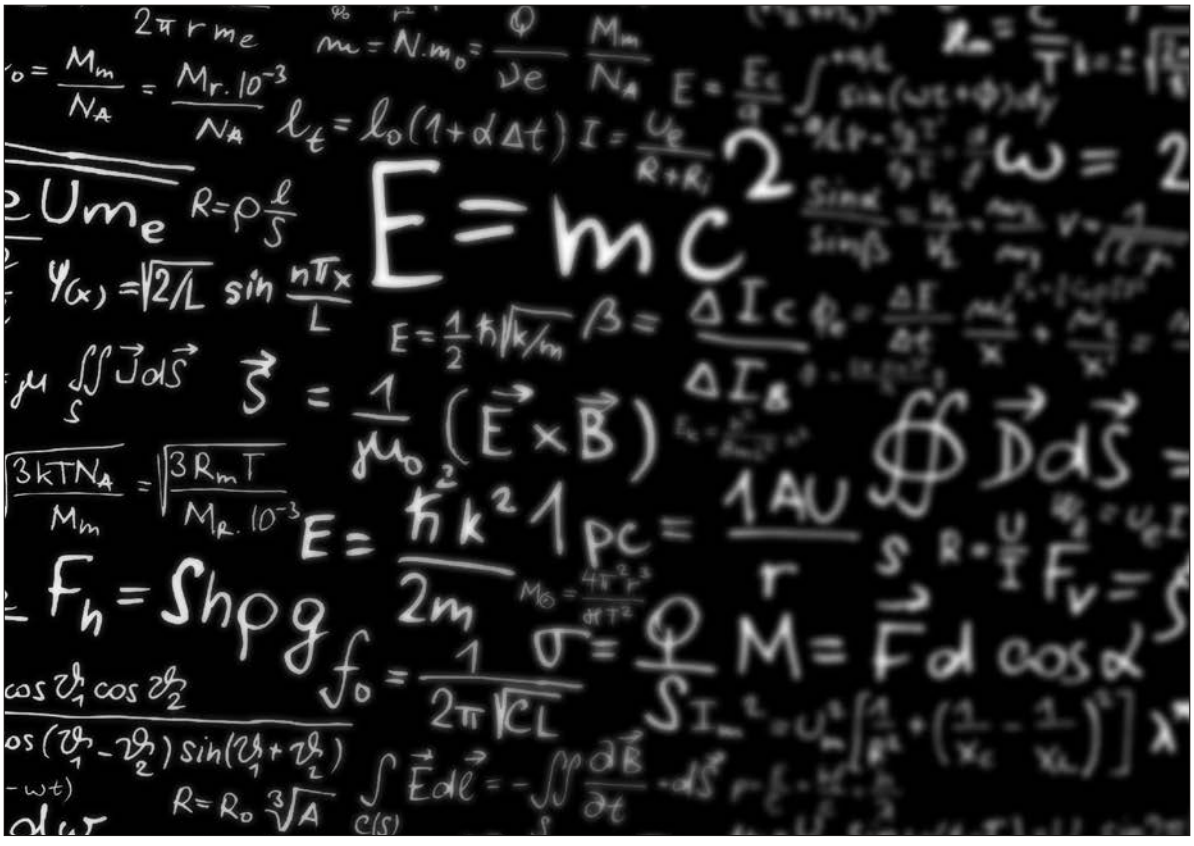

Asian Journal of Mathematics & Statistics

ISSN 1994-5418

science alert

ANSInet
an open access publisher
http://ansinet.com

Asian Journal of Mathematics and Statistics 3 (3): 139-146, 2010
ISSN 1994-5418

On the Estimation of Power and Sample Size in Test of Independence

[1]G.M. Oyeyemi, [1]A.A. Adewara, [2]F.B. Adebola and [3]S.I. Salau
[1]Department of Statistics, University of Ilorin, Ilorin, Nigeria
[2]Department of Mathematics,
Federal University of Technology Akure, Akure, Nigeria
[3]Misa Statistical Consulting and Education Management, Zaria, Nigeria

Abstract: In this study, power and sample size estimations in the context of test of independent between categorical variables were examined. The required sample size in an experiment is a function of the alternative hypothesis, the size of type I error and the variability of the population. Power of a test is the probability of rejecting a false null hypothesis and it depends on the effect size, type I error and the sample size. A priori power analysis is determination of minimum sample size to obtain a required power while post-hoc power analysis is calculating power of a test. A test with small effect size requires large sample size to achieve a power 80% or more while effect size of medium or large size needs small sample size to achieve that. Test with small degrees of freedom will attain higher power than the same test with larger degrees of freedom.

Key words: Effect size, contingency table, power, sample size, type I error

INTRODUCTION

Power analysis and sample size estimation are aspects of the design of experiments and other research studies in which data are collected. Determining the appropriate sample size for an investigation, whether it is clinical trial or field experiments, is essential step in the statistical design of the project (Cohen, 1988; Murphy and Myors, 1999). An adequate sample size helps ensure that the study will yield reliable information, regardless of whether the ultimate data suggest a clinically important difference between the treatments being studied, or the study is intended to measure the accuracy of a diagnostic test or the incidence of a disease (Foster, 2001; Nemec, 1991; Di Stefano, 2001). Generally, the researchers choose a sample size large enough to enhance chances of conclusive results while small enough to lower the study cost, constrained by limited budget and/or some medical consideration. The required sample size in an experiment (test) is a function of the alternative hypothesis, the probabilities of type I and type II errors and the variability of the population(s) under study (Kramer and Rosenthal, 1999).

The probabilities of type I and type II errors are always predetermined prior to the test. Type I error is also the level of significance of the test and it is the probability of rejecting a true null hypothesis. Type II error is the size of probability of accepting a false null hypothesis (Chow *et al*., 2003). The power of a test is therefore 1-β (β is size of type II error). Power of a test is the probability of rejecting a false null hypothesis and it depends on the effect size (which is defined by the alternative hypothesis), type I error rate and the sample size (Roger, 2000). In fact, considering these three parameters and the power of a test

Corresponding Author: G.M. Oyeyemi, Department of Statistics, University of Ilorin, Ilorin, Nigeria

together, fixing any three will allow the determination of the fourth. For example, once we define the effect size, type I error rate and the desired power, this definitely determines the required sample size. Similarly, if the effect size, type I error rate and the sample size are defined, then the power of the test is determined.

In any method for deriving a conclusion from experimental data carries with it some risk of drawing a false conclusion. There are two types of false conclusions that can be committed and they are known as type I error and type II error (Huck, 2000). A type I error occurs when one concludes that a difference exist between the treatment groups when, in fact, it does not. It is type of false positive. The risk of type I error, assuming that there is really no difference between groups is equal to α. A type II error occurs when one concludes that a difference does not exist between groups being compared when, a difference does exist. A type II error is a type of false negative. The risk of a type II error occurring is denoted by β. In a classical hypothesis of H_0 (null hypothesis) against the H_1 (alternative hypothesis), there are four possible outcomes, two of which are incorrect:

- Accept H_0 when H_0 is true
- Reject H_0 when H_0 is false
- Reject H_0 when H_0 is true (type I error)
- Accept H_0 when H_0 is false (type II error)

To construct a test, the distribution of the test statistic under H_0 is used to find the critical region which will ensure that the probability of committing a type I error does not exceed some predetermined level. This probability is typically denoted by α. The power of the test is its ability to correctly reject the null hypothesis, which is based on the distribution of the test under H_1. The required sample size will be a function of:

- The effect size (alternative hypothesis)
- The size of type I error
- The desired power to detect H_1

Current available methods for power analysis include paired and pooled t-test, fixed effect ANOVA and regression models, binomial proportion comparison, bioequivalence, correlation and simple survival analysis models and even in multivariate analysis (Oyeyemi, 2007). Numerous mathematical formulae have been developed to calculate sample size for various scenarios in different researches based on objectives, designs, data analysis methods, power, type I and type II errors and effect size (Chow *et al*., 2003).

Contingency Table

A contingency table is a cross classification of two or more categorical variables, the simplest being a 2×2 contingency table. The contingency table test is a common method of analyzing categorical data. One of its applications is to test whether two or more categorical variables are independent of one another (Lebart *et al*., 2000; Clausen, 1998). Suppose X and Y are 2 categorical variables with r and c categories, respectively. If o_{ij} is the observed count/frequency in the cell ij (i = 1, 2, 3, , , r; j = 1, 2, 3, , , c). Then $n_{i\bullet}$ is the marginal total for the ith row and $n_{\bullet j}$ is the marginal total for the jth column. Then the sample size n is given as:

$$\sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j} = n$$

The expected count/frequency e_{ij} of cell ij is the obtained as:

$$e_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

In testing, the hypothesis of independence between variables X and Y at α level of significance, a test statistic Q is obtained as follows:

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

If the null hypothesis is true, the test statistic follows a chi-square distribution with (r-1)(c-1) degrees of freedom. The hypothesis of independence is rejected if:

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} > \chi^2_{(r-1)(c-1);\alpha}$$

where, $\chi^2_{(r-1)(c-1);\alpha}$ is the critical value of the χ^2 distribution at a significance level α with (r-1)(c-1) degrees of freedom. If the null hypothesis is not true, Q has the limiting non-central χ^2 distribution, with the non-centrality parameter λ and (r-1)(c-1) degrees of freedom.

In general, the following is valid for the non-centrality parameter λ (Lachin, 1977):

$$\lambda = nf(\theta^0, \theta^1) \tag{1}$$

where, n is the sample size and f is the function of the vectors of parameters θ^0 and θ^1, which are involved in the test statistic Q under the H_0 and H_1, respectively. From different perspective, f can be considered as the observed result's degree of deviation from the condition stated through H_0 and therefore is a function of the statistical test's corresponding effect size. From Eq. 1, we can show that:

$$n = \frac{\lambda}{f(\theta^0, \theta^1)} \tag{2}$$

Therefore, if the parameter λ and its corresponding effect size are estimated, then Eq. 2 can be used to calculate the minimum sample size required, at a significance level α and power, for the chi-square test of independence.

Power Analysis

Generally Power Analysis is Used to Determine

- The minimum sample size n required to implement statistical test to detect an effect as statistically significant at a significance level α and power
- The power of a statistical test, given the sample size, the level of significance and the observed effect size

The first task is known as a priori approach to power analysis while the second is the post-hoc approach to power analysis. Therefore post-hoc power analysis of a statistical test obtains the power while a priori power analysis determines the sample size required to detect a true significant effect for a test.

Using the type II error which is estimated as follows:

$$\beta = P\left(Q < \chi^2_{(r-1)(c-1),\alpha} / H_0 \cdot \text{true}\right) = P\left(\chi^2_{nc(r-1)(c-1)}(\lambda) < \chi^2_{(r-1)(c-1),\alpha}\right)$$

where, $\chi^2_{nc(r-1)(c-1)}(\lambda)$ is the value of the no-central χ^2distribution with parameter λ and (r-1)(c-1) degrees of freedom. The power of the Chi-square test is then:

$$\text{power} = 1 - \beta = P\left(\chi^2_{nc(r-1)(c-1)}(\lambda) \geq \chi^2_{(r-1)(c-1),\alpha}\right)$$

In order to estimate the power, it is necessary to have an estimate of the parameter λ. According to Cohen (1988):

$$\lambda = nw^2 \tag{3}$$

where, n is the sample size and w is an estimate of the effect size given as:

$$w = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j}}$$

It can be easily shown that:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}} = \frac{Q}{n}$$

Therefore, $w^2 = Q/n$, this implies that the non-centrality parameter λ. We can then obtain the power as:

$$\text{power} = P\left(\chi^2_{nc(r-1)(c-1)}(Q) \geq \chi^2_{(r-1)(c-1),\alpha}\right)$$

Effect Size

It is possible to make an a priori calculation of the minimum sample size required, when an estimate of the non-centrality parameter ë and effect size are given. Using Eq. 3 $n = \lambda/w^2$. For the χ^2 distribution, the values of non-centrality $\lambda(\alpha, \beta, df)$ that correspond to significance level α, power $(1-\beta)$ with degrees of freedom df, can be found in tables (Haynam *et al*., 1970; Pearson and Hartley, 1972) or can be calculated using relevant software. The only problem lies in providing a predetermined estimate of effect size that is significance within the framework of the hypothesis.

The determination of effect size could be achieved either through pilot research project or from previous related studies on the same research subject. Cohen's convention can also be used in relation to what can be considered as a small, medium, or large effect size within the framework of Pearson Chi-square test of independence (Table 1).

Table 1: Cohen's convention of classification of effect size

Effect size	Classification
Small	w = 0.10
Medium	1.0<w≤0.30
Large	0.30<w≤0.50

Table 2: Epidemiological data on 535 children

	Race		
Risk	Black	White	Others
At risk	185	140	90
Not at risk	80	17	23

Power and Sample Size Determination

Table 2 shows the epidemiological data on 535 children as contained in Nelson *et al.* (2005). The children were cross-classified according to their race {Black, White and Others} and risk of becoming obesity. Based on Table 2, we want to test whether race of the children is independent of being at risk of becoming obesity or not.

The test statistic:

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(o_{ij} - e_{ij}\right)^2}{e_{ij}}$$

where, Q is chi-square distributed with 2 degrees of freedom. The above test can be performed using R-language as follows:

```
x = matrix (c (185, 80, 140, 17, 90, 23), ncol=3)

chisq.test (x, correct = FALSE)
```

The following summary statistics were obtained as:

$$Q = 21.5947, \quad df = 2, \quad p\text{-value} = 0.000, \quad n = 535$$

With p-value of 0.000, we can therefore conclude that there exist relationship between the race of a child and risk of becoming/developing obesity at 0.05 level of significance. The reliability of the above conclusion can be verified with computation of the power of the test through power analysis. As discussed earlier, the power of a chi-square test depends on its non-centrality parameter Q. The power of a test is the right tail probability under the alternative hypothesis characterized by Q (Bergerud and Sit, 1992). The power can be obtained with the same R-language as follows:

```
c = qchisq(1-α, df)

power = 1-chisq(c, df, Q)
```

At $\alpha = 0.05$, the power of the above test is 0.9905. The interpretation is that, if race and risk of becoming obesity were indeed related to the extent suggested by the data in Table 2, the test would be able to detect that 99.05% of the time. The high power so obtained makes the conclusion to be more reliable.

Table 3: Computed power values for different sample (n) and effect sizes (w) and when degrees of freedom, df = 2

	Power		
Sample size (n)	w = 0.10	w = 0.2009	w = 0.30
100	0.1327	0.4187	0.7706
150	0.1783	0.5881	0.9186
200	0.2255	0.7217	0.9745
250	0.2735	0.8191	0.9927
300	0.3215	0.8861	0.9988
350	0.3690	0.9302	0.9995
400	0.4154	0.9583	0.9999
450	0.4604	0.9756	0.9999
500	0.5037	0.9859	1.0000

Table 4: Modified epidemiological data on 535 children presented in Table 2

	Race	
Risk	Black	Non-black
At risk	185	230
Not at risk	80	40

The high value of power in the test is as a result of high value of the non-centrality parameter (Q) and sample size (n). Using the Cohen's classification of effect size, which is a function of non-centrality parameter and sample size. The above test gives effect size w = 0.2009 which is classified as medium according to Cohen (1988). For this effect size, power was calculated for different sample sizes of 100, 150, 200, 250, 300, 350, 400, 450 and 500. Likewise, for the same set of sample sizes, the effect sizes of 0.10 (small effect) and 0.30 (large effect) were used to obtain the power values and the results were presented in Table 3.

The degrees of freedom for chi-square test for the data in Table 2 was modified by collapsing two categories (white and others) as non-black and the modified table is presented in Table 4. The same hypothesis is tested and the test statistic and the p-value were obtained.

$$Q = 18.1677, \quad df = 1, \quad p\text{-value} = 0.000, \quad n = 535$$

The same conclusion of relationship between race and risk of becoming obesity is established. At $\alpha = 0.05$, the power of the test is 0.9893 though with effect size w = 0.1843. Table 5 gives the computed power for different sample sizes for this effect size and when effect size is small (0.10) and large (0.30).

RESULTS AND DISCUSSION

The power of a test increases as the sample size increases irrespective of the non-centrality parameter value or effect size as shown in Tables 3 and 5 for the 2×3 and 2×2 contingency tables respectively. From Table 3 with degrees of freedom of 2, for small effect size, sample size of more than 500 is required to obtain power of 0.80 while sample sizes of 250 and 150 are required to attain the same power for the same test with medium and large effect sizes, respectively.

When the degree of freedom is 1 as shown in Table 5, the test attains higher power than when the degree of freedom is 2. For instance, when for large effect size of 0.30 when the degrees of freedom is 1, sample size 100 gave power value of 0.8508 while the same sample size gave power value of 0.7706 when the degrees of freedom is 2. Also for small effect size (w = 0.10), when the degrees of freedom is 2, the power value is 0.5037 for sample size of 500 while the same sample size gave 0.6088 when the degrees of freedom is 1.

Table 5: Computed power values for different sample (n) and effect sizes (w) when degree of freedom, df = 1

	Power		
Sample size (n)	w = 0.10	w = 0.1843	w = 0.30
100	0.1701	0.4534	0.8508
150	0.2318	0.6168	0.9568
200	0.2930	0.7409	0.9888
250	0.3526	0.8299	0.9973
300	0.4100	0.8910	0.9994
350	0.4646	0.9316	0.9999
400	0.5160	0.9578	0.9999
450	0.5641	0.9743	0.1000
500	0.6088	0.9846	1.0000

CONCLUSION

In test of independence between two categorical variables, apart from the size of non-centrality parameter (effect size) which determines the power and sample size of the test, the number of categories of the variables also affect the sample size and the power of the test.

REFERENCES

Bergerud, W. and V. Sit, 1992. Power analysis workshop notes. Ministry of Forests Research Program.

Chow, S., J. Shao and H. Wang, 2003. Sample Size Calculation in Clinical Research. Marcel Dekker Inc., USA.

Clausen, S.E., 1998. Applied Correspondence Analysis: An Introduction. Sage Publication Inc., Oakes, CA.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences. 2nd Edn., Lawrence Erlbaum, Hillsdale, New Jersey, ISBN: 0-8058-6283-5, pp: 128.

Di Stefano, J., 2001. Power analysis and sustainable forest management. For. Ecol. Manag., 154: 141-153.

Foster, J.R., 2001. Statistical power in forest monitoring. For. Ecol. Manage., 151: 211-222.

Haynam, G.E., Z. Govindarajulu and F.C. Leone, 1970. Tables of the Cumulative Non-central Chi-Square Distribution. In: Selected Tables in Mathematical Statistics, Harter, H.L. and D.B. Owen (Eds.). Vol. 1, Markham Publishing Co., Chicago.

Huck, S., 2000. Reading Statistics and Research. Addison Wesley Longman Inc., New York.

Kramer, S.H. and R. Rosenthal, 1999. Effect Sizes and Significance Levels in Small-Sample Research. In: Statistical Strategies for Small Sample Research, Hoyle, R. (Ed.). Sage Publications Inc., Oakes, California.

Lachin, J., 1977. Sample size determinations for rxc comparative trials. Biometrics, 33: 315-324.

Lebart, L., A. Morineau and M. Piron, 2000. Statistique Exploratoire Multidimensionelle. Dunod Publisher, Paris.

Murphy, K.R. and B. Myors, 1999. Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. J. Applied Psychol., 84: 234-248.

Nelson, P.R., P.S. Wludka and A.F. Copeland, 2005. The Analysis of Means: A Graphical Method of Comparing Means, Rates and Proportions. SIAM Publisher, Philadelphia, pp: 179.

Nemec, A., 1991. Power analysis handbook for the design and analysis of forestry trials. Biometrics Handbook No. 2. B.C. Min. For., Res. Br., Victoria, B.C.

Oyeyemi, G.M., 2007. Computing power and sample size for hotelling T test. Global J. Math. Sci., 6: 71-73.

Pearson, E.S. and H.O. Hartley, 1972. Biometrika Tables for Statisticians. Vol. 2, Cambridge University Press, Cambridge.

Roger, J.L., 2000. Power analysis and sample size determination: Concepts and software tools. Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine (SAEM), San Francisco, California.