Black Sea Journal of Engineering and Science
Open Access Journal
e-ISSN: 2619-8991

BSPublishers

Research Article
Volume 2 - Issue 1: 16-22 / January 2019

DETERMINING SAMPLE SIZE IN LOGISTIC REGRESSION WITH G-POWER

Aysel YENİPINAR[1]*, Şeyma KOÇ[1], Demet ÇANGA[2], Fahrettin KAYA[3]

[1]Kahramanmaraş Sütçü Imam University, Faculty of Agriculture, Department of Animal Science, 46040, Onikişubat, Kahramanmaraş, Turkey

[2]Osmaniye Korkut Ata University, Department of Food Processing, 80050, Bahçe, Osmaniye, Turkey

[3]Kahramanmaraş Sütçü Imam University, Andırın Vocational School, Computer Technology, Department of Animal Science, 46410, Onikişubat, Kahramanmaraş, Turkey

***Received:** October 11, 2018; **Accepted:** November 23, 2018; **Published:** January 01, 2019*

Abstract

There are several methods used to determine the sample size. Investigator; because of the insufficient precious resources such as time, labor, money, tools and equipment, it works by pulling the sample with a suitable sampling method from the population it is examining. According to the statistics obtained from the sample, he will make comments about the population and make decisions. The correctness of the decisions made is closely related to the size of the sample. For this reason, the problem of determining sample size is one of the first and important problems of an investigator. A small sample of information causes loss of information and misjudgments. A very large sample is contrary to the purpose of sampling and resources are wasted. The calculation of the sample size can now be done very easily via free programs.

Keywords: Sample size determination, G-Power, Logistic regression

***Corresponding author:** Kahramanmaraş Sütçü Imam University, Faculty of Agriculture, Department of Animal Science, 46040, Onikişubat, Kahramanmaraş, Turkey
E mail: ayselyenipinar46@gmail.com (A. YENIPINAR)

Aysel YENİPINAR	https://orcid.org/0000-0001-2345-6789
Şeyma KOÇ	https://orcid.org/0000-0001-5708-9905
Demet ÇANGA	https://orcid.org/0000-0003-3319-7084
Fahrettin KAYA	https://orcid.org/0000-0003-1666-4859

Cite as: Yenipinar A, Koc S, Canga D, Kaya F. 2019. Determining sample size in logistic regression with G-power. BSJ Eng Sci, 2(1): 16-22.

1. Introduction

The general problem of many researchers is the problem of determining sample size in their studies. In addition to the fact that logistic regression analysis is a difficult method of analysis, there is no common acceptance in determining sample size. Different views and formulas were developed by the authors to determine the sample size in logistic regression analysis. Hsieh et al. (1998) proposes to compare averages or compare ratios to calculate a simple sample size for linear and logistic regression. Moineddin et al. (2007) conducted a simulation study of sample size for multi-level logistic regression models. In the study, it was determined that the sample size of the individual and group level, multilevel logistic regression models parameters and variance components to evaluate the effect on the accuracy of estimates. Demidenko (2007) derived the general Wald-based power and sample size formulas for Logistic regression in the sample size determination study for the revised logistic regression and then applied them to binary exposure and agitator to obtain closed form expression. Herrera and Gomez (2008) conducted a Monte Carlo simulation study. They examined the different sample sizes of reference and focal groups in their studies and determined the effect of logistic regression on power and type I error. In their logistic regression analysis, Nemes et al. (2009) reported that the odds ratio was more predicted in small to medium sample size studies, whereas small sample size was systematic bias and deviation from zero.

Different package programs have been produced to determine sample size with the help of the findings and opinions of the researchers in logistic regression analysis. G* Power is one of the programs that calculations of the sample size. It is a program of 17 Mb sizes, easy to install and use. G*Power (Erdfelder et al., 1996) was designed as a general stand-alone power analysis program for statistical tests commonly used in social and behavioral research. However, it can also be used for science and medicine.

The purpose of this article is to introduce the researchers to the calculation of the sample size problem for logistic regression analysis. The results of logistic regression analysis can be interpreted easily for medical data as it is easy to interpret for binary variables (yes-no, not- none or limit values up- down, etc.). Cost, labor and time are very important for the researcher when calculating sample size. These factors are becoming more important, especially because of the high cost in the field of medicine and the difficulty in the application of some medical procedures according to the ethical rules. In logistic regression analysis, G-power can be used as an alternative to complex formulas and simulation applications in determining sample size and it is a program that everyone can easily apply to solve the problem of determining sample size.

2. Material and Method

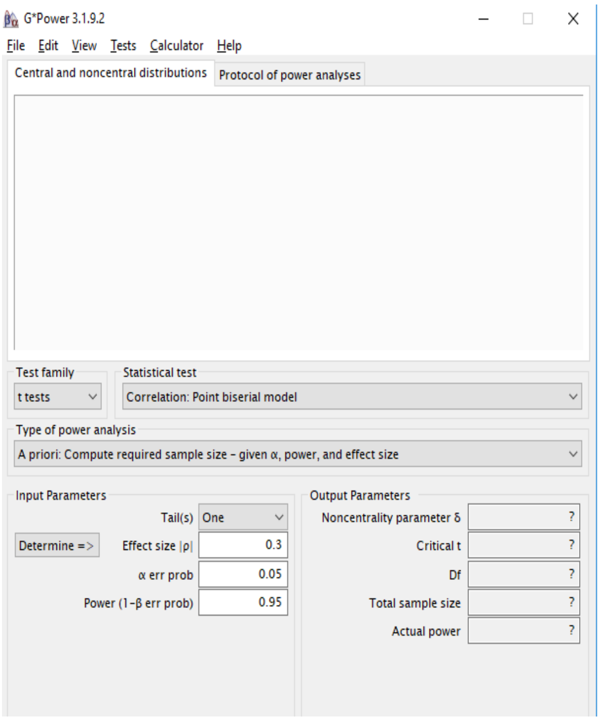
2.1. Material

In this study, a numerical example is used to understand the use of G power. In our example, 45% male X = 1, 55% female and X = 0.70% of the girls and 80% of the boys were taken as normal birth weight. These numerical sample values were obtained randomly from the births of Kahramanmaraş Maternity Hospital between January-June 2018.Probability of error (significance level) α = 0.05 and the power of the test was taken as 1-β = 0.95.Impact magnitude (calculated in one way or look at previous work).

2.2. Method

G* Power is a major extension of, and improvement over, the previous versions. It runs on widely used computer platforms (Windows XP, Windows Vista, and Mac OS X 10.4) and covers many different statistical tests of the t, F, and chi2 test families. In addition, it includes power analyses for z tests and some exact tests. G*Power provides improved effect size calculators and graphic options, supports both distribution-based and design-based input modes, and offers all types of power analyses in which users might be interested.

G*Power (Figure 1 shows the main window of the program) covers statistical power analyses for many different statistical tests of the F test, t test, χ2-test and z test families and some exact tests. G*Power provides effect size calculators and graphics options. G*Power supports both a distribution-based and a design-based input mode. It contains also a calculator that supports many central and noncentral probability distributions. G*Power is free software and available for Mac OS X and Windows XP/Vista/7/8.


G*Power 3.1.9.2
File
Edit
View
Tests
Calculator
Help
Central and noncentral distributions
Protocol of power analyses
Determine =>

Test family: t tests

Statistical test: Correlation: Point biserial model

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

Input Parameters

Tail(s): One

Effect size |ρ|: 0.3

α err prob: 0.05

Power (1–β err prob): 0.95

Output Parameters

Noncentrality parameter δ: ?

Critical t: ?

Df: ?

Total sample size: ?

Actual power: ?

Figure 1. The main window of G*Power

Types of analysis; G*Power offers five different types of statistical power analysis:

1. A priori (sample size *N* is computed as a function of power level 1−β, significance level α, and the detected population effect size)
2. Compromise (both α and 1−β are computed as functions of effect size, *N*, and an error probability ratio q = β/α)
3. Criterion (α and the associated decision criterion

are computed as a function of 1−β, the effect size, and N)

4. Post-hoc (1−β is computed as a function of α, the population effect size, and N) 5. Sensitivity (population effect size is computed as a function of α, 1−β, and N).

Program handling; Perform a Power Analysis Using G*Power typically involves the following three steps:

1. Select the statistical test appropriate for your problem.
2. Choose one of the five types of power analysis is available
3. Provide the input parameters required for the analysis and click "Calculate".

Plot parameters; In order to help you explore the parameter space relevant to your power analysis, one parameter (α, power (1−β), effect size, or sample size) can be plotted as a function of another parameter.

A logistic regression model describes the relationship between a binary response variable Y (with Y = 0 and Y = 1 denoting non-occurance and occurance of an event, respectively) and one or more independent variables (covariates or predictors) Xi. The variables Xi are themselves random variables with probability density function f X(x) (or probability distribution f X(x) for discrete X).

In a simple logistic regression with one covariate X the assumption is that the probability of an event P = Pr (Y = 1) depends on X in the following way:

$$P = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \qquad (1)$$

For $\beta_1 \neq 0$ and continuous X this formula describes a smooth S-shaped transition of the probability for Y = 1 from 0 to 1 ($\beta_1 > 0$) or from 1 to 0 ($\beta_1 < 0$) with increasing x. This transition gets steeper with increasing β_1. Rearranging the formula leads to: log(P/(1−P)) = $\beta_0 + \beta_1 X$. This shows that the logarithm of the odds P/(1−P), also called a logit, on the left side of the equation is linear in X. Here, β_1 is the slope of this linear relationship. The interesting question is whether covariate X_i is related to Y or not. Thus, in a simple logistic regression model, the null and alternative hypothesis for a two-sided test are:

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$ (2)

The procedures implemented in G*Power for this case estimates the power of the Wald test. The standard normally distributed test statistic of the Wald test is:

$$z=\frac{\widehat{\beta_1}}{\sqrt{var(\widehat{\beta_1})/N}}=\frac{\widehat{\beta_1}}{SE(\beta_1)} \qquad (3)$$

where $\hat{\beta}_1$ is the maximum likelihood estimator for parameter β_1 and var($\hat{\beta}_1$) the variance of this estimate.

2.2.1. Effect size index

In the simple logistic model the effect of X on Y is given by the size of the parameter β_1. Let p_1 denote the probability of an event under H_0, that is

$exp(\beta_0) = p_1/(1-p_1)$ (4)

and p_2 the probability of an event under H_1 at X=1, that is

$exp(\beta_0 + \beta_1) = p_2/(1-p_2)$. (5)

Then

$exp(\beta_0 + \beta_1)/exp(\beta_0)$

= $\exp(\beta_1) = [p_2/(1-p_2)]/[p_1/(1-p_1)]$

= odds ratio OR, (6)

which implies

β_1 = log[OR] (Şahin, 1999) . (7)

Given the probability p_1 (input field Pr(Y=1|X=1) H_0) the effect size is specified either directly by p_2 (input field Pr(Y=1|X=1) H_1) or optionally by the odds ratio (OR) (input field Odds ratio). Setting $p_2 = p_1$ or equivalently OR = 1 implies $\beta_1 = 0$ and thus an effect size of zero. An effect size of zero must not be used in a priori analyses. Besides these values the following additional inputs are needed. (Faul et al., 2009).

R^2 other X:

In models with more than one covariate, the influence of the other covariates $X_2,...,X_p$ on the power of the test can be taken into account by using a correction factor. This factor depends on the proportion $R^2 = \rho^2_{1,2,3...p}$ of the variance of X_1 explained by the regression relationship with $X_2,...,X_p$. If N is the sample size considering X_1 alone, then the sample size in a setting with additional covariates is:

$N| = N/(1-R^2)$ (8)

This correction for the influence of other covariates has been proposed by Hsieh et al. (1998). R^2 must lie in the interval [0,1].

X distribution:

1 Bimonial

$P(k) = \binom{N}{k}\pi^k(1-\pi)^{N-k}$, (9)

where k is the number of successes (X = 1) in N trials of a Bernoulli process with probability of success π, 0 < π < 1

2 Exponential

$f(x) = (1/\lambda)e^{-1/\lambda}$, (10)

exponential distribution with parameter $\lambda > 0$.

3 Lognormal

$f(x) = 1/(x\sigma\sqrt{2\pi})\exp[-(\ln x - \mu)^2/(2\sigma^2)$ (11)

lognormal distribution with parameters μ and $\sigma > 0$.

4 Normal

$f(x) = 1/(\sigma\sqrt{2\pi})\exp[-(x - \mu)^2/(2\sigma^2)$ (12)

normal distribution with parameters μ and $\sigma > 0$.

5 Poisson

$(P(X = k) = (\lambda k/k!)e^{-\lambda}$, (13)

Poisson distribution with parameter $\lambda > 0$.

6 Uniform

$(f(x) = 1/(b-a)$ for $a \le x \le b$, $f(x) = 0$ (14)

otherwise, continuous uniform distribution in the interval [a,b], a < b).

7 Manual (Allows to manually specify the variance of $\hat{\beta}$ under H_0 and H_1)
G*Power provides two different types of procedure to calculate power: An enumeration procedure and large sample approximations. The Manual mode is only available in the large sample procedures.
The odds ratio is found when H1 and H0 values are entered and calculated.When it transfers to the other window it looks like these transfer. Type 1 error is 0.05 and power is 0.95. Take R as 0, and enter the X distribution binominal.

Pr(Y=1| X=1)= H_1 =0.7 factor + probability of occurrence
Pr(Y=1| X=1)= H_0 =0.3 factor - probability of occurrence

And after calculating chose calculate and transfer to main window.
R^2other =X In the presence of other variables, the variance of the main locator is '0' if there is no other locator. If there is correlation, it is squared according to Low / medium / high correlation. For example R = 0.20 then R^2 = 0.04
X parm μ: Factor + events available in reference studies. If factor + and factor - equals 0.50 must be entered. In this example + factor % 45 so that it must be used %45.

2.2.2. Options

In Input mode, you can choose between two input modes for the effect size: The effect size may be given by either specifying the two probabilities p_1, p_2 defined above, or instead by specifying p1 and the odds ratio OR.
Procedure G*Power provides two different types of procedure to estimate power. An "enumeration procedure" proposed by Lyles et al. (2007) and large sample approximations. The enumeration procedure seems to provide reasonable accurate results over a wide range of situations, but it can be rather slow and may need large amounts of memory. The large sample approximations are much faster. Results of Monte-Carlo simulations indicate that the accuracy of the procedures proposed by Demidenko (2007) and Hsieh et al. (1998) are comparable to that of the enumeration procedure for N > 200. The procedure base on the work of Demidenko (2007) is more general and slightly more accurate than that proposed by Hsieh et al. (1998). It is therefore recommended to use the procedure recommended by Demidenko (2007) as a standard procedure. The enumeration procedure of Lyles et al. (2007) may be used to validate the results (if the sample size is not too large). It must also be used, if one wants to compute the power for likelihood ratio tests.
The enumeration procedure provides power analyses for the Wald-test and the Likelihood ratio test. In logistic regression analysis, Wald test can be used to test whether the coefficients are important. The distribution of the test statistic of the Wald test approaches the standard normal distribution. For each variable, the Z test is performed using standard errors in the list. The Wald test is meaningful if the sample volume is large. (Buse, 1982).
The highest probability estimator of the slope parameter is compared to the estimated value of the standard error. β_1=0, while the distribution of the test statistic is suitable for standard normal distribution. The test statistic of this test;

$$W=\frac{\widehat{\beta_1}}{\widehat{SE}\widehat{\widehat{\beta_1}}} \quad (15)$$

is obtained by the formula (Buse, 1982).
The general idea is to construct an exemplary data set with weights that represent response probabilities given the assumed values of the parameters of the X distribution. Then a fit procedure for the generalized linear model is used to estimate the variance of the regression weights (for Wald tests) or the likelihood ratio under H_0 and H_1 (for likelyhood ratio tests). The size of the exemplary data set increases with *N* and the enumeration procedure may thus be rather slow (and may need large amounts of computer memory) for large sample sizes. The procedure is especially slow for analysis types other then "post hoc", which internally call the power routine several times. By specifying a threshold sample size *N* you can restrict the use of the enumeration procedure to sample sizes < *N*. For sample sizes ≥ *N* the large sample approximation selected in the option dialog is used. Note: If a computation takes too long you can abort it by pressing the ESC key. 2. G*Power provides two different large sample approximations for a Wald-type test. Both rely on the asymptotic normal

distribution of the maximum likelihood estimator for parameter β_1 and are related to the method described by Whittemore (1981). The accuracy of these approximation increases with sample size, but the deviation from the true power may be quite noticeable for small and moderate sample sizes. This is especially true for X-distributions that are not symmetric about the mean, i.e. the lognormal, exponential, and Poisson distribution, and the Binomial distribution with $\pi \neq 1/2$. The approach of Hsieh et al. (1998) is restricted to binary covariates and covariates with standard normal distribution. The approach based on Demidenko (2007) is more general and usually more accurate and is recommended as standard procedure. For this test, a variance correction option can be selected that compensates for variance distortions that may occur in skewed *X* distributions (see implementation notes). If the Hsieh procedure is selected, the program automatically switches to the procedure of Demidenko if a distribution other than the standard normal or the binomial distribution is selected (Faul at al., 2007).

3. Results and Discussion

Select the statistical test appropriate for your problem In Step1, the statistical test is chosen using the distribution based or the design-based approach.

Distribution-based approach to test selection First select the family of the test statistic (i.e., exact, F, t, χ2, or z test) using the Test family menu in the main window. The Statistical test menu adapts accordingly, showing a list of all tests available for the test family.

Is there a relationship between birth weight and gender? Is it a predictor for gender *X* addiction?

In this example, 45% male *X* = 1, 55% female and X = 0. 70% of the girls and 80% of the boys were taken as normal birth weight. Probability of error (significance level) $\alpha = 0.05$ and the power of the test was taken as $1-\beta = 0.95$.Impact magnitude (calculated in one way or look at previous work).

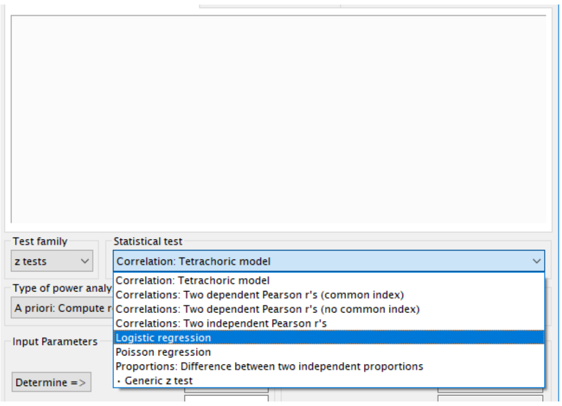

Test family
z tests
Statistical test
Correlation: Tetrachoric model
Correlation: Tetrachoric model
Correlations: Two dependent Pearson r's (common index)
Correlations: Two dependent Pearson r's (no common index)
Correlations: Two independent Pearson r's
Logistic regression
Poisson regression
Proportions: Difference between two independent proportions
· Generic z test
Type of power analy
A priori: Compute r
Input Parameters
Determine =>

Figure 2. Choice of Logistics regression from the windows

In order to select the logistic regression menu from the statical test, the test family Z test selected first.

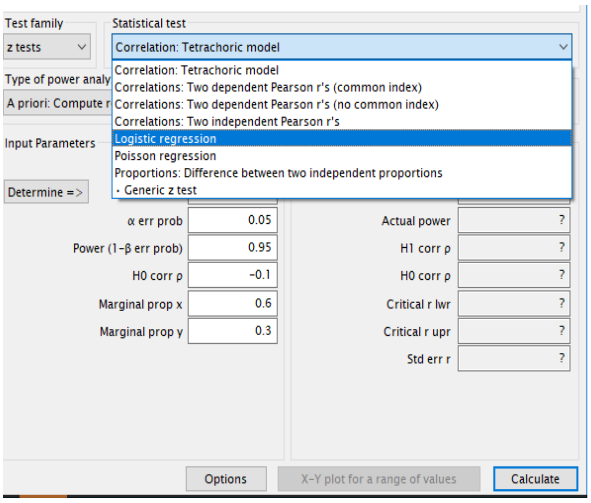

Test family
z tests
Statistical test
Correlation: Tetrachoric model
Correlation: Tetrachoric model
Correlations: Two dependent Pearson r's (common index)
Correlations: Two dependent Pearson r's (no common index)
Correlations: Two independent Pearson r's
Logistic regression
Poisson regression
Proportions: Difference between two independent proportions
· Generic z test
Type of power analy
A priori: Compute r
Input Parameters
Determine =>
α err prob 0.05
Power (1-β err prob) 0.95
H0 corr ρ -0.1
Marginal prop x 0.6
Marginal prop y 0.3
Actual power ?
H1 corr ρ ?
H0 corr ρ ?
Critical r lwr ?
Critical r upr ?
Std err r ?
Options
X-Y plot for a range of values
Calculate

Figure 3. Choice of options menu from the windows

On the pre-pop-up screen there is a section called option, from which can calculate the sample size from odds ratio or two probabilities from two different locations. Especially with variance correction, the sample size increases slightly. Because other features are selectively closed, no changes are made.

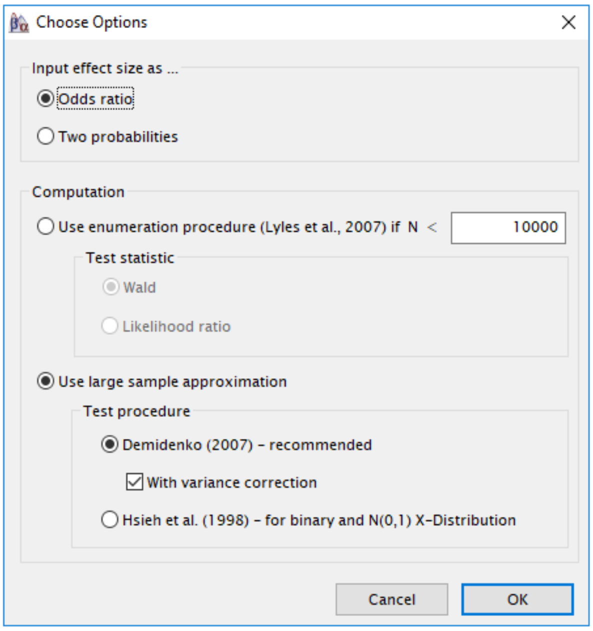

Choose Options
Input effect size as ...
● Odds ratio
○ Two probabilities
Computation
○ Use enumeration procedure (Lyles et al., 2007) if N < 10000
Test statistic
○ Wald
○ Likelihood ratio
● Use large sample approximation
Test procedure
● Demidenko (2007) – recommended
☑ With variance correction
○ Hsieh et al. (1998) – for binary and N(0,1) X-Distribution

Figure 4. It can be checked with variance correction option

Define the following values;

Probability of error (significance level) $\alpha = 0.05$

The power of the test was taken as $1-\beta = 0.95$.

Impact magnitude (calculated in one way or look at previous work).

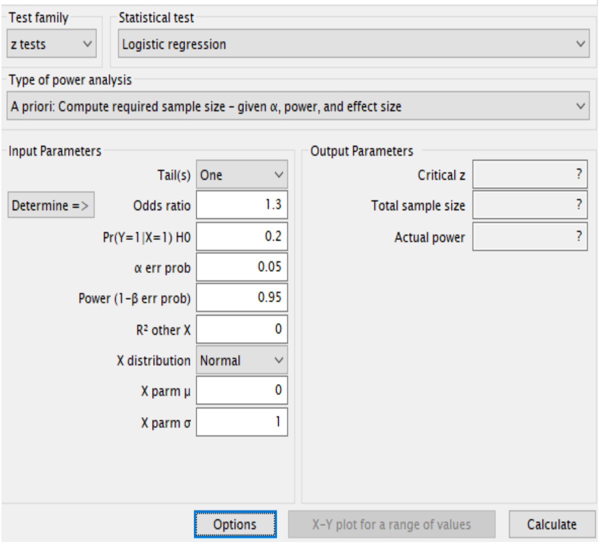

Test family: z tests
Statistical test: Logistic regression
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size
Input Parameters
Tail(s): One
Odds ratio: 1.3
Pr(Y=1|X=1) H0: 0.2
α err prob: 0.05
Power (1−β err prob): 0.95
R² other X: 0
X distribution: Normal
X parm μ: 0
X parm σ: 1
Output Parameters
Critical z: ?
Total sample size: ?
Actual power: ?

Figure 5. Probability of error and power values

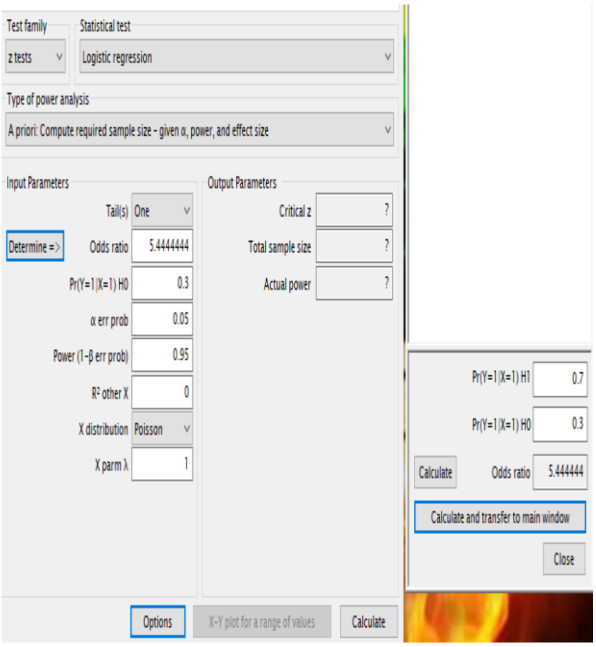
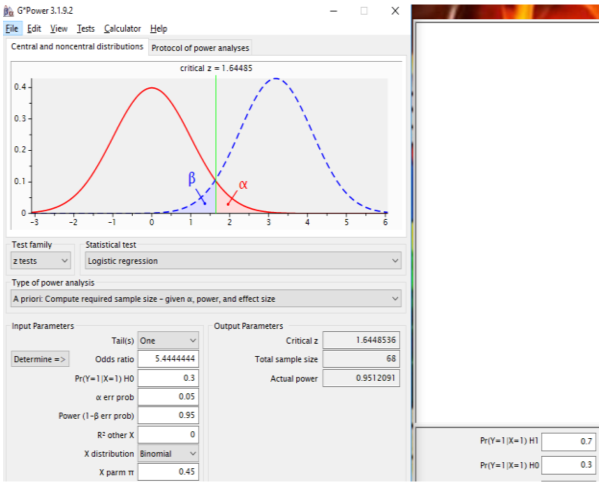

Test family: z tests
Statistical test: Logistic regression
Type of power analysis: A priori: Compute required sample size - given α, power, and effect size
Input Parameters
Tail(s): One
Odds ratio: 5.444444
Pr(Y=1|X=1) H0: 0.3
α err prob: 0.05
Power (1-β err prob): 0.95
R² other X: 0
X distribution: Poisson
X parm λ: 1
Output Parameters
Critical z: ?
Total sample size: ?
Actual power: ?
Pr(Y=1|X=1) H1: 0.7
Pr(Y=1|X=1) H0: 0.3
Odds ratio: 5.444444
Determine =>
Options
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

Figure 6. Choose Determine and write next window H_0 and H_1values

Enter the *X* distribution binominal.



Test family: z tests
Statistical test: Logistic regression
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size
Input Parameters
Tail(s): One
Odds ratio: 5.4444444
Pr(Y=1|X=1) H0: 0.3
α err prob: 0.05
Power (1−β err prob): 0.95
R² other X: 0
X distribution: Binomial
X parm π: 0.45
Output Parameters
Critical z: 1.6448536
Total sample size: 68
Actual power: 0.9512091
Pr(Y=1|X=1) H1: 0.7
Pr(Y=1|X=1) H0: 0.3

Figure 7. Write next window σ,1-β and x- parm values.

Specified and calculated σ, 1-β and x- parm values are written in the pop-up window. After than calculate.

3. Conclusion

In this study, how to calculate the sample size with g-power in logistic regression analysis is explained. As a result, a sample size of 68 in numerical sample and the probability of correctly rejecting the H0 hypothesis indicating that there is no relation between the main predictor variable and result variable is 95%.

The limit of the work is more than one argument. The limitation of sample size calculation with G * power is that there are multiple independent variables in multiple logistic regression analysis and no separate sample size can be calculated for each. So experts advise to calculate the power for the most striking variable.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgements

This research was presented as an oral presentation at the International Congress on Domestic Animal Breeding Genetics and Husbandry (ICABGEH-2018) held on 26-28 September 2018 in Antalya.

References

Buse A. 1982. The likelihood ratio, wald and lagran ge multiplier tests: an expository note. American Stat, 36(3): 1.

Demidenko E. 2007. Sample size determination for logistic regression revisited. Statist Med, 26: 3385-3397.

Erdfelder E, Faul F, Buchner A. 1996. Gpower: A general power analysis. Behav Res Met Instrumen Comput, 28: 1-11.

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

Herrera AN, Gomez J. 2008. Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. Qual Quan, 42: 739–755

Hsieh FY, Bloch DA, Larsen MD. 1998. A simple method of sample size calculation for linear and logistic regression. Statis Med, 17: 1623-1634.

Lyles RH, Lin HM, Williamson JM. 2007. A practial approach to computing power for generalized linear models with nominal, count, or ordinal responses. Statis Med, 26: 1632-1648.

Nemes S, Jonasson JM, Genel A, Steineck G. 2009. Bias in odds ratios by logistic regression modelling and sample size. BMC Medical Res Methodol, 9(56): 1-5.

Moineddin R, Matheson FI, Glazier RH. 2007. A simulation study of sample size for multilevel logistic regression models. BMC Medical Res Methodol, 4(34): 1-10.

Şahin M. 1999. Lojistik Regresyon ve Biyolojik Alanlarda Kullanımı. Yüksek Lisans Tezi. Kahramanmaraş Sütçü İmam Üniversitesi Fen Bilimleri Enstitüsü Zootekni Ana Bilim Dalı. Kahramanmaraş.

Whittemore AS. 1981. Sample size for logistic regression with small response probabilities. JASA, 76: 27-32.