American Journal of Mathematics and Statistics 2018, 8(6): 171-178
DOI: 10.5923/j.ajms.20180806.02

Factors Determining the Power of a Statistical Test for the Difference between Means and Proportions

Habib Ahmed Elsayir

Dept. of Mathematics, Al Qunfudha University College, Umm Al-Qura University, Al-Qunfudha, Saudi Arabia

Abstract One of the most essential issues in research problems design is statistical power of a test is. The problem motivating this topic is to identify the factors and relationships among the components of power analysis for a study. In this paper, we presented testing procedures of hypothesis for means and proportions in different sample situations. Hypothesis testing requires several stages, including specifying the null and alternative or research hypothesis, selecting and computing an appropriate test statistic, setting up a decision rule to reach a conclusion. Some related concepts such as sample size and confidence intervals were demonstrated, and illustrations on theoretical data would be carried. Results and conclusions on the basis of the discussions reflected the relationship among power analysis components and factors that influence the statistical power of a test would be shown.

Keywords Effect size, Power of a test, Sample size, Significance test

1. Introduction

Power is defined as "the probability that a statistical significance test will reject the null hypothesis for a specified value of an alternative hypothesis" (Robin High, 2000). It is the ability of a test to detect an effect, given that the effect actually exists". Type II error is the compliment of power.

Some questions can arise when conducting power analysis, These such as: What is needed to take into account when considering statistical power analysis? How powerful is underlined study test? What sample size needed to carry out the study? When giving the answer to these questions, we must take into account a number of factors, including the study objective, target population size, the risk of selecting a sample, and the tolerance of sampling error. Adequate power can't be achieved to detect the effect you're looking for without a sufficient number. Choosing too many observations, may be using valuable resources inefficiently. A study with too little or too much power does not spend time and resources economically; and may be seen as unfavorable scientific behavior. The power analysis and sample size determination objectives is to provide researcher with the statistical methods to respond to these issues.

Depends on statistical inference, the subject of power of a test, has been covered in many statistical studies. Earlier, Alan G. Sawyer (1982) has reviewed the factors that determine statistical power and illustrated how the use of the appropriate statistical test can improve power and even sometimes change the statistical conclusion. Most of studies has covered the issue from different points of view. However most of these studies has focused in sample size which has a relation with effect size, which mostly called "d", (Muller & Lavange, 1992). Power analysis for the behavioral sciences was introduce by Cohen (1988). Zodpey (2004) on an online article has described the common used terms in sample size estimation and power analysis. Models and tests concerning power could be found in Murphy (1998), Rudolf (1998), and Jeeshem and Kucc (2004). Plotting and the shape of the test statistic distributions for sample size and statistical power were demonstrated by Blake (2001). Some calculations of power and sample size for some distributions is partly introduced by Bret Hanlon and Bret Larget (2011). A set of software programs in power analysis were found to enable evaluation of the factors affecting power and sample size(see for instance power analysis, Electronic Text Book, Statsoft, 1984-2003).

* Corresponding author:
Habibsayiroi@Yahoo.com (Habib Ahmed Elsayir)
Published online at http://journal.sapub.org/ajms

2. Components of Statistical Power of a Test

The components which are to be considered when conducting a statistical power analysis beside the model (the test) include: standardized effect size, Sample size (n) (significance level α) and the Power of the test (1–ß).

Let us consider a normal distribution with unknown parameter μ but known variance σ^2 from which is a sample

of size n is drawn. The sample size n is given by the formula:

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{d^2}, \text{ and}$$

$$n = \frac{z^2_{1-\alpha/2}P(1-P)}{d^2}$$

Where:

P: is the estimated proportion of the population, d is absolute precision and $1-\alpha/2$ is the desired level of significance.

Then we put H_0 and H_1 as below:

$$H_0: \mu = \mu_0 \; against \; H_A: \mu > \mu_0$$

Let $\mu_1 > \mu_0$.

The decision possibilities on the test of hypothesis is as in the following table:

	H_0 Acceptance	**H_0** Rejection
H_0 is True	Correct decision	Type I Error
H_0 is false	Type II Error	Correct decision

Power= 1-β =P (rejecting H_0 when the means are equal, that is $\mu = \mu_1$)

$$= P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq Z_a \; when \; \mu = \mu_1\right)$$

$$= P\left(\bar{x} \geq \mu_0 + Z_a\left(\sigma/\sqrt{n}\right) when \; \mu = \mu_1\right)$$

$$= P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq \left(\frac{(\mu_0 - \mu_1) + Z_a(\sigma/\sqrt{n})}{\sigma/\sqrt{n}}\right)\right)$$

$$= P\left(Z \geq Z_a - \frac{(\mu_1 - \mu_0)}{\sigma/\sqrt{n}}\right)$$

So, the power of a test can be increased by increasing the value of $\mu_1 - \mu_0$, or by increasing the sample size n and it can also be increased by decreasing the σ. The power of a test can also be increased by decreasing Z_a, that is increasing a.

The power explanation has relation with hypothesis testing process. For instance, if a researcher is conducting a testing of hypothesis of significance of a difference of proportions, then the null hypothesis may look as follows:

$$H_0: P_1 - P_2 = 0 \; against \; H_A: P_1 - P_2 = D > 0$$

However, in the complex sample set-up, we can define the variance of $\widehat{P_1} - \widehat{P_2}$ as

$$var\left(\widehat{P_1} - \widehat{P_2}\right) = \frac{DEFF}{n}\left(1 - \frac{n}{N}\right)(P_1Q_1 + P_2Q_2)$$

Where DEFF is the design effect that collects the inflation of variance due to complex sampling design. If $P_1 - P_2 = 0.5$ (larger value), then:

$$var\left(\widehat{P_1} - \widehat{P_2}\right) = \frac{DEFF}{n}\left(1 - \frac{n}{N}\right)$$

The formula of the effect size is:

$$\text{ES} = \frac{\text{difference between the means of two comparable groups}}{\text{population standard deviation}}$$

Following Lisa Sullivan (2017), we have:

In continuous variable observation, one Sample: H_0: $\mu = \mu_0$

$$z = \frac{\bar{X} = \mu_0}{s/\sqrt{n}}$$

- Continuous variable observation, two independent samples: H_0: $\mu_1 = \mu_2$

$$z = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

and $S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

- Continuous variable observation, two matched samples: H_0: $\mu_d = 0$

$$z = \frac{\bar{X}_d - \mu_d}{s_d/\sqrt{n}}$$

and

$$s_d = \sqrt{\frac{\Sigma\text{Differences}^2 - (\Sigma\text{Differences})^2/\text{n}}{n-1}}$$

- For dichotomous variable, one sample: H_0: $P = P_0$

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- For dichotomous variable, two independent samples: H_0: $P_1 = P_2$, RD=0, RR=1, OR=1

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Any research design may consist of certain or specific models (tests) with different forms for computation of test statistics on which their statistical powers are established. Formulas vary according to the type of models (tests) to compute test statistics. For instance, the T-test depends on the T distribution to determine its statistical power, while ANOVA depends on the F distribution. A standardized effect size, a test statistic (e.g., T and F scores) is computed by combining the effect size and variation. An effect size in actual units of the response is the "degree to which the phenomenon exists" (Cohen 1988). Alternatively, an effect size is the deviation of hypothesized value in the alternative hypothesis from the baseline in the null hypothesis.

The test size (α), or the significance level is the probability of rejecting the true H_0 hypothesis. The power of the test (1–ß) is the probability of correctly rejecting a false null hypothesis, where ß is the type II error. It measures the test's ability to reject the null hypothesis when it is actually false - that is, to make a correct decision. The maximum power a test can have is 1, the minimum is 0. Ideally we want a test to have high power, close to 1 (Valerie & John (1997).

The Sample size and variance are the main factors that affect power, since power depends on sample size. Other things being equal, larger sample size yields higher power and power also depends on variance since smaller variance yields higher power.

3. Power and Experimental Design

Power may sometimes be increased by conducting a different design of experiment that has lower error variance. For example, clustering, stratification approach sampling can usually tends to reduce error variance and hence can increase power level. However, the power computation will depend on the type of the adopted experimental design. For more on designs that may increase power, see: Lipsey, MW (1990). *and* McClelland, Gary H. (2000).

The statistical power of a contemplated research design should be determined before the beginning of data collection. If a design has unacceptably low power to detect the effect of interest, the design ought to be changed to improve the power. If limited resources preclude a satisfactory level of power and if statistical significance at a low Type I error rate is desired, the research is probably not worth the time, cost, and effort and should be abandoned. A researcher who decides to conduct a study with low statistical power should be aware of the result of the rejection of the null hypothesis.

Cohen (1988) and his colleagues research studies illustrate how the use of the appropriate statistical test can improve power and sometimes even alter the statistical conclusion.

After becoming aware of the magnitude of effect size, (often very low), a researcher can often increase power by "developing insights which lead to research procedures and tools which make effects measurably large enough to be detected by experiments of reasonable size. Stronger and better controlled techniques of explanatory variables also can improve power via increased effect size.

Another statistical procedure to increase power is to combine several studies in a type of "meta-analysis", see Elsayir (2015). It could be shown that the analysis of several replications results in a rejection of the null hypothesis when the replications are combined into a replication effects.

4. Statistical Approach

There are several approaches for performing power analysis which depend primarily on the study design and the main outcome measure of the study (Zodpey, 2004). For example, one can specify the desired width of confidence interval and determine the sample size that achieves that goal, or a Bayesian approach can be used where we optimize some utility function. One of the most popular approaches for studying the power of a test of hypothesis involves specifying (Russel, 2001):

1. A parameter for hypothesis test.
2. Significance level of the test.
3. Effect size that reflects an alternative of scientific interest.
4. The values or estimates of other parameters needed to compute the power function of the test.
5. A target value of the power of the test.

5. Performing Power

It should be said that the power of any statistical test depends mainly on: the actual population mean μ, the sample size n, the significance level (α) and the population variance σ^2. Using G. Power software, values are set for the sample size, the population standard deviation and the significance level. when performing a statistical power analysis, the following important components should be considered: Significance level (α) or the probability of a type I error, Power to detect an effect, which is expressed as (1-β), where β is the type II error, and effect size the researcher wants to detect, variation and the sample size. These components of power analysis are not independent. Hence, any four of them automatically determines the fifth one.

We used t test for difference between dependent means (matched pairs) to compute required sample size (n=54), given α=0.05, power =0.95, and effect size (ES) =0.05. The input and output parameters are as summarized in table 1 (model 1). The plot of values is in figure (1), while the x-y plot for a range of values is presented in figure (2). The figure demonstrates that a larger sample size yields higher power. Similar input parameters has been made for greater sample size procedure (model 3) in the above mentioned table, but for independent means (two groups). The one tailed t test for independent means (two groups) is presented in figure 4. The same procedure has been conducted as seen in (figure (5)) and figure (6) for two tailed test. In figure (7), power and total sample size has been plotted for unequal proportions (two independent groups using two tailed Fisher's exact test. Figure (8) shows plot (on y axis) α err prob. as a function of power (1–ß err prob.) and total sample size at 2000 and proportional p1 at 0.5. Figure (8) illustrates that a larger (α) level gives higher power.

Hence, when the α level, the effect size or the sample size increases, the power level increases. generally, the larger the sample size n, the smaller the sampling error and higher power. If we are to make accurate decision about a parameter, we need to have an n large enough so that error will tend to be" reasonably small ". If n is too small, there is not much point in gathering the data, because the results will tend to be imprecise to be of much use. Once n is large enough to produce a reasonable level of accuracy, making it larger simply wastes time and money.

Table (1). Power Input and Output Parameters Summary for Means

#	Model(statistic Test)	Input Parameters	Output Parameters	Note
1	T test for difference between two dependent means (matched pairs)	ES d=0.05, α err prob =0.05, power(1–ß)=0.95	Critical=2.0057460 Df=53 Total sample size 54 Actual power=0.9502120	Two tails (figure 1)
2	T test for difference between two dependent means (matched pairs)	ES d=0.05, A err prob. =0.05, power(1–ß)=0.95	Critical=2.0057460 Df=53 Total sample size 54 Actual power=0.9502120	Two tails (figure 2)
3	T test for difference between two independent means (two groups)	ES d =0.05, α err prob. =0.05, power(1–ß)=0.95	Critical t=1.6536580, Df=174, Sample size group 1=88, Sample size. Group 2=88. Total sample size 176, Actual power=0.9514254	Two tails (figure 3)
4	T test for difference between two independent means (two groups)	ES d=0.5, α err prob. =0.05, power(1–ß)=0.95	-	One tail (figure 4) Allocation ratio =$N_2/N_1 = 1$
5	T test for difference between two independent means (two groups)	ES d=0. 5, α err prob. =0.05, power(1–ß)=0.95	-	One tail, (figure 5), Allocation ratio $N_2/N_1 = 1$
6	T test for difference between two independent means (two groups)	ES d=0.5, α err prob. =0.05, power(1–ß)=0.95	sample size =210	Two tails, (figure 6), Allocation ratio =$N_2/N_1 = 1$

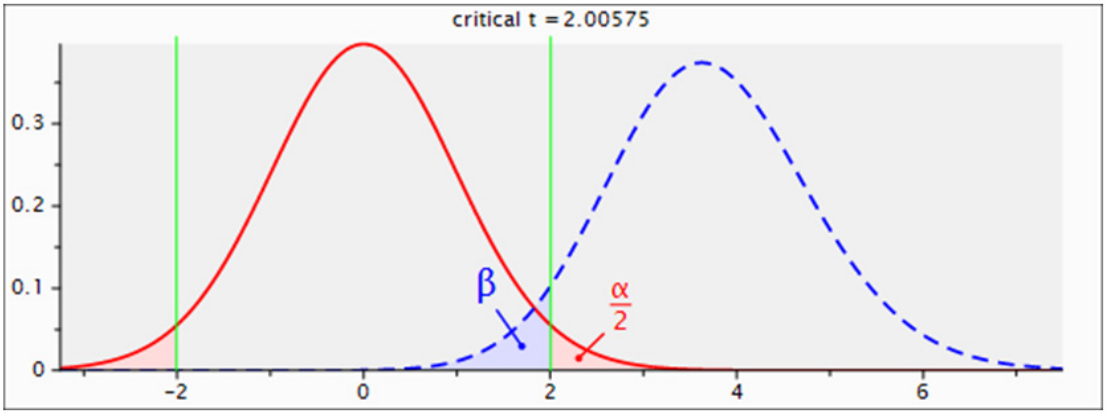

Figure (1). Power Graph for Test of Difference between means for Matched Pairs (Two Tailed Test)

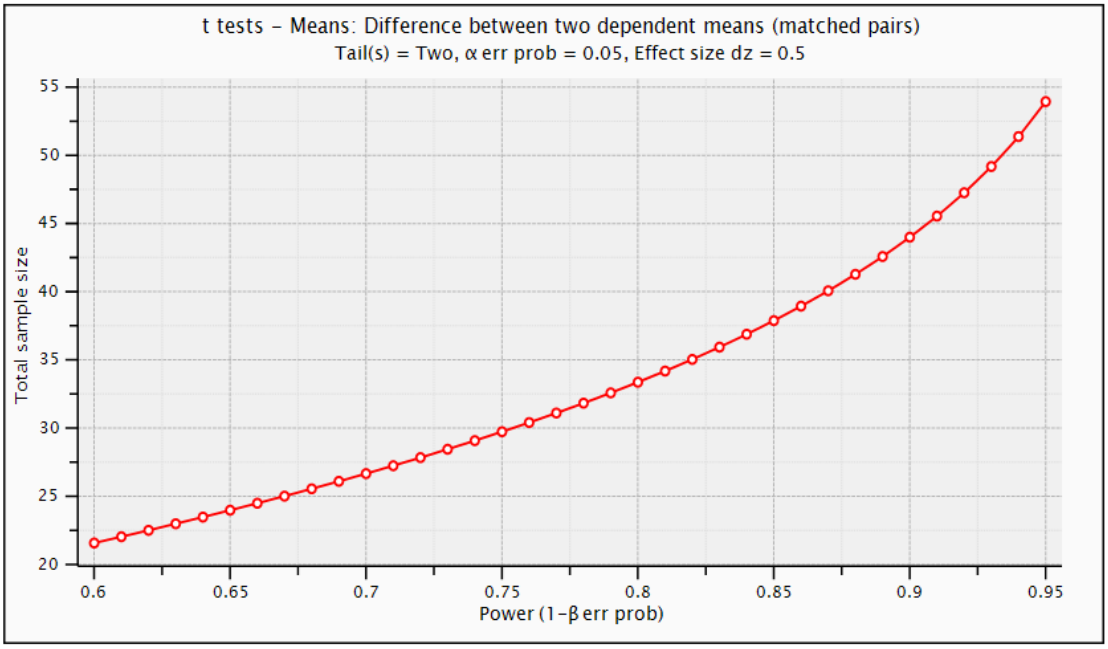

Figure (2). T tests-Means: Difference between Two Dependent Means (Matched Pairs) (Two Tailed Test)

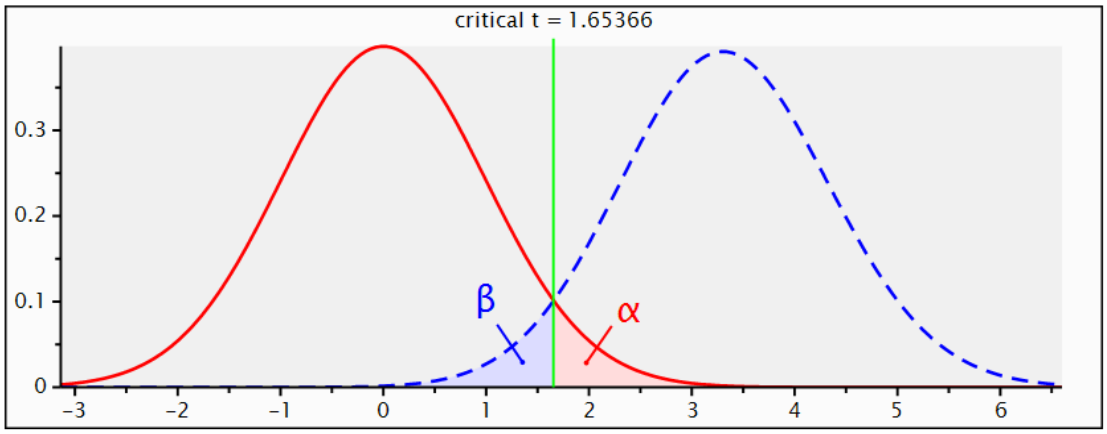

Figure (3). Power Graph for Test of Difference between means for Matched Pairs (One Tailed Test)

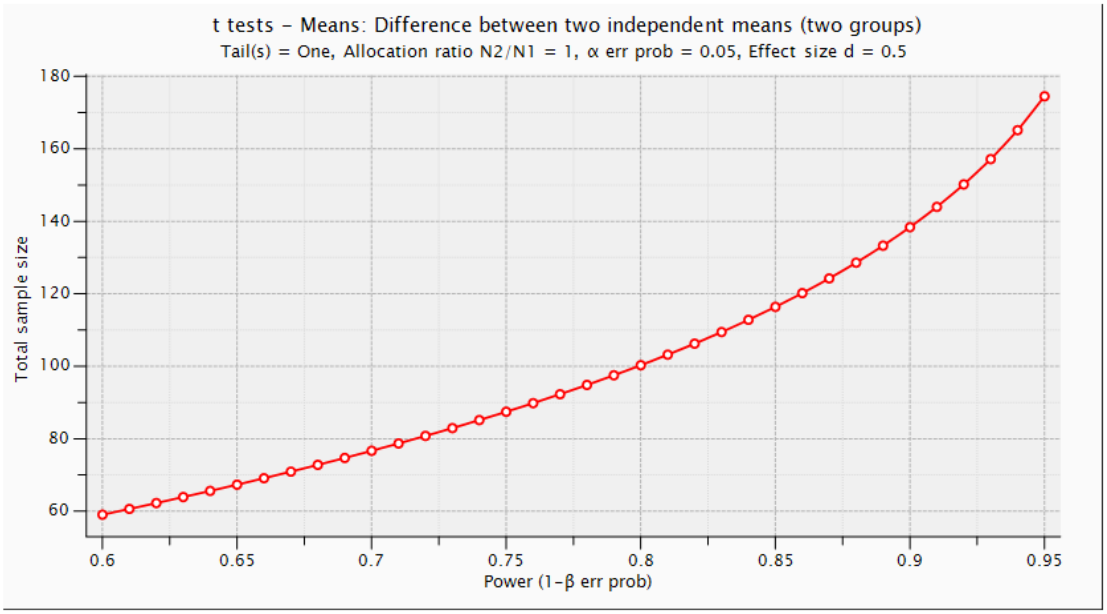

Figure (4). T tests-Means: Difference between Two Independent Means (Two Groups) (One Tailed Test)

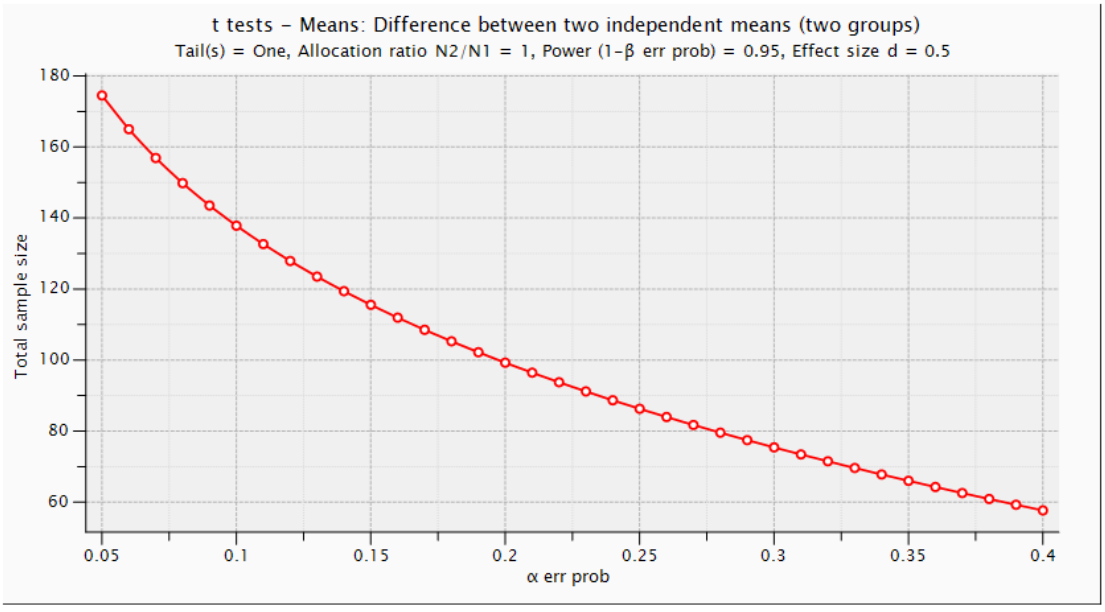

Figure (5). T tests-Means: Difference between Two Independent Means (Two Groups) (One Tailed Test)

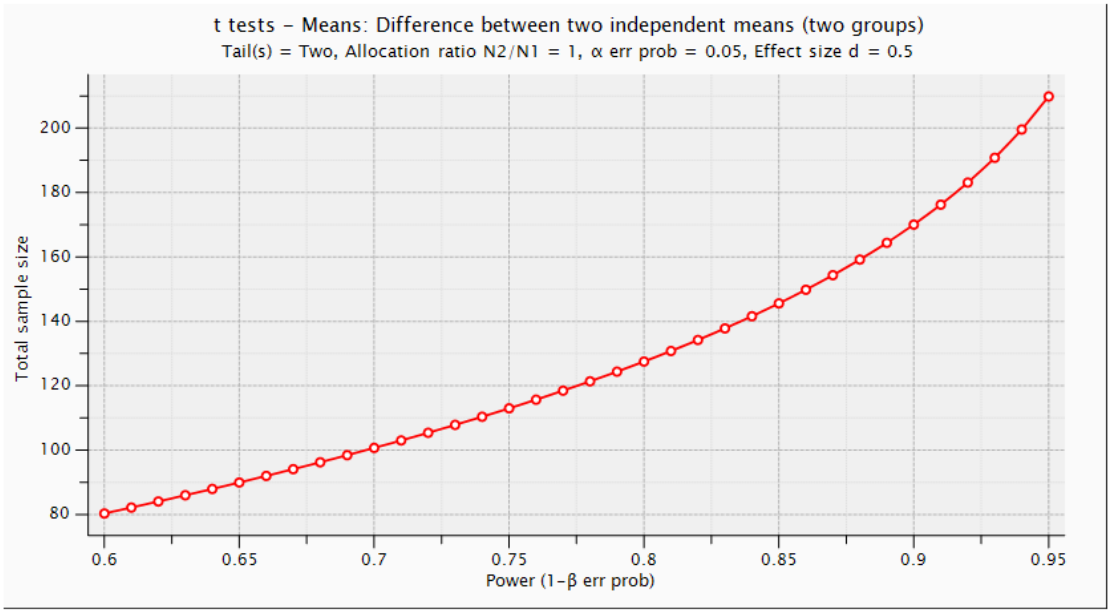

Figure (6). Power and Total Sample Size: Difference between Two Independent Means (Two Groups) (Two Tailed T Test)

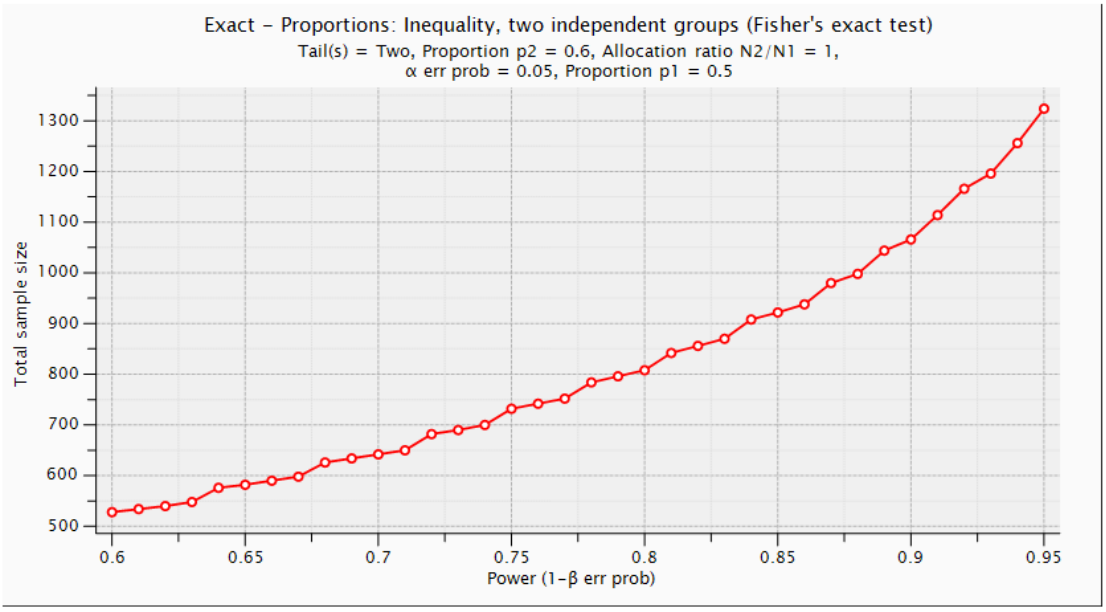

Figure (7). Power and Total Sample Size: Unequal Proportions, Two Independent Groups (Two Tailed Fisher's Exact Test)

Table (2). Power Input and Output Parameters Summary for proportions

#	Model(statistic Test)	Input Parameters	Output Parameters	Note
1	Proportions: Inequality, two independent groups (Fisher's Exact Test)	Proportion p1=0.5 Proportion p2=0.6 α err prob =0.05 power(1–ß)=0.95	Sample size group 1=662 Sample size Group 2=662 Total sample size 1324 Actual power=0.9502923 Actual α =0.0443789	Two tails (figure 7) Allocation ratio $=N_2/N_1$
2	Proportions: Inequality, two independent groups (Fisher's Exact Test)	Proportion p1=0.5 Proportion p2=0.6 α err prob =0.05 power(1–ß)=0.95	Total sample size 2000	Two tails (figure 8) Allocation ratio $=N_2/N_1$

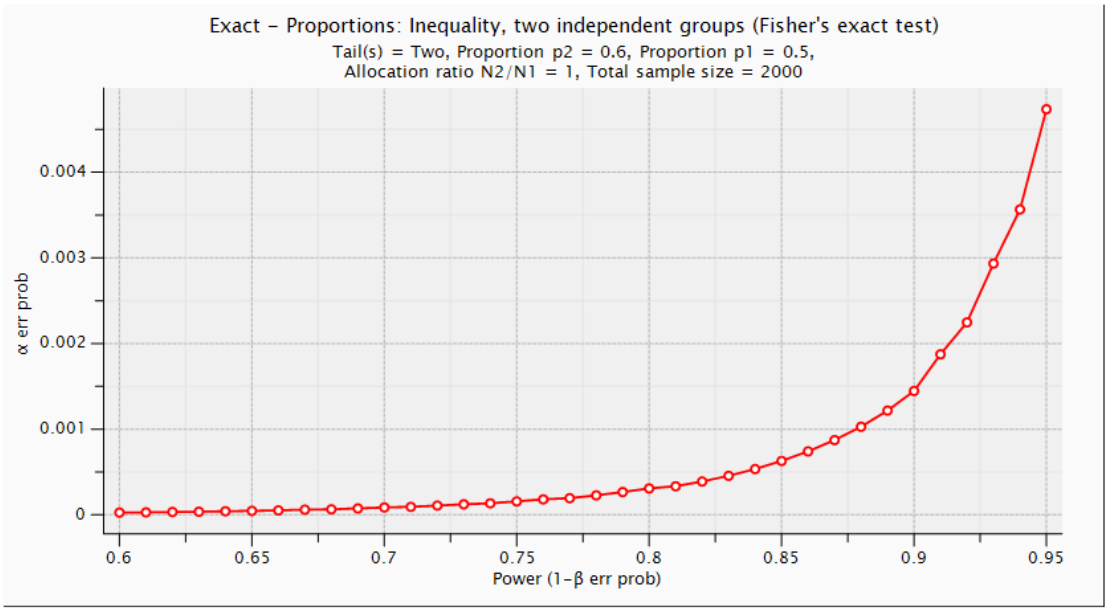

Figure (8). Power and α Prob: Unequal Two Independent Groups (Two Tailed Fisher's Exact Test)

6. Discussion

This article reviews the analyses of factors that affect power and examines the sensitivity of power and sample size to other components, enabling researchers to efficiently use the research resources. Power analysis is a procedure to balance between Type I and Type II error. It is the probability of detecting a true difference. If Without an adequate power, a significant result might not be reached. In addition, if too many observations are used (or if a test is too powerful with a large sample size), even a very small effect will be mistakenly detected as a significant one. However, if too few subjects are used, the hypothesis test will result in low statistical power and, thus there is little chance to detect a significant effect. A study with low power will have inconclusive results, even if the investigated phenomenon is real. Stated differently, the effect may well be there, but without adequate power, you won't find it. It can be stated that the most important component affecting the statistical power is the sample size. In fact, there is a little space to change a test size (significance level). It is also difficult to control effect sizes in many cases. It is costly and time-consuming to get more observations, of course. But the frequently asked question in practice is how many observations need to be collected.

It was suggested to follow an informal rule that alpha is set to 0.05 and beta to 0.2. In other words, power is expected to be 0.8. This rule implies that a Type I error is four times as costly as a Type II error. There are challengers to this "0.05 and 0.2 rule." For example, for a simple study a Type I error rate of 0.05 is acceptable. However, pushing alpha to a more conservative level should be considered when many variables are included. One can argue that for a new experiment a 0.05 level of alpha is acceptable. But to replicate a study, the alpha should be as low as 0.01 However, low power does not necessarily make a study a poor one if you found a significant difference. Even if the null is rejected, the power may still be low. But this can be interpreted as a strength rather than as a weakness.

7. Conclusions

We have presented the definitions of power concept and significance level. We also have explained how to determine sample sizes for desired sizes of parameters for both means and proportions, as well as examining and interpreting the power curve and its changes as n changes. The issue of factors affecting power and the related issues is discussed here. Sample size justification is intimately tied with power analysis. Therefore, to understand sample size justification, understanding of power analysis is needed. These techniques are related to confidence interval estimation which is useful in implementing the above objectives and in evaluating the size of experimental effects in practice. There is some benefits of sample size increase which include a greater likelihood of correctly rejecting a false null hypothesis and more accurate estimation of effect size. Several factors affect the power of a statistical test. Some of the factors are under the control of the experimenter, whereas others are not.

REFERENCES

[1] Alan G. Sawyer (1982), "Statistical Power and Effect Size in Consumer Research", in NA - Advances in Consumer Research Volume 09, eds. Andrew Mitchell, Ann Abor, MI: Association for Consumer Research, Pages: 1-7.

[2] Bret Hanlon and Bret Larget (2011). Power and Sample Size Determination. Department of Statistics. University of Wisconsin Madison November 3[8, 2011].

[3] Cohen, Jacob. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nded. Hillsdale, NJ: L. Erlbaum Associates.

[4] G*Power-2000 (http://www.psychologie.uni-trier.de.8000/ projects/gpower.html).

[5] Elsayir, H.A. (2015). Significance Test in Meta-analysis Approach. A Theoretical Review. American Journal of Theoretical and Applied Statistics. Vol. 4 no. 6, pp. 630-639.doi:10.11648/j.ajtas.20150406.35.

[6] Jeeshim and Kucc (2004). Understanding the Statistical Power of a Test: 1UITS Center for Statistical and Mathematical Computing http://mypage.iu.edu/kuucc625.

[7] John Blake (2001). Sample Size and Statistical Power "A Lecture of Dr Eric Rexstad, Dept of Biology and Wildlife.

[8] Lisa Sullivan (2017). Biostatistics. Boston University School of Public Health. Hypothesis Testing for Means & Proportions in http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTest-Means-Proportions/BS704_HypothesisTest-Means-Proportions_print.html. Date last modified: November 6, 2017.

[9] Lipsey, MW (1990). Design sensitivity: Statistical power for experimental research. Newbury Park, CA: Sage.

[10] McClelland, Gary H. (2000). Increasing statistical power without increasing sample size, American Psychologist 55(8), 963 – 964.

[11] Muller, K.E. & Lavange, L.M. (1992). Power calculations for general linear multivariate models including repeated measures applications. Journal of the American Statistical Association, 87, 1209-1216.

[12] Murphy, Kevin R. (1998). Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests.

[13] Power Analysis, Electronic Text Book, Stat soft, Inc 1984-2003(online).

[14] Robin High (2000). "Important Factors in Designing Statistical Power Analysis". Computing Centre Home Page (Available in robin@darkwing.uoregon.edu).

[15] Rudolf K Bock.. (7 April 1998). Hypothesis Testing in htt:/rkb.home.cern.ch/rkb/AN16pp/html.

[16] Russel V. Lenth (2001). "Some Practical Guidelines for Effective Sample Size Determination". Dept of Statistics, University of Iowa.

[17] Valerie J. Easton & John H. McColl STEPS Glossary Web version revised and updated Sep 97 by Stuart G. Young., STEPS, Statistics Glossary, vol 1, httpllwww.stats.gla.ac.uk/steps glossary/index.html.

[18] Zodpey (2004). Sample Size and Power Analysis in Medical Research. IJDVL Indian Journal of Dermatology, Venerology and Leprology Vol 70 Issue 2 P 123-128.