

Examining the Impact of Sample and Effect Sizes on the Power of One-Way Analysis of Variance F-Test

G.M. Oyeyemi and B.L. Adeleke

Department of Statistics, University of Ilorin, Nigeria
E-mail: gmoyeyemi@gmail.com, matanmi@unilorin.edu.ng
Department of Statistics, University of Ilorin, Nigeria
E-mail: lbadeleke@yahoo.com, bladeleke@unilorin.edu.ng

Abstract

The power of a test is a function of three factors; sample size, the size of detectable effects and the level of significance. The paper focused on determining the non-centrality parameter in one-way analysis of variance (ANOVA) test which by implication utilizes the completely randomized design. To examine the contributing effect of sample size and size of detectable effects, data were simulated for varying levels of treatments, replicates and effect sizes for several experiments of the completely randomized design type. The results showed that power of the F test increased monotonically with size of non-centrality parameter λ , where non-centrality parameter is a function effect size. An increase in sample size also brings proportionate increases in power irrespective of the number of treatments under consideration.

Keywords: Non-centrality parameter, Power, Effect size, Replicate, Treatment

Introduction

Experimental design is a planned interference in the natural order of events by the researchers, Mason et al (2003). The importance of experimental design stems from the quest of inference about causes or relationship as opposed to simply description. Researchers are rarely satisfied to simply describe the events they observe, rather they want to make inferences about what produced, contributed to, or caused events. To gain such information without ambiguity, some forms of experimental designs depending on the nature of the available experimental materials. Design of experiment was first developed in 1930, refers to the process of planning the experiment,

collecting data, analyzing the data using different methods and drawing conclusions about the system being analyzed (Montgomery, 2005). It can also be defined as the design of all information-gathering exercises where variation is present, whether under full control of the experimenter or not. Design of experiment is a structured, organized method that is used to determine the relationship between the different factors affecting a process and the output of that process. The importance of experimental design includes but not limited to the following;

1. To investigate a phenomenon in order to understand or improve performance.
2. It is widely used in research and development, where a large proportion of the resources go towards solving optimization problems.
3. To identify optimal conditions, the factors that most influence the expected outcome and those that do not, and some other details.
4. It also requires only a small set of experiments and thus helps to reduce costs.

Building a design means carefully choosing a number of experiments that are to be performed under controlled conditions. The basic principles of experimental design include randomization, replication and blocking (see e.g. Williams and Gertrude, 1957). Randomization is the process by which experimental units are allocated to treatments by random process and not by any element of subjectivity. Treatments are allocated to units in such a way that each treatment is equally likely to be applied to each unit. Randomization is an important tool in carrying out experiment because it averages out the effect of all other lurking variables (Williams, 1949).

Replication is an important tool in carrying out an experiment because it makes it possible to assess the mean square error required for testing the presence of treatment effect. It therefore gives an estimate of natural variation required to know if any observed difference can be attributed to the factor or just random change. It also enables the experimenter to have sufficient power to detect an effect of a meaningful size.

Blocking usually refers to categories of subjects with a treatment group. It is the arrangement of experimental units into homogenous groups (blocks) that are similar to one another. It reduces known but irrelevant sources of variation between units and thus allows greater precision in the estimation of the source of variation under study, Kempthorne (1952).

Analysis of Variance (Anova)

Analysis of variance (ANOVA) is a collection of statistical model and their associated procedures in which the observed variance is partitioned into components due to different explanatory variables (Kirk, 1995). It is also a technique for testing the hypothesis that sample means of three or more groups are derived from the same population. ANOVA presents a way to test for significant differences among sample means when the independent(predictor) variable is a set of discrete categories, and the dependent variables is continuous. The simplest version of experimental design which is the complete randomized design (CRD) is adequately represented by the one-way ANOVA model. Indeed one-way ANOVA model is used when there exists a single

systematic source of variation arising from differences among different groups. The one-way ANOVA model is given in equation 2.1

$$y_{ij} = \mu + \alpha_j + \ell_{ij} \quad (2.1)$$

Where y_{ij} is the observation i arising from treatment j

μ is the overall mean

α_j is the effect of treatment j

ℓ_{ij} is the error term which is assumed to be identical and normally distributed with mean of 0 and σ^2 . Where $i = 1, 2, 3, \dots, r$; and $j = 1, 2, 3, \dots, t$; $\alpha_j = \mu_j - \mu$. The appropriate null hypothesis (H_0) and alternative hypothesis (H_1) can be expressed symbolically as follows;

$H_0 : \alpha_j = 0$ against

$H_1 : \alpha_j \neq 0$ for at least one $j, j = 1, 2, 3, \dots, t$.

When the null hypothesis is true, the F statistic: $F = \text{MST}/\text{MSE}$ has an F distribution with $t-1$ and $t(r-1)$ degrees of freedom. When the null hypothesis is false, the F statistic follows a non-central F distribution with $t-1$ and $t(r-1)$ degrees of

freedom, and the non-centrality parameter $\lambda = t \sum_{j=1}^t \left(\frac{\alpha_j}{\hat{\sigma}} \right)^2$

The Non-Centrality Parameter

A non-centrality parameter of a non-central F distribution is a precision weighted sum of squares of treatment means, which is closely related to effect size, Cohen (1977). The F-distribution has two parameters namely, the numerator and the denominator degrees of freedom. Whereas the non-central F-distribution has the two aforementioned plus a non-centrality parameter. However, the non-central F-distribution is identical to F-distribution, so it includes F-distribution as a special case. What makes the non-centrality parameter especially important is that it is closely related to the truth or falsity of the null hypothesis that the distribution is typically used to test. The larger the absolute value of the non-centrality parameter for a given alpha and sample size the higher the likelihood of rejecting the null hypothesis (Cohen, 1988). The non-centrality parameter of the F-distribution is usually denoted by lambda (λ), when $\lambda=0$, the non-central F-distribution becomes standard F-distribution. The non-centrality parameter for the F-statistics is given as;

$$\lambda = t \sum_{j=1}^t \left(\frac{\alpha_j}{\hat{\sigma}} \right)^2,$$

where $\alpha_j = \mu_j - \mu, j = 1, 2, 3, \dots, t$ and $\hat{\sigma}$ is the mean square error. The power of the F test is a monotonically increasing function of the parameter λ .

Power of a Test

The power of a statistical test is the ability to detect an effect if there is one. It is the probability that the test will reject a false null hypothesis ($1-\beta$). As power increases, the size of type II error also decreases. Power analysis can either be done before (a priori) or after (post hoc) data is collected. A priori power analysis is conducted prior to the research study, and it is typically used to determine an appropriate sample size to achieve adequate power. Post hoc power analysis is conducted after a study has been completed, and uses the obtained sample size and effect size to determine what the power was in the study, assuming the effect size in the sample size is equal to the effect size in the population, Cohen and Nee (1987).

Power of a test depends on the variability of the data, the statistical significance criterion in the test and the size of the difference or the strength of the similarity (i.e. the effect size) in the population. Power of a test is essential to an experimental study in assessing a potentially effective treatment producing evidence about treatment effects. Power analysis is appropriate when the concern is with the correct acceptance or rejection of a null hypothesis, Hallahan and Rosenthal (1990). Power depends in part on effect size, because the smaller an effect the more observations you need to establish its existence. In calculating power, it requires first specifying the effect size that hypothesis expects to detect, the greater the effect size the greater the power. Power analysis is also used when the concern is with the correct acceptance or rejection of a null hypothesis (Tang, 1938).

Power computation and discussion of results

Power computation

Data were simulated, using R.2.6.1 package, to confirm the contribution of effect size, number of treatments and sample sizes on power of the ANOVA test. A completely randomized design model was used in the simulation and each simulation was repeated 1000 times to attain stability in the computation of non-centrality parameter and the power. The differences in means were used to measure effect size and the absolute difference in means considered in respect of standard error ($s_{\bar{x}}$) are; no difference, $2s_{\bar{x}}$, $5s_{\bar{x}}$ and $10s_{\bar{x}}$. Number of treatments considered are; 3, 4 and 5 respectively while different levels of replicate considered are; 3, 5, 8, 10, 12, 15 and 20. The results of the power computation for the treatment t , t in the range $3 \leq t \leq 5$ are presented in Tables 1, 2 and 3 respectively. The multiple line charts for number of treatments and difference in means were shown in Figures 1 and 2 respectively.

Table 1: Lambda and Power of varying effect sizes and replicates for 3 treatments

replicate	effect	lambda	power
3	0	2.9807	0.1932
5	0	2.4431	0.2181
8	0	2.1385	0.2232
10	0	2.1244	0.2080
12	0	2.1481	0.2201
15	0	2.1182	0.2076
20	0	2.0242	0.2191
3	2	4.0769	0.2523
5	2	4.1557	0.3094
8	2	4.9512	0.3916
10	2	5.5615	0.4466
12	2	6.1979	0.4860
15	2	6.9566	0.5457
20	2	8.6182	0.6238
3	5	12.0610	0.5040
5	5	14.9359	0.7102
8	5	19.4269	0.8787
10	5	24.5637	0.9327
12	5	27.2516	0.9621
15	5	33.4832	0.9849
20	5	44.0660	0.9968
3	10	40.4960	0.8811
5	10	49.6185	0.9870
8	10	74.1762	0.9997
10	10	87.5392	0.9999
12	10	104.5910	0.9999
15	10	128.1365	1.0000
20	10	167.2159	1.0000

Table 2: Lambda and Power of varying effect sizes and replicates for 4 treatments

replicate	effect	lambda	power
3	0	3.8889	0.2275
5	0	3.1868	0.2429
8	0	3.2657	0.2566
10	0	3.2332	0.2607
12	0	3.1706	0.2594
15	0	3.1306	0.2581
20	0	3.1762	0.2712
3	2	7.0109	0.3476

5	2	7.8102	0.4623
8	2	10.0153	0.6037
10	2	11.5666	0.6850
12	2	13.1449	0.7327
15	2	15.3927	0.8169
20	2	19.7521	0.8900
3	5	24.6798	0.7523
5	5	30.9456	0.9327
8	5	46.8416	0.9934
10	5	55.3187	0.9983
12	5	65.1685	0.9997
15	5	82.0896	0.9999
20	5	106.0168	1.0000
3	10	81.0274	0.9911
5	10	117.6800	0.9999
8	10	175.7369	1.0000
10	10	216.3266	1.0000
12	10	256.1862	1.0000
15	10	313.0088	1.0000
20	10	411.4553	1.0000

Table 3: Lambda and Power of varying effect sizes and replicates for 5 treatments

replicate	effect	lambda	power
3	0	4.8567	0.2550
5	0	4.6430	0.2772
8	0	4.1991	0.2872
10	0	4.3855	0.2898
12	0	4.2422	0.3007
15	0	4.1305	0.3124
20	0	4.2006	0.3081
3	2	11.5334	0.4752
5	2	13.3196	0.6405
8	2	17.0519	0.8113
10	2	21.0210	0.8669
12	2	23.7969	0.9190
15	2	28.3456	0.9596
20	2	36.0758	0.9876
3	5	42.1398	0.9214
5	5	60.1514	0.9949
8	5	90.1591	0.9991
10	5	108.4511	0.9999
12	5	128.3439	1.0000

15	5	156.8831	1.0000
20	5	207.9727	1.0000
3	10	157.8004	0.9998
5	10	237.1002	1.0000
8	10	363.8315	1.0000
10	10	438.2319	1.0000
12	10	522.0606	1.0000
15	10	651.6195	1.0000
20	10	861.5315	1.0000

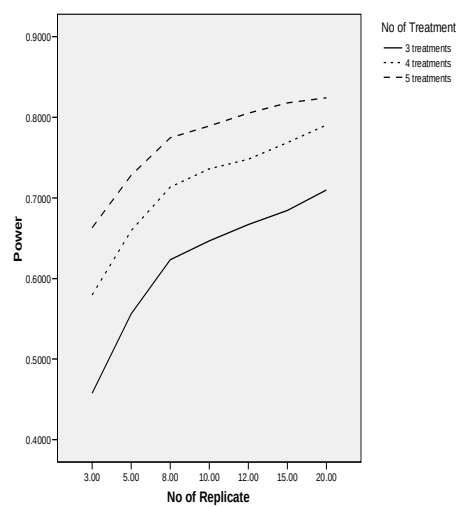


Figure 1: The computed power of the three different treatment levels for various replicates

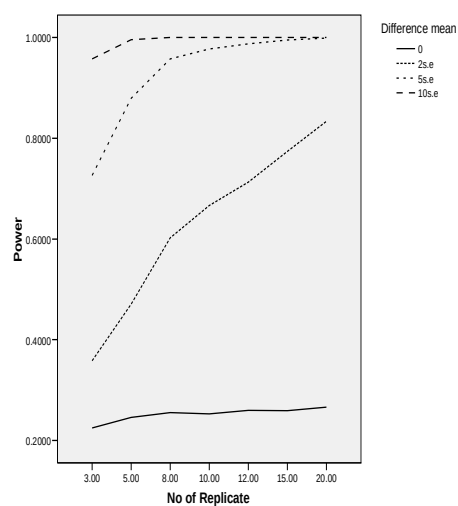


Figure 2: The computed power of the detectable effect sizes for various replicates

Discussion of Results

The results of the analysis of the simulated data for treatments 3, 4 and 5 are presented in Tables 1, 2 and 3 respectively. While the multiple line-charts for different treatment levels and detectable effect sizes are shown in Figures 1 and 2 respectively. From the Tables, irrespective of the number of treatments, the power of the F test increases with number of replicates (sample sizes). Likewise, increase in detectable effect size brings proportionate increase in power. Though an increase in sample size brings about increase in power, this is only valid if actually there is any detectable effect size in the experiment (see figure 2). However, the sample size does not have significant effect on the power if there is actually no effect to detect.

References

- [1] Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). Hillsdale, NJ: Erlbaum.
- [2] Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd.. ed.). Hillsdale, NJ: Erlbaum.
- [3] Cohen, J. & Nee, J. C. (1987). A comparison of two non-central F approximations, with applications to power analysis in set correlation. *Multivariate Behavioral Research*, 22, 483- 490.
- [4] Hallhan, M. & Rosenthal, R. (1996). Statistical power: concepts, procedures and applications. *Behavioral Research Therapy*, Pubmed, 34(6), 489 – 499.
- [5] Kirk, R. E. (1995). *Experimental Design: Procedures for the Behavioral Sciences (Psychology)*. 3rd. edition, Wadsworth publishing. ISBN: 978-0534250928
- [6] Mason, R. L., Gunst, R. F., & Hess, J. L. (2003). *Statistical design and analysis of experiments; with application to engineering and sciences*, (2nd ed.), John Wiley & Sons, Inc. Hoboken, New Jersey.
- [7] Montgomery, D. C. (2005). *Designs and analysis of experiments*. John Wiley, New York, NY.
- [8] Kempthorne, O. (1952). *The Design and analysis of experiments*. John Wiley, New York, NY.
- [9] Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power studies? *Psychological Bulletin*, 105(2), 309 – 316.
- [10] Tang, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their uses. *Statistical Research Memoirs*, 2, 126-149.
- [11] Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effect treatment. *Australian Journal of Scientific Research*, 2, 149 – 168.
- [12] Williams, G. C. & Gertrude, M. C. (1957). *Experimental designs*. John Wiley & Sons. New York.