for comparison, and draw conclusions based on these examples. This work aimed to challenge the stereotype that statistical tests are only useful in academic settings and demonstrate the strong connection between our world and the world of mathematics.

The study's result confirmed the theory that there is a connection between two initially distinct spheres of human activity.

Keywords: statistics, Mann-Whitney, human activity, data science

---

Theory and Algorithm:

The Mann-Whitney U-Test is a non-parametric criterion used to detect differences between samples. It serves as a non-parametric alternative to the t-test for independent samples.

Algorithm for the Mann-Whitney Criterion:

1) Formation of null (H0) and alternative (H1) hypotheses;

2) Calculation of empirical value:

a) Combining values and compiling a variation series from them;

b) Ranking of its value;

c) Determining the sample for each number (respectively each rank)

d) By rank (X, Y) division into 2 lines

e) Calculate separately the sum of ranks for the first and second samples ∑Rx and ∑ Ry;

f) Checking the correctness of calculating ranks through equality:

Rx + Ry = N(N+1)/2, where N- means the combined sample size;

g) Determination of the value of T as the greater of the sums of ranks (∑Rx and ∑Ry).

f) Calculation of the empirical value of Uem using the formula:

nx *ny + n(n+1)/2 - T, where nx is the volume of the first sample (X), ny is the volume of the second sample (Y), n is the volume of that sample with the larger sum of ranks, T is the larger of the sums ranks.

3) Definition of degrees of freedom as numbers nx and ny.

4) Determination of Ucr at the intersection of columns nx and ny in the table (the smaller of these numbers is taken in the row)

5) Plotting critical values on the significance axis and (significance areas are determined vice versa):

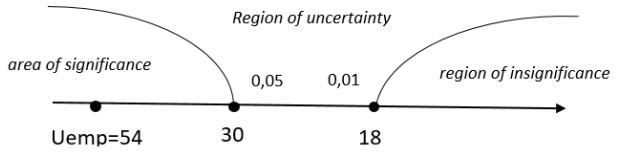

Figure 1.

Based on the belonging of the empirical value to the area, a statistical decision is made according to the rule:

If Uem ∈ OZ, hypothesis H1 is accepted.

If Uem ∈ OZ, then the hypothesis H0 is accepted.

If Uamp ∈ ONEOPR, it is considered that an incorrect decision may be made (it is necessary to increase the sample size or apply another criterion).

Comparison of statistical tests:

Each criterion used in statistics has its pros and cons. The main disadvantages are additional restrictions when compiling samples, which make it more difficult to come to the right conclusions. We propose to consider two criteria similar to **Mann-Whitney** to be able to compare them.

Comparison with Rosenbaum criterion:

The Rosenbaum Q test, like the Mann-Whitney test, is nonparametric and is used to study differences between two samples in terms of the level of any quantitatively measured trait.

This criterion has many restrictions on its application, for example:

each sample must contain at least 11 subjects; samples should have approximately the same size; there should be differences in the trait levels of each sample; Data should be presented on an ordinal scale. It is worth noting that this criterion may not be entirely accurate, and sometimes it may not reveal differences between samples, even when there are differences.

This means that after applying the Rosenbaum test, in some cases we have to use the Fisher test or the Mann-Whitney test, which are more powerful. For this reason, the time required to evaluate samples increases significantly, we have to re-analyze, and thus we spend additional resources to find the correct result.

Comparison with Student's t test:

The Student's t test is a parametric test that has slight similarities to the Mann-Whitney U test. It can be noted that these methods of data analysis are opposite.

However, it is quite difficult to compare parametric and non-parametric estimation methods, because each of them has its own disadvantages and advantages. In some situations it is more profitable to use nonparametric criteria, in others parametric ones. It all depends on the initial data and the task at hand.

Limitations of the Mann-Whitney test:

This Mann-Whitney U test has its limitations, the main ones being:

• the variables under consideration must be measured at least on an ordinal scale (ranked)

• calculated as the sum of indicators of pairwise comparison of elements of the first sample with elements of the second sample.

• At least 3 characteristic values are required in each sample.

• In the case where there are only two features in the first sample, it is necessary to balance the first sample with the second and therefore the second must have at least 5 features. Otherwise, the criterion will not be able to solve the problem.

Preface:

In applying the Mann-Whitney U-test algorithm, we decided to conduct two experiments:

Experiment 1: Comparing students' algebra knowledge levels at the beginning of the current academic year and the end of the previous one.

Experiment 2: Comparing the changes in the heights of our class students for the years 2020-2021 and 2021-2022.

Practical work #1:
Step 1: Null hypothesis H0 = the samples are identical. After a long vacation, students' knowledge has not significantly declined.

Alternative hypothesis H1 = there is a difference between the samples.

Step 2: Data collection (in this example, the results of students' assessments).

Algebra test results (for the 9th grade program) at the end of 9th grade:

Last name and first name of the student	Points (Maximum point - 15)
AA	7
PV	14
BN	7
YS	8
BD	14
AA	13
BA	10
DA	13
EB	10
SS	6

Figure 2.

Algebra test results (for the 9th grade program) at the beginning of 10th grade:

Last name and first name of the student	Points (Maximum point - 15)
AA	*8*
PV	*11*
BN	*10*
YS	*11*
BD	*8*
AA	*12*
BA	*8*
DA	*10*
EB	*10*
SS	*4*

Last name and first name of the student	*Results*	*Rank*

PV *BD*	*14*	*1,5*
DA *AA*	*13*	*3,5*
AA	*12*	*5*
PV *YS*	*11*	*6,5*
EB *DA* *BN* *BA* *EB*	*10*	*10*
YS *BA* *BD* *AA*	*8*	*14,5*
AA *BN*	*7*	*17,5*
SS	*6*	*19*
SS	*4*	*20*

Figure 3.

Step 3: Ranking the collected data in a single variation series.

Step 4: Finding the sum of ranks for the first and second samples.

Ranks for the previous academic year: 1, 5 * 2, 3.5 * 2, 10 * 2, 14.5, 17.5 * 2, 19 = 98.5

Ranks for the current academic year: 5, 6.5 * 2, 10 * 3, 14.5 * 3, 20 = 111.5

Step 5: Determining the larger sum of ranks from the two obtained values:

98.5 < 111.5; therefore, the rank sum of results for the current year is higher.

Step 6: Calculating the empirical value U_emp:

U_emp = 10 * 10 + 10 * (10 + 1)/2 - 111.5 = 43.5

Step 7: Determining degrees of freedom: nx = 10, ny = 10

Step 8: Selecting U_critical values (from the critical values table for a significance level of p = 0.05): U_critical1 = 23, U_critical2 (for p = 0.01) = 16.

Step 9: Plotting the obtained values U_emp, U_critical1, and U_critical2 on the significance axis, it becomes evident that the empirical value falls within the significance region. Therefore, we accept the alternative hypothesis, indicating that there is a difference between the two samples (results of the end-of-year assessments for the 9th grade program in the previous academic year and the current one).

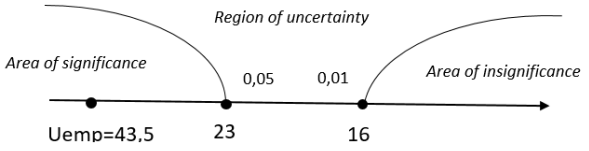

Figure 4.

Practical work #2:

As in the first practical work, we also compare the value (in this case, changes in student height) using the algorithm for applying the Mann-Whitney test.

H0=the height of students in our class has changed over these 3 years by the same value

H1=Changes in student height over these 3 years are not stable.

Last name and first name of the student	height in 2020, cm	height in 2021, cm	height in 2022, cm	ΔP 1	ΔP 2
KA	163	169	175	6	6
YS	163	167	170	4	3
MA	159	161	162	2	1
AA	179	182	182	3	0
BN	180	185	189	5	4
ЯА	173	176	179	3	3
БД	164	166	168	2	2
ЕБ	170	174	179	4	5
ПВ	177	180	184	3	4
ПД	168	168	168	0	0
CC	178	180	182	2	2

Фамилия и имя ученика	Δ	Ранг
KA	6	1,5
KA	6	1,5
EB	5	3,5
BN	5	3,5
YS	4	6,5
BN	4	6,5
EB	4	6,5
PV	4	6,5
YS	3	11
AA	3	11
YA	3	11
YA	3	11
PV	3	11
MA	2	16
BD	2	16
BD	2	16
SS	2	16
SS	2	16
MA	1	19

Figure 5.

Sum of ranks ΔP1=120;

Sum of ranks ΔP2=133.

120<133, the rank sum of changes in height for this year is greater than for the last.

Uamp=11*11+12*11/2-133=54.

Degrees of freedom: nx=11, ny=11.

Ucr1=30, Ucr2=18.

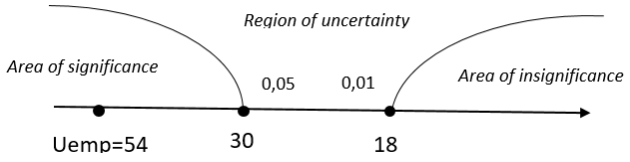

Figure 6.

Having analyzed the resulting significance axis, we come to the conclusion that H1 is significant, changes in student growth over three years are uneven.

Attachment (Critical Values of U, P=0.05)

n_1 \ n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2								0	0	0	0	1	1	1	1	1	2	2	2	2
3					0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
4				0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	13
5			0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
6			1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
7			1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
8		0	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
9		0	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
10		0	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
11		0	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
12		1	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
13		1	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
14		1	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
15		1	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
16		1	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
17		2	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
18		2	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
19		2	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
20		2	8	13	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127

Figure 7.

Conclusion:

At the beginning of our research, we hypothesized about the close connection between the world of statistics and our world. The result of the study consists of formulated hypotheses from two practical works. In the course of these works, we utilized the Mann-Whitney U-Test, conducted sampling, ranking, and determined whether the initial hypothesis was correct or not.

Based on our research and its results, we arrive at a general conclusion supported by the facts described above. Thus, we can confidently state that the "Mann-Whitney criterion is needed for our society and applicable in various areas of human activity."

Our research can be used as both theoretical and practical material. This is because our work includes examples of the criterion's application and an algorithm that anyone can use to rank and quickly formulate a hypothesis. Therefore, every member of society has the opportunity to apply the Mann-Whitney criterion in their work, simplify their lives, and have confidence in their assumptions.

In addition to theoretical application, my work can be used as a first step towards becoming familiar with the world of statistics or with statistical criteria of various types. However, given that our criterion began to be used not so long ago, we can say for sure that this topic will never exhaust itself, because statistics are a necessary part of our life, with the help of which we build a constantly progressing society. And people will always strive to develop, facilitate and simplify everything. Therefore, we hope that research in the field of statistical mathematics will not end and we believe that our work will become a “door” for someone into a new vast and very interesting world.

Acknowledgements:

The research was carried out with the help of mathematics teacher Lyazzat Bakytovna Kordabaeva and Aliya Musaeva, who helped in conducting the experiments

References:

1. The Wilcoxon-Mann-Whitney Test – An Introduction to Nonparametrics –: – With Comments on the R Program wilcox.test – by Frederick Ruland
2. Testing statistical criteria, E. Lehman
3. Mann H. B., Whitney D. R. On a test of whether one of two random variables is stochastically larger than the other. // Annals of Mathematical Statistics. - 1947. - No. 18. - P. 50-60.
4. Gubler E. V., Genkin A. A. Application of nonparametric statistics criteria in biomedical research. - L., 1973.
5. Sidorenko E. V. Methods of mathematical processing in psychology. - St. Petersburg, 2002.
6. Halmos and Savage, Application of the Radon - Nikodym theorem to the theory of sufficient statistics, Ann. Math. Stat. 20 (1948), 225-241.