Musings, Opinions, Tips and Experiences

The Challenges of Interpreting ANOVA by Dermatologists

Kamal Kishore, Vidushi Jaswal[1], Rahul Mahajan[2]

Department of Biostatistics, Postgraduate Institute of Medical Education and Research, [1]Department of Psychology, Mehr Chand Mahajan DAV College for Women, [2]Department of Dermatology, Venereology, and Leprology, Postgraduate Institute of Medical Education and Research, Chandigarh, India

Problem Statement/Abstract

Cutaneous wart is a common dermatological condition in the pediatric age group with a prevalence ranging between 22 and 33%.[1] Decision-making is challenging as 60–70% of the patients have spontaneous recovery by approximately 12 months without any intervention.[2,3] Recommendations regarding the intervention are uncertain as multiple available options lead to similar outcomes. The Cochrane review comparing several treatment modalities also does not offer a decisive answer.[4] Many physicians prescribe salicylic acid (SA) and cryotherapy (CRT) to clear warts. However, the literature does not unanimously favor any intervention over the other (SA vs. CRT vs. placebo). The outcomes of assessing the effectiveness of different treatment regimen for warts is of crucial importance for a practicing physician. Keeping this problem statement central to our analysis, we decided to hypothesize data of three treatment modalities (SA vs. CRT vs. placebo) which can be analyzed by the analysis of variance (ANOVA). The focus of this article is to appraise the readers about when and how to apply ANOVA rather than the effectiveness of any one intervention. Therefore, the readers are strongly encouraged to learn the application and interpretation of the ANOVA technique rather than drawing any conclusion regarding the effectiveness of the treatments. Initially, we will discuss a list of possible challenges [Figure 1] faced by the investigator at different stages of data analysis. Subsequently, we will define and discuss the use of the ANOVA technique to find out the difference in the effectiveness among the three treatment groups.

Wide and Long Format Data Entry

Usually, researchers begin by defining the study design, sample size, missing criteria, measurement scale, variance, level of significance, power, and type of primary outcome variable. However, in our study, to demonstrate the ANOVA technique, we artificially generated the data for 30 participants in each group. The ANOVA applies to the groups with unequal sample sizes but the power of the test decreases with the increase in the variation of sample sizes between the groups. The immediate task after collecting the records of the participants is to decide about the structure of data entry in the software. The wide and long formats are two broad data entry mechanisms, and the same are also known as multivariate and univariate formats, respectively. The responses of the patients under the columns heading placebo, SA, and CRT are known as the multivariate format [Table 1]. There will be 30 rows of data in a wide layout. This layout is preferred by books and faculty to teach ANOVA in the class due to the facilitation of manual calculation and the requirement of less space. However, it is not an ideal method to enter data in a wide setup when it comes to the application of ANOVA in statistical software.

In contrast to a multivariate layout, the data in the long format need clarity of data in terms of the independent and dependent variables. The independent variable representing group categories such as placebo, SA, and CRT makes one column, and the continuous dependent variable such as the number of days to heal cutaneous warts (response) will form the second column. There will be 90 data rows in the long layout. Table 1a and b displays the subset of data for both the

Address for correspondence:
Dr. Rahul Mahajan,
Department of Dermatology,
Venereology, and
Leprology, Postgraduate
Institute of Medical
Education and Research,
Chandigarh - 160 012, India.
E-mail: drrahulpgi@yahoo.com

How to cite this article: Kishore K, Jaswal V, Mahajan R. The challenges of interpreting ANOVA by dermatologists. Indian Dermatol Online J 2022;13:109-13.

Received: 18-May-2020. **Revised:** 28-Jul-2021.
Accepted: 29-Jul-2021. **Published:** 24-Jan-2022.

Access this article online

Website: www.idoj.in

DOI: 10.4103/idoj.idoj_307_21

Quick Response Code:

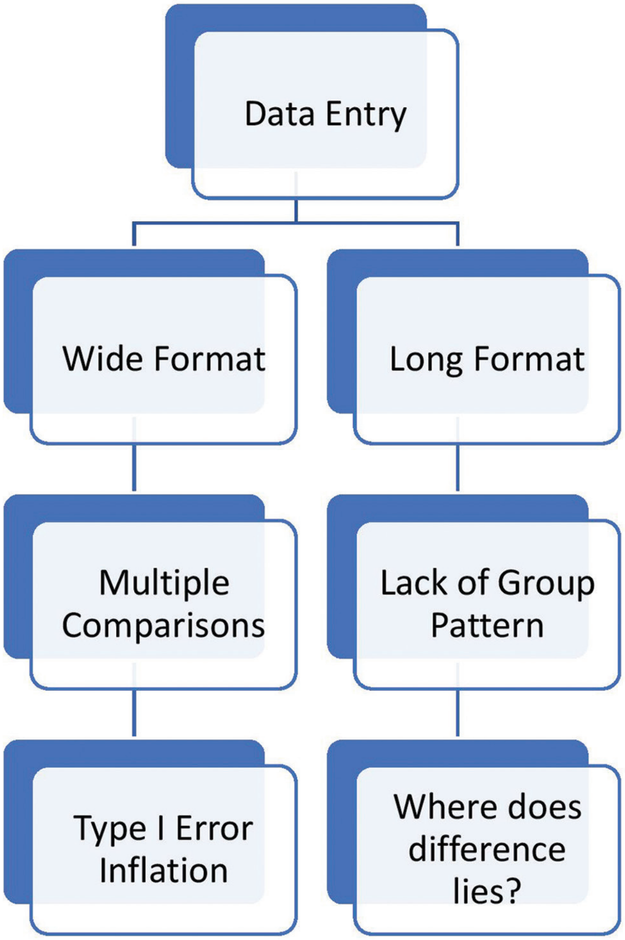

Figure 1: A flowchart depicting the possible challenges at each stage

Table 1: Depicting the subset of data in wide and long formats, respectively

Table a			Table b	
SA	CRT	Placebo	Drugs	PGA3
84	98	85	SA	84
80	90	85	SA	80
75	90	70	SA	75
70	85	70	CRT	98
68	80	105	CRT	90
64	80	98	CRT	90
62	75	100	PCB	85
60	70	98	PCB	85
55	56	84	PCB	70

wide ($n_1 = n_2 = n_3 = 9$) and long ($n_1 = n_2 = n_3 = 3$) data entry formats for the readers perusal. It is important to note that different software may need a different data structure for the application of the same statistical techniques. Thus, the researcher should carefully select the data analysis software and then enter the data in the required format. The Microsoft Excel® which is primarily a spreadsheet needs data for the application of ANOVA in the wide-format as compared to the long-form in the SPSS®, Stata®, and R-software.

Multiple Comparisons

The data arrangement in the wide-format for three or more than three groups might mislead the researcher to go for multiple t-tests such as placebo versus SA, placebo versus CRT and SA versus CRT for three groups. The application of several t-tests will increase the chance of making incorrect decisions. Moreover, multiple comparisons can substantially affect the power of the study. Thus, it is essential to know beforehand about the experiment-wise and comparison-wise error rates. A majority of the researchers fix experiment-wise error rate at $\alpha = 0.05$ (level of significance) while calculating the sample size for studies. However, investigators rarely adjust sample size calculations to control comparison-wise error for subgroup analysis or multiple comparisons.

The long-form of data is also not without limitations. The fundamental idea is to visualize group differences. Usually, the researcher takes all the data in contrast to group membership while visualizing data or validating assumptions. Figure 2a depicts the position of all the individuals in the overall study rather than in the groups. When we plotted the same data with group membership in Figure 2b, the differences are evident. Thus, the researcher needs to plot the participants' data with group identification carefully. The failure to visualize and consider group differences at the planning stage may hamper the identification of the relevant discovery or increase the chance of a wrong conclusion.

Type-I error inflation

The multiple comparisons lead to an inflation of α (type-I error) which increases the chances of false positive, and thus, contributes to the replicability crisis.[5-8] Table 2 displays the effect of multiple comparisons on the interpretation of the results. The formula to calculate the number of multiple comparisons for ANOVA is n (n-1)/2, where n is the number of groups. Three, four, and five groups have 3, 6, and 10 multiple comparisons, respectively. Typically, researchers consider type-I error more severe and fix the same first at 0.05 or 0.01 or 0.001 level or any other level of importance. The investigator unknowingly inflates type-I error during multiple comparisons at an exponential rate [Table 2].

The data in the long format are ideal for ANOVA. Typically, the ANOVA technique provides information about the significant differences among the groups. However, the same does not tell whether all the groups or subset of groups are significantly different from each other. Still, from a clinician's perspective, it is essential to know which intervention or set of interventions are substantially different from each other. The researcher must apply *post hoc* tests to identify the difference between the groups. The *post hoc* tests control both experiment-wise and comparison-wise error rates. There are a total of 18 *post hoc* tests under equal variance (n = 14) and unequal (n = 4)

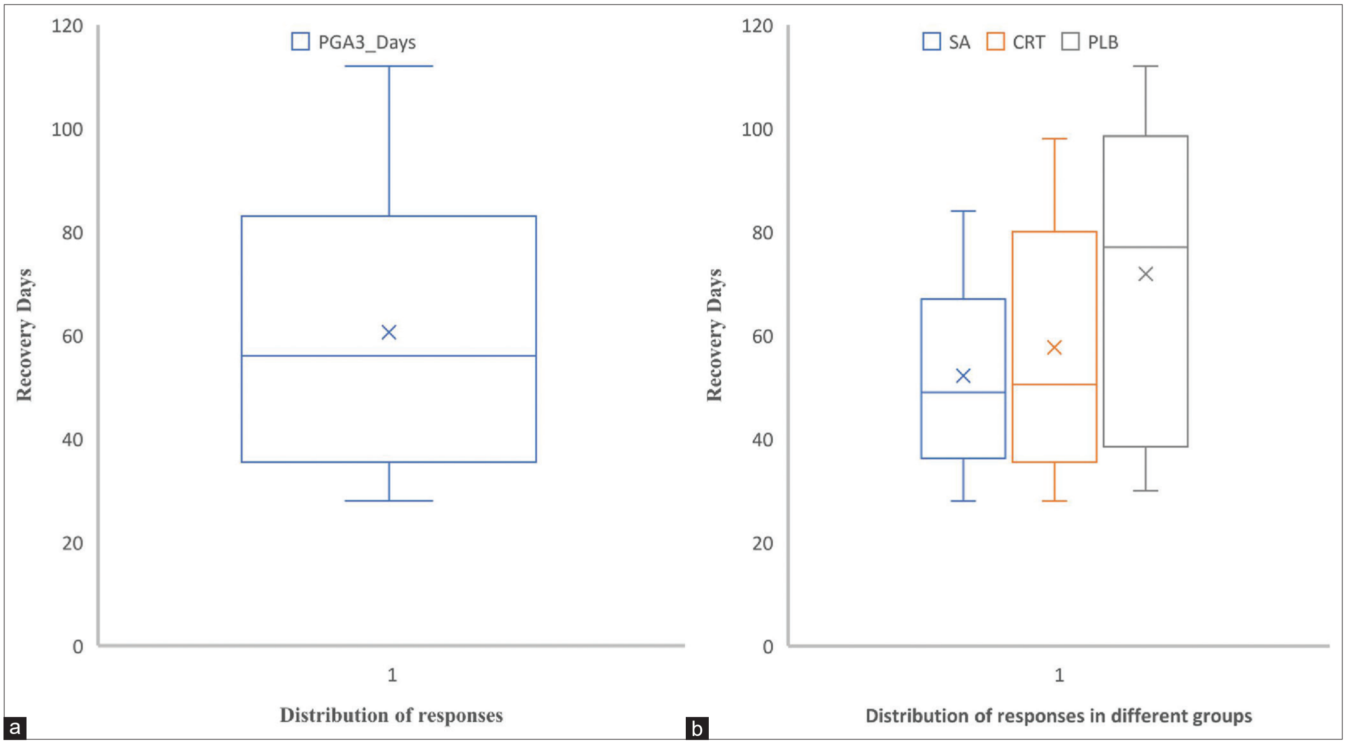

Figure 2: (a and b) Display the distribution of complete and segregated data in groups, respectively

Table 2: A table highlighting the consequences of multiple comparisons

Groups	Comparisons	Type-I error	Interpretation
2	1	0.05	Making the wrong decision of rejecting a null hypothesis on an average is one out of 20 tests.
3	3	0.143	Making the wrong decision of rejecting a null hypothesis on an average is one out of 7 tests.
4	6	0.265	Making the wrong decision of rejecting a null hypothesis on an average is one out of 4 tests.
5	10	0.401	Making the wrong decision of rejecting a null hypothesis on an average is one out of 2-3 tests.

variance options in the SPSS package. The readers can read in detail about the *post hoc* tests in excellent articles written by Sauders *et al*.[9]

What is ANOVA

The purpose of ANOVA is to ascertain the variability that an investigator can attribute to the difference between groups in comparison to within groups. Thus, ANOVA is defined as the statistical technique which divides the total variance into known (drug types such as placebo, SA, and CRT) and unknown factors (such as environmental conditions, human nature, and nurture conditions). ANOVA is used to compare the significant differences between three or more groups. When there is only one factor of interest, such as drug types with three or more categories, it is known as one-way ANOVA. A two-way ANOVA that may affect the outcome consists of two factors such as drug type (placebo, SA, and CRT) and severity of immune suppression (low, moderate, and high). The groups such as placebo, SA, and CRT are known as levels of a factor (drug type). The response, such as remission rate may change as per the levels of factors. Therefore, factors and responses are known as the independent and dependent variables, respectively. The factors and responses are categorical and continuous, respectively.

Null and Alternative Hypothesis for ANOVA

A well-defined hypothesis is an essential component of any study. A hypothesis is a testable statement. Many researchers are aware of the formulation of the hypothesis for two groups. The hypothesis for three or more groups is a straightforward generalization of two groups for the null hypothesis. However, generalization to alternative hypothesis is not straightforward for three and more groups. The null hypothesis for ANOVA states that “the effectiveness does not differ among the types of drug.” An alternative hypothesis for three groups states that the “effectiveness significantly differs among groups.” However, the null hypothesis will be rejected even if the effectiveness is different between any two groups. Thus, the correct way to state the alternative hypothesis is that the effectiveness significantly differs between at least two groups.

Assumptions

The parametric tests make certain assumptions about the parameters of the population distribution from which the samples are drawn. However, many researchers do not take assumptions seriously. These assumptions are like gatekeepers, and it is essential to test them before applying parametric tests. In other words, assumptions are like diagnostic criteria which a physician assesses before prescribing any drug to the patient. The patients may feel that they are diseased, but a physician/dermatologist validates the presence of the disease through physical and verbal examination and laboratory tests. The fulfillment of the set of diagnostic criteria helps the physician to prescribe the best set of prescriptions from the multiple set of prescriptions.

Many tests are robust to violation of assumptions to a certain degree. However, it is challenging to decide the degree of violation of assumptions. Moreover, the violation of assumptions may render any statistical analysis useless. Despite the availability of software, many researchers are in a hurry to analyze their datasets without validating assumptions. Hence, it is crucial and best to consult a statistician at the planning stage. Parametric tests are more powerful compared to non-parametric tests, but they also make more assumptions. Table 3 summarizes a broad set of assumptions for parametric tests. Various checklists such as consolidated standards of reporting trials (CONSORT) and strengthening the reporting of observational studies in epidemiology. (STROBE) suggest verifying assumptions before applying appropriate tests.

Table 3: The table of assumptions and tests for their validation

Assumptions	Definitions	Tests
Independence	The selection of a participant must be random and independent of the selection of the other participants in the group.	Wald–Wolfowitz run test
Normality	The response variable is normally distributed.	Graphical: p-p$^{\#}$ plot, q-q* plot, histogram, and boxplot Test: Shapiro–Wilk test ($N \leq 50$) and D'Agostino skewness test ($N > 50$)‡
Homoscedasticity	The variances are similar in all the groups.	Graphical: Boxplot Test: Levene test and Brown–Forsythe test
Group ratio	The number of participants in any of the group should not exceed the 1:4 ratio.	Check the sample size in each group

ANOVA Output and Interpretation

The ANOVA table displays whether the difference between the group means is statistically significant or not. The assumption of normality for our data was met by only one group. However, ANOVA is robust against violation of non-normality for a large sample. The violation of homogeneity of variance is crucial, and it affects the output from the routine ANOVA technique. When data are heterogeneous across groups, it is better to apply the "Welch test" or "Brown–Forsythe" test. We reported the P value from the Welch test as our data did not meet the assumption of homogeneity of variance. The P value with the routine test was 0.03 against 0.046 with the Welch test. Therefore, investigators need to be careful as a result may change from significant to non-significant or vice-versa in the absence of a correct statistical procedure. Table 4 gives the results of the ANOVA output.

The output from ANOVA indicates that there is a significant difference between the groups. However, it does not tell whether all or a subset of groups is different. Thus, the researchers need to apply *post hoc* tests to determine which pairs of groups are significantly different? The researchers need to make a note that there are multiple *post hoc* tests available in the literature. It is essential to carefully study the properties of the *post hoc* tests and then select the appropriate test to identify group differences.

Conclusions

The ANOVA is a frequently used statistical technique to compare the outcome between three or more groups. Understanding the application and interpretation of ANOVA by the clinicians is crucial as they often come across situations either during their postgraduate training or later as clinical researchers and educators which mandate its use. The validation of assumptions plays a vital role in the generalization of the results. This aspect is often neglected either intentionally or due to ignorance while generalizing the results of a study from controlled conditions (as in a clinical trial).

Definitions

Placebo

A placebo is a pharmacologically inert substance used to treat patients as if it is an active substance.

Type-I error (false positive)

The probability of rejecting a null hypothesis when it should not be rejected. In other words, it is declaring a drug as effective, which is not effective in reality.

Table 4: Output from ANOVA

Variance	Sum of Squares	Degree of Freedom	Mean Sum of Square	*F*-ratio	*P* value*
Between Groups	4304.5	2	2152.3	3.7	0.046
Within Groups	33271.7	57	583.7		
Total Variance	37576.2	59			

P-value – Welch test

Type II error (false negative)

The probability of not rejecting a null hypothesis when it should be rejected. In other words, it is declaring a drug as not effective, which is effective in reality.

Experiment-wise error

It is defined as the probability of making a type-I error in the entire family of comparisons in a study. In other words, the cumulative α is kept at 0.05 or 0.01 or 0.001 by adjusting α for each comparison.

Comparison-wise error

It is defined as the probability of making a type-I error for a particular comparison in the study. Typically, researchers take $\alpha = 0.05$ or 0.01 or 0.001 for the main outcome of the study. However, the value of α is not adjusted for subset analysis.

Null hypothesis

It is the hypothesis of no difference. The null hypothesis assumes that there is no significant difference between groups. In other words, the difference among groups is due to chance.

Alternative hypothesis

It is the hypothesis of interest and is also known as the research hypothesis. The alternative hypothesis assumes that there is a significant difference between groups. In other words, the difference among groups is not due to chance.

Parametric test

Parametric methods assume the distribution of the data. It can be used only for variables measured on Interval or Ratio Scale. The *t*-test, F-test, Z-test, ANOVA, and Regression are parametric tests.

Non-parametric test

These are also known as distribution-free methods as they do not make assumptions about the distribution of the data. They are mostly used for variables measured on nominal and ordinal scales. The Chi-square, Mann–Whitney, and Freidman tests are some of the examples of non-parametric tests.

Power

It is defined as the probability of correcting rejecting a null hypothesis. Usually, a study with 80% or more power is acceptable.

Normality

The continuous outcome variable in each group is distributed as per the normal distribution.

Homogeneity of variance

The spread of the distribution in each group follows the same pattern.

Post hoc test

The pairwise comparisons of outcomes between different groups after finding statistically significant differences with the application of ANOVA is known as a *post hoc* test.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. van Haalen FM, Bruggink SC, Gussekloo J, Assendelft WJ, Eekhof JA. Warts in primary schoolchildren: Prevalence and relation with environmental factors. Br J Dermatol 2009;161:148-52.
2. Bruggink SC, Eekhof JA, Egberts PF, van Blijswijk SC, Assendelft WJ, Gussekloo J. Natural course of cutaneous warts among primary schoolchildren: A prospective cohort study. Ann Fam Med 2013;11:437-41.
3. Kilkenny M, Merlin K, Young R, Marks R. The prevalence of common skin conditions in Australian school students: 1. Common, plane and plantar viral warts. Br J Dermatol 1998;138:840-5.
4. Kwok CS, Gibbs S, Bennett C, Holland R, Abbott R. Topical treatments for cutaneous warts. Cochrane Database Syst Rev 2012;2012:CD001781.
5. Miłkowski M, Hensel WM, Hohol M. Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. J Comput Neurosci 2018;45:163-72.
6. Hengartner MP. Raising awareness for the replication crisis in clinical psychology by focusing on inconsistencies in psychotherapy research: How much can we rely on published findings from efficacy trials? Front Psychol 2018;9:256.
7. Johnson VE, Payne RD, Wang T, Asher A, Mandal S. On the reproducibility of psychological science. J Am Stat Assoc 2017;112:1-10.
8. Knoke JD, Anderson CM, Koch GG. Analyzing repeated measures marginal models on sample surveys with resampling methods. J Stat Softw 2006;15:1-13.
9. Sauder DC, DeMars CE. An updated recommendation for multiple comparisons. Adv Methods Pract Psychol Sci 2019;2:26-44.