European Journal of Sport Sciences
www.ej-sport.org

Reality of Common Mistakes When Using Mathematical Methods of Statistics in Analyzing and Processing Data in the Field of Sport Science

Nguyen Van Phuc, Ta Huu Hieu, and Nguyen Duc Thanh

ABSTRACT

Based on conventional research methods in the field of Sport Education, the results of the article have been systematized with formulas and algorithm systems commonly used in analysis and processing Sport Science Data, thereby identifying the errors that are often made when applying Mathematical Methods of Statistics, to be the basis for offering the application of mathematical theory of methods in analyzing and processing data, contributing to improving the quality of scientific research in school.

Keywords: Analyzing and Processing Data, Common Mistakes, Mathematical System, Scientific Research, Sports, Statistical Methods.

Published Online: May 29, 2023

ISSN: 2796-0048

DOI: 10.24018/ejsport.2023.2.3.79

N. V. Phuc
Bac Ninh Sport University of Viet Nam, Viet Nam
(e-mail: dhtdttbn@upes1.edu.vn)
T. H. Hieu
Bac Ninh Sport University of Viet Nam, Viet Nam
(e-mail: Hieulldc@gmail.com)
N. D. Thanh*
Director of Physical and Defense Education Center, HCMC of University of Technology and Education, VietNam
(e-mail: thanhnd@hcmute.edu.vn)

Corresponding Author

I. Introduction

In the research process, many methods are applied; it is impossible not to mention the Mathematical Methods of Statistics, a field of mathematics that is applied in most human activities, where: “data, summary and conclusions appear more and more in our daily work than any other form of mathematical analysis”; “in practical terms, probability theory and statistics have risen to be the subject that has the most applications and have become an essential tool for a wide variety of science and engineering careers”.

At Bac Ninh Sport University of Vietnam, the scientific research work is highly focused, it is set as one of the prerequisites to successfully carry out the tasks of physical education, sports training, etc.

Through a preliminary survey of a number of small-scale enterprise-level scientific research schemes, doctoral theses and master theses show that the application of Mathematical Methods of Statistics in the research process has many limitations: it has not yet followed any scientific procedure; with the same type of topic, there are different ways of applying the Mathematical Methods of Statistics or even making unfortunate mistakes. Therefore, it is necessary to propose the application of Mathematical Methods of Statistics in data analysis for some basic types of scientific research schemes in the field of sports education.

II. Methodology

In the process of researching the topic, the following methods are used: analysis and synthesis of documents; discussion interviews; pedagogical observations, Mathematical Methods of Statistics.

III. Results and Discussion

A. Identifying Statistical Formulas and Algorithms Commonly Used in Researching Sports Science

The necessity of Mathematical Methods of Statistics is confirmed based on the survey results of 72 grassroots-level scientific researches, 89 doctoral theses and 189 master theses. The survey results show that all schemes use this method in the research process, but the level of application is different.

Master's theses usually only focus on calculating characteristic parameters, comparing 2 observed average numbers, self-comparing, determining the level of reliability and announcing when making tests, calculating the rate of growth before and after the experiment, etc.

With grassroots-level scientific research schemes of lecturers and doctoral theses, besides using the formulas, the above algorithm also exploits many other algorithms, more deeply about how to use them such as the level of evaluation, conclusion of the problem. In terms of algorithms, the commonly used Mathematical Methods of Statistics formulas are presented in Table I.

TABLE I: Formulas and Algorithms Often Used in Grassroots Scientific Research Projects of Lecturers, Graduate Students and Postgraduate Students of Bac Ninh Sport University

No.	Formulas, algorithms commonly used	Grassroots-level scientific researchs (n=72)		Doctoral theses (n=89)		Master theses (n=189)	
		Quantity	%	Quantity	%	Quantity	%
1	Calculate the average number	52	72.22	89	100	152	80.42
2	Calculate the variance	52	72.22	89	100	152	80.42
3	Calculate the standard deviation	52	72.22	89	100	152	80.42
4	Calculate the coefficient of variation	29	40.28	54	60.67	107	56.61
5	Calculate the proportional error of the average number	18	25.00	35	39.33	70	37.04
6	Compare 2 observed average numbers	55	76.39	70	78.65	140	74.07
7	Use variance analysis method	12	16.67	39	43.82	58	30.69
8	Compare 2 collated average numbers	34	47.22	55	61.80	110	58.20
9	Calculate the level of absolute growth, level of relative growth, level of average growth	57	79.17	72	80.90	145	76.72
10	Compare two observation ratios using ‘t’ test, “F” test	11	15.28	32	35.96	64	33.86
11	Compare multiple observation ratios using “t” test	7	9.72	24	26.97	48	25.40
12	Compare 2 observation ratios using “2” test	25	34.72	54	60.67	108	57.14
13	Compare mutiple observation ratios using “2” test	15	20.83	50	56.18	100	52.91
14	Calcultate pair-correlation coefficient, multiple correlation coefficient.	42	58.33	63	70.79	117	61.90
15	Identify the regression line	5	6.94	22	24.72	44	23.28
16	Identify the proportion of influence	5	6.94	22	24.72	44	23.28
17	Identify the accuration of a distribution	9	12.50	26	29.21	52	27.51
18	Determine the reliability of the scale and test the suitability of factor analysis	15	20.83	24	26.97	27	14.28
19	Other statistical Algorithms	6	8.33	23	25.84	46	24.34

Table I shows that most of the research schemes of lecturers, graduate students and postgraduate students use the Mathematical Methods of Statistics.

Within 89 doctoral schemes, there are 89 schemes using the Mathematical Methods of Statistics, all of 89 schemes use the formulas to calculate the average number, variance and standard deviation. This accounts for 100% of all the schemes; 54 schemes using the formula for calculating variance accounts for 60.67%; 70 schemes used the “t” formula to compare two observed average numbers accounts for 78.65%; 24 schemes using the “t” formula for the purpose of comparing two observation rates (PA, PB) accounts for 26.97%, 55 schemes using the method of comparing two self-comparing average numbers accounts for 61.8%, there are 54 schemes using the -calculating formula, which accounts for 60.67%, 63 schemes uses the formula to calculate pair-correlation coefficient, multiple correlation coefficient, accounting for 70.79%. The same goes for grassroots-level scientific research of lecturers and graduate students, and in general, all projects use at least 2 of the above formulas

B. Analyzing the Current Situation of Common Mistakes When Applying the Mathematical Methods of Statistics in Analyzing and Processing the Data of Sport Science

We can easily see that not all of the above research schemes is applied the Mathematical Methods of Statistics correctly and scientifically, but only a very few topics are applied efficiently. This efficiency is reflected in the way of analyzing, the orientation to find the algorithm, the accuracy in using the algorithm, the caution in the conclusion, ... in order to prove the correctness and superiority of the methods and problems that the author proposed. The remaining schemes through the research we found still have some mistakes and limitations. The results of the survey are presented in Table II.

Table II shows that in many research schemes, there are still some shortcomings in the process of using the Mathematical Methods of Statistics methods, both in scientific research projects of lecturers, graduate students and postgraduate students. The mistakes and their causes are explained more specifically as follows:

1) Not define analytic variables clearly

The characteristic of science is weighing, measuring, counting; However, in the process of studying quantitative characteristics, some authors do not clearly explain what the measurands are nor how to interpret those variables to convince readers.

TABLE II: The Reality of Common Mistakes When Applying the Mathematical Methods of Statistics in Analyzing and Processing Data in the Field of Sport Science

No.	Type of Common Mistakes	Grassroots-level scientific researchs (n=72)		Doctoral theses (n=89)		Master theses (n=189)	
		Quantity	%	Quantity	%	Quantity	%
1	Not define analytic variables clearly	7	9.72	3	3.37	31	16.40
2	Not provide a specific measurement scale for each variable	4	5.56	5	5.62	24	12.70
3	Divide continuous variables into groups without explaining	4	5.56	3	3.37	14	7.41
4	Use the mean and standard deviation to describe a continuous variable that does not follow the normal distribution	2	2.78	0	0.00	13	6.88
5	Use the mean and standard error instead of the mean and standard deviation	3	4.17	2	2.25	10	5.29
6	Only conclude the difference based on the probability threshold P instead of testing by t test, F test, etc.	6	8.33	4	4.49	19	10.05
7	Not test the hypotheses in data analysis	7	9.72	2	2.25	23	12.17
8	Interpret the wrong results which is not statistically significant	3	4.17	0	0.00	14	7.41
9	Not report adjustment for multiple comparisons	4	5.56	1	1.12	20	10.58
10	Confusio between statistical meaning and actual meaning	2	2.78	1	1.12	17	8.99
11	Not determine if the data in the analysis of variance (ANOVA) and the t-test meet the statistical assumptions or not	3	4.17	3	3.37	-	-
12	Not describe the method used to analyze the difference between the two groups in the analysis of variance	4	5.56	3	3.37	-	-

2) *Not provide a specific measurement scale for each variable*

The scale of measurement is an important piece of information in statistical analysis. In the measurement process, some authors do not clearly distinguish three types of variables: nominal, hierarchical and continuous.

We must see that, at the lowest level, data are nominal, i.e. variables consisting of two or more categories (male or female, gifted or not, etc.), categorical but not having hierarchy (such as career characteristics), Hierarchical data includes high-low-order and rankable categories; for example, an individual athlete may win the title of grandmaster, level 1, level 2, ...or may not know the exact height of the athlete, but can know that the athlete belongs to the tall, medium or low group.

3) *Divide continuous variables into groups without explaining*

Sometimes, to simplify statistical analysis, researchers can divide continuous variables into several groups. For example, BMI can be divided into 4 groups: obese, pre-obese, normal and underweight. But there are also many cases where researchers divide groups arbitrarily and without following any convention, such as dividing age into group of 5 (6-10, 11-15, ...), when some other times dividing age into groups of 10 (10-19, 20-29, ...). Dividing a continuous variable into a discontinuous variable by clustering as above is an unscientific way.

It can be affirmed that the continuous variable is the variable that has the highest exact value compared to the categorical and identifier variables. Once the continuous variable is cut into several segments, it also means reducing the accuracy of the variable. There have been many theoretical and practical studies showing that such subgroups are unscientific and it can give results that are difficult to interpret, even causes erroneous.

4) *Use the mean and standard deviation to describe a continuous variable that does not follow the normal distribution*

Unlike nominal and funding variables that can be described by frequency or rate for each group, continuous variables can be described by a histogram. For variables that follow the normal distribution, there are two main parameters: the mean and the standard deviation. According to the definition of the normal distribution, about 67% are within ±1σ of the mean and approximately 95% is within ±2σ.

However, it is not well understood by many researchers that if a variable does not follow the normal distribution, the mean and standard deviation will not be significant. For variables that do not follow the normal distribution, the inferences about 67% and 95% are no longer true. In this case, the median and spread (which is a numerical measure of the dispersion of a data set) should be used.

5) Use the mean and standard error as descriptive statistics instead of the mean and standard deviation

Mean and standard deviation are statistical indicators that describe a sample of research (provided that the variables follow the rules of normal distribution). Standard error is an indicator that measures the accuracy of a population characteristic. Standard deviation reflects the variability or difference between individuals in a sample, while standard deviation reflects the variation in an index such as the mean between samples.

The standard error can be estimated from the standard deviation, by dividing the standard deviation by the square root of the sample size. Therefore, the standard error is always lower than the standard deviation. Many researchers do not understand the meaning of standard error, so they use it as a measure instead of standard deviation and make the variable to fluctuate less than normal.

6) Only report the results through the P-value instead of testing with the "t" test, "F" test, etc.

Many researchers, when reporting results via P-values, didn't analyze interpretatively and specify whether the difference is statistically significant at the threshold of probability greater or less than 5%, 2%, etc. This fact causes unnecessary misunderstandings.

7) Not test the hypothesis in the analysis

Any statistical analysis model is based on a number of assumptions. For example, the "t" test assumes that the variable must follow the rule of normal distribution, the variances of the two comparing-groups must be equal, the values in the variable must be independent of each other and so on. Similarly, in a linear regression model, in addition to assumptions such as t-test, there is also an assumption that the relationship between two dependent and independent variables must follow a linear function. In survival analysis, the assumption of proportionality is also important and if this assumption is not true, the results may also be false. If the variable does not meet these assumptions, the analysis results may be unreasonable or even wrong. Therefore, testing the hypothesis in analysis is very important.

8) Interpret the wrong results which is not statistically significant

Suppose a researcher compares the in situ jumping performance between two control and experimental groups at the time before the experiment and the results are not statistically significant ($P > 0.05$). The researcher must decide whether the non-difference means the two groups are similar (equivalent) or the data are incomplete in order to reach a more definite conclusion. It should be said that a result that is not statistically significant does not mean that the two groups are the same, but only that the null hypothesis cannot be rejected. The null hypothesis is the hypothesis that the two groups are equal.

9) Not report adjustments for multiple comparisons

Most empirical researches report multiple P-values, because authors test multiple hypotheses or make multiple comparisons in the same study, sometimes with the same data. Multiple hypothesis testing or multiple comparisons occurs when the researcher establishes equivalence between groups or makes comparisons between groups for several points in time.

10) Confusion between statistical meaning and practical meaning

There are some authors who interpret a result as statistically significant ($P < 0.05$) where the difference between two or more groups is important. Statistical significance is necessary, but not sufficient enough to conclude if the relationship or effect is real.

11) Not determine if the data in the analysis of variance (ANOVA) and the t-test meet the statistical assumptions or not

Analysis of variance, or a simpler version: "t-test", is based on the assumption that the data must follow the rule of normal distribution, independent of each other and that the variances between groups must be similar. But in reality, many variables do not meet these assumptions. If the researcher disregards the assumptions and analysis, the results may not be correct. When the data do not follow the normal distribution or do not meet the above assumptions, researchers need to transform the data to meet the above general assumptions before analysis. If the data cannot be converted, the researcher can apply non-parametric analysis methods such as the Wilcoxon rank-sum test, instead of using the analysis of variance method.

12) Not describe the method used to analyze the difference between the two groups in the analysis of variance

Analysis of variance is used to compare more than two groups. If there are three groups A, B and C, we can compare A to B, A to C and B to C. Analysis of variance usually provides two important results: the F-test and P-value. The P-value tells the researcher that there are at least two different groups (within the analyzed groups) that have statistical meanings but do not specify which groups! To know which groups are really different, the researcher needs to perform the second step in the analysis process: that is, post-test analysis, such as Fisher's least significance test, Tukey, Student-Neuman-Keuls, etc. These methods sometimes give different results because of different assumptions.

Therefore, in reporting the results of the analysis of variance, the researcher must clearly state which method was used in detecting the differences and the assumptions behind the analytical method.

With the master's theses, in the research process, there are some other mistakes such as writing the wrong formula; misinterpret; write the correct formula but explain it wrongly; write the correct formula, correct calculation but wrong conclusion; the applicable conditions are not strict and even the calculation is wrong.

C. Causes of Common Mistakes Made in the Process of Applying the Mathematical Methods of Statistics

From the analysis of the above mistakes, we give the following common causes:

First, some researchers do not fully and accurately understand the properties of statistical concepts; sometimes even lack of some necessary knowledge of logic;

Second, the difficulty in analysis and orientation to find algorithms to solve selection problems, such as selection of control groups, experimental groups, comparison of characteristic parameters, finding the relationship between two or more quantitative characteristics…

Third, the calculation skills are not fluent enough, researchers do not know how to effectively use calculation tools such as calculators, computers and some specialized software in analyzing and processing data such as SPSS, SPLUS, R, etc.

Fourth, the lack of caution in checking and reviewing the presented terms.

Fifth, a few lecturers who direct and guide the researchers are not really thorough and have not mastered the method of applying the Mathematical Methods of Statistics in scientific research, so they are sometimes afraid to check the contents related to this method in the researching process.

IV. Conclusion

The research results have determined the basic statistical algorithms, formulas and pointed out 12 common mistakes when applying the Mathematical Methods of Statistics in analyzing and processing scientific research data in the field of sport science. At the same time, the research results also find out the main causes of mistakes as a basis for proposing the application of Mathematical Methods of Statistics in scientific research, contributing to improving the ability to apply Mathematical Methods of Statistics for staff, lecturers, graduate students and postgraduate students, contributing to improving the quality of scientific research in Bac Ninh Sport University.

Conflict of Interest

No potential conflict of interest was reported by the author(s).

References

Blekman, T. T., Mưskix, A. D., Panovko, Ia. G. (1976). *Applied mathematics, objects, logic, features of method, science and technics*. Publishing House, Ha Noi.

Duong Nghiep Chi, Tran Duc Dung, Ta Huu Hieu, Nguyen Duc Van. (2004). *Sport Measurement Curriculum*. Sport Publishing House, Ha Noi.

Do Ngoc Dat (1994). *Mathematical Applied Statistics in Educational Scientific and Sociological Research*. University of Education Publishing House, Ha Noi.

Dao Huu Ho, Nguyen Thi Hong Minh (2002). *Processing Data by mathematical statistics on computers*. Vietnam National University Publishing House, Ha Noi.

Duong Thieu Tong (2001). *Applied Statistics in Scientific Educational Research*. Vietnam National University Publishing House, Ha Noi.

Nguyen Xuan Sinh, Le Van Lam, Pham Ngoc Vien, Luu Quang Hiep (1999). *Scientific Research Method Curriculum in Sport Science*. Sport Publishing House, Ha Noi.

Nguyen Huu Chau (10/2004). Educational scientific research in the coming period. *Journal of Education*, 98, 1-3.

Polya, G. (1975) *Mathematical and Rational Reasoning. Part 1.* (Translators: Ha Si Ho, Hoang Chung, Le Dinh Phi). Educational Publishing House, Ha Noi.

Polya, G. (1976). *Mathematical and Rational Reasoning*. Part 2. (Translators: Ha Si Ho, Hoang Chung, Le Dinh Phi). Educational Publishing House, Ha Noi.

Polya, G. (1976). *Mathematical and Rational Reasoning*. Part 3. (Translators: Ha Si Ho, Hoang Chung, Le Dinh Phi). Educational Publishing House, Ha Noi.

Article source: The article is extracted from a grassroots-level scientific research project in 2018 with the title: "Application of statistical mathematical methods in data analysis for some types of basic scientific research topics in the field of sports"

Assoc. Prof. PhD Nguyen Van Phuc was born in 1976 in Vietnam. In 2008. He is the Director of Bac Ninh Sports University, Vietnam.

Assoc. Prof. PhD Ta Huu Hieu was born in 1975 in Vietnam. In 2008. He is Head of Olympic Sport gifted School, Bac Ninh Sport University of Viet Nam.

Assoc. Prof. PhD Nguyen Duc Thanh was born on July 19, 1971 in Dong Thap province, Vietnam. Associate Professor PhD, Manager of Physical and Defense Education Center, Ho Chi Minh University of Technology and Education, Vietnam.