Review

Two Common Mistakes in Applying ANOVA Test: Guide for Biological Researchers

Fateme Azizi[1], Rasoul Ghasemi[2] and Maryam Ardalan[3]*

[1] School of advanced medical technologies, Tehran university of medical sciences, Tehran, Iran
[2] Department of Physiology, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3] Centre for Perinatal Medicine and Health, Institute of Neuroscience and Physiology, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
*Corresponding author: maryam.ardalan@gu.se

Abstract: The importance of statistics in biological research is inevitable. The appliance of statistics is the most powerful tool to support the scientific hypothesis and to give credibility to biological research methodology, to interpret convoluted explanations and conclusions based on the research findings. The evidences from literatures point the persistent statistical errors, selection of non-proper statistical test with the consequence of misinterpretation of the scientific results which are published in international journals. One of the most extensively used statistical tests in the field of biology (preclinical and clinical) is analysis of variance (ANOVA). ANOVA test is in place of multiple T-tests to compare the means of more than two groups at a time; and there are some important points which biologists should be aware of them to avoid possible misinterpretation of results with the consequence of wrong conclusion. Accordingly, the aim of the current review is to help biologists to understand the basic concepts of ANOVA test and reach to a more valid interpretation of achieved results.

Keywords: ANOVA; homogeneity of variance; Levene’s test; multiple comparison analysis; post hoc test

1. Background

Statistic is the grammar of science, and we need to use the language of science with proper grammar [1], therefore, it is important that biological researchers achieve basic statistical knowledge about the application of proper statistical tests and interpretation of the results. Statistical errors, at any step, shadow the communication of an investigator’s results with the further impact on the research outputs [2,3].

Analysis of variance (ANOVA) is the most effective and widely used parametric statistical test for analyzing biological data in which several independent factors such as sex, treatment and/intervention are represented [4]. This model extends of the independent t-test for comparing the means of more than two (independent) groups. Statisticians refer to the ANOVA test as an omnibus test, as it tests the whole set of means at once (omnibus means “for all” in Latin[5]). In fact, an omnibus test provides overall results of the data. If this test does not become significant, there is no evidence of the null hypothesis rejection, which means no difference between group means. If the null hypothesis is rejected, one then proceeds to the next step of post-hoc pairwise group comparisons to determine sources of difference.

The application of statistical tests especially ANOVA in biological researches is increasing. A study considered 1128 original published papers in peer-reviewed journals which 800 (70.92%) studies used inferential statistical tests. Among those 800 papers, 203 (25.37%) papers used ANOVA and 92 (45.23%) used post hoc tests. Out of the 203 papers, 175 (86.21%) studies applied one-way ANOVA test, 20 (9.9%) studies used two-way ANOVA test and 8 (3.94%) studies used repeated measure ANOVA test. Out of the total

92 original papers, which used post hoc test, 40 (43.48%), applied the proper post hoc test and 11 (11.96%) didn`t define the name of the applied post hoc test[6]. Accordingly, ANOVA is one of the most applied statistical tests in biological research studies; however, the common errors in its application, lead to the misinterpretation of the results. These observations encouraged us to survey some of the most common mistakes in ANOVA test application and to provide some possible solutions for non-statistician biologists to avoid errors and misinterpretations of the presented data.

2. Type I and type II errors

In the medical research that also is known as an experimental medicine, it is usually impossible to collect data from an entire population, therefore, the researcher needs to collect a random sample from the entire population. Researchers need confidence that the samples accurately reflect the population, therefore, they apply statistical tests to draw conclusions from a sample and generalize them to the entire population.

Descriptive and inferential statistics are two broad categories in the field of statistics. Descriptive statistics describe the properties of the sample, while inferential statistics examine hypotheses and make conclusions about the characteristics of a population from the characteristics of the samples. The majority of statistical protocols include one or more statistical hypotheses to test. In other words, through testing hypotheses and producing estimates, inferential statistical analysis infers attributes of a population.

A hypothesis testing evaluates two exclusive expressions about a population to find which one is supported by the sample data. The hypothesis is called null hypothesis (all group means are equal) and the alternative hypothesis (all group means are not equal). Testing the hypothesis uses a random sample to draw conclusion about entire population, so, it is not 100% accurate. In the other word, for every hypothesis testing, there is a chance to reject a true hypothesis. When conducting a hypothesis test, there are two types of errors which lead to an incorrect conclusion. Type I error occurs when a null hypothesis which is true, statistically is rejected (false positive), whereas type II error occurs when the test fails to reject a null hypothesis which is not correct (false negative) (Fig 1).

Are two types of errors important at the same level? Testing Hypothesis is created by statisticians to control Type I errors, whereas Type II errors are less well characterized. There is a presumption that Type I errors are more serious. For the individual test, it is important to evaluate the consequences of each type of error inaccuracy.

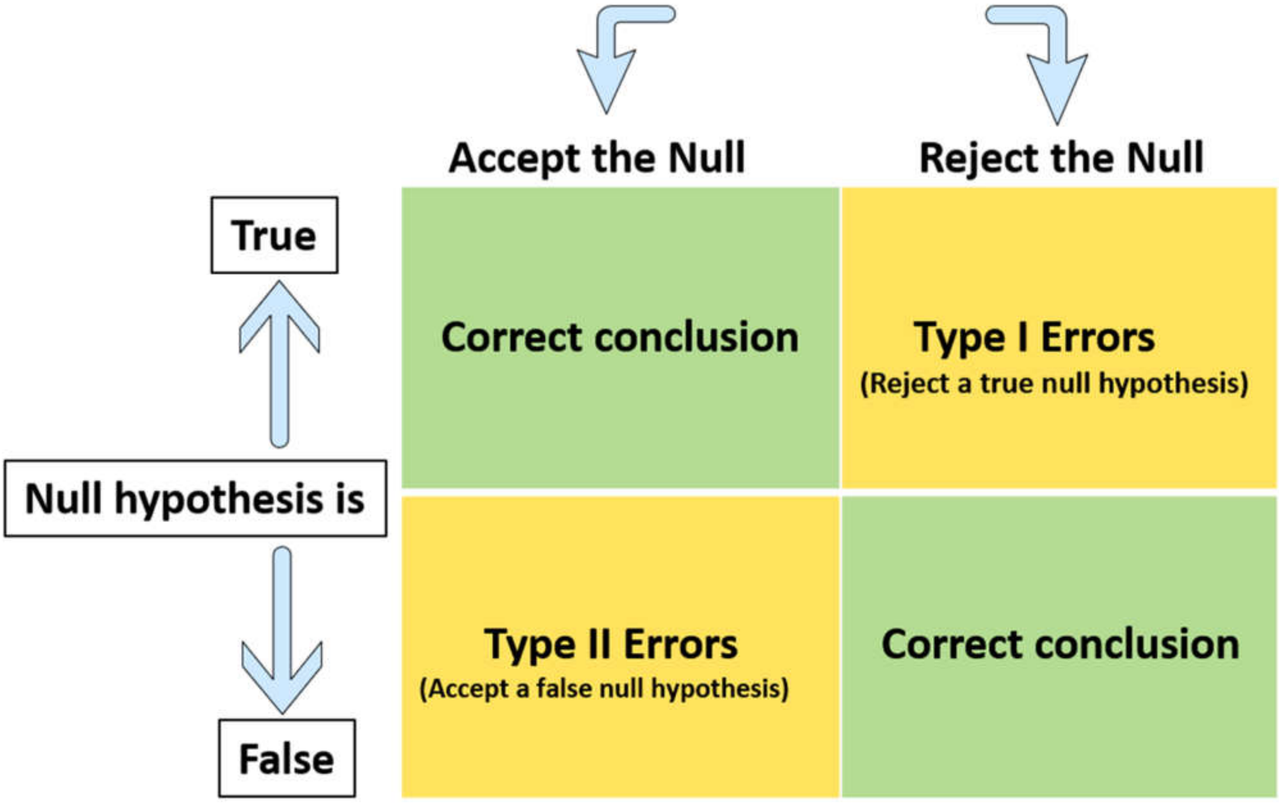

Figure 1. Types of Error in Statistical Hypothesis Testing.

3. Assumptions of ANOVA test

When we plan to determine differences in a scale-level dependent variable by a nominal variable, comparison of mean values is required. The independent t-test is used to compare the means between two independent groups. However, in biological research, more often, there are experiments which need to determine mean differences between more than two independent groups. For testing the differences between a large number of groups, we could simply use a series of t-tests instead of the ANOVA test. However, every time we run a t-test, we add up the type I error rates. If we compare three groups, we will have to make K(K-1)/2 comparisons (3 in this case). If each comparison is tested at the p=0.05 level, the final alpha (probability of Type I error) will be 0.15, i.e. a 15% Type I error rate. Therefore, ANOVA test is used to prevent the increase of errors.

ANOVA test is one of the most frequently used statistical test for such determinations and makes it possible to simultaneously compare several means [7]. Therefore, the issue that persuades researchers to use ANOVA test is avoiding the error of alpha level inflation, which increases Type 1 error probability (false positive) and emerges when repetitive paired comparisons are made [8].

There are three primary assumptions in ANOVA test, that violation of any of these assumptions undermines the reliability of the ANOVA output. These assumption are; all data show normally distribution; all data have equal variance and are independent of each other [9].

3.1. First Mistake: Classic ANOVA or Welch's ANOVA test

Homogeneity of variance (homoscedasticity) is an important assumption for ANOVA test that allows the comparison of the variances within the groups. Homoscedasticity assumes different samples have the same variance, even if they come from different populations. When this assumption is violated, the problem is known as heteroscedasticity [10].

Clinical and preclinical data heteroscedasticity are expected with theoretical and empirical explanations. In the biological science, treatment may have various impacts on different individuals/cases, making it more beneficial to certain clients than others, or even potentially harmful to others [11]. For example, in pharmacological research, drug response can be influenced by different factors (diet, comorbidities, age, weight, drug-drug interactions, and genetics), and resulting in variations in drug response due to inter-individual differences [11]. In psychiatric field, there is an evidence of negative outcomes for patients receiving psychotherapy [12,13].

Given the fact that most of the biomedical data come from animal or human samples, it is important to check homogeneity of data and report them in the statistical section of manuscript, a prerequisite which is not reported in most of the biological published studies and may lead to incorrect conclusions. Homogeneity of variance can be determined by the following approaches: 1. Comparison of graphs (esp. box plots) 2. Comparison of variance, standard deviation, and IQR statistics and 3. Statistical tests (Fligner Killeen test, Bartlett's test, O'Brien's test, Levene's test) [14]. Among statistical tests for examining the homogeneity of variances, Levene's test [15] is one of the most popular ones and can be conducted in the SPSS Explore procedure (Analyze → Compare means → One-way ANOVA → Options → Homogeneity of variance tests). If Levene's test does not become statistically significant (groups have equal variances), the assumption of homogeneity of variances is met, therefore, we can trust and report the classic ANOVA test outputs.

Possible solutions for first mistake: As mentioned above, the classic one-way ANOVA test assumes that all groups have equal variances (or standard deviation) even when their means are different. If Levene's test is statistically significant, the assumption of homogeneity of variances is violated. When the homogeneity of variances assumption is not verified, you might not trust the results, as in this situation, classic ANOVA test produces Type I error. In this case, Welch's ANOVA can be a good option as it is not affected by unequal variances. Therefore, if data violates the assumption of variances homogeneity but the assumptions of normality and data independence have been supported by data, Welch's test would be considered as an appropriate statistical test. It can be conducted in the SPSS Explore procedure (Analyze → Compare means → One-way ANOVA → Options → Welch).

On the other hand, Welch's ANOVA is not sensitive to unequal variances. It eliminates the need to be concerned with the assumption of uniform variances.

3.2. Second Mistake: Multiple comparison test selection

The null hypothesis in ANOVA test always points that there is no difference in the mean of data between the groups, therefore when null hypothesis (H_0) is rejected; there is sufficient evidence to conclude that not all the means are equal. If the p-value from ANOVA or Welch's test (in case of unequal variance) becomes significant, you can reject the null hypothesis.

However, ANOVA test does not provide detailed information of pairwise comparison between the groups. So, the next question is that how the researcher investigates the differences between the various subgroups which are tested with ANOVA test? The first answer is to run a series of t-tests between each pair of interests. This strategy is not the correct approach due to two reasons: First, performing repeated statistical tests on the same data causes alpha error [16]. Second, individual t-test examines only two independent groups at a time, therefore the results will remain unclear [17]. Accordingly, there is a need to perform additional analyses to indicate differences between each pair of experimental groups [18]. Differences between the pairs of groups can be performed by applying post hoc tests, which points to "the analysis after the fact" and it comes from the Latin word for "after that" [8]. A category of post hoc tests that provides this type of detailed information for ANOVA test results in "multiple comparison analysis" (MCA) tests. Each of the MCA tests has its own set of advantages and disadvantages. Therefore, selection of proper MCA test, should be based on the specific research questions [19].

When comparing four groups (A, B, C, D) in the context of ANOVA test, we make six comparisons. The term 'family' refers to a pair for these comparisons. In fact, the 'family-wise error (FWE) (sometimes called experiment-wise) refers to the type I error that occurs when each comparison is performed within each family. The likelihood of false positive increases as more tests are run; in the other words, even if the null hypothesis is true, the probability of rejection is high[18].

A multiple comparison analysis was developed to appropriately adjust the FWE and serves two critical functions: indication of which group mean differs significantly from another group mean. Importantly, they also, control the experiment- or family-wise error rate [18].

Selection of the appropriate post hoc test gives researchers the most detailed information by also limiting type 1 error caused by alpha inflation. The most common statistical errors found in biological experiments are caused by issues with multiple comparisons. This is due to testing multiple hypotheses in a single experiment at the same time [3]. According to the importance of multiple comparisons, an increasing number of reviewers considers the issue whether multiple comparisons are appropriately being used for the experimental data. In this regard, a study reviewed the appropriateness of multiple comparisons of published papers in three medical journals over the last 10 years. Of 142 papers which were reviewed, 47 (33%) papers did not use multiple comparison correction, 86 papers (61%) used correction without rationale and only 9 papers (6.3%) used appropriate correction methods [20].

The most commonly stated problem with ANOVA test is the selection of proper post hoc test for analysis of the data. Several types of post hoc tests are available and selection of them is depending on the objectives of the experiment. In many circumstances, different post-hoc tests may lead to the same conclusion; It is critical that we define the methodology in advance and select one post-hoc test that we will use at the time of designing the study. It will be biased approach to run the experiment with different methods and then select the one that yields the desirable results.

According to the importance of this subject, in the following section, we reviewed differences between post-hoc tests based on the reliability and situation that they are used.

To better understand MCAs, they can be divided based on different approaches.

1) ANOVA test is generally conducted with equal group sizes and variances; however, group sizes vary in practice for a variety of reasons, including restrictions of sampling. Many MCAs fail to account for the possibility of unequal group sizes and variances, therefore as a result, may not function properly when group sizes and variances are unequal. Therefore, depending on the homogeneity of variance, various MCAs can be conducted. There are two categories of post hoc tests: one includes tests which are applicable when equal variance assumed (e.g., Tukey, Bonferroni, Scheffé's, LSD, Sildak, SNK tests) and the other group includes tests which are applicable when this assumption is not accomplished (Tamhane's T2, Dunnett's T3, Games-Howell, and Dunnett's C) [8]. The second group accounts for assumption violation by using a sample-size-weighted, pooled variance for only two groups being compared and by adjusting the degrees of freedom for the statistical test [21]. (Table 1)

Table 1. Classification of different multiple comparison analysis tests based on equality variance assumption.

Equal Variance Assumed	Equal Variance Not Assumed
LSD (least significant difference)	Tamhane's T2
Bonferroni	
Sidak	
Scheffe	
R-E-G-W F (Ryan-Einot-Gabriel-Welsch F test)	Dunnett's T3
R-E-G-W Q (Ryan-Einot-Gabriel-Welsch range test)	
S-N-K (Student-Newman-Keuls)	Games-Howell
Tukey's HSD (honestly significant difference) test	
Tukey's b	
Duncan	
Hochberg's GT2	
Gabriel	Dunnett's *C*
Waller-Duncan	
Dunnett	

2) Another group of MCAs is in terms of the type of comparison. Accordingly, there are two categories of MCA: post hoc range tests and pairwise multiple comparisons. Range tests identify homogeneous subsets of means that are not different from each other. In fact, non-statistically significant comparisons are placed in homogeneous subsets. If the means of two groups are in different subsets, they differ statistically.

Pairwise multiple comparisons test the difference between each pair of means and provide a matrix that indicates significantly different group means at an alpha level of 0.05. (Table 2)

Table 2. Classification of different multiple comparison analysis tests based on the type of comparison.

Multiple Comparison Tests Only	Range Tests Only	Multiple Comparison Tests AND Range Tests
Bonferroni	Tukey's b	Tukey's HSD (honestly significant difference) test
	S-N-K (Student-Newman-Keuls)	
Sidak	Duncan	Hochberg's GT2
Dunnett	R-E-G-W F (Ryan-Einot-Gabriel-Welsch F test)	Gabriel
	R-E-G-W Q (Ryan-Einot-Gabriel-Welsch range test)	
LSD (least significant difference)	Waller-Duncan	Scheffe (confidence intervals that are fairly wide)

3) Another group of MCAs is based on the tests used to handle type I errors. MCTs are thus divided into two types: single-step type and stepwise type. The single-step procedure, as the name implies, assumes one hypothetical type I error rate. Almost all pairwise comparisons (multiple hypotheses) are performed under this assumption. In the other word, each comparison is independent from others. This category, includes Fisher's least significant difference (LSD), Bonferroni, Sidak, Scheffé, Tukey, Tukey-Kramer, Hochberg's GF2, Gabriel, and Dunnett tests.

Stepwise tests are further classified as step-up and step-down ones. The stepwise tests type I error according to the previously selected comparison results. Indeed, it performs pairwise comparisons in a predetermined order, and each comparison is performed if the result of the previous comparison is statistically significant. Ryan-Einot-Gabriel-Welsch Q (REGWQ), Ryan-Einot-Gabriel-Welsch F (REGWF), Student-Newman-Keuls (SNK), and Duncan tests are included in this group of tests. The SNK test, for example, is a Stepwise MCT that compares means using the step-down procedure. As a result, the means are ordered, with the largest and smallest being compared first. If the largest and smallest means are statistically significant different, the next smallest is compared to the largest, and the smallest is compared to the next largest, and so on until all comparisons are completed. As a result, this method is known as the step-down method because the extents of the differences decrease as the comparisons progress [18].

3.3. Common Types of Multiple comparison tests

Tukey, Newman-Keuls, Scheffe, Bonferroni, LSD, and Dunnet are some of the most often used multiple comparison statistical tests.

34. Tukey's

Tukey's HSD (Honestly Significant Difference) (also called Tukey's A and, sometimes, wholly significant difference [WSD] test is the most commonly used statistical test for comparing all possible group pairings [18] . Tukey test has the advantages of testing all pairwise differences, being simple to compute, and lowering the likelihood of making a Type I error [17]. Tukey's HSD was designed with equal variance and sample size. If these assumptions are violated, the test may become too conservative or too liberal. Tukey's HSD controls Type I error when assumptions are met, with power that is about average, being greater than some and less than other MCTs. Tukey's HSD, on the other hand, does not strictly control Type I error when assumptions are not met. Its main drawbacks are that it is less powerful than some tests and also is not intended to test complex comparisons [22].

3.5. Student-Newman-Keuls (SNK)

The Student-Newman-Keuls (SNK) test is named after three papers published by Student (1927), Newman (1939), and Keuls (1952). This test is a step-down test, which means that extreme differences are tested first, and they are ordered from highest to lowest. As it performs more pairwise comparisons than Tukey, this is a more powerful test than Tukey. As a result, it is more likely that some differences will be statistically significant. As a result, some differences are more likely to be statistically significant. It does, however, frequently come with an increased family-wise error level. SNK should be used in studies with equal group size [23].

3.6. Dunnett

Dunnett's test shows a large power. Dunnett test can be used when a researcher plan to compare the treatment groups only to a single control group. The probability of rejecting H_0 and the inflation of probability of type I error increases as the number of comparisons increases [18].

When we apply Dunnett's, we benefit from the extra power gained by making fewer comparisons. Therefore, this test is a particularly powerful statistical test which can detect relatively small but significant differences between groups or combinations of groups. Accordingly, when the following conditions are met, performing Dunnett's test is a good option: 1. researcher plans to compare control group with all the other groups. 2.According to the study plan, it is not necessary to compare the treatment groups to each other [17].

3.7. Scheffee

This test is the most conservative post hoc test. Scheffee developed a test for computing all feasible comparisons at the same time (not just pairwise comparisons). Scheffee test has the advantage of allowing a researcher to conduct any post-hoc comparison that prefers. The scheffee test tends to be conservative and underpowered, therefore it is the more appropriate test to utilize the consequences of a Type II error outweigh.

In circumstances that the theoretical background for differences between groups is unavailable or there is insufficient information prior research to test the theory, scheffee is a good exploratory statistical test as it tests all possible comparisons.

3.8. Bonferroni

Bonferroni test is less susceptible to Type I errors than the Scheffee test, but similar to the Scheffee test, can be used to compare complex pairs. However, Bonferroni test is not a tool for exploratory data analysis.

It necessitates that the researcher specifies all contrasts to be investigated. In order to determine which contrasts to specify, the researcher needs to have strong hypothesis about the phenomena of interest. As a result, the Bonferroni approach is more appropriate to hypothesis testing studies that confirm a hypothesis about the experimental group’s results [24].

As Bonferroni test outputs have reduced power when the number of comparisons is large, it may be preferable for comparisons of a small number of groups means or preplanned comparisons of selected groups. The limitation of the Bonferroni test is that the output tends to be too conservative and does not have enough power when the set of tests is large, therefore, this test often fails to detect real differences [25].

3.9. Possible Solutions for Second Mistake

As previously stated, post hoc tests are for controlling the family-wise error rate. However, post hoc tests control the family-wise error rate by reducing the statistical power of the comparisons. Power is the ability to detect a difference in the population. Power in means comparison reflects the likelihood of corrected identifying a difference between two groups' means. Conservative tests go to great lengths to ensure that the user does not make a Type I error. They use stricter criteria to determine significance. Although these tests provide protection against Type I error, they have some limitations. You lose power as the tests become more stringent. Therefore, the choice of procedure should be directed by a desire to control family-wise Type I error while also providing high power [21] (Fig 2).

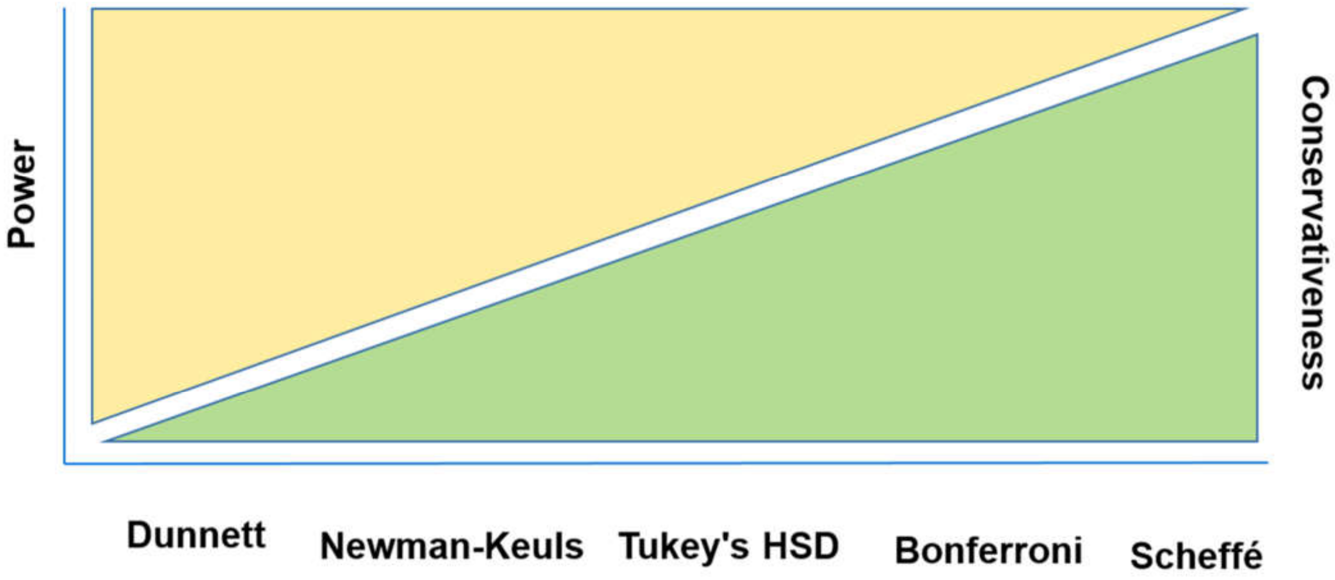

Figure 2. Comparative chart of multiple comparison tests base on power and conservative.

There is no set of rules for determining which test to be used, but it is critical to consider the specific situation. Power can be an issue, so we should select tests with more power than others, such as SNK. For example, when there were no drugs to treat a disease, even minor differences between treatment groups were important. In this case, a Type I error is not as dangerous as rejecting an effective drug for an unavoidably fatal disease when there are no other options. Another application of the SNK test is sorting all treatment means into treatment subsets. These subsets will be homogeneous in the sense that they will not differ from each other, but they will differ from other subsets. The NK procedure establishes a higher confidence bound for the number of best treatments [26].

In general, the SNK statistic is appropriate for studies in which relatively small but significant differences are sought, as well as where the consequences of a Type II error are worse than those of a Type I error. As a result, it is a useful tool in new areas of science where little is known about the phenomena of interest [17].

The other issue is that groups may have different sample sizes or variance. Several multiple comparison analysis tests (such as Tamhane's T2, Dunnett's T3, Games-Howell, and Dunnett's C) was specifically developed to handle this problem.

For comparisons of a small number of groups mean or preplanned comparisons of selected groups, Bonferroni test may be the best choice, but Bonferroni is not a good choice for studies that groups are unequal in size.

In the studies in which it is not necessary to compare the treatment groups with each other and only a control group is compared with other experimental groups, Dunnett's test may be of choice.

Each test has its own set of applications, advantages, and limitations, and the ability of certain statistics to address the questions of interest and the types of data to be analyzed should be used to guide the selection of numerous comparative statistical tests. As a result, it's critical that researchers choose the tests that best fit the data, the types of information regarding group comparisons, and the analyses' required power.

4. How improve the quality of reporting?

Many papers do not provide enough information to determine why a specific test was chosen, which type of test was used, or how the test result was verified. Statistical method reporting (Report the reason for selection the statistical test) is required for study evaluation and leads to a better understanding of studies. It also allows for the detection and correction of errors prior to publication. This information helps readers to understand why the authors chose a specific statistical test.

A systematic review investigated the quality of reporting for two statistical tests, t-tests and ANOVA test, in papers which are published in physiology journals in June 2017. They found that95% of papers that used ANOVA test were lacking the information needed to determine which type of ANOVA test was used, and 26.7% of papers did not determine which post-hoc test was applied [27].

The observation report is another issue that can help to improve the report's quality. We start the discussion with a question. How reporting the details of the statistical outputs are critical? Many studies are restricted to reporting statistical findings. While many times the biological studies do not fit within the framework of statistical assumptions. For example, we found that the Morris water maze is one of the most commonly used tests in behavioral neuroscience for investigating the psychological processes and neural mechanisms of spatial learning and memory. The criterion for scoring the animal's behavior in this behavioral test is reaching the platform. However, the interpretation of animal behavior cannot be limited to this scoring. The pattern of animal behavior from Armon's beginning to the platform can reveal a lot of information. Therefore, from our biological point of view, one of the possible solutions to improve reporting is to dedicate a part of the paper to report observations that may not fit in the form of statistical frameworks.

One of the limitations that appears to be one of the reasons for not mentioning the selection criteria of statistical tests is the limitation of words for the main text of paper for publishing in the journal. To address this issue, journals may count words without accounting for statistical section as the same as references.

5. Conclusion

The ANOVA test is used widely in different areas of biological science. Classic ANOVA test performs the best when data is homogeneous, normal, and balanced/unbalanced and the Welch test performs the best when data are heterogeneous normal, and balanced/unbalanced.

Regarding the selection of MCT, each MCT is appropriate for some situations, the standard of choice is the ability to control type 1 error and the degree of power detecting the significant differences between groups.

References

1. Kucuk U, Eyuboglu M, Kucuk HO, Degirmencioglu G. Importance of using proper post hoc test with ANOVA. *International journal of cardiology.* 2016;209:346.
2. Karadeniz PG, Uzabacı E, Kuyuk SA, et al. Statistical errors in articles published in radiology journals. *Diagnostic and Interventional Radiology.* 2019;25(2):102.
3. Lee S. Avoiding negative reviewer comments: common statistical errors in anesthesia journals. *Korean journal of anesthesiology.* 2016;69(3):219.
4. Armstrong RA, Eperjesi F, Gilmartin B. The application of analysis of variance (ANOVA) to different experimental designs in optometry. *Ophthalmic and Physiological Optics.* 2002;22(3):248-256.
5. Abdi H, Williams LJ. Newman-Keuls test and Tukey test. *Encyclopedia of research design.* 2010:1-11.
6. Patel S, Naik V, Patel P. An analysis of application of multiple comparison tests (post-hoc) in Anova in recently published medical research literature. *National Journal of Community Medicine.* 2015;6(1):117-120.
7. Lee DK. Alternatives to P value: confidence interval and effect size. *Korean journal of anesthesiology.* 2016;69(6):555.
8. Kim TK. Understanding one-way ANOVA using conceptual figures. *Korean journal of anesthesiology.* 2017;70(1):22.
9. Scheffe H. *The analysis of variance.* Vol 72: John Wiley & Sons; 1999.
10. Kim J, Cribbie RA. ANOVA and the Variance Homogeneity Assumption.
11. Grissom RJ. Heterogeneity of variance in clinical data. *Journal of consulting and clinical psychology.* 2000;68(1):155.
12. Cuijpers P, Reijnders M, Karyotaki E, de Wit L, Ebert DD. Negative effects of psychotherapies for adult depression: a meta-analysis of deterioration rates. *Journal of affective disorders.* 2018;239:138-145.
13. Mohr DC. Negative outcome in psychotherapy: A critical review. *Clinical psychology: Science and practice.* 1995;2(1):1-27.
14. Conover WJ, Johnson ME, Johnson MM. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics.* 1981;23(4):351-361.
15. Brown MB, Forsythe AB. Robust tests for the equality of variances. *Journal of the American Statistical Association.* 1974;69(346):364-367.

16. Ilakovac V. Statistical hypothesis testing and some pitfalls. *Biochemia Medica.* 2009;19(1):10-16.
17. McHugh ML. Multiple comparison analysis testing in ANOVA. *Biochemia medica.* 2011;21(3):203-209.
18. Lee S, Lee DK. What is the proper way to apply the multiple comparison test? *Korean journal of anesthesiology.* 2018;71(5):353.
19. Marusteri M, Bacarea V. Comparing groups for statistical differences: how to choose the right statistical test? *Biochemia medica.* 2010;20(1):15-32.
20. Armstrong RA. When to use the B onferroni correction. *Ophthalmic and Physiological Optics.* 2014;34(5):502-508.
21. Sauder DC, DeMars CE. An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science.* 2019;2(1):26-44.
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological).* 1995;57(1):289-300.
23. Sauder D. Examining the type I error and power of 18 common post-hoc comparison tests. *Graduate Psychology: James Madison University https://commons lib jmu edu/cgi/viewcontent cgi.* 2017.
24. McDonald JH. *Handbook of biological statistics.* Vol 2: sparky house publishing Baltimore, MD; 2009.
25. Chen S-Y, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. *Journal of thoracic disease.* 2017;9(6):1725.
26. Wu SS, Wang W, Annis DH. On identification of the number of best treatments using the Newman-Keuls test. *Biometrical Journal: Journal of Mathematical Methods in Biosciences.* 2008;50(5):861-869.
27. Weissgerber TL, Garcia-Valencia O, Garovic VD, Milic NM, Winham SJ. Meta-Research: Why we need to report more than'Data were Analyzed by t-tests or ANOVA'. *Elife.* 2018;7:e36163.