Black Sea Journal of Health Science

doi: 10.19127/bshealthscience.866556

Open Access Journal
e-ISSN: 2619 – 9041

Research Article
Volume 4 - Issue 3: 221-227 / September 2020

SAMPLE SIZE IN CLINICAL RESEARCHES: POWER OF THE TEST AND EFFECT SIZE

Adnan ÜNALAN[1]*

[1]Niğde Ömer Halisdemir University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, 51240, Niğde, Turkey

Abstract: The approval of local ethics committees is required for clinical researches. In order to obtain approval, how the sample size is determined, whether power analysis is done or not and under what assumptions these analyses are made, are important questions/problems. In hypothesis tests, it is possible two types of errors (type 1 error denoted by α and type 2 error denoted by β), of which α is the probability of rejecting the null hypothesis that is actually true and is the probability of accepting the actually false null hypothesis. These errors also determine the reliability of the test (1-α) and the power of test (1-β). While α is directly determined by the researchers and generally as taken 0.05 (in some cases 0.01), β cannot be determined directly. Because β, hence the power of test (1-β) depends on the α (negatively correlated with β) the variation in the population (positively correlated with β) and sample size (n; negatively correlated with β). In clinical researches, it is required that β does not exceed 0.10 (in some cases 0.05) so the power of test should be at least 0.90 and above. In this study, the sample sizes required for some statistical tests (independent sample t-test, one-way ANOVA and Chi-square) which are widely used in clinical research, were calculated with the G*Power program and some evaluations were made. As a result, as expected in the statistical tests, it was observed that decreasing both α and effect size and increasing the power of the test significantly increased the required sample size. However, it was also observed that increasing effect on the sample size of increasing the power of test decreased (5-11%) in the smaller values of α in the independent sample t-test, decreased (nearly 5%) when increasing the number of compared groups in one-way ANOVA and decreased (10-15%) when increasing degree of freedom of Chi-square test.

Keywords: Clinical researches, Test of hypothesis, Sample size, Power analysis, Effect size

***Corresponding author:** Niğde Ömer Halisdemir University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, 51240, Niğde, Turkey
E mail: aunalan@ohu.edu.tr (A. ÜNALAN)
Adnan ÜNALAN https://orcid.org/0000-0002-9187-9713

Received: January 22, 2021
Accepted: March 17, 2021
Published: September 01, 2021

Cite as: Ünalan A. 2021. Sample size in clinical researches: power of the test and effect size. BSJ Health Sci, 4(3): 221-227.

1. Introduction

From the past to the present, many studies have been carried out on primarily the protection of individual's health, hence public health or on the diagnosis and treatment of various health problems faced over time. When the subject is evaluated in this respect; in order to potential drugs, medical devices, other diagnostic/therapeutic products and methods to be made available to the public, the safety and effectiveness of these products/methods must be proven by a number of studies (Anonymous, 2020a).

One of the most important research in the field of health is clinical research. Clinical research is briefly defined as "scientific study conducted with the participation of volunteers and aimed at obtaining medical knowledge" (Anonymous, 2020b). In addition to this definition, clinical research can be conducted with the aim of more effective use of an existing diagnosis/treatment method/product or to provide more information about them.

As in many types of research, in order to obtain more accurate, reliable and effective results from clinical research, which are becoming more important day by day, it is extremely important to plan every stage of the research with the necessary financial support and to conduct it with a study team with sufficient knowledge and experience. As understood from the explanations given above, the first thing to do for a clinical research; of course, the research idea/subject can be determined realistically and correctly. Then subsequent stages of the research are if the research results are to be used for a population (if the research is not only a descriptive study) exactly determination of this population or limitations, if any, the establishment of the research hypothesis, determination of the sample size that can accurately represent the population, collection of accurate and unbiased data from research units/subjects with appropriate tools, preparing the raw data for statistical analysis, selecting and analysing the suitable statistical test for the collected data, interpreting of the results and finally reporting of the research.

In this study, the effects of some factors such as the error types in hypothesis tests, the power of test and effect size which is much more prominent in clinical studies on the sample size for some basic statistical tests were calculated with the G*Power 3.1 program (Cohen, 1988; Faul et al, 2007; Faul et al, 2009) and some comments were made on the results obtained.

2. Material and Methods

2.1. Establishment of Hypothesis in Research

In the hypothesis set of a scientific research; there are two hypotheses containing opposing judgments:

H_0: Null hypothesis

H_1: Research/alternative hypothesis

Alternative hypothesis may be more than one depending on the type of the researches. In scientific research hypothesis, it is generally examined the means, proportions or relationships between/among the groups or variables. For example, H_0: "There is no statistically significant difference between the means of the groups to be compared (two groups or more)" (mathematically $\mu_1=\mu_2$ or $\mu_1-\mu_2=0$ or the mean of two groups, where μ represents the population mean), opposite this the alleged situation, that is, the alternative hypothesis is put. For example, H_1: "There is a statistically significant difference between the means of the two groups to be compared" (if the hypothesis is two-tailed $\mu_1\neq\mu_2$ or $\mu_1-\mu_2\neq0$) or if the hypothesis is one-tailed "the mean of the first group is statistically significant and greater than the mean of second group" right tailed test: $\mu_1>\mu_2$) or "the mean of the first group is statistically significant and smaller than the mean of the second group" (left tailed test: $\mu_1<\mu_2$)". When the number of groups is more than two, the null hypothesis is established as "there is no statistically significant difference between the means of the groups", while the alternative hypothesis will be "there is a statistically significant difference between the means of at least one of the groups to be compared".

2.2. Types of Errors and Their Effects in Hypothesis Tests

Statistical decision because of hypothesis test; by looking at the resulting value of probability (P) of test statistics: it is given as whether the null hypothesis (H_0) cannot be rejected (in other words, it is accepted, $P>\alpha$) or it is rejected ($P<\alpha$). Here α indicates the significance level of the test. Naturally, if the null hypothesis is accepted as a result of the statistical test the alternative hypothesis will be rejected, and if the null hypothesis is rejected the alternative hypothesis will be accepted. Any decision made as a result of hypothesis testing is either a truly correct or incorrect decision.

It is possible to face two types of errors (type 1 and type 2 error) in the decision made at the end of the hypothesis tests. Type 1 error is denoted by α (this is also the significance level of the test) and indicates the probability of rejecting the null hypothesis, which is actually true, as a result of the statistical test, while the type 2 error is denoted by β, and is the probability of accepting the null hypothesis that is actually false. For example, finding a significant difference between the effects of two drugs with the same active ingredient if only the box labels given different indicates that type 1 error was made, while the active ingredients were different and the effect of one was really better, there was no significant difference between the effects of the two drugs indicates a type 2 error. These errors also determine the reliability level of the test (1-α) and power of the test (1-β). While the probability of type 1 error (α) is determined by the researcher and is usually taken as 0.05, but the power of the test cannot be determined directly. Because β, hence the power of test (1-β) depends on α (α is correlated negatively with β) the variation in the population (variation is correlated positively with β) and sample size (denoted by n; it is correlated negatively with β).

2.3. Sample Size

Today, ethics approvals are required by both authorized local ethics committees at the application phase of clinical research and the journal editors at the publication of the results of research. In order to obtain approval from the ethics committees, issues such as how the sample size projected in the research is determined, whether power analysis have been made for the statistical test to be used, and under what assumptions these analyses are made are important questions/problems.

The sample size, in other words, the number of volunteers/subjects used in the study; it is extremely important in terms of showing both whether the results of the research are scientifically valid and whether the research meets the ethical principles. Because the use of more than necessary subjects in the research will cause economic losses by bringing more time, labor and cost, as well as bringing serious ethical problems, and the fact that the sample size is less than necessary will cause the decisions made at the end of the study to be wrong and thus the research to lose its scientific validity. When this situation is evaluated clinically, the use of fewer subjects than necessary in the study may cause a significant clinical effect not to be seen, while using a larger number of subjects may result in a statistically significant but not actually clinically significant effect.

2.4. Effect Size

Although the effect size is calculated in different ways according to the statistical tests used in the analysis, simply; it can be defined as the difference between the means of the groups to be compared (e.g. control/placebo group and experimental group). This difference is usually expressed in terms of standard deviation. Effect size is an extremely important criterion for clinical significance in clinical researches. That is to say, a statistically significant result may not be clinically significant (Kalacıoğlu and Akhanlı, 2020). For example, in a study conducted on too many subjects than it should have been, even if the difference between group means is very small, this difference may be statistically significant ($P<\alpha$). Therefore, giving effect sizes as well as statistical significance in clinical studies will make the research more valid.

We know that the sample size is determined at the beginning of the study and this value is significantly affected from the effect size selected for the study. Effect size in a study; it should not be manipulated in order to reduce the sample size (by increasing the effect size bias), it should be determined in accordance with the effect sizes derived from the results of previous research on the subject or the results of the pilot study. The effect sizes were defined by Cohen (1988) as Cohen's *d* for the independent sample *t*-test, Cohen's *f* for one-way ANOVA and Cohen's *w* for the Chi-square test. The researcher has also classified effect sizes as small, medium and large (0.20, 0.50 and 0.80 for independent *t*-test; 0.10, 0.25 and 0.40 for one-way ANOVA test; 0.10, 0.30 and 0.50 for and Chi-square test).

The effect size (*d*) formula (equation 1) for the independent sample *t*-test is given below;

For the population;

$$d = \frac{\mu_1 - \mu_2}{\sigma} \tag{1}$$

Where, μ_1 and μ_2 are the population means, σ is the population standard deviation. σ formula (equation 2) is given below;

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{2}$$

Cohen argued that the standard deviation of either group could be used when the variances of the two groups are homogeneous (equation 3 and 4).

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}} \tag{3}$$

$$\sigma_{pooled} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}} \tag{4}$$

In practice, the effect size (*d*) is calculated from differences between the compared samples means by dividing standard deviation (*S*). Formula (equation 5) is given below;

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S} \tag{5}$$

Where, $\bar{X}_1$ and $\bar{X}_2$ are the compared sample means, *S* is the pooled within sample estimate of the population standard deviation. The formula (equation 6) of pooled *S* is given below;

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \tag{6}$$

Where, n_1 and n_2 are compared sample sizes and S_1^2 and S_2^2 are variances of the compared samples.

The effect size (*f*) formula (equation 7) for the one-way ANOVA test is given below;

$$f = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MSE\left(\frac{n_1 + n_2 - 2}{n_1 + n_2}\right)}} \tag{7}$$

Where, $\bar{X}_1$ and $\bar{X}_2$ are the sample means, n_1 and n_2 are the sample sizes and MSE is mean square error.

The effect size (*w*) formula for the Chi-square test is given below (equation 8);

$$w = \sqrt{\sum_{i=1}^{m} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}} \tag{8}$$

Where, P_{0i} is the proportion in cell *i* posited by the null hypothesis P_{1i} is the proportion in cell *i* posited by the alternate hypothesis and reflects the effect for the cell, *m* is the number of the cell.

3. Results

There is much free software available on the web for the calculation of sample size. One of them is the G*Power program that helps researchers to calculate the sample size. In this study, sample sizes for different power (1-*β*) and effect sizes (Cohen's *d*, *f* and *w*) for independent sample *t*-test, one-way ANOVA and Chi-square tests, which are frequently used in clinical research, were determined using the G*Power 3.1 program (Foul et al, 2007; Foul et al, 2009) were calculated, summarized in tables, and some comments were made.

3.1. Sample Size for Independent Sample *t*-test

The t-test, one of the parametric tests, is used to test whether there is a statistically significant difference between the means of two independent groups (e.g. control/placebo and experimental/treatment groups). Here, it is assumed that the data to be used meet the parametric test assumptions (normal distribution and homogeneity of variances). The sample sizes required for this test were calculated in G*Power 3.1 program and summarized in Table 1.

From the data in Table 1, it is seen that increasing the sample size as expected to increase the power of the test and increasing the effect size significantly reduces the sample size. When the results in the table are evaluated in terms of the effect of α on the sample size; for example, when *α* = 0.05 and effect size *d* = 0.5 (medium), rising up the test power from 0.80 to 0.95 which increases increases the sample size by about 73% (from 102 to 176); when the value of *α* at the same level (0.5) effect size is reduced to 0.025, it is seen that the sample size increases by 64% (from 128 to 210). This shows that increasing the power of the test on the sample size has a less enhancing effect (5-11%) at smaller values of *α*.

3.2. Sample Size for One-way Analysis of Variance (ANOVA)

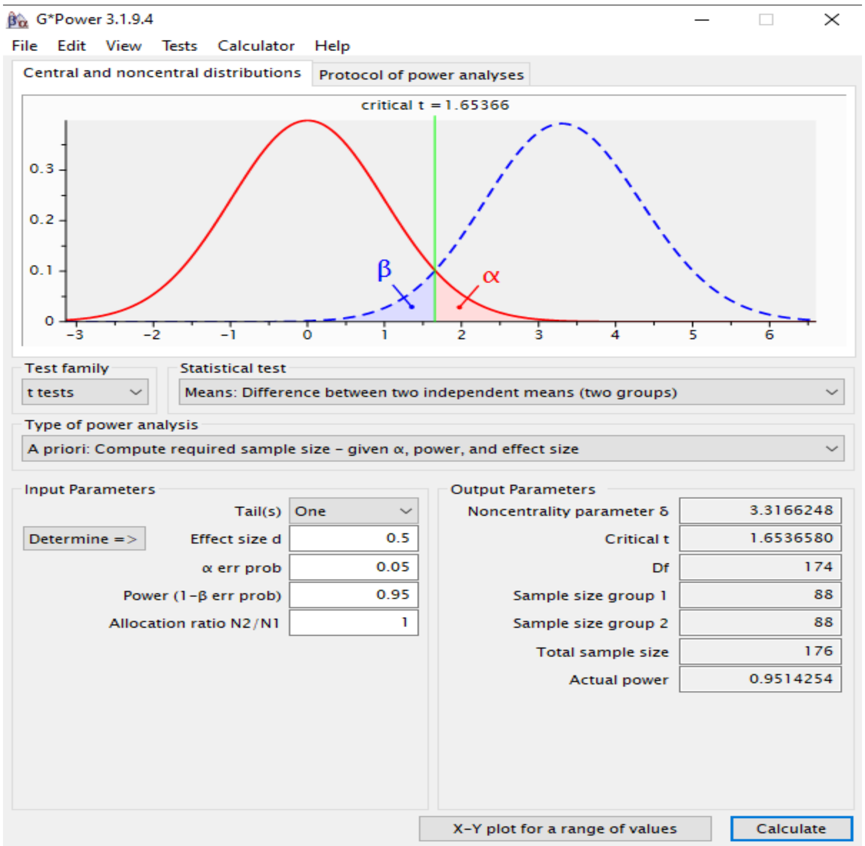
One-way ANOVA; it is a parametric test used to test whether there is a statistically significant difference between the means of more than two independent groups. Here, it is assumed that the data to be used meet the parametric test assumptions (such as normal distribution of errors and homogeneous variances).

Table 1. Sample sizes (n) for the independent sample t-test (α=0.05 one-way / two-way; $\alpha/2$=0.025 in two-way test)

Power of Test	Effect Size (Cohen's d)						
(1-β)	0.2 (Small)	0.3	0.4	0.5 (Medium)	0.6	0.7	0.8 (Large)
0.80	620/788	278/352	156/200	102/128	72/90	52/68	42/52
0.85	722/900	322/402	182/228	118/146	82/105	62/76	48/60
0.90	858/1054	382/470	216/266	140/172	98/120	72/88	56/68
0.95	1084/1302	484/580	272/328	176**/210	122/148	90/110	70/84
Increasing of n (%*)	75/65	74/65	74/64	73/64	69/64	73/62	67/62

* When the power of the test is increased from 0.80 to 0.95.

** The screenshot of G*Power 3.1 program is given in Figure 1.



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s)	One
Effect size d	0.5
α err prob	0.05
Power (1−β err prob)	0.95
Allocation ratio N2/N1	1

Output Parameters

Noncentrality parameter δ	3.3166248
Critical t	1.6536580
Df	174
Sample size group 1	88
Sample size group 2	88
Total sample size	176
Actual power	0.9514254

Figure 1. G*Power 3.1 program screenshot for independent samples t-test.

The sample sizes required for this test were calculated in G*Power 3.1 program and summarized in Table 2.

From the data in Table 2; in one-way ANOVA, it is seen that generally increasing the number of groups and the power of the test which increases the sample size as expected and increasing the effect size significantly decreases the sample size. When the subject is evaluated together in terms of the number of groups and the power of the test; for example, for α=0.05 and group number 3, the effect size f=0.25 (medium), while rising up the test power from 0.80 to 0.95; it increased the sample size by about 58% (from 159 to 252); while the number of groups with the same α and effect size was 4, the sample size increased by 56% (from 180 to 280); when the number of groups is 5, it is seen that the sample size increases by 53% (from 200 to 305). These results show that increased number of groups to be compared decreases sample size slightly (approximately 5%).

3.3. Sample Sizes for Chi-Square Test

Chi-square (χ^2) test; it is a test frequently used in the analysis of categorical data. In this test, the cross table consisting of rows and columns is created and it is investigated whether the observed and expected frequencies (number of subjects/units) in each cell of this table are compatible. In a single row or single column Chi-square table (homogeneity), the degree of freedom is determined as the total number of cells-1, while the degree of freedom of the Chi-square table consisting of rows and columns is calculated as (row number-1) x (column number-1). The sample sizes required for this test were calculated in G*Power 3.1 program and summarized in Table 3.

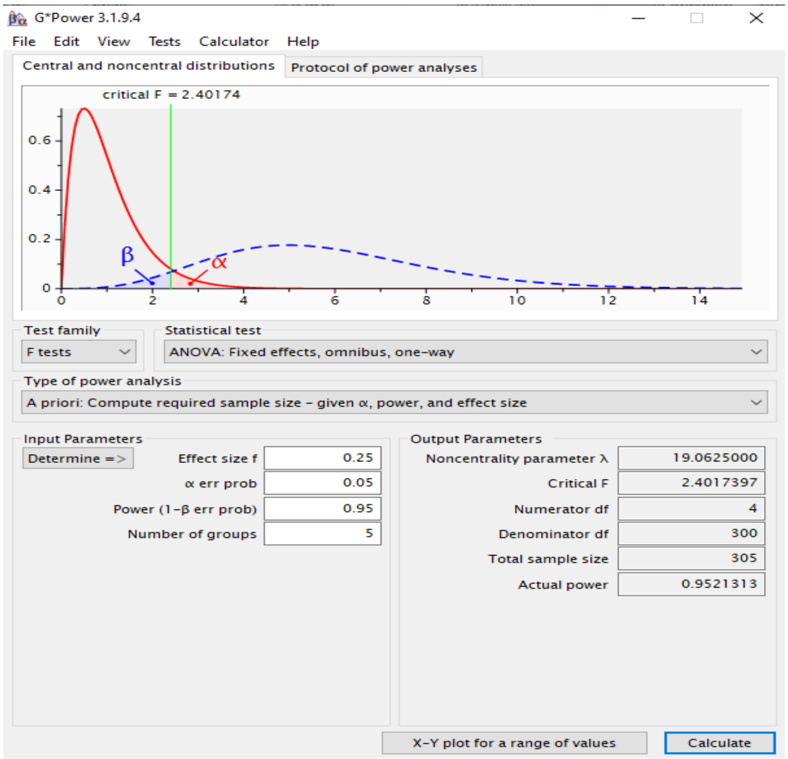
From the data in Table 3, it is seen that increasing the power of the test in the Chi-square test which increases the sample size as expected, the sample size increases again with the increase in the degree of freedom and increasing the effect size significantly reduces the sample size. When the results are evaluated together in terms of the degree of freedom and the power of the test; for example, for α=0.05 and degree of freedom 1, when the effect size w=0.30 (medium), subtracting the power of test from 0.80 to 0.95; it increased the sample size approximately 65% (from 88 to 145); it is seen that while the degree of freedom is 5 at the same α and effect size, the sample size increases by 53% (from 143 to 220). This shows that as the degree of freedom that increases in Chi-square tests, the increasing effect of increasing the power of the test on the sample size decreases (approximately 10-15%).

Table 2. Sample sizes (n) for one-way ANOVA (α=0.05)

		Effect Size (Cohen's f)						
No. of Groups	Power of Test (1-β)	0.10 (Small)	0.15	0.20	0.25 (Medium)	0.30	0.35	0.40 (Large)
3	0.80	969	432	246	159	111	84	66
	0.85	1098	489	279	180	126	93	72
	0.90	1269	567	321	207	144	108	84
	0.95	1548	690	390	252	177	132	102
Increasing of n (%*)		60	60	59	58	59	57	55
4	0.80	1096	492	280	180	128	96	76
	0.85	1236	552	312	204	144	108	84
	0.90	1424	636	360	232	164	120	96
	0.95	1724	768	436	280	196	148	112
Increasing of n (%*)		57	56	56	56	53	54	47
5	0.80	1200	540	305	200	140	105	80
	0.85	1350	605	345	220	155	115	90
	0.90	1550	690	390	255	180	135	105
	0.95	1865	835	470	305**	215	160	125
Increasing of n (%*)		55	55	54	53	54	52	56

* When the power of the test is increased from 0.80 to 0.95.

** The screenshot of G*Power 3.1 program is given in Figure 2.



Test family: F tests

Statistical test: ANOVA: Fixed effects, omnibus, one-way

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f: 0.25
α err prob: 0.05
Power (1–β err prob): 0.95
Number of groups: 5

Output Parameters

Noncentrality parameter λ: 19.0625000
Critical F: 2.4017397
Numerator df: 4
Denominator df: 300
Total sample size: 305
Actual power: 0.9521313

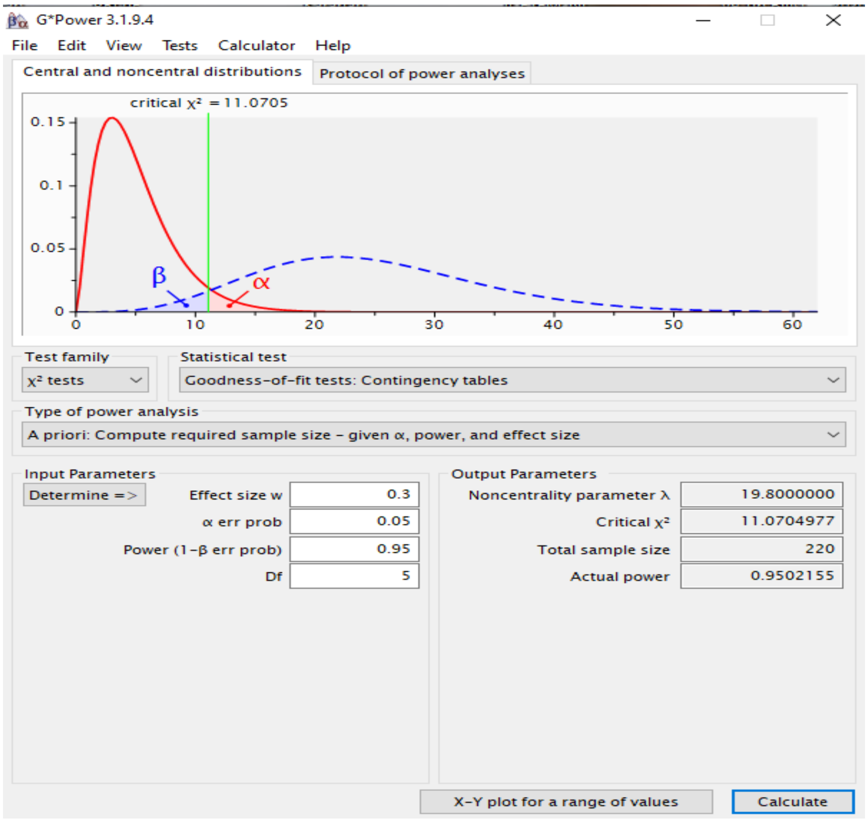
Figure 2. G*Power 3.1 program screenshot for one-way ANOVA test.

Table 3. Sample sizes for the Chi-square test (α=0.05)

		Effect Size (Cohen's *w*)				
Degrees of Freedom	Power of Test (1-β)	0.10 (Small)	0.20	0.30 (Medium)	0.40	0.50 (Large)
1	0.80	785	197	88	50	32
	0.85	898	225	100	57	36
	0.90	1051	263	117	66	43
	0.95	1300	325	145	82	52
Increasing of n (%*)		66	65	65	64	63
2	0.80	964	241	108	61	39
	0.85	1093	274	122	69	44
	0.90	1266	317	141	80	51
	0.95	1545	387	172	97	62
Increasing of n (%*)		60	61	59	59	59
3	0.80	1091	273	122	69	44
	0.85	1231	308	137	77	50
	0.90	1418	355	158	89	57
	0.95	1717	430	191	108	69
Increasing of n (%*)		57	58	57	57	57
4	0.80	1194	299	133	75	48
	0.85	1343	336	150	84	54
	0.90	1541	386	172	97	62
	0.95	1858	465	207	117	75
Increasing of n (%*)		56	56	56	56	56
5	0.80	1283	321	143	81	52
	0.85	1440	360	160	90	58
	0.90	1647	412	183	103	66
	0.95	1979	485	220**	124	80
Increasing of n (%*)		51	54	53	54	51

* When the power of the test is increased from 0.80 to 0.95.

** The screenshot of G*Power 3.1 program is given in Figure 3.



Test family: χ² tests

Statistical test: Goodness-of-fit tests: Contingency tables

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size w: 0.3

α err prob: 0.05

Power (1-β err prob): 0.95

Df: 5

Output Parameters

Noncentrality parameter λ: 19.8000000

Critical χ²: 11.0704977

Total sample size: 220

Actual power: 0.9502155

Figure 3. G*Power 3.1 program screenshot for Chi-square test.

4. Discussion

Correct determination of the sample size in clinical research is of great importance both in terms of ethics rules and for the results to be obtained from statistical tests to be scientifically acceptable and valid. One of the most important criteria when determining the sample size in clinical studies is the effect size. Keeping the effect size high by manipulating it in order to reduce the sample size is not an acceptable approach. For this reason, the effect sizes should be determined in accordance with the results of the previous studies on the subject, if not, the pilot study results, and if this is not possible, the medium effect size value should be used. In this study, the effects of the significance level of the test (α), the power of test (1-β) and the effect size on the sample size were investigated for some statistical tests that are frequently used in clinical research.

When the subject is evaluated in general, increasing the effect size significantly reduces the sample size (for example, in the independent sample t-test for α=0.05 and the test power=0.95, the effect sizes are 0.2 (small), 0.5 (medium) and 0.8 (large), the sample sizes are 1084, 176 and 70, respectively), so reducing α increases also the sample size as expected, so using the generally accepted value of 0.05 instead of 0.01 or 0.025 (α=0.05 two-tailed) for α. It was observed that it would be a more accurate approach, besides, the effect of increasing from 0.80 to 0.95 the power of the test on the sample size in the independent sample t-test has a lower effect (5-11%) at smaller values of α. Increasing the number of groups to be compared in the one-way ANOVA test increases the sample size. However, the effect of increasing from 0.80 to 0.95 the power of the test when the number of compared groups increase, the sample size decreases slightly (about 5%). In Chi-square tests, the sample size increases by increasing of degree of freedom. However, the effect of increasing from 0.80 to 0.95 the power of the test when the degree of freedom increase, the required sample size decreases slightly (10-15%).

Author Contributions

All tasks were done by the single author and author reviewed and approved the manuscript.

Conflict of Interest

The author declared that there is no conflict of interest.

Ethical Approval/Informed Consent

Ethics committee approval is not required for this study and was not provided.

References

Anonymous 2020a URL: https://www.titck.gov.tr/faaliyetalanlari/ilac/klinik-arastirmalar (access date: September 20, 2020).

Anonymous 2020b. URL: https://ikua.saglik.gov.tr/TR,259729/klinik-arastirma-nedir.html, (access date: September 20, 2020).

Cohen J. 1988. Statistical power analysis for the behavioral sciences. Hillsdale, Lawrence Erlbaum Associates, New Jersey, USA, 567 pages.

Faul F, Erdfelder E, Lang AG, Buchner A. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods, 39(2): 175-191.

Faul F, Erdfelder E, Buchner A, Lang AG. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Methods, 41(4): 1149-1160.

Kalaycıoğlu O, Akhanlı SE. 2020. Sağlık araştırmalarında güç analizinin önemi ve temel prensipleri: Tıbbi çalışmalar üzerinde uygulamalı örnekler. Turk J Public Health, 18(1): 103-112.