Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 5): pp. 5819-5831
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.656

The Statistical Significance, Practical Significance and Power of the Statistical Test For Master's Theses at Amman Arab University

Mo'en Salamn Alnasraween, Amman Arab University

Abstract- This study aimed at identifying the statistical power tests in educational and psychological research in master's theses at Amman Arab University, as the study included all master's theses that used statistical tests (*P*) and (t) for the period from (2010) to (2019). The number of these studies reached (1670) dissertations, the study sample consisted of (80) dissertations, which contained (725) statistical tests, of which (313) were for (f) test, (113) for (t) test, and (218) for (r), and (100) for multiple analysis of variance and multiple regression analysis.The necessary datawerecollected to calculate the effect size and the statistical power test. The results showed that (27)% of the hypotheses contained a small effect size, and the average effect size for the (t) test reached (0.43) while the (*P*) test reached (0.52). The study showed that about (56.4%) of the hypotheses that were tested had a weak effect, and about (21%) of the hypotheses had a moderate statistical power test and (24.2%) of the hypotheses had a high impact.

Key words: Statistical Methods, Scientific Research, Power Test, Effect Size, Statistical Significance, Practical Significance.

I. INTRODUCTION:

For decades, the fields of education and psychology have had a close relationship with the statistical sciences, as researchers in the field of psychology and education have been keen on verifying the accuracy of the attained results. Theyneed to confirm the validity of the hypotheses that they have constructed. Besides, important developments have led the field of psychometrics and research designs to take advantage of statistics and integrate these various statistical improvements in favour of the development of psychology.

Statistical methods are used to conduct educational research, especially quantitative research, which has become necessary at the level of research projects that are implemented by faculty members or postgraduate students at various stages. Accordingly, any of this research is hardly devoid of using some descriptive or inferential statistical methods. Knowing statistical methods may improve the quality of educational research outputs, increase its efficiency, and increase the ability to conduct it properly. Test with statistical significance is one of the most used methods for evaluating hypotheses and theories. They usually build two types of hypotheses. The first one is known as the null hypothesis and the second is called the alternative hypothesis. They have also adopted several statistical concepts, including statistical significance, It is the symbolized value of the probability between accepting or rejecting the statistical hypotheses of the research of indication that is denoted by (α) and read Alpha. Educators usually use the indication level of either (0.05) or (0.01), and the calculated value of the test is compared with the tabulated one. If the calculated value is greater than the tabular value, the null hypothesis is rejected and the alternative hypothesis is accepted. That is why the concept of scientific research has been linked to the use of statistical methods to accept or reject the hypotheses to generalize the findings of the researcher. Therefore, a look at educational and psychological research is sufficient to reveal the extent of using this contemporary research (Uttley, 2019; Stang & Rothman, 2011).
Some researchers attribute the shortcomings of using statistical methods to the weakness in educational curricula that target the initial and postgraduate stages in educational disciplines. Since this is the case, statistical sciences are important to be given in the curricula of courses and lectures designed for scientific research at the level of master and doctoral programs. In addition, researchers need to learn the method of processing research data using appropriate statistical packages, to use evaluating methods for the student who are studying statistical science courses and related materials as well asmeasuring the evaluation students' learning quality of educational sciences (Ibrahim, 2013, Sansanwal,2015).

II. STATISTICAL METHOD:

The most commonly used method is the statistical one since it helps the researcher to decide in a case of uncertainty, that is, the probability condition, so the researcher formulates the statistical method, the statistical hypotheses according to the research pre-study, and the researcher has two types of possibilities:

The first: the presence of the phenomenon in the sample that does not have an actual presence in the original population, this possibility is symbolized by (α) Alpha.

The second: the absence of the phenomenon in the sample, however, it is present in the original population, and it is symbolized by the symbol Beta, which is atype II error. The Alpha and Beta, express doubts about the attained results, and the Alpha error are related to the so-called "statistical significance", which expresses confidence while Beta error is related to what is called "statistical power test" (Nazik,2015).

Interpretation of statistical significance

Statistical significance simply is the difference between the assumed value of the study population phenomena and its calculated value from the sample data, which is large to the extent that it is difficult to attribute it to the errors of the samples. The researcher usually determines the Alpha value before he engages in collecting his study data. There are several interpretations of statistical significance. Sometimes, it is the probability of obtaining results by chance where the small indications designate that the results are not due to chance, while the large significant value indicates that the results are due to chances. The significance value (1-p) is attributed to the reliability of the result, i.e. obtaining the same result when repeating the experiment, this result may indicate a statistically significant difference between the two compared groups (experimental and control), or between the pre and post measurements. On the other hand, this significant difference does not indicate a real improvement for the experimental group in examining this phenomenon (Uttley, 2019).

Levels of statistical significance:

Statistical significance levels indicate the amount of confidence in the achieved results regardless of the differences in size or the power of the correlations with these results where they are directly affected by the sample size. The effect size (or practical significance) measurement is not affected by the sample size but indicates the amount of the differences or the power of the correlations among the variables, regardless of the amount of confidence. Statistical significance tests, perhaps the most criticized, are strongly dependent on the sample size, as the greater the sample size, the greater the probability of rejecting the null hypothesis. Nevertheless, the effect size is also affected by what is called the size of the study, which includes the number of its members and its groups together. There is a direct linear relationship between the values of the Eta-squared (η2) as a measure of the effect size and the value of the statistic (f) as a measure of statistical significance, so both indicators imply each other. If the null hypothesis is rejected based on the value of calculated (f), it will be rejected as well based on the value of Computed Eta-squared and vice versa (Knottnerus & Tugwell, 2020).

Statistical significance and practical significance

The presence of statistical significance does not necessarily mean the existence of a practical significance. It may be the result of the sample large size or exploiting the statistical indicators to prove the existence of a statistically significant difference between the groups being compared (Das , Mitra & Mandal, 2016)

The main purpose of any research is to detect the real difference between the two groups and to present the estimated difference. For this reason, researchers should previously estimate the sample size, before carrying out the study. Due to the adequate sample size, the random error is less and nothing happens by chance. If the sample is too small it may fall short to answer, the research question, and the results could be of questionable validity. On the other hand, if the sample is too large the outcomes may answer the question but it could be unethical as the resource is intensive. Hence, researchers should have simplicity in the calculation of sample size in their research as a result it can be justified and replicated while treatment (Paterson, Harms, Steel & Credé, 2016; Olivier, May & Bell,2017).

As for the practical significance, it is the numeric value of the differences between the real and unknown population phenomena and the values determined through the null hypothesis. This difference represents one of the most important elements of the "practical significance" indicator or effect size. It is closely related to the power test indicator, where the greater the size of the effect is, the greater the test power.

Through practical significance, the test has meaningful interpretations for the research results; the size of the resulting difference can be applied and interpreted. It shows the amount of variation in the dependent variable that can be explained by the associated independent variable. The statistical significance is not sufficient for decision-making, it is a statistical indicator to determine whether there is a difference between two or more groups or the relationship between two or more variables due to real differences and not by chance. The value of p (*P*-value) helps in making a judgment regarding the statistical significance of the results, but the *P*-value alone is not sufficient in helping the readers to understand the practical significance of the research, because *P*-values are affected by many factors and therefore cannot be used to decide the treatment effect (Amrhein, Greenland& McShane, 2019).

For this reason, it is recommended that the significance test should be followed by calculating the effect size in order to decide the results of practical significance. The statistical significance and the effect size are two sides of the same coin, complementing each other but not replacing each other. Hence, researchers must bear in mind both sides.

Among the positive aspects of resorting to the practical significance test, is that it gives an estimation of the phenomenon's level in the population, through specific values that describe the effect size, whether it is small, medium, or large, where the effect size results can also be used to compare the results of more than one study. The power analysis estimates the sample size needed for a study, there are many measures of practical significance, which are known as measures of effect size. It helps in making decisions about whether the statistically significant difference between programs can be interpreted into a difference that is sufficient to justify adopting one program over the other (Paterson, Harms, Steel & Credé, 2016; Albers & Lakens, 2018)

Statistical Power Test

The statistical power analysis is one of the complementary tests for statistical significance tests, and it estimates the probability of type II error, which is not to reject the null hypothesis when it is false. Any researcher who accepts the null hypothesis without knowing the statistical power test is responsible for the occurrence of a large error rate of type II. Researchers who support mistakenly the null hypothesis, might not notice any statistical significance without considering the statistical power, thus, they do not make the right decision. So the low-power test will not enable the researcher to feel and capture the statistical significance, while the high-power test, even if the size of the differences is small it will be statistically significant. Despite the many criticisms of the statistical significance, researchers still use it inappropriately, misinterpreting and misunderstanding, unconcerned to the large number of controversies that result from the use of this test most of the time (Cohen, 2013; Bakker, Hartgerink, Wicherts & Van 2016).

Criticisms of statistical significance:

First: its sensitivity to sample size:

When the sample size is small, the effects are strong, but they may not be statistically significant, and here the researcher falls into the type II error and when the sample size is large, the effects may be of no value but are statistically significant. Therefore, the statistical significance in the small and large samples may be contradictory, which means that it is not possible to obtain consistent, objective, and meaningful values because they depend only on the size of the sample. Thus, any researcher can reject the null hypothesis by increasing the sample size sufficiently. Many researchers rely on the statistical significance alone without taking into account the severity of the relationship or its power, and (the statistical significance and the severity of the relationship) do not meet, so the values of the correlation coefficients are desirable by squaring it to facilitate their reading (Sabag, 2019).

Significant Level

The statistical significance test is carried out after determining the level of significance (α) where the level of (α) shows the probability that the given result is taken due to sampling errors. This means the probability of committing a type I error, as it is the rejection of the null hypothesis, which is correct. Moreover, the levels of significance of (α) whether (0.05) or (0.01) are traditionally used by the researchers are random imitation. When it was difficult to calculate the exact value of the probability (P) of a statistic test, people instead used the table of statistic test values corresponding to a few, selected, random values of (P) of 0.05, it is 0.01. These values have become known as a threshold limit for the values that define the significance (Suresh, 2012; Lakens et al, 2018).

Alternative solutions

First: Confidence Intervals

Intervalmeans the set of values that fall between two values and what is meant here in the interval that includes the value of the unidentified parameter with a known probability. For example, μ can be estimated with a period accompanied by a known amount of confidence (95% or 99%). Confidence is the amount of the probability that the researcher trusts and is called the confidence factor such as the confidence of 99%. It means that there is a chance of 99 out of 100 that the interval includes the true mean value of the community μ. Because the sample is a small part of the population, it is difficult to determine 100% of the interval validity.Therefore, the confidence interval range of calculation depends on the coefficientof confidence. For example, the 95% confidence coefficient means the expectation that in 95 % of the cases, the study population parameter will be located between the limits of the lower and upper interval, while 5% fall outside it, which is another way to describe the data(Feinstein,1998, Quintana, 2017, Albers &Lakens, 2018).

The confidence interval demonstrates how descriptive measures such as mean, standard deviation, and standard error of the mean can represent the real community. Studies make assumptions about a population by using a sample of the same population. The probability P-value determines whether one group differs from the other, depending on the sample's mean. Nevertheless, it fails to show the extent of this difference, and if another sample is drawn from the same population then a different mean will be obtained. So the confidence interval gives an estimate of a greater range of values that represent the true mean of the population. The statistical significance tests only provide us with information about the explanation or lack of interpretation of the observed differences by chance. Using confidence limits is considered a suitable alternative for statistical significance tests, and the reason is that both methods provide the researcher with the same result (rejecting or not rejecting the null hypothesis). It should be noted that confidence limits help the researcher with additional information to interpret the best results that the statistical significance tests do not provide (Streib & Dehmer, 2019; Das, Mitra, Mandal, 2016).

Third: The Statistical Power Test

The power of the statistical test lies in its ability to reject the hypothesis in question when in reality it is false, and its value (power test) depends directly on the probability of committing type II error as the test power is (1- Beta). Therefore, it is important to remember that the power of the test does not mean absolutely rejecting the hypothesis, but rather the ability to examine the attained data and the efficiency of its determination to reject the hypothesis of the study (Cohen, 2013; Bun, Scheer, Guillo, Tubach, Dechartres, 2019).

Fourth: The Effect Size

The effect size is a term used to describe a group of statistical indicators that measure the size and power of the treatment effect. The effect size is different from the significance tests because its measurements focus on the importance of the practical result. Besides, it allows a comparison between studies through the researchers' ability to judge on the presented results level of the practical significance and the effect size influence. Researchers in educational, social, psychological, and statistics sciences can acknowledge the importance of the scientific results of their research and studies. It is particularly concerned with showing the amount of impact that the independent or dependent variables have on the research based on design. Classifications of the effect size level have been developed when calculated using the (d) index of Cohen's three categories: small, medium, and large. Table (1) shows the values of the effect size; small, medium, and large when using three statistical tests: the F and the t-test (tr henceforth) to test a hypothesis about whether the Pearson coefficient is for the correlation between two variables in the population which is not zero. It should be noted that the lower limits of the effect size levels are equal distances from each other, while the upper limits of these levels are open (Ioannidis, 2019, Albers & Lakens, 2018). Cohen (1988) elucidated the standard judgment of the impact size using Cohen indicator (d) as it appears in the following table.

Table (1) Cohen's classification ofimpact size according to the used statistical test type

Effect Size	Using t-tests	Using f- tests	Using t- test
Small	0.29 – 0.1	0.24 – 0.1	0.49 – 0.2
Medium	0.49 – 0.3	0.39 – 0.25	0.79 – 0.5
Large	and more – 0.5	and more – 0.40	and more 0.8

This study could be seen as a continuation of the efforts that tried to identify indicators related to the effect size, the statistical power test of studies, and research published in periodicals or master and doctoral theses. In fact, it is hoped that it will arouse researchers' interest in issues related to the statistical power test that is used in testing null hypotheses and the effect size that is associated with those hypotheses instead of relying only on statistical significance. Hence, the researcher who follows specific and known statistical procedures does not necessarily mean his decision is well-founded about the null hypothesis although he may know nothing about its validity or incorrectness.

Problem of the Study:

Dealing with the necessary statistical methods to carry out various researches is a very important matter that entails making decisions that may be wrong in many cases due to the misuse of statistical tests. Therefore, the problem needs treatment and diagnosis for appropriate scientific methods. In order to analyze this problem, we should identify its elements; suggest solutions and ways that prevent the researcher from falling into such problems.

The use of postgraduate students to calculate the statistical significance in educational research is not sufficient to demonstrate the correlational relationships or the amount of difference between the arithmetic averages and whether these differences have an impact and value, The biggest criticism for using statistical significance arises from the American Statistical Association’s call for researchers to use this tool properly and immediately to stop using it as the sole decision-making tool, and in particular, this study comes to find out the extent to which researchers use the practical significance as well as the statistical significance for the tests used in master's thesis they prepared to complete the master's degree.

This study sought to answer the following questions:

The first question: What is the percentage of statistical hypotheses in which the statistical significance is associated with a practical significance (small, medium, and large) for a master's thesis at Amman Arab University?

The second question: What is the percentage of using the practical significance concept besides the statistical significance in the used statistical tests in a master's thesis at Amman Arab University?

The third question: How are the statistical power tests distributed according to the different levels of power?

Thesignificance of the study:

This study derives its importance from the value of its topic, especially in light of the currently existing numerous criticism and controversies around it. It is expected for this study to diagnose the causes and results related to research achievement and reduce its economic and social consequences. This study also seeks to enrich Arab libraries in the field of educational research specialized studies for its importance. It alsotackles some of the problems that may arise by presenting the role that can be used to avoid what is possible through proactive work and leadership of managing the future of educational research in such sensitive and priority topics. Furthermore, the expression of researchers’awareness of the futility of using statistical significance, but rather it can be considered a positive signal to find alternative solutions that are more decisive and important.

Research limitations:

Time limits: This study was carried out during the first semester of the academic year (2020-2021).

III. Review of literature

Hamadneh (2015) conducted a study that aimed at knowing the statistical power and the size of the impact on educational and psychological research that was published in the Al-Manara Journal for Research and Studies. The study aimed at assessing the statistical power that was used in testing the null hypothesis and the effect size in educational and psychological research. The study sample consisted of (87) articles and research. The results revealed that (63%) of the specified hypotheses had weak statistical power test, about (11%) medium, and (26%) strong, and that the use of statistical power test is better than relying on only statistical significance. Similarly, Muhammad (2013) sought to find out the effect of the statistical powertest and its relationship to the level of general significance and the effect size. The study sample consisted of (90) research published in the Journal of Research in Education and Psychology for numbers issued from (2010) to (2012) at the Faculty of Education, University of Mina. The results showed that (72%) of the hypotheses' average showed the statistical powerwas (0.22%), and (0.28%) of the hypotheses were accepted. The average power of the test was (0.92%), the effect size percentages with the significance level of the "*t*" test (20%) and analysis of variance (5%). The levels of the effect of the hypotheses indicating the "*t*" test reached (17%) weak, (11%) medium, (72%) large, while the non-indicative hypotheses (95%) were weak: (1.2) medium, (3.6). It was also found that there is a statistically significant direct correlation relationship between the statistical power test and both the effect size and the level of significance.

AbuJarad (2013) performed a study that wanted to find out the powerful effect of the statisticalpower test in published educational research in the Journal of Al-Quds University for Research and Studies. The study included all educational research that used statistical tests (F) (t) in all issues of the Al-Quds Open University Journal for the period between (2002 -2010). The number of these studies reached (74) studies that contained (445) statistical tests, of which (226) the statistical test (t) was used and (219) the statistical test (f) was used. The attained data were collected to calculatethe effect size and the power of the statistical test. The results indicated that about (27%) of the hypotheses contained a small effect size, and the average effect size for (t) test was (0.38), while the average effect size for (f) test was (0.12). In addition, the results indicated that about (71%) of the hypotheses that were selected had the statistical power test was weak, about (6%) of the hypotheses, the statistical powertest was moderate, and about (23%) of the hypotheses were high statistical power test.

Sabah (2019) elucidated in his study the statistical significance in psychological and educational research: methodological issues, the study showed that the statistical significance was adopted 300 years ago and served an important purpose in advancing research in the social sciences.However,many controversies prevailed about the misusing and interpreting of this test. The convincing criticisms of the futility of using the statistical significance test can be found in almost all areasof scientific fields, which need to be taken seriously. The researcher pointed out that after half a century of convincing arguments and calls for the adoption of alternative practices in some disciplines such as psychology and educational sciences, the question remains is why psychological researchers and educationalists do not want to abandon these statistical practices.

Al-Sharifienifain (2017) conducted a study aimed at evaluating the statistical significance, the practical significance, and the statistical power tests for all studies published in the Jordanian Journal of Educational Sciences. The study sample consisted of (1363) statistical tests. The practical significance indicators (effect size) were calculated, and then the statistical power tests were extracted from Cohen's tables. The results of the study indicated that (75.79%) of the hypotheses were associated with a small practical significance, (10.86%) with a medium practical significance, and the findings also showed that (54.86%) of the hypotheses selected were associated with the low or medium power test. The outcomes revealed that there is no statistical significance independence for both the practical significance and the statistical power test.

In the same vein, Jaradat&Jawdeh(2004) performed a study aimed at evaluating the statistical power test, the effect size, and the sample in the studies published in (16) volumes of the Yarmouk Research Journal at the Human and Social Sciences Series. The study sample contained (785) statistical tests of type (P) used in the variance analysis and (f) for the independent samples and (t) for the correlation coefficient, and the value of (d) for Cohen's measure of the effect size calculated, and the statistical power test was extracted from the Cohen's tables. The study concluded that the main factor behind the high percentage of the null hypothesis rejection is the size of the large samples. The effect size is not associated with the null hypotheses that were tested.

A long study has been done by Mahsneh & Shrafeeien (2020) to make a meta-analysis of university theses results which dealt with the effectiveness of the constructivist based approach in Jordan during (2010 - 2017) using dependent variables: (achievement, thinking skills, learning concept) and independent variables (the field subject, the period of applying the tool, achieving the goals). The researchers use the Meta-analysis and coding model for data. The sample consists of (105) theses. The results proved that the effect sizes were not homogeneous. Researchers use a random-effects model in which the value of the homogeneity Test (Q-values) reached (2100.811) in (198) degree of Freedom. The findings also showed the overall average size reached (1.549) in an amount of (201) and (0.063) is the Standard Error. The value was perfect according to Cohen's Classification. This illustrated the effectiveness of using the constructivist model in the educational process. The average of achievement reached (1.654) as (74) each, (1.709) for thinking skills (89) each, learning concept also reaches (1.239) each (65). The results also illustrated that there are statistical differences in values of the study field especially language and the period of application study tool especially the implementation period from (1-4) weeks.

Kp & Srikantiah (2012) examined the impact of the sample size for conducting a successful study with a statistical significance. The size of the sample plays a vital role as too many participants in a study are expensive and reveal the procedures to a large number of subjects. Their study provides some essentials in calculating power and sample size for various practical study designs. Calculations of the sample size could be for one or two-group mean and proportions or rates, correlation studies, and many other types. The use of an adequate sample size together with high-quality data collection will have reliable, valid, and generalizable outcomes. Thus, researchers conduct high-quality research.

Sabri& Gyateng(2015) wanted to help researchers interpret their results by understanding control group analysis and how the statistics and assumptions have been made. They encouraged researchers to ignore the inconclusive results. In fact, they suggested that a bigger sample is needed to have valid results.

Knottnerus & Tugwell, (2018) demonstrated how researchers often make their analytical choices in a context of judgments where their choices are not explicitly stated or even used anymore. Furthermore, making decisions is not an easy task even for highly skilled methodologists. They highlighted the importance of distribution methods and choosing suitable techniques, and statistical tests as well as sample size estimations. Moreover, they explained that traditionally two-sided testing is the preferred default option more than one-sided testing. Eventually, researchers should use the appropriate tool to achieve the required results.

Later on, Knottnerus & Tugwell(2020) explained the importance of statistical significance for concluding the accepting or rejecting research hypotheses. The term ‘statistically significant’ is used in many research papers based on predefined P-value thresholds such as (0.05) Classifying the population samples helps in deciding in yes- or no- statements and is associated with risks of false-negative and false-positive conclusions. The more the researchers put effort to employ trial participants the more they have higher attrition, but they could not display bias in effect estimation. The findings revealed that behavioural intervention estimation could not be generalized for a small sample, thus a study should seek the relation to other population groups, interventions, and settings.

Commentary on previous studies:

It is noted through the review of the previous studies that they differed among themselves in terms of their goals and the samples they dealt withthe terminology of study:

Statistical significance:

The significance or valuelevelrefers to the relationship of the sample with the original population, i.e. the existence of a relationship between the independent and dependent variable, which is a real relationship not due to chance factors. The degree of confidence in this relationship is determined by the level of significance (α) alpha.

Practical significance:

The word significance means that something is important and refers to the power of the relationship or connection between the dependent and the independent variable. It is possible to know the size of the difference or the size of the relationship between two or more variables when using it. It is sometimes called the 'effect size' or the 'power of the effect', and it refers to the scientific and practical importance of the phenomenon in the population under study. Furthermore, it has a number of indicators that are used to challenge it after determining the statistical difference for multiple statistical tests, and these indicators are ETA square, Epsilon square, and Omega square.

Statistical Power Test:

It means the ability to reject the null hypothesis when it is actually false, as the statistical power indicates the probability that the statistical test will produce results of a statistically significant function, and the statistical power test is related totype II error or beta error, and the statistical powertest is calculated in the equation:

power = 1- β: (Streib & Dehmer, 2019)

Method and procedures

Study methodology: The researchers used a meta-analysis method, which is a structured statistical method through which the results of a large number of different quantitative studies are collected, dealing with a set of research hypotheses, and then analyzed intending to obtain a common measure such as the effect size for evaluation and assessment. It is one of the applied aspects of evaluating studies and research (Patty, Qi, Frank, and Ronald, 2010).

Study population: The study population consisted of all master's theses in which the (t) and (P) test were used, and the number of these theses reached (800) during the period (2010-2019).

The study sample:(175) studies were randomly selected from the study population, which contained statistical (t) and (f) tests.

Study procedures:

The master's theses for the period (2010-2019) were reviewed. The necessary data were classified according to the statistical test type used and the level of statistical significance that was adopted, the result of the statistical test and the data for calculating the effect size, and then extracting the statistical power test from Cohen's special tables (Cohen, 1992). Cohen's (d) values were calculated as indicators of the effect size. The statistical test differences were used to examine the null hypothesis, as (d) was considered equal to the value of the difference between the mean of two independent samples divided by the standard deviation as a measure of the effect size in the (t) test of the difference between the means of two independent samples. The ETA value (by dividing the sum of the squared numbers of the deviations between groups by the sum Squared number of total deviations) is a measure of the effect size in the test condition (f).

IV. RESULTS:

The results of the first study question, which states: What is the percentage of statistical hypotheses in which the statistical significance is associated with practical significance (small, medium, and large) for master's thesis at Amman Arab University?

To answer this question, the researchers calculated the frequencies and percentages for each of the categories of practical significance (effect size) and the statistical significance of tests used in hypothesis examination of the selected theses. The results are shown in Table (2) below:

Table (2)
Frequencies and percentages for each of the categories of practical significance (effect size) and statistical significance of tests used in hypothesis examination.

Test	Category of practical significant and Effect Size	Sig				total	
		Significant		Not significant			
		Frequency	percentage	Frequency	Percentage	frequency	Percentage
t-test	small - 0.49) (0	64	%43	49	%66	113	%51
	moderate -0.79) (0.50	25	%17	14	%19	39	%17
	large (0.80- 1)	60	%40	11	%15	71	%32
	Total	149	%67	74	%33	223	%100
F	small - 0.24) (0	127	%52	55	%80	182	%58
	moderate -0.25) (0.39	55	%23	7	%10	62	%20
	large (0.40- 1)	62	%25	7	%10	69	%22
	Total	244	%78	69	%22	313	%100
R	small - 0.29) (0	110	%83	48	%56	158	%72
	moderate -0.30) (0.49	15	%11	4	%5	19	%9
	large (0.50- 1)	8	%6	1	%1	9	%4
	Total	133	%61	85	%39	218	%100
Multiple regression	small - 0.14) (0.00	18	%50	1	%33	19	%49
	moderate -0.15) (0.34	8	%22	2	%67	10	%26
	large (0.35- 1)	10	%28	0	%0	10	%26
	Total	36	%92	3	%8	39	%100
Manova	small - 0.49) (0.00	25	%54	9	%60	34	%56
	moderate -0.50) (0.95	15	%33	4	%27	19	%31
	large (0.96- 1)	10	%22	2	%13	12	%20
	الكلي	46	%75	15	%25	61	%100

Total	479	**%66**	**246**	%34	725	%100

It is noticed from Table (2) results that the percentage of statistically significant hypotheses has reached (66%), while the percentage of statistical hypotheses that are not statistically significant is (34%). Concerning the (t) test, it is observedfrom Table (2) that (0.67%) of the assumptions were statistically significant compared to (33%) that were not statistically significant, and that the highest percentage was for the effect size category (0 - 0.49).

Table (2) also shows that the highest percentage for the (*P*) test was for the category (0 - 0.49), as it reached (52%), as well as for the correlation coefficient test, the percentage was (83%) for the category (0 - 0.29). The multiple regression analysis percentage reached (50%) for the category (0 -0.14), and for the multiple variance analysis, the percentage was (54%) for the effect size category (0 -0.49). It was noted from the results of Table (2) that the statistical significance was associated with a small practical significance. The findings of the answer to this question were in agreement with the AbuJarad study (2013) and the Hammadena study (2015), while it differed with the results of the study of Muhammad (2013).

The second question: What is the percentage of the use of the practical significance concept besides the statistical significance in the statistical tests used in master's theses at Amman Arab University?

To answer this question, the frequencies and percentages of the hypotheses in which the statistical significance was only calculated, for those, which only the practical significance was calculated, and for those which both significances were calculated, and Table (3) shows that.

Table (3)
Frequencies of statistical significance and practical significance of hypothesis tests in Master's theses at Amman Arab University

Statistical Hypothesis		Hypothesis with statistical significant		Hypothesis with practical significant		Hypothesis with statistical and practical significant	
Frequencies	Percentage	Frequencies	percentage	Frequencies	percentage	Frequencies	Percentage
725	%100	691	%95	0	%0	34	%5

Results of Table (3) shows the number of hypotheses for the research statistical significance was calculated only (691) at a percentage of (95%), while the number of hypotheses for which both meanings were calculated reached (34) with a percentage of (5%). This result may be attributed to many faculty members the importance of extracting the practical significance and the statistical significance. Another reason is the lack of awareness of the statistical analysts that students turn to for conducting theirs. It is worth noticing the hypothesis with a practical significant percentage was zero. It may demonstrate the faculty members and graduate students do not depend on using it in their theses.

The outcomes of the answer to this question agreed with the results of Muhammad's study (2013) which showed that (20%) of the tests were associated with a practical significance. It also agreed with the study of Sharifien(2017), which showed that (53.78%) of the studies had no calculation for the practical significance.

The third question:

How are the statistical power tests distributed according to the different levels of power?

To answer this question, the statistical powerwas divided into four categories, the frequencies, percentages of the statistically significant and non-statistically significant tests were found, and Table (4) shows that.

Table (4)

Frequencies and percentages for each of the categories of statistical power test and statistical significance of the tests used in hypothesis examination

Used test	**Categories of test power**	**Statistical Significant**				**Total**	
		Significant		**Not significant**			
		Frequencies	percentage	Frequencies	Percentage	Frequencies	percentage
t-test	- 0)low (0.39	**23**	**%15**	**54**	**%73**	**77**	**%35**
	moderate (0.59 -0.4)	**25**	**%17**	**10**	**%14**	**35**	**%16**
	- 0.6) high (1	**70**	**%47**	**10**	**%14**	**80**	**%36**
	Total	**149**	**%67**	**74**	**%33**	**223**	**%100**
F	low (0.39- 0)	**70**	**%29**	**55**	**%80**	**125**	**%40**
	moderate -0.4) (0.59	**55**	**%23**	**7**	**%10**	**62**	**%20**
	- 0.6) high (1	**62**	**%25**	**7**	**%10**	**69**	**%22**
	total	**244**	**%78**	**69**	**%22**	**313**	**%100**
R	low (0.39- 0)	**65**	**%49**	**73**	**%86**	**138**	**%63**
	moderate -0.4) (0.59	**15**	**%11**	**4**	**%5**	**19**	**%9**
	high (1- 0.6)	**53**	**%40**	**8**	**%9**	**61**	**%28**
	total	**133**	**%61**	**85**	**%39**	**218**	**%100**
Multiple regression analysis	low (0.39- 0)	**7**	**%19**	**0**	**%0**	**7**	**%18**
	moderate -0.4) (0.59	**8**	**%22**	**0**	**%0**	**8**	**%21**
	high (1- 0.6)	**21**	**%58**	**3**	**%100**	**24**	**%62**
	total	**36**	**%92**	**3**	**%8**	**39**	**%100**
Manova	low (0.39- 0)	**7**	**%15**	**10**	**0.67**	**17**	**%28**
	moderate -0.4) (0.59	**17**	**%37**	**2**	**0.13**	**19**	**%31**
	high (1- 0.6)	**22**	**%48**	**3**	**0.20**	**25**	**%41**
	total	**46**	**%75**	**15**	**%25**	**61**	**%100**
Total		**479**	**%66**	**246**	**%34**	**725**	**%100**

It is noticed from Table (4) that the statistically significant hypotheses that were associated with a low power level of the (t) test were (15%) and those that were associated with a high power level were

(47%). As for the (f) test, the related hypotheses were linked statistically with the low power level of (29%), followed by those associated with a high level of power at a percentage of (25%). Concerning testing the correlation coefficient, the hypotheses that were associated with a low power level came at the highest percentage of (49%), followed by those associated with a high level which reached (40%).

As for the regression analysis, the hypotheses that were associated with a high power level (0.6-1) were (58%). Regarding the multiple variance analysis, the percentage of hypotheses that were associated with a high power level reached (48%), followed by those that were associated with an average power level of (37%). It is noted from the results of table (4) that the multiple regression analysis was associated with a high power level of (58%). This result may be attributed to the reliability of the design used, and the results were more accurate due to suitable decisions taken.

The provided interpretations were more persuasive in rejecting or accepting the Null hypothesis. This result may be attributed to the lack of the statistical hypothesis that used this test. Actually, some researchers neglect or pass over the statistical interpretation as something secondary in conducting a study. For this reason, numeral studies have a week impact or are not analyzed well. Sometimes the findings are important but due to bad interpretation, the research fails to achieve its aim. The outcomes of this study were compatible with the results of the AbuJarad study (2013), and the study of Sabah (2019) and Knottnerus &Tugwell(2020).

V. Conclusion:

In order to generalize the results of studies and become reliable, researchers must pay attention to the practical significance and statistical power tests used in addition to the statistical significance. It is worth noticing that studies with high statistical power testsand practical significance can contribute to the interpretation of the study results. In return, this may constitute a strong justification for generalizing results of the research, providing information on the psychometric properties of the used measures and the used analyses quality.

VI. Recommendations

In light of the results of the study, researchers recommend the following:

- Improving the content of undergraduate courses related to statistics for postgraduate students in order to prepare students for postgraduate studies to understand statistical skills, statistical methods, associated software, and enable them to conduct research properly.

- Teaching modern educational research methods courses and scientific research mechanisms for faculty members in universities as well as enhancing and distributing statistical knowledge among them.

- Faculties should employ a specialist faculty member in statistics or measurement and evaluation in all university dissertation discussions.

References:

1. AbuJarad, H.(2013). The power of the statistical tests and the size of effect of the educational researches. *Al Quds Open University Journal for research and studies*,14(2), 351-368.
2. Albers, C., Lakens, D. (2018). When power analyses based on pilot data are biased: inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*. 74,187–95.
3. Amrhein, V. , Greenland, S. , McShane, B.(2019). *Scientists rise up against statistical significance*. Nature, 567:305e7.
4. Bakker, M, Hartgerink CH., Wicherts, JM., Van der Maas, HL. (2016). Researchers' Intuitions about Power In Psychological Research. *Psychological Science*. 27(8):1069–77.
5. Bun, R., Scheer, J. , Guillo, S., Tubach, F. &Dechartres, A., (2019). Meta-analyses frequently pooled different study types together: a meta-epidemiological study,*Journal Of Clinical Epidemiology*, 118, P18-28, https://doi.org/10.1016/j.jclinepi..10.013.

6. Cohen, J. (1988). *StatisticalPower Analysis for the Beahavioral Sciences*. New York (NY): Lawrence Erlbaum Associates.
7. Cohen, J. (1992). A power primer. *Psychological Bulletin*. 112 ,(1):155–59.
8. Cohen, BH. (2013). Explaining Psychological Statistics. 4th Ed. New York (NY): Wiley.
9. Das, S., Mitra, K., Mandal, M.(2016). Sample size calculation: Basic principles. *Indian J Anaesth*, 60:652-6., DOI: 10.4103/0019-5049.190621
10. Feinstein, AR.(1998). P-values and confidence intervals: two sides of the same unsatisfactory coin. *Journal of clinical Epidemiogyl*, 51:355-60.
11. Hamadneh, E. (2015) .statistical power and the size effect in educational and psychological research published in Al-Manara Journal for Research and Studies, *Al-Manara Journal for Research and Studies*, 21(2), 14-39.
12. Ibrahim, A, (2013).Using Statistical Methods to Measure the Quality of the Assessment Process of Science Students' Learning of Academic Courses in the College of Education - Aden - for The Academic Year (2009-2010), *the Arab Journal for Quality Assurance of University Education,* No. 11, Republic of Yemen.
13. Ioannidis, J. (2019).Retiring significance: A free pass to bias. Nature, 567, 461–461.*Journal of Human Reproductive Sciences*, 5(1),7-13.
14. Knottnerus, J., Tugwell, P.(2020).Thresholds and innovation: discussion on statistical significance, *Journal Of Clinical Epidemiology*, 118, 5-7.
15. Kp, S.& Srikantiah, C. (2012). Sample Size estimation and Power Analysis for Clinical research studies. *Journal of human reproductive sciences*. 5. 7-13. 10.4103/0974-1208.97779. DOI: 10.4103/0974-1208.97779
16. Lakens D., Adolfi, FG., Albers, CJ., Anvari, F., Apps, MA., Argamon, SE.,Buchanan,EM. (2018). Justify your alpha. *Nature Human Behaviour*. 2(3):168–71.
17. Mahsneh, N. and Shrifien, N. (2020).A Meta-analysis of Results of University Theses Which Dealt With Effectiveness of the Constructivist Based Approach in Jordan During (2010-2017).*Journal of Educational and Psychology Sciences (Islamic University of Gaza)*. 28(5), 588 -609.
18. Muhammad, I. (2013). Anylzing test power and its relationship with in significant level and size effect in educational research,*Journal of Educational Arab Union*, 12(5), 33-49.
19. Nazik, G. (2015).Methodological Issues in Educational Research, *University News*, 53(17), 20-23.
20. Olivier J, May WL, Bell ML. (2017). Relative effect sizes for measures of risk. and *Communications in Statistics-Theory Methods*. 46(14):6774–81.
21. Paterson, TA., Harms, PD., Steel, P.& Credé, M. (2016). An assessment of the magnitude of effect sizes: evidence from 30 years of meta-analysis in management.*Journal of Leadership & Organizational Studies*. 23(1):66–81.
22. Patty, W., Qi, S., Frank, b. & Ronald, M.(2010). Meta-analysis of prospective cohort studies evaluating the association of saturated fat with cardiovascular disease. *American Journal of Clinical Nutrition*, 91(3), 535-546.
23. Quintana, DS. (2017). Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology*. 54(3):344–49.
24. Sabag, A.(2019). Statistical Significance in psychological and educational research: methodological issues, *International for Research in Education*. 43(2), 86-101.
25. Sabri, F. & Gyateng, T. (2015).*Understanding statistical significance: A short*
26. *guide*. (NPC) New Philanthropy Capital.
27. Sansanwal, D. (2015).Methodological Issues in Experimental Research, *University News*, 53(10), 12-16.
28. Sharifien, N. (2017). Beyond the analysis of research published in the Jordanian Journal of Educational Sciences: practical significance and test force.*University Union Journal*.15 (3). p 130-160
29. Shehata, H., (2001).*Scientific and Educational Research Between Theory and Practice*, Arab Library for Book, (First Edition), Cairo, Egypt.
30. Stang A., Rothman, KJ.(2011) *That confounded P-value revisited*. Journal of Clinical Epidemiology, 2011;64:1047e8.
31. Streib, f., & Dehmer, M.(2019).Understanding Statistical Hypothesis Testing: The Logic of Statistical Inference Machine learning knowledge extraction , 1, 945–961; doi:10.3390/make1030054.
32. Uttley, J.(2019). Power Analysis, Sample Size, and Assessment of Statistical Assumptions—Improving the Evidential Value of Lighting Research, LEUKOS 15(2–3),143–162, https://doi.org/10.1080/15502724.2018.1533851.