Economical Model of Selecting an Appropriate Sample Size of Experiments

Ful-Chiang Wu, Chi-Hao Yeh, and Cheng-Hsiung Chen

***Abstract*—The orthogonal array and sample size determine the experimental cost. Selection of an appropriate sample size is one the most important aspects of any experimental design problem. The choice of sample size and the probabilities of type I error α and type II error β are closely connected in traditional experimental design. We can specify the α and β, and then decide the standard deviation σ based on the prior experience to calculate the ratio, Δ, of absolute value for difference in means dividing by standard deviation. The required sample size *n* is obtained from operating characteristic curve. The purpose of this paper is to present economical models for selecting an appropriate sample size based on the testing cost of observations and loss by adopting a wrong level and probability of factor. Some tables are provided that can be used to select sample size for the variances of two populations are known or unknown.**

***Index Terms*—Economical model, sample size, type I error α, type II error β.**

I. Introduction

The Taguchi method has been widely applied to optimize the industrial parameter design, including static and dynamic problems. Robust design is where a product or process to be optimized has several control factors that directly decide the target or desired value of the output. The optimization involves determining the best control factor levels so that the output is at the target or desired value. The static problem is defined so that the desired output of product or process has a fixed target. In the dynamic problem, the desired output of the system depends on the signal factor setting, that is, the dynamic system is the one without a single target but a response, which is a function of a signal. To conduct the optimization experiments by using the orthogonal array, the experimental cost is decided by the orthogonal array and sample size. Selection of an appropriate sample size is one the most important aspects of any experimental design problem. A statistical hypothesis is a statement concerning one or more populations. Suppose that we are interested in testing the equality of means for the two populations. Thus, the null hypothesis (H_0) and alternative hypothesis (H_1) are stated as

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 \\ &H_1 : \mu_1 \neq \mu_2 \end{aligned} \tag{1}$$

Two kinds of errors may be committed when testing hypotheses. Rejection of the null hypothesis when it is true is called a type I error. Acceptance of the null hypothesis when it is false is called a type II error. The probabilities of these two errors are given special symbols:

$$\begin{aligned} \alpha &= P(\text{type I error}) = P(\text{reject } H_0 | \mu_1 = \mu_2) \\ \beta &= P(\text{type II error}) = P(\text{accept } H_0 | \mu_1 \neq \mu_2) \end{aligned} \tag{2}$$

The choice of sample size and the probabilities of type I error α and type II error β are closely connected in traditional experimental design. If the means are not equal so that $\delta = \mu_1 - \mu_2$. Since the H_0 is not true, the probability β of type II error depends on the true difference in means δ and is also a function of sample size. A graph of β versus δ for a particular sample size is called the operating characteristic curve or O.C. curve for the test [1]. We can specify the α and β, and then decide the standard deviation σ based on the prior experience to calculate the ratio, Δ, of absolute value for difference in means dividing by standard deviation. The required sample size *n* is obtained from operating characteristic curve yielded n^* by $n^* = 2n - 1$.

Several researchers have studied the sample size determination for experiments. Gould [2] provided a procedure for incorporating the uncertainty explicitly into the sample size determination on the basis of joint confidence distributions obtained from the pilot or prior information. Hand, Stamey and Young [3] developed methods for sample-size determination for hypothesis testing in a Bayesian context. Rao [4] derived sample size equations for studies with a continuous exposure. Dunnett, Horn and Vollandt [5] addressed the problem of sample size determination in multiple comparisons of k treatments with a control for step-down and step-up testing, assuming normal data and homogeneous variances. Rahme and Joseph [6] provided a new approach for calculating sample size is developed by combining Bayesian and frequentist idea when a hypothesis test between two binomial proportions is conducted. Thode [7] studied the power and sample size requirements for tests of differences between two Poisson rates. Gordon and Watson [8] studied the sample size

Manuscript received March 5, 2013; revised June 28, 2013. This research was supported by National Science Council of Republic of China under Contract NSC 101-2221-E-238-008.

Ful-Chiang Wu is with the Department of Management and Information Technology, Vanung University, Tao-Yuan, Taiwan (tel.:+886-3-4515811-61647; fax: +886-3-4613715; e-mail: shogo@vnu.edu.tw).

Chi-Hao Yeh is with the Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, Taiwan (e-mail: chyeh@ntut.edu.tw).

Cheng-Hsiung Chen is with the Department of Information Management, Hsing Wu University, New Taipei City, Taiwan (e-mail: 101051@mail.hwu.edu.tw).

DOI: 10.7763/IJIET.2013.V3.336

determination for comparison of small probabilities. Lachin [9] introduced the sample size determination and power analysis for clinical trials. Schulz and Grimes [10] provided a method to calculate the sample size in randomized trials. Lui [11] used the full likelihood for calculating sample sizes. Lui and Cumberland [12] used the Monte Carlo simulation to estimate the corresponding type I error and power with the given resulting sample sizes in a variety of situations. Heilbrun and McGee [13] derived a formula for the required number of subjects in a comparison group to test the equality of two Normal means when one sample size is fixed. Gail [14] described a procedure to calculate the total sample size, 2n, required to attain specified power against the null hypothesis. Tang, Tang and Carey [15] proposed two kinds of approximate sample size formulas, based on rate ratio, for comparison of the marginal and conditional probabilities in a correlated 2×2 table with structural zero. Becker [16] studied the sample size determination in case-control studies. Murthy and Haywood [17] provided an additional consideration when plausible effect sizes and error tolerances on hypothesis tests are balanced against feasibility of accruing various sample sizes. Hale [18] presented the equations for calculating the sample size necessary to determine, within a given probability, the upper and lower confidence limits of the geometric mean of a log normal distribution. Zhang, Cutter and Belin [19] developed a Bayesian approach for calculating sample sizes for clinical trials under the framework of hypothesis tests. Taguchi [20] considered the economical view to determine an appropriate sample size n of experiments. Suppose that factor A has two levels A_1 and A_2. The means of two populations A_1 and A_2 are μ_1 and μ_2, respectively. If μ_1 equals to μ_2, there is no loss whichever level is selected. The loss increases as the sample size increases under this situation. If μ_1 and μ_2 have large difference, we can almost identify which level is better only using one sample. The optimal sample size is one for this situation.

The purpose of this paper is to present the economical models for selecting an appropriate sample size based on loss function. We consider the standard deviation of population is known or not to decide the optimal sample size based on the testing cost of observations and loss by adopting a wrong level and probability of factor.

II. Economical Models for Sample Size

Consider the difference in the means of two populations from normal distributions. Population 1 has mean μ_1 and variance σ_1^2, while population 2 has mean μ_2 and variance σ_2^2. The sample sizes from the two populations are equal; that is, $n_1 = n_2 = n$. $X_{11}, X_{12}, \ldots, X_{1n}$ is a random sample of n observation from population 1, and $X_{21}, X_{22}, \ldots, X_{2n}$ is a random sample of n observation from population 2. The testing cost of observations is denoted as a and the loss is denoted as b for adopting a wrong level of factor when the true difference in means 1σ. Since the probability β for adopting a wrong level of factor depends on the true difference in means. The loss function, $I_n(\Delta)$, for sample size n is expressed as a quadratic model.

$$I_n(\Delta) = 2an + b\Delta^2 \cdot \beta \tag{3}$$

A. Two Population Variances σ_1^2 and σ_2^2 Are Known and Equal

Suppose that random variable **X** presents Let the random variable **X** present the difference in sample means $\overline{X}_1 - \overline{X}_2$ and has mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n + \sigma_2^2/n$. We consider the mean of population is the larger-the-better characteristic and the true difference in means $\Delta\cdot\sigma$ for population 1 is larger than population 2. Under this situation, the loss function is expressed as

$$\begin{aligned} I_n(\Delta) &= 2an + b\Delta^2 \cdot \beta \\ &= 2an + b\Delta^2 \cdot P(\text{adopt population 2}) \\ &= 2an + b\Delta^2 \cdot P_r(X<0) \end{aligned} \tag{4}$$

Let $\lambda = b/a$ and $I_n(\Delta)/a$ substitute for $I_n(\Delta)$, and then equation is modified as

$$\begin{aligned} I_n(\Delta) &= 2n + \lambda\Delta^2 \cdot P_r(X<0) \\ &= 2n + \lambda\Delta^2 \cdot \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sqrt{2/n}\ \sigma} e^{-\frac{1}{2}\left[\frac{x-(\mu_1-\mu_2)}{\sqrt{2/n}\ \sigma}\right]^2} dx \end{aligned} \tag{5}$$

Let $Z = \left(X - (\mu_1 - \mu_2)\right)/\left(\sqrt{2/n}\ \sigma\right)$, and then equation (5) is transformed as

$$I_n(\Delta) = 2n + \lambda\Delta^2 \cdot \int_{-\infty}^{-\sqrt{n/2}\cdot\Delta} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{6}$$

When Δ is known

From equation (6), the optimal sample size for a particular Δ is tabulated in Table I.

TABLE I: The Optimal Sample Size When Δ Is Known

Δ	Optimal sample size		
	λ=100	λ=500	λ=1000
0.0	1	1	1
0.1	1	1	1
0.2	1	1	1
0.3	1	1	3
0.4	1	4	10
0.5	1	8	14
0.6	2	10	15
0.7	3	11	15
0.8	4	11	14
0.9	4	10	13
1.0	5	10	12
1.2	5	8	10
1.4	5	7	9
1.6	4	6	7
1.8	4	6	6
2.0	3	5	6
2.5	3	4	4
3.0	2	3	3
3.5	2	2	3
4.0	2	2	2
4.5	1	2	2
5.0	1	2	2

When Δ is unknown

Similarly, we use the Minimax principle to find the upper bound of optimal sample size for the quadratic model. The economical model in equation (6) is rewritten as:

$$\underset{n}{\text{Min}}\,\underset{\Delta}{\text{Max}}\left\{2n+\lambda\Delta^2\cdot\int_{-\infty}^{-\sqrt{n/2}\cdot\Delta}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz\right\} \tag{7}$$

Differentiating $I_n(\Delta)$ with respect to Δ, and setting the derivative equal to zero, we have

$$\begin{aligned}\frac{\partial I_n(\Delta)}{\partial\Delta}&=2\lambda\Delta\cdot\int_{-\infty}^{-\sqrt{n/2}\cdot\Delta}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz\\&+\lambda\Delta^2\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\sqrt{n/2}\cdot\Delta\right)^2}\left(-\sqrt{n/2}\right)=0\end{aligned} \tag{8}$$

$$\sqrt{n/2}\cdot\Delta=\frac{2\int_{-\infty}^{-\sqrt{n/2}\cdot\Delta}e^{-z^2/2}dz}{e^{-\frac{1}{2}\left(\sqrt{n/2}\cdot\Delta\right)^2}} \tag{9}$$

Let $\sqrt{n/2}\cdot\Delta=y$, and then we have

$$y=\frac{2\int_{-\infty}^{-y}e^{-z^2/2}dz}{e^{-y^2/2}} \tag{10}$$

Solving equation (10), we can obtain $y\doteq 1.1906$. Therefore, we have

$$\begin{aligned}&2n+\lambda\left(\sqrt{2/n}\cdot 1.1906\right)^2\cdot\int_{-\infty}^{-1.1906}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz\\&=2n+0.3314\cdot\lambda/n\end{aligned} \tag{11}$$

Differentiating $2n+0.3314\cdot\lambda/n$ with respect to *n*, and setting the derivative equal to zero, we have

$$n=\sqrt{0.1657\cdot\lambda} \tag{12}$$

From equation (12), the upper bound of optimal sample size for a particular λ is TABULATED in Table II.

TABLE II: The UPPER BOUND of OPTIMAL SAMPLE SIZE WHEN Δ IS UNKNOWN

Range of λ	Sample size *n*	Range of λ	Sample size *n*
1~12	1	2064~2293	19
13~36	2	2294~2534	20
37~72	3	2535~2787	21
73~120	4	2788~3053	22
121~181	5	3054~3330	23
182~253	6	3331~3620	24
254~337	7	3621~3922	25
338~434	8	3923~4236	26
435~543	9	4237~4562	27
544~663	10	4563~4899	28
664~496	11	4900~5249	29
797~941	12	5250~5611	30
942~1098	13	5612~5986	31
1099~1267	14	5987~6372	32
1268~1448	15	6373~6770	33
1449~1641	16	6771~7180	34
1642~1846	17	7181~7603	35
1847~2063	18	7604~8037	36

B. Two Population Variances σ_1^2 and σ_2^2 Are Unknown but Equal

TABLE III: THE OPTIMAL SAMPLE SIZE WHEN Δ IS KNOWN

Δ	Optimal sample size		
	λ=100	λ=500	λ=1000
0.0	1	1	1
0.1	1	1	1
0.2	1	1	1
0.3	1	1	3
0.4	1	4	10
0.5	1	8	14
0.6	2	10	15
0.7	3	11	15
0.8	4	11	14
0.9	4	10	13
1.0	5	10	12
1.2	5	8	10
1.4	5	7	9
1.6	4	6	7
1.8	4	6	6
2.0	3	5	6
2.5	3	4	4
3.0	2	3	3
3.5	2	2	3
4.0	2	2	2
4.5	1	2	2
5.0	1	2	2

TABLE IV: THE UPPER BOUND OF OPTIMAL SAMPLE SIZE WHEN Δ IS UNKNOWN

Range of λ	Sample size *n*	Range of λ	Sample size *n*
1~5	2	2143~2376	20
6~36	3	2377~2622	21
37~77	4	2623~2880	22
78~131	5	2881~3150	23
132~196	6	3151~3433	24
197~274	7	3434~3727	25
275~363	8	3728~4034	26
364~464	9	4035~4352	27
465~578	10	4353~4683	28
579~703	11	4684~5025	29
704~841	12	5026~5380	30
842~990	13	5381~5747	31
991~1152	14	5748~6126	32
1153~1326	15	6127~6517	33
1327~1512	16	6518~6920	34
1513~1710	17	6921~7335	35
1711~1919	18	7336~7762	36
1920~2142	19	7763~8202	37

Since the two population variances σ_1^2 and σ_2^2 are unknown but equal, we assume that S_1^2 and S_2^2 represent the sample variances of population 1 and population 2. The *T* statistic [21] is expressed as

$$T=\frac{\left(\bar{A}_1-\bar{A}_2\right)-\left(\mu_1-\mu_2\right)}{S_p\sqrt{1/n_1+1/n_2}}=\frac{\left(\bar{A}_1-\bar{A}_2\right)-\left(\mu_1-\mu_2\right)}{S_p\sqrt{\frac{2}{n}}} \tag{13}$$

where the pooled estimator of σ^2, denoted by S_p^2, is defined by

$$S_p^2=\frac{(n_1-1)S_1^2+(n_1-1)S_1^2}{n_1+n_2-2}=\frac{S_1^2+S_1^2}{2} \tag{14}$$

We consider the mean of population is the larger-the-better characteristic and the true difference in means $\Delta\cdot S_p$ for population 1 is larger than population 2. Under this situation, the loss function is expressed as

$$I_n(\Delta) = 2n + \lambda\Delta^2 \cdot \int_{-\infty}^{-\sqrt{n/2}\cdot\Delta} \frac{\Gamma\left((2n-1)/2\right)}{\Gamma(n-1)\cdot\sqrt{\pi(2n-2)}} \cdot \left(1+\frac{t^2}{2n-2}\right)^{-\frac{(2n-1)}{2}} dt \qquad (15)$$

When Δ is known

From equation (15), the optimal sample size for a particular Δ is tabulated in Table III.

When Δ is unknown

The upper bound of optimal sample size for a particular λ is tabulated in Table IV.

III. Result

From the above Tables, we can obtain the following results for the selection of an appropriate sample size.

1) When variances σ_1^2 and σ_2^2 are known and equal, if Δ is known, from Table I, the sample size $n > 15$ is disadvantageous for the quadratic model at $\lambda = 1000$
2) When variances σ_1^2 and σ_2^2 are known and equal, if Δ is unknown, from Table II, the upper bounds of optimal sample size are 4, 9 and 13 for the quadratic model at $\lambda = 100$, $\lambda = 500$ and $\lambda = 1000$, respectively.
3) When variances σ_1^2 and σ_2^2 are unknown and equal, if Δ is known, from Table III, the sample size $n > 16$ is disadvantageous for the quadratic model at $\lambda = 1000$
4) When variances σ_1^2 and σ_2^2 are unknown and equal, if Δ is unknown, from Table IV, the upper bounds of optimal sample size are 5, 10 and 14 for the quadratic model at $\lambda = 100$, $\lambda = 500$ and $\lambda = 1000$, respectively.

IV. Conclusion

The determination of sample size is a common task for many experiments. Inappropriate, inadequate, or excessive sample sizes continue to influence the accuracy and cost of experiments. This paper describes the economical models of selecting an appropriate sample size of experiments. Some tables are provided that can be used to determine sample size for experiments based on the testing cost of observations and loss by adopting a wrong level and probability of factor. The sampling issues for the variances of two populations are known or unknown are addressed.

References

[1] D. C. Montgomery, *Design and Analysis of Experiments*, 3rd edition, New York: Wiley, 1991.

[2] A. L. Gould, "Sample sizes required for binomial trials when the true response rates are estimated," *Journal of Statistical Planning and Inference*, vol. 8, no. 1, pp. 51-58, 1983.

[3] A. L. Hand, J. D. Stamey, and D. M.Young, "Bayesian sample-size determination for two independent Poisson rates," *Computer Methods and Programs in Biomedicine*, vol. 104, no. 2, pp. 271-277, 2010.

[4] B. R. Rao, "Sample size determination in case-control studies: The influence of the distribution of exposure," *Journal of Chronic Diseases*, vol. 39, no. 11, pp. 941-943, 1986.

[5] C. W. Dunnett, M. Horn, and R. Vollandt, "Sample size determination in step-down and step-up multiple tests for comparing treatments with a control," *Journal of Statistical Planning and Inference*, vol. 97, no. 2, pp. 367-384, 2001.

[6] E. Rahme, and L. Joseph, "Exact sample size determination for binomial experiments," *Journal of Statistical Planning and Inference*, vol. 66, no. 1, 1998, pp. 83-93, 1998.

[7] H. C. Thode, "Power and sample size requirements for tests of differences between two Poisson rates," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 46, no. 2, pp. 227–230, 1997.

[8] I. Gordon and R. Watson, "A note on sample size determination for comparison of small probabilities," *Controlled Clinical Trials*, vol. 15, no. 1, pp. 77-79, 1994.

[9] J. M. Lachin, "Introduction to sample size determination and power analysis for clinical trials," *Controlled Clinical Trials*, vol. 2, no. 2, pp. 93-113, 1981.

[10] K. F. Schulz and D. A. Grimes, "Sample size calculations in randomised trials: mandatory and mystical," *The Lancet*, vol. 365, no. 9467, pp. 1348-1353, 2005.

[11] K. J. Lui, "Sample size determination in case-control studies," *Journal of Clinical Epidemiology*, vol. 44, no. 6, pp. 609-611, 1991.

[12] K. J. Lui and W. G. Cumberland, "Sample Size Determination for Equivalence Test Using Rate Ratio of Sensitivity and Specificity in Paired Sample Data," *Controlled Clinical Trials*, vol. 22, no. 4, pp. 373-389. 2001.

[13] L. K. Heilbrun, and D. L. McGee, "Sample size determination for the comparison of normal means when one sample size is fixed," *Computational Statistics & Data Analysis*, 3, pp. 99-102, 1985.

[14] M. Gail, "The determination of sample sizes for trials involving several independent 2 ×2 tables," *Journal of Chronic Diseases*, vol. 26, no. 10, pp. 669-673, 1973.

[15] M. L. Tang, N. S. Tang, and V. J. Carey, "Sample size determination for 2-step studies with dichotomous response," *Journal of Statistical Planning and Inference*, vol. 136, no. 3, pp. 1166-1180, 2006.

[16] S. Becker, "Sample size determination in case-control studies," *Journal of Chronic Diseases*, vol. 40, no. 12, pp. 1141-1143, 1987.

[17] V. K. Murthy and L. J. Haywood, "Sample size determination for comparing two therapies taking both the α (type I or false positive) error and the β (type II or false negative) error into consideration," *Applied Mathematics and Computation*, vol. 3, no. 2, pp. 169-174, 1977.

[18] W. E. Hale, "Sample size determination for the log-normal distribution," *Atmospheric Environment*, vol. 6, no. 6, pp. 419-422, 1972.

[19] X. Zhang, G. Cutter, and T. Belin, "Bayesian sample size determination under hypothesis tests," *Contemporary Clinical Trials*, vol. 32, no. 3, pp. 393-398, 2011.

[20] G. Taguchi, *Design of Experiments*, 3rd edition, Tokyo: Maruzen, 1977, (in Japanese).

[21] W. W. Hines, D. C. Montgomery, D. M. Goldsman, and C. M. Borror, *Probability and Statistics in Engineering*, 4th edition, New York: Wiley, 2003.

Ful-Chiang Wu is a professor in the Department of Industrial Management at Vanung University, Taiwan. He received his B.S. in industrial engineering from Chung-Yuan Christian University and M.S. and Ph.D. in industrial engineering and management from Yuan-Ze University in Taiwan. His current research activities include quality engineering, quality management, and statistics in industrial applications.

Chi-Hao Yeh is a professor in the Department of Industrial Engineering and Management, National Taipei University of Technology, Taiwan. He received his MS degree from Iowa State University, in 1993, and PhD degree from Yuan-Zen University, on 2001. His research includes automatic optical inspection, product R&D process management, TRIZ-based innovative design in engineering and management.

Cheng-Hsiung Chen is a lecturer in the Department of Information Management at Hsing Wu University, Taiwan. He received his B.S. in Industrial Engineering from Yuan-Ze University in Taiwan. His current research activities include information management, quality management, and statistics in industrial applications.