P-Value Demystified

Abstract

Biomedical research relies on proving (or disproving) a research hypothesis, and *P* value becomes a cornerstone of "null hypothesis significance testing." *P* value is the maximum probability of getting the observed outcome by chance. For a statistical test to achieve significance, the error by chance must be less than 5%. The pros are the *P* value that gives the strength of evidence against the null hypothesis. We can reject a null hypothesis depending on a small *P* value. However, the value of *P* is a function of sample size. When the sample size is large, the *P* value is destined to be small or "significant." *P* value is condemned by one school of thought who claims that focusing more on *P* value undermines the generalizability and reproducibility of research. For such a situation, presently, the scientific world is inclined in knowing the effect size, confidence interval, and the descriptive statistics; thus, researchers need to highlight them along with the *P* value. In spite of all the criticism, it needs to be understood that *P* value carries paramount importance in "precise" understanding of the estimation of the difference calculated by "null hypothesis significance testing." Choosing the correct test for assessing the significance of the difference is profoundly important. The choice can be arrived by asking oneself three questions, namely, the type of data, whether the data is paired or not, and on the number of study groups (two or more). It is worth mentioning that association between variables, agreement between assessments, time-trend cannot be arrived by calculating the *P* value alone but needs to highlight the correlation and regression coefficients, odds ratio, relative risk, etc.

Keywords: *Confidence interval, hypothesis testing, non-parametric data, null hypothesis, null hypothesis significance testing, parametric data, P value*

Amrita Sil, Jayadev Betkerur[1], Nilay Kanti Das[2]

Department of Pharmacology, Rampurhat Government Medical College, Rampurhat, Birbhum, West Bengal, [1]Department of Dermatology, Venereology, Leprosy, JSS Medical College, JSS Academy of Higher Education, Mysuru, Karnataka, [2]Department of Dermatology, Bankura Sammilani Medical College, Bankura, West Bengal, India

Introduction

Research begins with a research question and every researcher tries to answer the research question by framing a research hypothesis or null hypothesis or H_0 i.e., assuming there is no difference between two or more study groups. The researcher refutes H_0 if there is a "significant" difference between the groups and accepts the alternate hypothesis or H_1 which means that there exists a difference.[1] This forms the basis of hypothesis testing and the definition of "significance" in a dichotomous pattern of "yes or no" by having a cut-off, which is defined be *P*-value. The scientific community is indebted to R.A. Fischer (1890–1962) who is thought to be the "father of modern statistical inference" who introduced *P*-value and the idea of "significance levels"; and to Jerzy Neyman (1894–1981) and Egon Pearson (1895–1980) who developed the theory of hypothesis testing.[2] The article will try to delve into elaborating of use and misuse of *P* value and what lies ahead.

P-value

Every researcher has faced the question of *P*-value and its implication in "significant" results. This article will take its readers to look at *P*-value not only for its association with a significant result but how to utilize *P*-value and not just its face value.

Definition

P-value is the maximum probability of getting the observed outcome by chance. In any test, be it a laboratory test, screening test, or a clinical diagnosis, there are chances of a false positive result. It is up to the experts in the field to decide how much error is acceptable. Similarly, for a statistical test, this margin of error has been decided to be <5%, and this cut-off value of allowable error is termed as *P*-value. For a statistical test to achieve significance, the error by chance must be <5%.

For easy understanding, $P < 0.05$ means if there was truly no effect, then

How to cite this article: Sil A, Betkerur J, Das NK. *P*-value demystified. Indian Dermatol Online J 2019;10:745-50.

Received: August, 2019. **Accepted:** August, 2019.

Address for correspondence:
Prof. Nilay Kanti Das, Department of Dermatology, Bankura Sammilani Medical College, Bankura, West Bengal, India.
E-mail: drdasnilay@gmail.com

Access this article online

Website: www.idoj.in

DOI: 10.4103/idoj.IDOJ_368_19

Quick Response Code:

one would expect to see a positive result less than 5% of the time.

Value of P-value

The cut-off of this chance has been agreed upon by statisticians as 1 in 20 or 0.05 (5%). Since then, the level of statistical significance has been determined at <0.05. More stringent *P*-values can be taken such as 0.01, where the chance factor is further reduced to 1% instead of 5%. In no case, a more relaxed *P*-value is unacceptable. It is worth mentioning that *P*-value is not negotiable e.g., $P = 0.051$ cannot be expressed as "near to significance" or "almost significant!" There is a dichotomized decision as to whether it is "significant" or "not significant."

Expression of P-value

Owing to the availability of statistical software's, the exact *P*-value can be determined to many places of decimal, but it is prudent to express the *P*-value up to 3 decimal places, e.g., $P = 0.002$.

Null hypothesis

To understand the concept of *P*-value, at first, we must understand null hypothesis, hypothesis testing, and errors.

The null hypothesis (H_0) is the assumption that there is no difference between the study groups. If "A" and "B" are two study groups, null hypothesis states that A = B or no difference between A and B. The null hypothesis is what we are trying to disprove. The aim of any research study is to find any difference that might exist between group A and B and is regarded as the alternative hypothesis (H_1). H_1 states A ≠ B. To test this alternative hypothesis, there are a few steps known as the steps of hypothesis testing. At the end of hypothesis testing, we arrive at a *P*-value. If the *P*-value is less than <0.05 (or in some cases <0.01), then the null hypothesis is rejected and the alternative hypothesis is accepted, i.e., A ≠ B.

Hypothesis testing

The following steps are to be followed:

a. Formulation of a research question and selection of appropriate research design
b. Calculation of sample size suitable to the hypothesis to be tested
c. Apply the test of statistical significance fitting to the hypothesis (stated later in the article)
d. Determine *P*-value from the results
e. Compare the obtained *P*-value with the critical value of P (either <0.05 or <0.01, as defined in the research protocol)
f. If *P*-value < critical value, reject null hypothesis and accept the alternative hypothesis (difference detected). If *P*-value > critical value, accept null hypothesis (no difference is detected).

Errors in hypothesis testing

Two types of errors can occur:

a. Type I error: Incorrectly rejecting null hypothesis. This gives rise to the chances of finding a false-positive result or detecting a difference when no such difference exists. The probability of Type I error is denoted as α. Usually, α is taken at 0.05
b. Type II error: Incorrectly accepting null hypothesis. This gives rise to the chances of finding a false negative result or inability to detect a difference when such a difference exists. The probability of Type I error is denoted by β. β error should not be more than 20%.

Power of the study: The probability of detecting a real difference when it does exist is the power of the study. It is denoted by (1-β). The accepted power of the study is set at 80%.

The 2 × 2 table below shows schematically the concept of errors [Table 1].

For example, a new drug B has come in the market for the treatment of psoriasis, which the researcher wants to test against the existing drug A. The researcher has done a study with the null hypothesis (H_0) that there is "No difference in the effectiveness between Drug A and Drug B." If in reality, there is no difference between A and B (H_0 is true), and the study has found some difference (thus rejecting the H_0) then the researcher is committing Type I error. However, if in reality there exists some difference between A and B (H_0 is false), but the study has found no difference between them (thus accepting the H_0); the researcher is committing Type II error. It is more grievous to err in terms of showing the better result when there is none; thus, Type I error is kept as 5%, and there is some relaxation with Type II error (which by convention is taken as 20%). Type I error can introduce an ineffective drug into the market causing more harm to the patients.

What does a P value < 0.05 mean?

Let us start with an example as shown in Table 2.

For the parameter "Age," comparing the age by Students' *t*-test between the groups A and B, the *P*-value is 0.442,

Table 1: 2 × 2 table showing schematically the concept of errors

	Researcher's decision	
	Fail to reject null hypothesis	Reject null hypothesis
Reality		
Null is true	Correct decision	Type I error (α)
Null is false	Type II error (β)	Correct decision, Power (1-β)

Table 2: The concept of *P*

Parameter	Group A	Group B	*P*
Age			
Mean±Standard deviation	31.06±13.98	33.02±12.05	0.442
Urticaria activity score (UAS)			
Mean±Standard deviation	4.81±3.63	6.92±4.05	0.009

which is "not significant." The null hypothesis is the age of Group A = Age of Group B. The alternative hypothesis is A ≠ B. Because *P*-value is not significant, we have to accept the null hypothesis that A = B. *P*-value = 0.442 means that the chances of having a false-positive result (that there exists an age difference between two groups when actually there is none) are 44.2%, which is very high compared to the chance factor set at 5% (or the critical value of *P*-value <0.05). Thus, the age in both the groups "A" and "B" are comparable.

In the next parameter, "UAS," *P*-value comes as 0.009, which is highly significant. We reject the null hypothesis and accept the alternative hypothesis that A ≠ B, or UAS is significantly less in Group A compared to Group B. The chance of finding a false positive result (that there exists a difference in UAS between both groups when actually there is no difference) is 0.09%, which is much below the critical value of *P* at 5%.

Pros and cons of P-value

1. The pros are that *P*-value gives the strength of evidence against the null hypothesis. We can reject a null hypothesis based on a small *P*-value.
2. The value of *P* is a function of sample size. When the sample size is large, the *P*-value is destined to be small or "significant".
   For such a situation, the confidence interval (CI) should be mentioned along with the *P*-value to arrive at a more "precise" understanding of the estimation of difference. The concept of CI is provided later in the article.
3. A large effect size can give a small *P*-value.
   The effect size is the size of the smallest clinically important effect to be detected. Preferably, the size of the effect should be based on clinical reasoning. It should be large enough to be clinically important but not so large that it is implausible. Further notes on effect size are described shortly.
4. If the *P*-value is above the critical value of *P* (say 0.05), we usually conclude that the null hypothesis is not rejected. Nonetheless, it does not mean that null is true. The safer interpretation is that there is insufficient evidence to reject null. *"Absence of evidence is not evidence of absence."*

Life beyond P

P-value has its own limitations, and at times, there is a liability of it being misused and also there are concerns raised over the fact that *P*-value is used poorly by people not properly trained to perform data analysis. In recent times, the scientific community is furthermore deeply concerned on the issues of reproducibility and replicability of scientific conclusions drawn on the basis of *P*-value. The concern was of that extent that some journal banned the use of *P*-value (null hypothesis significance testing).[3]

Understanding the fallacies of *P*-value, various other approaches are introduced to eliminate the errors introduced by null hypothesis testing. Statisticians have argued in favor of introducing "confidence interval" and "strong descriptive statistics including measures of central tendencies and variation and effect size" to help the reader understand the result of any study more comprehensively and could make a rational choice in clinical practice.

Confidence interval

CI is a measure of the precision of the results. By convention, we usually quote 95% CI. By 95% CI, we can be 95% confident that the interval will contain the true population value. The two values that define the interval are called the "confidence limits" [Figure 1]. A wide CI means that the results are imprecise, whereas a narrow CI indicates the estimate is precise. The upper and lower limits provide the means of assessing whether the results are clinically important. The CI can be calculated for mean, proportion, odds ratio, relative risk, correlation coefficient, and regression coefficient. The CI of mean is expressed as

95% CI of mean = mean ± 1.96 × Standard error of mean.

The *P*-value must be considered with 95% CI. In the diagram Figure 2, let us consider the following:

The dotted line represents 0 or a position of no difference between groups. The zone between 0 and 0.1 represents the "zone of scientific or clinical indifference." The area beyond 0.1 is the "zone of clinical relevance." The big dot represents the test statistic (in this example, let us take it as "mean"), and the straight lines from the big dot on either side represent the 95% CI of the mean. The bracket denotes the confidence limits. If the confidence limits lie on either side of 0, then the change may not be clinically relevant [Figure 2 and Table 3].

Effect Size

Effect size is the measure of magnitude of the difference between groups, which can be either absolute effect

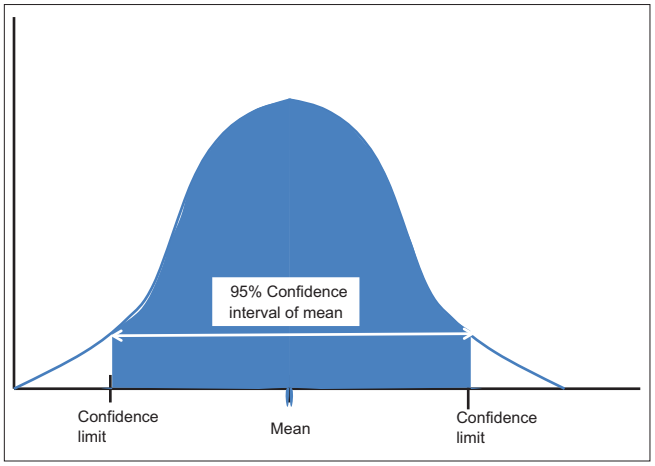

Figure 1: Graphical representation of 95% confidence interval of mean in a normally distributed (bell-shaped curve) population. The white both ways arrow area represents the 95% CI. The ends of the interval are the "confidence limits"

size (raw difference between the average, or mean, outcomes in two different intervention groups where variable understudy has intrinsic meaning e.g., age, body weight, etc) or calculated indices of effect size (which are useful when the measurements have no intrinsic meaning, e.g., physicians' global improvement score on a Likert scale). While *P*-value report statistical significance, effect size reports substantive significance.

The common indices of effect size include Cohen's *d*, odds ratio, relative risk to measure between groups; to report measure of association, Pearson's *r* correlation, r^2 coefficient of determination are used.[4]

Cohen's *d* is a widely used measure of effect size between two independent groups. It is calculated by the formula = mean (group A) – mean (group B)/pooled estimate of standard deviation. Cohen classified effect sizes as having *small practical effect* if $0.2 \leq d < 0.5$; *medium practical effect* if $0.5 \leq d < 0.8$ and large practical effect if $d \geq 0.8$.

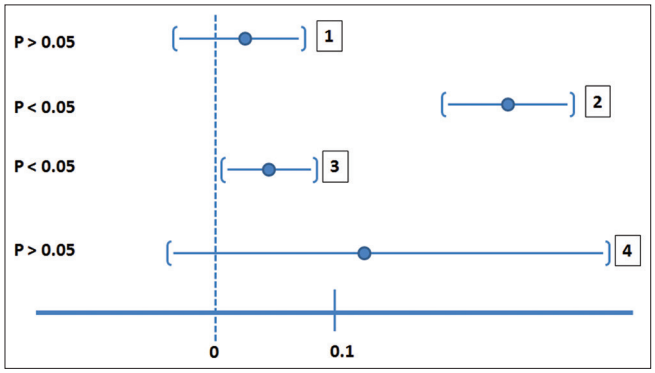

Figure 2: Concept of *P*-value with 95% confidence interval

Table 3: Explanation of Figure 2

Serial no	Scenario	Explanation	Interpretation
Scenario 1	$P > 0.05$, mean in the zone of clinical indifference, CI crossing 0 in one limit and more than 0 in the other (one on either side of 0)	Not statistically significant, Not clinically relevant, Imprecise	Rejected
Scenario 2	$P < 0.05$, mean in the zone of clinical relevance, both confidence limits greater than 0 (on the same side)	Statistically significant, Clinically relevant, precise	Clinically relevant difference → can be accepted
Scenario 3	$P < 0.05$, mean in the zone of clinical indifference, confidence limits greater than 0 (on the same side)	Statistically significant, not clinically relevant, precise	Observed change not clinically relevant though *P*-value is significant Rejected.
Scenario 4	$P > 0.05$, mean in the zone of clinical relevance, CI crossing 0 in one limit and more than 0 in the other (one on either side of 0)	Not statistically significant, clinically relevant, imprecise	The *P*-value does not reflect the clinically relevant change, also the CI is wide. Rejected

The concept of solely relying on effect size is also criticized and a study after analysis of publication after the ban on "null hypothesis significance testing" found that results of those articles were seemingly being overstated beyond what the data would support if *P*-values (or some other form of statistical inference) had been used.

Thus, it cannot be overemphasized that both inferential statistics and descriptive statistics have their own place, and it will be apt to quote "*problem lies not so much with P* values *in themselves as with the willingness of researchers to lurch casually from descriptions of data taken from poorly designed studies, to confident generalisable inferences."*[5]

Null Hypothesis Significance Testing

The test to be selected to find the significance of "Null hypothesis" is guided by various parameters, including the type of data, the number of groups, and of course the setting in which the data is acquired. In the present time, when statistical software's are available to help in calculating the results a bio-medical researcher should choose the right test to use in the right setting.

A. When the difference between groups is to be tested, ***"3 + 1 question approach"*** can be adopted [Figures 3-5]:
Example: Difference in PASI score when psoriasis patients are treated with two treatment modalities, methotrexate, and apremilast or difference in urticaria activity score (UAS) when urticaria patients are treated by levocetirizine and olopatadine.[6]

The approach can be enumerated as follows:

Question 1. Is the data "numerical" or "categorical"?

Categorical or qualitative data do not require measurement. The object to be studied is grouped into categories according to qualities. "Scores" are used in several clinical settings (e.g., PASI score and UAS score) when we cannot measure a quantity and are taken as "qualitative data"

Numerical or quantitative data have measurable data expressed in numbers. It can be continuous (take any range of value such as fractions or decimals), discrete (only integrate value), percentages, ratios, and rates

If the data is "numerical" than additional question needs to be asked.

Question 1a. Is the data "parametric" and follows a normal distribution or "non-parametric"?

Normal distribution is a unimodal distribution represented by a bell-shaped curve, which flattens symmetrically on both ends. The data are considered

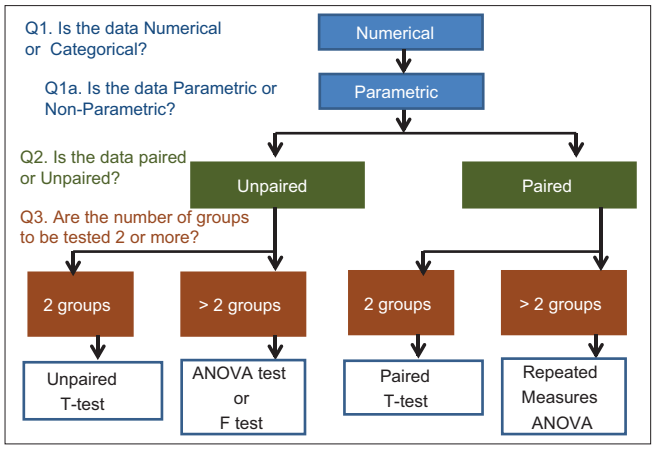

Figure 3: Null hypothesis significance tests to be used while testing the difference between groups of numerical parametric data

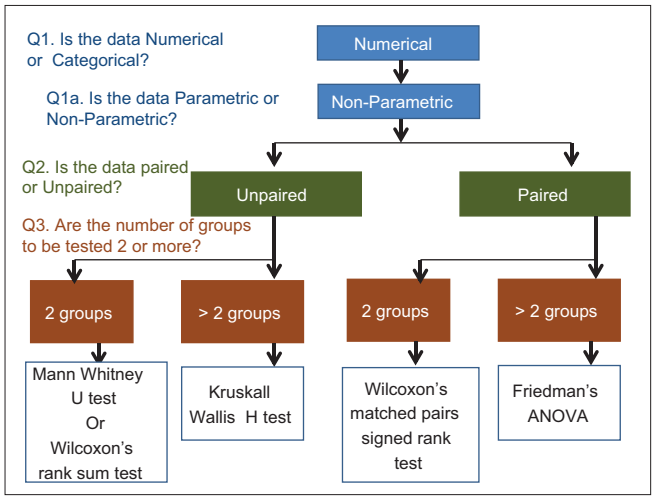

Figure 4: Null hypothesis significance tests to be used while testing the difference between groups of numerical non-parametric data

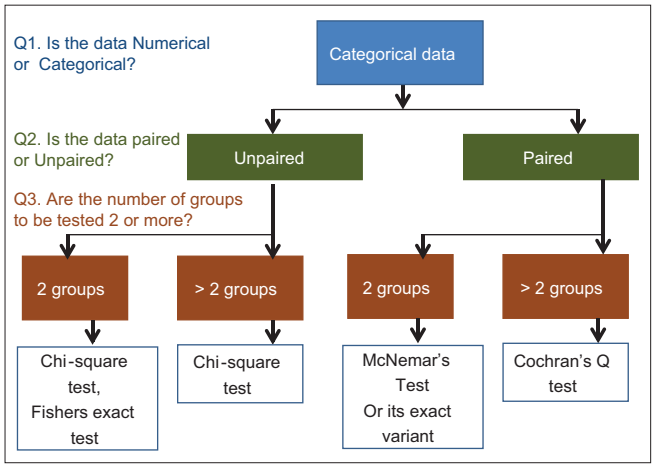

Figure 5: Null hypothesis significance tests to be used while testing the difference between groups of categorical data

to be parametric when its distribution in the underlying population can be represented by the normal distribution curve. Non-parametric data are those which follow a distribution other than a normal distribution, skewed distributions, or does not follow any distribution or follows an unknown distribution

Usually, a sample size of 100 or above follows a normal distribution. We can apply the tests of normality for a sample size <100 and see whether it is normally distributed or not. The tests for normality are easily available in all statistical software. The data can be tested for normality using Kolmogorov–Smirnov test, D' Augustino Pearson test, or Shapiro–Wilk test.

Question 2: Is the data "paired" or "unpaired"?

Paired data means when the result of one influences the result of another. Examples: Crossover studies, before-after tests, duplicate or triplicate, or repeated measurements of the same set, twin studies, right-left body part/eye

Question 3: Are the number of groups to be tested 2 or more than 2?

Multiple group (>2 groups) comparison tests are followed by "post–hoc" tests to find where the significance lies when there is a significant result. For parametric tests (ANOVA, Repeated measures ANOVA), Tukey's test or Dunnet's test is used. For non-parametric tests (Kruskal–Walis ANOVA, Friedman's ANOVA), Dunn's test is used.

B. For the testing association between 2 variables:

1. Numerical parametric data → Pearson's product-moment correlation coefficient (r)

   Correlation quantifies the strength of the linear relationship between two random variables. The correlation takes a value between +1 and –1. The positive sign indicates a positive correlation (if the value of one variable increases, the value of the other also increases), whereas the negative sign indicates a negative correlation (inverse relation where the value of one decreases with increase in the other). The more the test parameter is near to +1 or –1, more strong is the correlation

   E.g., The research question of "Correlation of dermoscopy and histopathological characteristics in actinic keratosis" was addressed by Pearson's product-moment correlation coefficient (r) to correlate the orthokeratosis and parakeratosis[7]
2. Numerical non-parametric data à Spearman's rank correlation coefficient (ρ) or Kendall's rank correlation coefficient (τ)

   E.g., The research question of "Correlating the changes in Dermatology Life Quality Index (DLQI) with change in PASI," was addressed by Spearman's rank correlation coefficient (ρ).[8]
3. Categorical data in 2 × 2 table → Odds ratio, risk ratio, or relative risk

   E.g., The research question of "Assess the association between psoriasis and metabolic syndrome" was addressed using the odds ratio.[9]

4. Categorical data other than 2 × 2 table → logistic regression, chi square for trend
   E.g., The research question of "Association between vitiligo extent and distribution and Quality-of-Life Impairment" was addressed using the logistic regression model.[10]

C. For testing agreement between assessments:
   These tests are used for screening tests; diagnostics tests; and validation of rates, scales, and scores.
   Example: Dermoscopy as a diagnostic tool, MASI score for melasma
   1. For quantitative data → Intra-class correlation coefficient and Bland–Altman plot
   2. For qualitative data → Cohen's kappa statistics and Kendall's coefficient of concordance.

D. For determining survival or time trends:
   Survival data are always non-parametric. Example: Survival time after suffering from toxic epidermal necrolysis
   1. For 2 groups → Logrank test, Cox–Mantel test, and Gehan's test
   2. For >2 groups → Logrank test, Peto and Peto test.

Conclusion

P-value should be presented with 95% CI as together they give a better understanding of the results. Being medical statisticians, the increased availability of statistical software aid us in using the various tests of statistical significance. The proper understanding of the nature of the data set, whether quantitative or not, normally distributed or not is the primary requirement for selection of the test of significance.

It is important to understand that it is the responsibility of individual researcher to understand that all the interpretations have got their own biases and limitations; thus, it is onto them to choose the statistics well and to do good to science.

"*Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalized, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary"*[11]

……. *Francis Galton*

Acknowledgment

The Authors would like to express their thanks to Dr Pramit Ghosh, MD (Community Medicine), Department of Community Medicine, Purulia Government Medical College and Mr Tuhin Kanti Das, MBA (Duke University-Fuqua School of Business), Accenture Strategy, North America for extending their valuable opinion and reviewing the manuscript.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. De D, Singh S. Basic understanding of study type and formulating research question for a clinical trial. Indian Dermatol Online J 2019;10:351-3.
2. Biau DJ, Jolles BM, Porcher R. Value and the theory of hypothesis testing an explanation for new researchers. Clin Orthop Relat Res 2010;468:885-92.
3. Wasserstein RL, Lazar NA. The ASA Statement on *P*-values: Context, process, and purpose. Am Stat 2016;70:129-133.
4. Sullivan GM, Feinn R. Using effect size—Or Why the *P*-value is not enough. J Grad Med Educ 2012;4:279-282.
5. Matthews RAJ, Wasserstein R, Spiegelhalter D. The ASA's *P*-value statement, one year on. Significance2017;14:38-41.
6. Sil A, Tripathi SK, Chaudhuri A, Das NK, Hazra A, Bagchi C, *et al.* Olopatadine versus levocetirizine in chronic urticaria: An observer-blind, randomized, controlled trial of effectiveness and safety. J Dermatolog Treat 2013;24:466-72.
7. Lee DW, Kim DY, Hong JH, Seo SH, Kye YC, Ahn HH. Correlations between dermoscopic and histopathologic findings in actinic keratosis. J Investigative Dermatol 2017;34:137.
8. Hesselvig JH, Egeberg A, Loft ND, Zachariae C, Kofoed K, Skov L. Correlation between dermatology life quality index and psoriasis area and severity index in patients with psoriasis treated with ustekinumab. Acta Derm Venereol 2018;98:335-9.
9. Singh S, Young P, Armstrong, AW. An update on psoriasis and metabolic syndrome: A meta-analysis of observational studies. PLoS One 2017;12:e0181039.
10. Silverberg JI, Silverberg NB. Association between vitiligo extent and distribution and Quality-of-Life impairment. JAMA Dermatol 2013;149:159-64.
11. Fricker RD, Burke K, Han X, Woodall WH. Assessing the statistical analyses used in basic and applied social psychology after their *P*-value ban. Am Stat 2019;73:374-84.