**Science Journal of Applied Mathematics and Statistics**
2017; 5(1): 41-48
http://www.sciencepublishinggroup.com/j/sjams
doi: 10.11648/j.sjams.20170501.16
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

# The American Statistical Association (ASA) Statement of 2016 on Statistical Significance and P-value: A Critical Thought

**Silas Memory Madondo**

Department of Research, Mount Meru University, Arusha, Tanzania

**Email address:**
silasethan@gmail.com

**To cite this article:**
Silas Memory Madondo. The American Statistical Association (ASA) Statement of 2016 on Statistical Significance and P-value: A Critical Thought. *Science Journal of Applied Mathematics and Statistics.* Vol. 5, No. 1, 2017, pp. 41-48. doi: 10.11648/j.sjams.20170501.16

**Received**: December 6, 2016; **Accepted**: January 4, 2017; **Published**: January 24, 2017

**Abstract:** A study on American Statistical Association (ASA) policy statement on statistical significance testing and p-value of 2016 was carried out in Tanzania. The purpose of the study was to explore the feelings and reactions of university statistics tutors towards the American Statistical Association policy statement on statistical significance testing and p-value of 2016. A sample of 9 statistics tutors from different disciplines were selected from public and private universities via heterogeneous purposive sampling to participate in the study. Respondents had mixed feelings towards ASA policy statement of 2016. The ASA policy statement was criticized for being shallow in depth, subjective and failing to answer the core problems raised against the use of Null Hypothesis Significance Testing (NHST) and p-value. The ASA policy statement was dismissed as a non event with nothing new to offer. However, despite being shallow, the ASA policy on NHST and p-value is likely to trigger a health debate on the shortfalls of NHST and p-value and the debate will eventually lead to a breakthrough.

**Keywords:** American Statistical Association, Heterogeneous Purposive Sampling, Null Hypothesis Significance Testing, P-value, Null and Alternative Hypotheses

## 1. Introduction

The policy statement availed by ASA in 2016 brought a lot of mistrust and low levels of confidence on statistical significance and p-value among the positivists. Statisticians raised a lot of queries just soon after the introduction of p-value in 1925 by Sir Ronald Fisher [6]. The ASA team's move to offer a policy statement on significance testing and p-value was a popular event to be remembered for years in the history of inferential statistics. The aim of this paper is to explore the feelings and reactions of university statistics tutors towards the ASA policy statement of 2016 on NHST/p-value.

The ASA Executive Board formed a taskforce led by Wasserstein in 2015. The taskforce was assigned to deliberate on the queries raised against NHST/p-value and come up with a policy position. In October 2015, a taskforce of 20 experts met in USA at ASA offices in Alexandria, Virginia for two days. Regina Nuzzo chaired the meeting. The debate set in motion the development of ASA policy statement with six key principles regarding NHST/p-value which was published in 2016 [5].

This paper is divided into the following sections; the origin of inferential statistics, controversies associated with null hypothesis testing/p-value, events/processes leading to the development of ASA's policy statement of 2016 on significance tests/p-value, a critique of the ASA policy statement, methodology and results/discussions. The policy statement by ASA exposed the trustworthiness of null hypothesis significance testing and p-value. The popularity of inferential tests will definitely suffer a serious blow in the near future.

### *1.1. The Origin of Inferential Statistics*

According to reference [2], the first known hypothesis test was the trial of the Pxy, a periodic ritual of the Royal mint which was introduced in London in 1279. It was an acceptable procedure of testing the manufacturing standards of coins. Astronomers were said to have been involved in hypothesis testing since 1700 when they tried to determine the position of the moon. According to references [8] and [6], Sir Ronald Fisher proposed the p-value in 1925.

Fisher was a British statistician and biologist [8] He taught mathematics and physics in public schools before proceeding to Rothamsted Experimental Station where he worked as a statistician [2]. While at Rothamsted Experimental Station, Fisher conducted and analyzed experiments in agronomy and biology as a statistician. The experience of working as a statistician at the Rothamsted Experimental Station exposed him adequately to the field of inferential statistics.

There is also a view that Sir Ronald Fisher was influenced by Sealy Gosset, Gosset played a key role in the history of inferential statistics when he introduced the concept of 'the probable error of the mean' [2]. Whether it was Sealy Gosset's influence or the experience from Rothamsted Experimental Station, there is no doubt that Fisher was among founders of inferential statistics. However, Sealy Gosset should be credited for being among the pioneers/founders of inferential statistics and his role in influencing Fisher shouldn't be overlooked.

Reference [10] stated that Fisher was the first scholar to write a textbook in statistics. He published a textbook titled ' Statistical Methods for Research Workers' in 1925 and managed to introduce the concepts of p-value and null hypothesis. It is interesting to note that Fisher was the first to introduce the concept of null hypothesis but was subtle about alternative hypothesis. All in all, Sir Ronald Fisher is credited for introducing p-value concept and null hypothesis. He was the first person to write a textbook in statistics and provoked debate in the area of inferential statistics.

The textbook written by Fisher in 1925 provoked a serious debate in the field of inferential statistics. It gave rise to the birth of new scholars in hypotheses studies, a Polish scholar Jerzy Neyman with the help of a British scholar Egon Pearson challenged the work of Fisher. Fisher and Neyman/Pearson differed on philosophy and there was a cold war between their supporters. The cold war between the two camps gave rise to alternative hypothesis. Today, null and alternate hypothesis are widely used and it was the conflict between Fisher and Neyman/Pearson that gave birth to the two key types of hypotheses tests.

### *1.2. Controversies Associated with Null Hypothesis Significance Testing/P-value*

This section is based on the advice from references[6, 2, 1, 5, 4,8, 9, 7]. These scholars played a key role in exposing the dark side of p-value and null hypothesis significance testing. The work published by these intellectuals guided the ASA taskforce to come up with the 2016 policy statement on significance testing/p value and they either contributed directly or indirectly to the ASA policy statement.

In the first section of this paper, I pointed out that controversies on significance testing started soon after the introduction of p-value and null hypothesis by Fisher in 1925. According to reference [6], there was a cold war between Fisher and Neyman regarding the p-value and null hypothesis. Neyman termed Fisher's work *'worse than useless'* while Fisher labeled Neyman's approach *'childish and horrifying'* for intellectual freedom in the west. However, the cold war gave rise to the wide spread usage of both null and alternative hypotheses.

According to Goodman as cited by reference [6], p-value was never meant to be used the way it is being used today. Reference [6]stated that p-value was never meant to be permanent. She argued that *'Fisher coined p-value in 1920s but did not mean it to be a definitive test'.* Schmidt and Hunter (2002) as cited by reference [4] stated that significance tests are disastrous method for testing hypothesis. Reference [3] resisted the use of p-value and significance tests while supporting his work of effect sizes. Cohen saw the resistance against the use of p-value coming way back in 1990s but his call was a lone voice in the wilderness.

Reference [7] observed through empirical study that above 11% of the work guided by p-value and NHST published by medical journals is false and misleading. What it means is that some of the research findings have been misleading policy makers since 1925. It is an open secret that research findings can be expensive to institutions when it comes to implementation. The ASA's move on new policy direction is therefore recommendable.

### *1.3. Arguments for the Rejection of P-value/Significance Testing*

I read a lot of literature supporting the rejection of p-value and significance testing. The literature included the work cited by the ASA taskforce which was considered during their debates. I read the work of Cohen who is regarded as the founder of 'effect sizes'. I also managed to explore the American Psychological Association 2001 taskforce report which once recommended for the rejection of the null hypothesis significance testing and am also privy to the move by the Basic and Applied Social Psychology to reject the use of p-value by their authors in their journal.

The Basic and Applied Social Psychology (BASP) journal rejected the use of p-values by their authors because they are convinced that p-value supports poor quality research work. The threshold of $p<0.05$ is easy to pass for both poor and good quality researches/studies [11]. It is sad to note that p-values are not able to distinguish between poor and good quality work. According to Stephen Ziliak as cited by reference [6], p-values are neither reliable nor objective. NHST/p-value always fail in replication of experiments and does not give the probability that null hypothesis is false contrary to the beliefs of some scientists. Should we therefore trust p-value and NHST?

I am forced to believe beyond any doubts that the work of Doctor Andy Field which was published in 2005 summarized well the controversies surrounding the use of p-value and significance testing. There is no doubt that the work of Field inspired a lot of recent statistics scholars' perceptions on NHST/p-value. Reference [4] summarized evidence against the p-value and significance testing into five categories.

The first argument against p-value and significance testing is that null hypothesis is never true [3]. Cohen as cited by reference [4] argued that p-value is meaningless because it is based on the assumption that can never be true. I fully support Cohen's argument because p-values are based on researcher's beliefs and not the reality on the ground. It is therefore important to note that null hypothesis is incompatible with the research data or the reality on the ground. Null or alternative hypotheses are compatible with researcher's beliefs and are capable of misleading the consumers of the research output.

Reference [4] further stated that Null Hypothesis Significance Testing (NHST) is misunderstood and that resulted in misinterpretation and dissemination of false findings over the years. Majority of researchers created their own interpretations of p-value and they wrongly think that p-value can be used to measure the size of magnitude or effect. They wrongly think that p-value is the probability that the results would be replicated if the experiment was conducted a second time. They wrongly think that p-value is the probability that the results are due to chance, the probability that the null hypothesis is true [4].

Reference [4] demonstrated by examples from SPSS tables that null hypothesis significance testing depends upon sample size. Tables 1&2 below show the independent t-tests results from two cases based on the same scenario. In both cases, a mean difference of -2.21 is observed but it is shocking to note differences in t-values and significance. First case is showing a significant result while in second case a non-significant result despite having similar mean differences of -2.21. The conflict in conclusion is caused by different sample sizes per case.

***Tables 1.*** *Presents the first case from the SPSS data.*

| | Levene' Test for Equality of Variance | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | Sig. | t | df | Sig. (2-tailed | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | Lower | Upper |
| Equal Variance Assumed | 1.046 | 0.308 | -6.22 | 198 | 0.000 | -2.20699 | 0.35472 | -2.90650 | -1.50748 |
| Equal Variance not Assumed | | | -6.22 | 195.3 | 0.000 | -2.20699 | 0.35472 | -2.90656 | -1.50742 |

Source:[4]

***Tables 2.*** *presents the second case from the SPSS data.*

| | Levene' Test for Equality of Variance | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | Sig. | t | df | Sig (2-tailed | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | Lower | Upper |
| Equal Variance Assumed | 2.744 | 0.136 | -1.510 | 8 | 0.169 | -2.20736 | 1.46173 | -5.57810 | -1.16339 |
| Equal Variance not Assumed | | | -1.510 | 5.23 | 0.189 | -2.20736 | 0.46173 | -5.91660 | -1.50188 |

Source 4

Reference [4] gave another scenario where a mean difference of zero is observed between groups. The result shown in table 3 below proved to be significant basing on the connection of the p-value of 0.22 and the significance level of 0.05 ($p<0.05$). The expatiation was that the result could have been non-significant because the mean difference between the groups is zero. The confusion is caused by the differences in sample size.

***Table 3.*** *presents the confusion in the second scenario of NHST.*

| | Levene' Test for Equality of Variance | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | F | Sig | t | df | Sig (2-tailed | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | Lower | Upper |
| Equal Variance Assumed | 0.994 | 0.319 | -2.296 | 999998 | 0.022 | 0.00 | 0.00200 | -0.00851 | -0.00067 |
| Equal Variance not Assumed | | | -2.296 | 999997.4 | 0.022 | 0.00 | 0.00200 | -0.00851 | -0.00067 |

Source: [4]

It is therefore strange to note that NHST can produce parallel conclusions on the same event. Reference [8] argued that p-value should not change in replication if it represents the truths. The bottom line is, if p-value represents the truth, therefore the truth should remain constant or unchanged. It is true that the computations and conclusions based on NHST are affected by the sample size and not the reality of a phenomenon on the ground. Therefore, NHST is misleading researchers.

NHST according to reference [4] is illogical and based on probability, reasoning and syllogism. He further argued that $p< 0.05$ is completely arbitrary. Should we accept that $p=0.05$ because Fisher said so? According to reference [5], a professor of mathematics and statistics George Cobb raised critical questions during one of ASA forums. He asked;

a) Why do so many colleges and graduate schools teach p=0.05?

b) Why do so many people still use p=0.05?

Fisher's selection of 0.05 was not based on sound empirical evidence and a lot needs to be done in order ensure uniformity on the selection of the levels of significance. I strongly believe that a scientific computation of levels of significance based on the nature of the study, sample size and other variable should be developed in order to ensure uniformity of levels of significance selection across disciplines.

Statisticians managed to observe the dark side of NHST/p-value since time immemorial and are currently making a call for the replacement/rejection or complementation of NHST/p-values with other statistical tests. Reference [3] challenged the use of p-value and argued that 'effect sizes' are better than p-value. Reference [4] supported Cohen and wrote expensively against NHST/p-value and in support of 'effect sizes'. Reference [8] supported the use of effect sizes too, as well as bayes' rule and confidence interval.

The discussion on P-value and NHST above clearly exposed four major findings about their future use and credibility. The first observation is that p-value contributed with false findings in the field of research. It is therefore obvious that some of the national/international policies and positions formulated basing on the findings from p-value/NHST were misled. There is a strong nexus between policy and research and scholars should always aim at producing credible results.

It is also observed that p-value is misunderstood and misinterpreted by some researchers. Quite a number of researchers and journals do not understand the correct interpretation and use of p-value. This has resulted in several interpretations of p-value basing on the interests of individual researchers and scientists. The ASA of 2016 has nothing new to offer but aiming at highlighting the misinterpretations of NHST/p-values.

It is also observed that apart from being misinterpreted, p-value is also misleading researchers. Proper evidence should be given on the selection of p=0.05. The selection of significance levels should be verified by empirical evidence rather than rationality. The effect of sample sizes on the conclusions of NHST/p-value should not be ignored and it is the sample size that shape the conclusions basing on p-value rather than the reality on the ground.

It is also observed that p-value/NHST should be replaced or complemented with other statistical tools. Scholars suggested the use of effect sizes, bayes, confidence intervals and other tools to replace or complement NHST/p-value. This signifies that p-value/NHST are weak and cannot provide strong findings to lean on.

I still have a strong feeling that NHST/p-value will suffer a setback in popularity and usage. The general agreement from the literature is that p-value/NHST are weak, unreliable and not scientific. The dependability, credibility and conformability of the findings from these tests should be a serious concern in positivist paradigm. This led to development of a controversial ASA policy statement on p-values and significance testing.

### *1.4. ASA Policy Statement on Statistical Significance and P-values*

According to reference [5], the recent resistance on the use of NHST/p-value pushed the ASA board to provide a policy statement explaining their position. Reference [11] stated that the Basic Applied Social Psychology journal has already rejected the use of p-value because it supports poor quality research work.

The ASA formed a taskforce in 2015 led by Wasserstein to look on the controversies surrounding the NHST/p-value. The purpose of the taskforce was to help them to come up with a position on issues raised against p-value and significance testing. According to reference [5], a team of experts met in October 2015 at ASA offices Alexandria, Virginia in USA.

The team debated for 2 days and Regina Nizzo facilitated the meeting. There was a serious debate which gave rise to a policy statement that was later approved by the ASA board on 29 January, 2016. Twenty statistics experts were involved during the development of the policy statement. The ASA statement started by acknowledging that p-value is still important in data analysis but it's being misused and misinterpreted by researchers. The policy came up with 6 principles;

a) P-values can indicate how incompatible the data are with a specified statistical model.

b) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

c) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

d) Proper inference requires full reporting and transparency.

e) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

f) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

### *1.5. Critique of the ASA Policy Statement*

In as much as I appreciate the great work done by Wasserstein and his team, there are some loopholes which may likely to affect the acceptability of the ASA policy statement to the consumers of research output. No matter how genuine and good a policy is, it may face resistance if the process and means to the end were flawed. According to reference [5], the team of 20 experts met for 2 days in Alexandria, Virginia under the facilitation of Regina Nuzzo in October 2015.

The procedure of selecting the 20 experts was not disclosed, however, two days could have been insufficient to handle a complex problem of such magnitude. I agree with some of their conclusions but the duration of their debate is questionable and ASA could have improved on transparency. They should have involved wider network of stakeholders during the process. The report should have been preceded by an apology to policy makers and research output consumers who could have been misled by false research output that was disseminated to the consumers. It is not the duty of ASA to apologize but acknowledgment was highly called for.

They should have followed an example from Tuskegee Syphilis Study, 399 African Americans were denied syphilis treatment for 40 years by officials of public health services. The taskforce asked to look on the matter requested the then president of United States Mr. Bill Clinton to apologize to the victims and Americans at large[10].

I am strongly convinced that only one out of the six principles introduced by ASA is important and a breakthrough. Principal number three states that 'scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold'. The rest of the principles are a reminder of the misinterpretations of the p-value. It is not logical for any institution to give out a statement full of information/content already flooded in the academic arena. The policy statement acknowledged that researchers are misusing and misinterpreting p-value but the report is subtle about the argument against p-value and NHST, that they are misleading scientists. There was no solution given to the impact of sample size on the conclusions. Scholars could have expected the policy statement to handle the challenges related to the belief that p=0.05. I still believe that the selection of the level of significance should be based on practical evidence.

The policy statement stated that scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold but the policy still supports the position that p-value is relevant. How can p-values become relevant if they are not supposed to be considered for business and policy decisions? My feeling is that ASA's statement was a humble rejection of NHST/p-value.

The policy statement supports the view that p-values should be complemented with otherstatistical tools like effect sizes, bayes, confidence interval and others but did not tell us how they reached that agreement. A strong discussion could have been done to ascertain the methods to complement p-value and there is still a lot to be done.

# 2. Methodology

The study adopted a qualitative approach and the type qualitative design used was phenomenology. Phenomenology according to reference [12] is used to get information about the lived experiences from a group of individuals. The study collected data through structured interviews, soliciting information about the feelings and reactions of statistics university tutors towards American Statistical Association policy statement of 2016 on NUST/p-values. Nine lecturers of statistics in various disciplines from public and private universities in Tanzania were selected to be part of the study using heterogeneous method of purposive sampling. The interviews were recorded and manually transcribed. The data was analyzed using open coding and the discussion was captured into seven codes.

# 3. Results and Discussion

The discussion is based on the analyses of interview data from nine statistics professionals teaching at university level in various disciplines. The respondents were drawn from departments of science and mathematics. Three of the respondents were statistics tutors for engineering discipline, four were mathematics tutors while two were tutors for biology and physics with extensive experiences of teaching statistics at university level. The aim of the study was to explore the feelings and reactions of statistics university tutors towards the American Statistical Association 2016 policy statement on significance testing/p-value. The reactions and feelings of the respondents were categorized into 7 codes.

## *3.1. Nature and Credibility of P-Values*

The respondents had mixed reactions on the nature and credibility of p-value/NHST. Respondent number nine argued that p-value is not based on reality but rather on probability. This is what he had to say;

*'p-value is based on probability and probability is not a reality. It is therefore a grave mistake to base our conclusions on probability supported tools. We need to make use of data analysis tools that are close to reality like the effect sizes.......'*

The respondent number nine went on to say that probability is not different from beliefs/metaphysics and it is therefore unacceptable to rely of beliefs. NHST according to reference [4] is illogical and based on probability, reasoning and syllogism. Reference [4] does not trust probability measures as well and he supported the rejection of p-value/NHST.

Respondent number one was of the view that p-value and NHST are not scientific and should be therefore handled with extra care. He argued that p-value cannot be scientific because it is not consistent when it comes to replication of experiments. The view was also supported by Stephen Ziliak as cited by reference [6] who stated that p-values are neither reliable nor objective. They cannot be trusted because different conclusions can be reached in the event of replication of experiments.

All the respondents except researcher number 4 were not comfortable with the level of significance of 0.05 they said that the level of significance can be either higher or lower than 0.05 depending with several factors. Researcher number 3 suggested 3 factors that may be considered when determining the level of significance. The level of significance selection should be based on the nature of the study, previous studies and the perceived probability of occurrence of an event or phenomenon under the study. Researcher number 3 argued that the use of p-value is not a problem but the problem is on misinterpretation. This is what he had to say;

*'p-value cannot give you conclusive findings but it guides you to the conclusion...'*

According to respondent number 3 p-value is there to guide researchers to make desirable conclusions and researchers should know how to handle them in order to make acceptable decisions. Respondent number 4 claimed that p-value cannot tell the strength of the relationship between variables. He argued that p-value is incompatible with applied statistics and should be treated with caution when applied to statistics;

*'p-value originated from pure statistics and using p-value in different disciplines and applied statistics is a mistake. There are other several problems in statistics which should be solved, statisticians have not yet agreed on what we mean by large sample size, some say 30, others 50 and so on... challenges are in every discipline and not only restricted to p-values.'*

According to reference [5], a professor of mathematics and statistics George Cobb raised critical questions during the one of ASA forum. He asked;

a) Why do so many colleges and graduate schools teach p=0.05?

b) Why do so many people still use p=0.05?

It is interesting to note that the respondents are sharing the same feeling with George Cobb on the use 0.05 as the level of significance. The use of 0.05 as a parameter may change depending with the nature of the study and other factors. Reference [6] also argued that Fisher was not expecting researchers to use p-value the way they are using it today. Reference [4] questioned whether researchers are using p-value because Fisher said so.

### 3.2. Future Use of P-value

Respondents number 1 and 8 differed with the rest, they are of the view that p-value and NHST should be banned. They believed that researchers will eventually reject NHST/p-value because they are based on probability/beliefs. They also believed that NHST/p-value are not scientific and should be reject. The Basic and Applied Social Psychology (BASP) journal rejected the use of p-value by their authors because they are convinced that p-value supports poor quality research. P-value of $p<0.05$ is easy to achieve for both poor and good quality researches/studies [11].

The rest (7 respondents) argued that the basis of rejecting p-value/NHST is not strong and those pushing for their rejection should reconsider their positions. Respondents number 6 and 7 felt that statisticians should improve the p-value in order for it to meet the expected standards rather than rejecting them today after 91 years of existence. Respondents 2, 3, 4, 5 and 9 supported the idea of complementing p-value with other tests like effect sizes, confidence intervals, bayes etc. This position was also supported by reference [8] and the latest ASA policy statement (2016) on significance testing/p-value.

### 3.3. Credibility of ASA Process

Out of the nine respondents, eight expressed their dissatisfaction with the procedures used by ASA to develop the 2016 policy statement on significance testing/p-value. Respondent number 3 said;

*'ASA made a blunder by involving a paltry of 20 individuals and selecting 20 scholars only from the ocean of scholars is a mistake'*

Respondent number one questioned the process of sample selection and was of the view that the selection of the 20 experts was not clear. According to respondent number one, ASA could have used a representative scientific method of selecting experts. The approach for sample selection could have been objective and inclusive. Respondent number 2 argued that ASA should have gone beyond boarders to select statistics experts. Experts from various continents should have been involved in order to make the outcome appealing to the majority of scholars.

Respondent number 7 highlighted that *'no matter how good the policy is, if the process is not inclusive, people will definitely reject it.'* Respondent number seven emphasized the need to satisfy everybody to be affected by a policy in order for it to be acceptable. Respondents 2,5,6,7 and 8 questioned the two day meeting of 20 experts which was chaired by Regina Nuzzo. They said that 2 days may not have been enough to create an acceptable policy statement on p-value and NHST. Respondent number 4 was in support of the two day meeting of experts. He argued that;

*'... number of days does not matter but what matters is the content....'*

The respondent number 4 was satisfied with the policy content and urged scholars to concentrate on the content rather the policy process. Reference [5], believe that the process that led to the policy statement on p-value/NHST was above board. Wasserstein was the leader of the team and he claimed that the debate which led to the policy statement was hot but the document is subtle about the methods used to select experts, their geographical and career distribution.

### 3.4. Contradiction of ASA Policy Statement

Out of 9 respondents, 7 believed that ASA policy statement is contradicting itself. They supported their arguments basing on principle number three *'scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.'* respondent number 1 said;

*'ASA still believes that p-values are important in data analysis but they are discouraging institutions and individuals to make business and policy decisions basing on p-value alone. What does this mean?'*

Respondent number 7 believed that;

*'ASA policy statement sounds to be a polite rejection of p-value, telling consumers to stop using p-value for business decision making is a total rejection of p-value..........'*

I still feel that principle number three may be misinterpreted by scholars to mean a complete rejection of p-value. The principle is likely to force some researchers to develop negative attitudes towards using inferential statistics. Respondents 3 and 4 believed that there is no contradiction within the ASA policy content. Respondent 4 said;

*'.... no contradiction at all, ASA is just trying to put p-value in its correct position....'*

Respondent number 4 further argued that researchers have been misinterpreting p-value for a long time and coming up with a policy position was a noble move. Respondent number 3 could not grasp the contradiction within ASA policy. He believed that ASA was trying to deal with misuse and misinterpretations of p-value/NHST.

### 3.5. Depth of ASA Policy Statement

All the respondents agreed that ASA policy statement was not strange to them and ASA is trying to address facts which are known and that there is nothing new in the policy statement. Respondents 1,3,6,8 and 9 said that ASA statement is only addressing the problems related to misinterpretation of p-value/NHST and they gave examples of principles number 1,2,4,5 and 6. Respondent number 2 said;

*'ASA failed to address the core problems of p-value like the problems related to 0.05, replication and the effect of sample size on conclusions. They only spent much of their time addressing trivial issues like p-value misinterpretations.'*

Reference [4] came up with five reasons while proposing the rejection of NHST/p-value but only one reason on misunderstanding of p-value was answered by ASA policy statement of 2016. The rest (4 reasons) were not answered by the ASA policy statement of 2016. I also support the view that ASA policy statement is shallow in death because it only managed to address problems related to the misunderstanding and the policy should have directly handled other problems raised.

### 3.6. Complementing P-Value by Other Statistics Tools

The ASA policy statement on p-value and significance testing recommended the use of other statistical tools like bayes, confidence intervals, effect sizes etc to complement p-value. The respondents welcomed the move though they raised some questions against the procedure ASA used to reach the decision on the complementation of p-value. Respondent number 1 questioned the methods used to reach the decision. he said;

*'... one cannot wake up over night and tell researchers to complement p-value with other statistics tools without conducting any studies to verify and accept the decision... a lot of groundwork was needed before making a decision'*

Respondent number 6 said;

*'... there was need for a scientific study before a decision on the right statistics tools to complement the p-value was reached.'*

Respondents number 3 and 4 questioned the use of confidence intervals in complementing the p-value. Respondent number 4 argued that both p-value and confidence intervals are based on probability and complementing p-value with confidence intervals may be a mistake. They argued that serious research should be done before judging on the statistics methods to complement p-value. I agree with the view of respondents 3 and 4 because, there was no evidence to support that the selection of statistical methods was done after strong scientific research.

### 3.7. Overall Assessment of ASA Policy Statement on NHST/P-Value

There was a consensus among the 7 researchers that ASA policy statement of 2016 failed to address the core problems of p-value. Respondent number 2 agreed that the policy statement was shallow but it will provoke a health debate among the statisticians. He said that;

*'In as much as I agree that ASA policy statement was shallow, there is no doubt that the policy will provoke debate among the scholars of statistics and eventually a solution will be found.'*

Responded number 4 said that the ASA policy statement was a breakthrough and they managed to address the common misinterpretations of p-value. Respondent number 3 was in support of the policy statement but did not agree with the process that led to the policy statement. Respondent number 1 totally rejected the policy statement and labeled it a non event and was totally against the use of NHST/p-value.

## 4. Conclusion

The ASA policy statement of 2016 on significance testing and p-value has been hailed for dealing with the problems of misinterpretations of p-value. However, the policy was criticized for being shallow and failing to answer the key problems of NHST/p-value. The process of coming up with the ASA policy statement was criticized for being subjective and unscientific.

## Acronyms

APA - American Psychological Association
NHST - Null Hypothesis Significance Testing
USA - United States of America

## Acknowledgements

I want to express my gratitude to Mount Meru University for sponsoring this work. Special thanks goes to Engineer Tunzo Mnzava for his outstanding contribution towards this work.

## References

[1] American Statistical Association Releases Statement on Statistical Significance and P-values, March 7, 2016.[http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN].

[2] Curran-Everett, D. (2009). Explorations in Statistics: Hypothesis Tests and P-Values. *AdvPhysioEduc,33*, 81-86.

[3] Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

[4] Field, A. (2005). Effect Sizes-Statistics Hell. www.statisticshell.com>docs>effectsizes.

[5] Wasserstein, R. L. &Lazar, N. A. (2016). The ASA's Statement on P-values: Context, Process, and Purpose. The American Statistician. DOI:10.1080/00031305.2016.1154108.

[6] Nuzzo, R. (2014). Scientific Method: Statistical Errors. *Nature*, 506,150–152. Available at http://www.nature.com/news/scientific-methodstatistical-errors-1.14700. [129].

[7] Siegfried, T. (2014). To Make Science Better, Watch out for Statistical Flaws. *Science News Context Blog*, February 7, 2014. Available at *https://www.sciencenews.org/blog/context/make-science-betterwatch-out-statistical-flaws*. [129].

[8] Cumming, G. (2013). The Problem With *p* Values: HowSignificant are They, Really?. Available at *http://phys.org/wire-news/*EDITORIAL 131.*145707973/the-problem-with-p-values-how-significant-are-they-really.html*. [129].

[9] Gelman, A., and Loken, E. (2014). The Statistical Crisis in Science [online]. *American Scientist*, 102. Available at *http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science*.

[10] Madondo, S. M. (2016). *Essentials of Social Science Research.* Mount Meru University, Tanzania.

[11] Trafimow, D. &Marks, M. (2015). A Critical analysis of the Relevance of Inferential Statistics. *Basic and Applied Social Psycology, 37* (1).

[12] Creswell, J. W., & Clark, V. L. P., (2007). *Designing and Conducting Mixed Methods Research.* University of Nebraska-Lincoln: Sage Publications.