ORIGINAL ARTICLE

A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions

Thavatchai Vorapongsathorn[1], Sineenart Taejaroenkul[2] and Chukiat Viwatwongkasem[3]

Abstract

Vorapongsathorn, T., Taejaroenkul, S. and Viwatwongkasem, C.
A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions
Songklanakarin J. Sci. Technol., 2004, 26(4) : 537-547

This study compared the probability of Type I error and the power of three statistical tests (Bartlett, Levene and Cochran) by varying the sampling distribution, variances and sample sizes. Monte Carlo methods were used to generate responses based on sample sizes and distributions 1,000 times. The sample sizes were both equal and unequal: 15, 30 and 45. The data distributions were Normal, Gamma and Chi-square.

It was found that Bartlett's test was sensitive to the normality assumption whereas Cochran's test and Levene's test were robust when the normal assumption was violated. Moreover, Levene's test was quite good for both equal and small sample sizes. In the case of power, Bartlett's test had the highest power in all cases. When one variance was large, Cochran's test was the best test.

---

[1]Ph.D.(Research Design & Statistics), Assoc. Prof., [3]M.Sc.(Biostatistics), Assoc. Prof., Department of Biostatistics, Faculty of Public Health, Mahidol University, Bangkok 10400 Thailand [2]M.Sc.(Biostatistics), Statistician, Research Institute for Health Science, Chiang Mai University, Chiang Mai, 50200 Thailand.
Corresponding e-mail: phtvr@mahidol.ac.th
Received, 8 January 2004 Accepted, 15 March 2004

The recommendations from this study are that: Bartlett's test is the best test for homogeneity of variances since it is not affected by sample size. When data are non-normally distributed, Levene's test is a good choice for small equal sample sizes. Cochran's test is best when sample size is large and unequal and one variance is larger.

Key words : homogeneity of variance, Bartlett's test, Levene's test, Cochran's test, simulation

บทคัดย่อ

ธวัชชัย วรพงศธร[1] สินีนาถ แต่เจริญกุล[2] และ ชูเกียรติ วิวัฒน์วงศ์เกษม[1]

การเปรียบเทียบความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของสถิติ Bartlett's Test, Levene's Test และ Cochran's Test ภายใต้การละเมิดข้อตกลงเบื้องต้น

ว. สงขลานครินทร์ วทท. 2547 26(4) : 537-547

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบความน่าจะเป็นของความคลาดเคลื่อนประเภทที่ 1 และเปรียบเทียบอำนาจการทดสอบความแปรปรวนของสถิติ 3 วิธีคือ Bartlett's Test, Levene's Test และ Cochran's Test ภายใต้ผลกระทบของการแจกแจงตัวอย่าง ผลกระทบของความแปรปรวน และผลกระทบจากขนาดตัวอย่าง โดยใช้วิธี Monte Carlo จำลองข้อมูลจากขนาดตัวอย่างและการแจกแจงตัวอย่างที่แตกต่างกัน ซึ่งกระทำ 1000 ครั้ง ขนาดตัวอย่างมีแบบเท่ากันและไม่เท่ากันคือ 15, 30 และ 45 สำหรับข้อมูลมีการแจกแจง 3 แบบ ได้แก่ การแจกแจงแบบปกติ การแจกแจงแบบแกมมา (Gamma) และการแจกแจงแบบไคกำลังสอง (Chi-square)

ผลการศึกษา พบว่าสถิติที่มีความเหมาะสมกับการแจกแจงปกติ ได้แก่ Bartlett's Test Cochran's Test และ Levene's Test ตามลำดับ เมื่อข้อมูลไม่เป็นการแจกแจงปกติ พบว่า Cochran's Test และ Levene's Test มีความเหมาะสม โดย Levene's Test ค่อนข้างได้เปรียบเมื่อจำนวนตัวอย่างมีขนาดเล็ก ในด้านอำนาจการทดสอบพบว่า Bartlett's Test ยังคงให้อำนาจการทดสอบสูงเกือบทุกกรณี และเมื่อความแปรปรวนหนึ่งค่ามีขนาดใหญ่กว่าค่าอื่น พบว่า Cochran's Test ให้อำนาจการทดสอบสูงที่สุด

ข้อเสนอแนะสำหรับการใช้สถิติทดสอบความเท่ากันของความแปรปรวน ได้แก่ Bartlett's Test ยังคงเป็นวิธีที่ดีเมื่อข้อมูลมีการแจกแจงแบบปกติเนื่องจากไม่มีผลกระทบกับขนาดตัวอย่าง และเมื่อข้อมูลไม่เป็นการแจกแจงแบบปกติ Levene's Test สามารถนำไปใช้กับข้อมูลที่มีขนาดเล็กและจำนวนตัวอย่างเท่ากัน ส่วน Cochran's Test ใช้เมื่อตัวอย่างมีขนาดใหญ่และจำนวนตัวอย่างมีไม่เท่ากันและมีความแปรปรวนที่ใหญ่กว่าค่าอื่นหนึ่งค่า

[1]ภาควิชาชีวสถิติ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยมหิดล กรุงเทพฯ 10400 [2]สถาบันวิจัยวิทยาศาสตร์สุขภาพ มหาวิทยาลัยเชียงใหม่ อำเภอเมือง จังหวัดเชียงใหม่ 50200

One important problem in applied research is to decide whether sample differences in central tendency reflect true differences in parent populations. The analysis of variance (ANOVA) is the most powerful technique for testing hypotheses about this phenomenon when the assumptions of normality, homogeneity of variance and independence of errors are met. Failure of any assumption would impair the utility of the test, leading to wrong and invalid conclusions (Cochran, 1947; Bodhisuwan, 1991; Srisunsanee, 1998). Therefore, it is necessary to test the assumptions before using analysis of variance. Current literature recommends the use of several statistical procedures to test the assumption of homogeneity of variance. Among these statistics, Bartlett's test, Levene's test, and Cochran's test are widely used to check the ANOVA assumptions (Filliben *et al*., 2000a; Filliben *et al*., 2000b; Phil, 1999).

Bartlett's test (Filliben *et al*., 2000a) in-

volves computing a statistic whose sampling distribution is closely approximated by the chi-square distribution with *k*-1 degrees of freedom when the *k* random samples are from independent normal populations (Montgomery, 1997). Bartlett's statistic is designed to test for equality of variances across groups against the alternative that variances are unequal for at least two groups. The test statistic is $\chi_0^2 = 2.3026\frac{q}{c}$, where

$$q = (N-k)\log_{10} S_p^2 - \sum_{i=1}^{k}(n_i-1)\log_{10} S_i^2$$

$$c = 1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}(n_i-1)^{-1}-(N-k)^{-1}\right)$$

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{N-k}$$

where 2.3026 = a constant value
n_i = sample size of the i[th] group
S_i^2 = sample variance of the i[th] group
N = total sample size $\left(N=\sum_{i=1}^{k} n_i\right)$
k = number of groups
S_p^2 = pooled variance (weighting by degrees of freedom)

In testing the homoscedasticity of several populations by comparing a number of these statistics for power and for stability of error rates, Gartside (1972) found that Bartlett's statistic is very powerful in all the experimental cases. In Bartlett's test, the n_i's in each of the treatment classes need not be equal. However, no n_i's should be smaller than 3, and most n_i's should be larger than 5 (Winer, 1974).

Levene's test (Filliben *et al.*, 2000b) is used to test if k samples are from equal variance populations. Some statistical tests, for example, analyses of variance, assume that variances are equal across groups or populations. Levene's test can be used to verify that assumption. Given a variable Y with sample of size N divided into k sub-groups, where n_i is the sample size of the i[th] subgroup, Levene's test statistic is defined as:

$$W = \frac{(N-k)\Sigma_{i=1}^{k} n_i(\overline{Z}_{i.}-\overline{Z}_{..})^2}{(k-1)\Sigma_{i=1}^{k}\Sigma_{j=1}^{N}(Z_{ij}-\overline{Z}_{i.})^2}$$

where Z_{ij} can have one of the following three definitions:

1) $Z_{ij} = \left|Y_{ij}-\overline{Y}_{i.}\right|$, where $\overline{Y}_{i.}$ is the *mean* of the i[th] subgroup

2) $Z_{ij} = \left|Y_{ij}-\tilde{Y}_{i.}\right|$, where $\tilde{Y}_{i.}$ is the *median* of the i[th] subgroup

3) $Z_{ij} = \left|Y_{ij}-\overline{Y}'_{i.}\right|$, where $\overline{Y}'_{i.}$ is the 10% *trimmed mean* of the i[th] subgroup.

Here the 10% trimmed mean is the arithmetic mean calculated when the largest 10% and the smallest 10% of the cases have been eliminated. Eliminating extreme cases from the computation of the mean results in a better estimate of central tendency, especially when the data are non-normal (Neter *et al.*, 1996).

Cochran's test is a homogeneity of variances test. It is computationally simpler than Bartlett's test and also affected by non-normality (Phil, 1999).

Cochran's test is defined as: $c = \frac{\text{largest } S_i^2}{\Sigma S_i^2}$,

[where n (n = n_i) is number of replicates for each mean; if n's are unequal, (n n_i) use largest or harmonic mean of n's].

Here S_i^2 = sample variance of the i[th] group
k = number of groups
df = k-1; each of the k groups has n-1 degrees of freedom

Gartside (1972) found that Cochran's test performed very well in power for equal samples size of 16 and k varied as 3, 4, 5, 10. They were taken from populations whose variances increased in succession by a constant ratio, i.e., it showed

good sensitivity when departures from the equal variances were still small. Moreover, if one variance is larger, then Cochran's statistic would be a good choice as it maintained good power in this case (Cochran, 1947).

From literature reviews, it is seen that statistical tests have different methods to test data and they have some different weak points. Especially, Bartlett's test is sensitive to violation of the normality assumption. Cochran's and Levene's tests seem to be a good choice for checking homogenous variances if robustness against non-normality is needed. Yet, there are no studies reporting comparison of results when using these three statistical tests when assumptions are violated under different situations.

The objectives of this study were to compare the probability of Type I error and the power of Bartlett's test, Levene's test and Cochran's test in three situations; 1) when the data distributions were normal, and non-normal, 2) the sample sizes were equal and unequal, and 3) the sample variances were equal and unequal.

Methodology

There are two criteria to detect appropriate statistics under violation of assumptions, ***robustness*** and ***power. Robustness*** is the capability to control Type I error. In other words, it is the ability of the test to not falsely detect non-homogeneous groups when the underlying data is not normally distributed and the groups are in fact homogeneous. A statistical test is designated robust, if the departure of the empirical of Type I error ($\hat{\tau}$) from the nominal level of significance (α) does not exceed the predetermined value. In this study, robustness evaluation is based on the Cochran limit as follows:

at 0.01 significance level, τ value is between (0.007-0.015)

at 0.05 significance level, τ value is between (0.04-0.06)

τ = the true probability of a Type I error = Probability (H_0 is rejected when H_0 is true)

$\hat{\tau}$ = the empirical probability of a Type I error

$$\frac{\text{the number of } H_0 \text{ rejection when } H_0 \text{ is true}}{\text{the number of replications 1,000 times}}$$

α = the nominal level of significance or the theoretical alpha

The statistical test is called robust when its empirical alpha values lie within the Cochran limit (Peechawanich, 1992). If any actual probability of Type I error is over the limit, it shows that the test cannot control the error rate.

The power of the test is the probability of rejecting a null hypothesis when it is false and therefore should be rejected. In this study, the power of a test is calculated by subtracting the empirical probability of a Type II Error ($\hat{\beta}$) from 1.0. Type II Error is an error made by wrongly accepting or failing to reject a false null hypothesis.

$\hat{\beta}$ = the empirical probability of a Type II error

$$\frac{\text{the number of } H_0 \text{ failed to reject when } H_0 \text{ is false}}{\text{the number of replications 1,000 times}}$$

Power = 1 - the empirical probability of Type II error = $(1-\hat{\beta})$

The maximum total power of a test can have is 1.0; the minimum is zero.

Computations

The processes of calculating probability of Type I error and power of three statistical tests under different settings of three groups of population distributions were as follows:

1. ***Population Distribution from Defining Distribution***

A Fortran program was used to create the populations under three distributions, Normal, Gamma and Chi-square, by simulating the data from Monte Carlo method (Karlen, 1995). The simulation plans are shown in Table 1.

Normal distribution or ***Gaussian*** distribution has probability density of x, which is defined by $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\alpha < x < \alpha$

Which has two parameters, mean (μ) = 10 and variance (σ^2) = 26 (Peechawanich, 1992).

Gamma distribution has a formula as $f(x,\alpha,\beta)=\frac{1^{\alpha}}{\beta}\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)}$ for x > 0. It has two parameters, mean (μ) = $\alpha\beta$ and variance (σ^2) = $\alpha\beta^2$, which α = 2.67, β = 0.76 (Peechawanich, 1992).

Chi-square distribution has a formula as

$$f(x)=\frac{x^{\frac{n}{2}-1}e^{-x/2}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})},x>0$$

The mean and variance of the Chi-square distribution are n and 2n, respectively (Peechawanich, 1992).

2. ***Generate Groups of Populations***

The populations were generated in three groups with the same distribution as Normal (N N N) (Peechawanich, 1992); Gamma (G G G) (Kyle, 2001) and Chi-square (C C C) (Peechawanich, 1992). The researcher used the general rule of thumb that defined the sample size of 30 (ISixSigma, 2001). Therefore, the sample sizes in this study for three groups were less than 30, equal to 30 and greater than 30; these were 15, 30, and 45. The four sample size sets were 15,15,15; 30,30,30; 45,45,45 and an unequal one of 15,30, 45 (ISixSigma, 2001) (Table 1).

3. ***Define Proportion of Sample Variances***

The sample variances in each group of three populations were in the ratios 1:1:1 (under H_0); 1:1:2; and 1:2:4 (under H_1) (Box, 1953). In testing homogeneity of variances of the three statistical tests, theoretical alpha was defined as 0.01 and 0.05 (Table 1).

4. ***Compute the Values of Three Statistical Tests***

The data were generated in one situation for computing Bartlett's test, Levene's test and Cochran's test values. Then these values were compared with their critical region (Kingman & Zion, 1994). This was done a thousand times and

Table 1. Simulation plans for generating responses based on three distributions, equal and unequal sample sizes and unequal sample variances.

Distribution/ Level of Significance	Sample Size n_i	Under H_0 the ratio of variance = 1:1:1	Under H_1	
			the ratio of variance = 1:1:2	the ratio of variance = 1:2:4
Normal α = 0.01	15, 15, 15 30, 30, 30 45, 45, 45 15, 30, 45	Xi ~ N(μ, σ^2), $\forall$i $\mu_i = \mu$, $\forall$i $\sigma^2_i = \sigma^2_i$, $\forall$i	X_1 ~ N(μ, σ^2) X_2 ~ N(μ, σ^2) X_3 ~ N(μ, $2\sigma^2$)	X_1 ~ N(μ, σ^2) X_2 ~ N(μ, $2\sigma^2$) X_3 ~ N(μ, $4\sigma^2$)
α = 0.05	$\forall$i (i = 1,2,3)	$H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \sigma^2$ = 1:1:1	$H_1: \sigma^2_1 : \sigma^2_2 : \sigma^2_3 = \sigma^2_1 : \sigma^2_2 : 2\sigma^2_3$ = 1: 1:2	$H_1: \sigma^2_1 : \sigma^2_2 : \sigma^2_3 = \sigma^2_1 : 2\sigma^2_2 : 4\sigma^2_3$ = 1:2:4
Gamma α = 0.01	15, 15, 15 30, 30, 30 45, 45, 45 15, 30, 45	X_i ~ G(α, β), $\forall$i $\mu_i = \alpha\beta$ $\sigma^2_i = \alpha\beta^2$	X_1 ~ G(α, β) X_2 ~ G(α, β) X_3 ~ G(α/2, 2β)	X_1 ~ G(α, β) X_2 ~ G(α/2, 2β) X_3 ~ G(α/4, 4β)
α = 0.05	$\forall$i (i = 1,2,3)	$H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3 = \alpha\beta^2$ = 1:1:1	$H_1: \sigma^2_1 = \sigma^2_2 = \sigma^2_3$ = $\alpha\beta^2 : \alpha\beta^2 : 2\alpha\beta^2$ = 1:1:2	$H_1: \sigma^2_1 = \sigma^2_2 = \sigma^2_3$ = $\alpha\beta^2 : 2\alpha\beta^2 : 4\alpha\beta^2$ = 1:2:4
Chi-square α = 0.01	15, 15, 15 30, 30, 30 45, 45, 45 15, 30, 45	X_i ~ χ^2 (n) μ_i = n σ^2_i = 2n	X_1 ~ χ^2 (n), μ_1 = n, σ^2_1 = 2n X_2 ~ χ^2 (n), μ_2 = n, σ^2_2 = 2n X_3 ~ χ^2 (2n), μ_3 = 2n, σ^2_2 = 4n	X_1 ~ χ^2 (n), μ_1 = n, σ^2_1 =2n X_2 ~ χ^2 (2n), μ_2 = 2n, σ^2_2 = 4n X_3 ~ χ^2 (4n), μ = 4n, σ^2_2 = 8n
α = 0.05	$\forall$i (i = 1,2,3)	$H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3 = 2n$ = 1:1:1	$H_1: \sigma^2_1 = \sigma^2_2 = \sigma^2_3 = 2n : 2n : 4n$ = 1:1:2	$H_1: \sigma^2_1 = \sigma^2_2 = \sigma^2_3 = 2n : 4n : 8n$ = 1:2:4

the values that rejected null hypothesis were counted. In case of Type II Error, the values that failed to reject the null hypothesis were counted and the power of the test was calculated by subtracting the probability of a Type II Error from 1.0. The process of computation was repeated for all situations.

Results

1. Type I Error

1.1 Normal Distribution

Robustness at 0.01 and 0.05 Significance Levels

When the assumption of a normal distribution was met all three statistical tests could control Type I error for all of equal and unequal sample sizes at $\alpha = 0.01$ and .05. This evidence showed that when the distribution was normal, the sample size did not affect the robustness of the tests (Table 2).

1.2 Gamma Distribution

Robustness at 0.01 and 0.05 Significance Levels

Levene's test was the only test that could control Type I error for the sample sizes of 15 at $\alpha = 0.01$ and of 45 at $\alpha = 0.05$. The values of Type I Error were .013 for n = 15 and .036 for n = 45 respectively. For Gamma distribution, Bartlett's and Cochran's tests could not control Type I error for all sample sizes (Table 2).

1.3 Chi-square Distribution

Robustness at 0.01 and 0.05 Significance Levels

Type I error values of Barlett's test for equal sample size of 30, Cochran's test for sample sizes of 30 and 45 and Levene's test for sample sizes of 15, 30 and 45 were below the Cochran limit (0.015 at $\alpha = 0.01$ and 0.055 at $\alpha = 0.05$). For unequal sample sizes, there were no statistical tests that could control Type I error (Table 2).

2. Power of Test

2.1 Normal Distribution

For Variance Ratio of 1:1:2

For all of the equal sample size and unequal sample sizes at $\alpha = 0.01$ and 0.05, the powers of the three tests were lower than .50 or the tests could detect the faulty null hypothesis less than 50 % of the time. Only Cochran's test for equal sample size, n = 45 at $\alpha = 0.05$, had power

Table 2. Effect of empirical type I errors under equal variances hypothesis of the three statistical tests for various nominal significance levels with normal, gamma and chi-square distribution.

Nominal Level of Significance	Sample Sizes, n_i	Normal Distribution			Gamma Distribution			Chi-square Distribution		
		B	L	C	B	L	C	B	L	C
0.01	15, 15, 15	0.015*	0.002*	0.011*	0.165	0.013*	0.146	0.027	0.003*	0.029
	30, 30, 30	0.009*	0.004*	0.006*	0.290	0.062	0.243	0.013*	0.007*	0.011*
	45, 45, 45	0.009*	0.007*	0.008*	0.377	0.130	0.315	0.021	0.006	0.010*
	15, 30, 45	0.012*	0.013*	0.011*	0.195	0.039	0.364	0.356	0.161	0.485
0.05	15, 15, 15	0.047*	0.024	0.035	0.317	0.080	0.292	0.089	0.018	0.096
	30, 30, 30	0.054*	0.043*	0.041*	0.455	0.189	0.410	0.027*	0.048*	0.055*
	45, 45, 45	0.051*	0.032*	0.037*	0.563	0.036*	0.492	0.080	0.051*	0.050*
	15, 30, 45	0.052*	0.044*	0.055*	0.351	0.130	0.490	0.591	0.441	0.684

Note: B = Bartlett's Test
L = Levene's Test
C = Cochran's Test
*** Type I error in control**

as high as .70 (Table 3, Figure 1 and 2).

For Variance Ratio of 1:2:4

The power of the three statistical tests tends to get higher as the sample size increases. For all sample sizes, Bartlett's test had the highest power of .95 for n = 45 at $\alpha = 0.01$ and .99 at $\alpha = 0.05$. Levene's test had the lowest power in all sample size at $\alpha = 0.01$ but it increased to .96 for equal sample sizes of 45 at $\alpha = 0.05$. For unequal sample size, the power of Cochran's test is the highest being .91 at $\alpha = 0.05$ (Table 3, Figure 1 and 2).

2.2 Gamma Distribution

For Variance Ratio of 1:1:2

For all sample sizes, the three tests had low power, especially Levene's test had very low power being .01 at $\alpha = 0.01$ and .08 at $\alpha = 0.05$ (Table 3, Figure 1 and 2).

For Variance Ratio of 1:2:4

The three tests still had low powers and the same pattern as for variance ratio of 1:1:2. Bartlett's test had the highest power being .67 for n = 45 at $\alpha = 0.05$. For unequal sample size, Cochran's test had the highest power being .67 at $\alpha = 0.05$. (Table 3, Figure 1 and 2).

2.3 Chi-square Distribution

For Variance Ratio of 1:1:2

For all equal sample size sets, Bartlett's test had a little higher power than Cochran's test, (.45 vs. .44 at $\alpha = 0.01$ and .68 vs. .67 at $\alpha = 0.05$). For unequal sample size sets, Cochran's test had the highest power being .62 at $\alpha = 0.05$. (Table 3, Figure 1 and 2).

For Variance Ratio of 1:2:4

The data in Table 3 showed that, for equal sample sizes, Bartlett's and Levene's tests

Table 3. Effect of heterogeneous variances on power of the three statistical tests for various nominal significance levels with normal, gamma, and chi-square distribution.

Nominal Level of Significance	Ratio of σ^2, *S*	Sample Sizes, n_i	Normal Distribution B	L	C	Gamma Distribution B	L	C	Chi-square Distribution B	L	C
0.01	1:1:2	15, 15, 15	0.098	0.047	0.113	0.198	0.009	0.170	0.129	0.044	0.128
		30, 30, 30	0.249	0.162	0.249	0.303	0.034	0.192	0.247	0.149	0.272
		45, 45, 45	0.471	0.348	0.473	0.424	0.071	0.245	0.448	0.324	0.437
		15, 30, 45	0.245	0.238	0.258	0.310	0.026	0.347	0.318	0.150	0.467
	1:2:4	15, 15, 15	0.352	0.157	0.291	0.275	0.005	0.260	0.373	0.141	0.310
		30, 30, 30	0.767	0.553	0.576	0.436	0.002	0.388	0.754	0.575	0.584
		45, 45, 45	0.948	0.866	0.840	0.541	0.044	0.459	0.936	0.857	0.797
		15, 30, 45	0.604	0.374	0.784	0.442	0.009	0.550	0.632	0.392	0.774
0.05	1:1:2	15, 15, 15	0.246	0.154	0.274	0.355	0.080	0.295	0.275	0.146	0.287
		30, 30, 30	0.448	0.354	0.452	0.458	0.129	0.348	0.460	0.382	0.466
		45, 45, 45	0.688	0.594	0.700	0.607	0.228	0.429	0.676	0.571	0.674
		15, 30, 45	0.471	0.390	0.472	0.501	0.118	0.528	0.518	0.357	0.616
	1:2:4	15, 15, 15	0.599	0.407	0.509	0.446	0.040	0.390	0.621	0.393	0.525
		30, 30, 30	0.901	0.797	0.785	0.569	0.094	0.515	0.909	0.803	0.792
		45, 45, 45	0.988	0.963	0.929	0.672	0.153	0.591	0.986	0.963	0.923
		15, 30, 45	0.824	0.695	0.905	0.612	0.057	0.673	0.829	0.710	0.904

Note: B = Bartlett's Test
L = Levene's Test
C = Cochran's Test
The bold figure is the highest power value of that statistical test as compared with those of the other two tests.

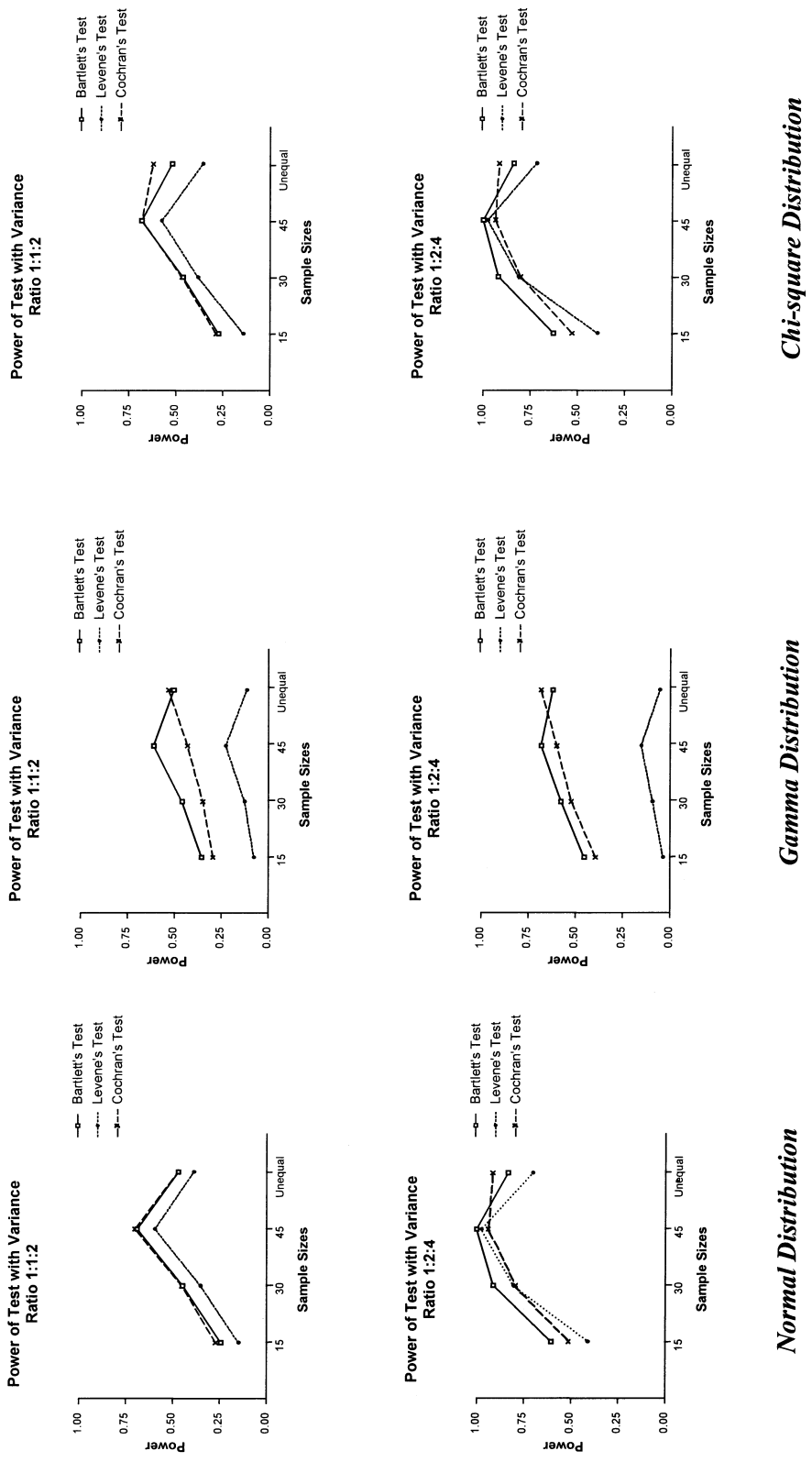

Figure 1. Power of the Three Statistical Tests as influenced by Variance Ratios and Sample Sizes with Normal, Gamma and Chi-square Distribution at 0.01 Significance Level.

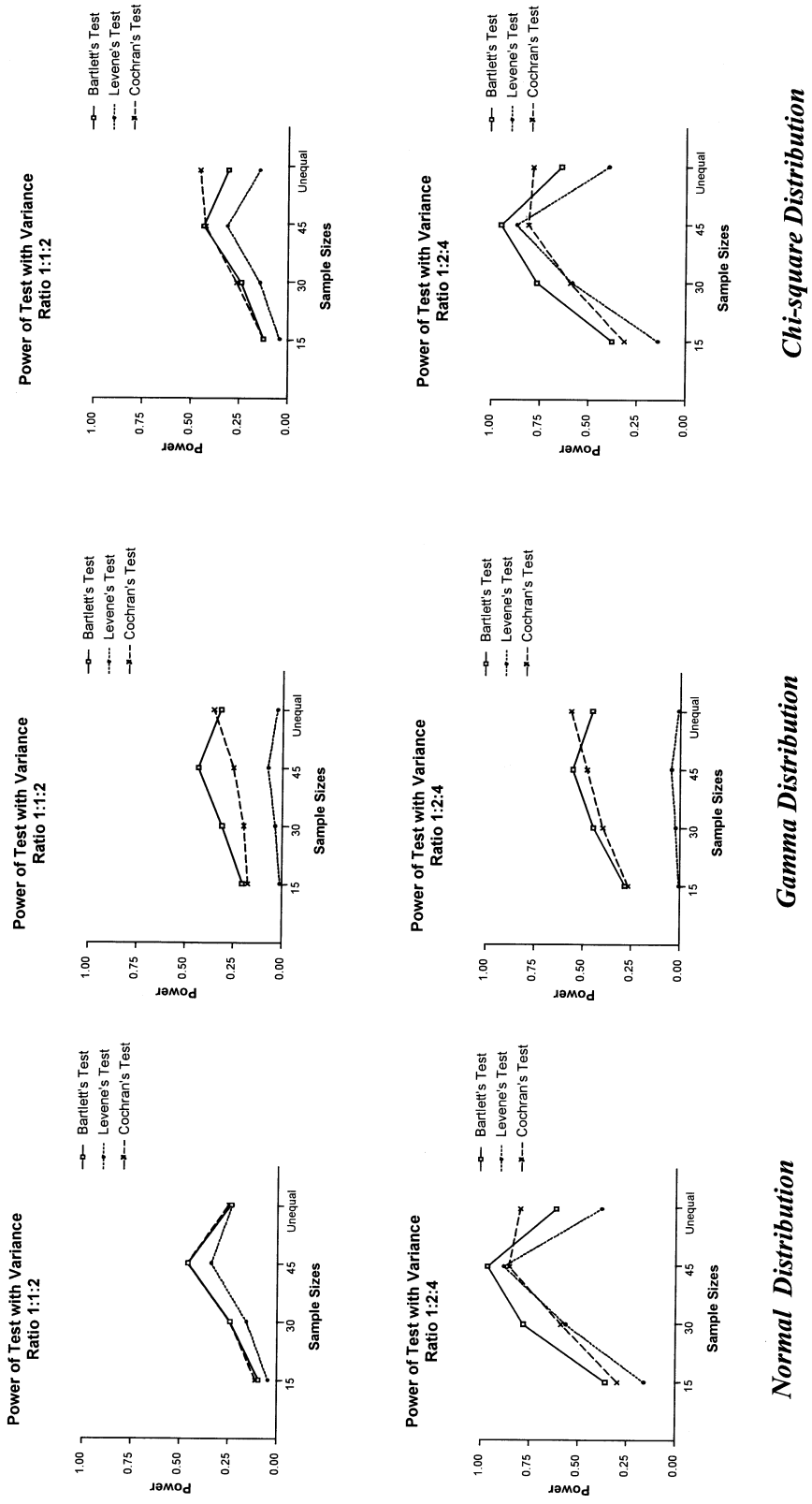

Figure 2. Power of the Three Statistical Tests as influenced by Variance Ratios and Sample Sizes with Normal, Gamma and Chi-square Distribution at 0.05 Significance Level.

had the powers as high as those for the normal distribution. For equal sample size of 45, the power of Bartlett's test was .94 as compared with .87 for Levene's test at $\alpha = 0.01$ and .99 vs. .96 at $\alpha = 0.05$. However, for unequal sample sizes, Cochran's test still had the highest power of .77 at $\alpha = 0.01$ and of .90 at $\alpha = 0.05$ (Table 3, Figure 1 and 2).

Discussion

Two important issues, the robustness and the power of three statistical tests were discussed under different situations of violation of the assumptions.

1. Type I Error

1.1 Normal Distribution

Under the normal assumption, homogeneity of variances, equal and unequal sample sizes, the Type I error probabilities of Bartlett's, Levene's and Cochran's tests were lower than Cochran limits at both $\alpha = 0.01$ and 0.05. In particular, the values of the Type I error of Levene's test were smallest for sample sizes of 15 and 30. This was the same as Brown and Forsythe (1974) found from their research, when the Gaussian distribution indicated that Levene's test with median (W_{50}) was for small sample sizes. Also Gartside (1972) found that Bartlett's test should be used when the alternative hypothesis was unknown and normality could be relied on. This study showed that unequal sample sizes had not affected the probability of Type I error of the three statistical tests.

1.2 Gamma Distribution

When the homogeneity of variances assumption and equality of sample sizes were met but the normal assumption was violated, Levene's test was the only test that was robust. Furthermore, if the homogeneity of variance assumption was met, the normal assumption and unequal sample sizes were violated, there were no statistical tests that were robust. These results correspond to all theories of the three statistical tests. Bartlett's test was not a good test when the distribution was doubtful. Cochran's test could not control Type I error when the data are non-normally distributed. Meanwhile, Levene's test was found to be robust under non-normality and small equal sample sizes (ISixSigma, 2001).

1.3 Chi-square Distribution

When chi-square distribution violates normal assumption, Bartlett's test could control Type I error for only sample size of 30. This is corresponding to the general rule of thumb that defines the sample size of 30 (ISixSigma, 2001). Cochran's test was robust in sample size of 30 and 45. Among the three tests, Levene's test was robust in various equal sample sizes of 15, 30 and 45. These results are similar to Brown and Forsyth's study (1974) that made sampling from chi-square with four degree of freedom.

2. Power of Test

2.1 Normal Distribution

When the normality and equality of sample sizes assumption were met but homogeneity of variance was violated as one variance is larger (variance ratio is equal to 1:1:2), Cochran's test was a good choice as it maintained good power (Gartside, 1972). However, when the variance ratio was increased to 1:2:4, and sample sizes were equal, Bartlett's test gave the highest power (Forsythe, 1974).

2.2 Gamma Distribution

When the equality of sample size is met, but normal assumption and homogeneity of variances are violated, Bartlett's test has the highest power for all sample sizes. This result is different from the hypothesis that the violation of normal distribution assumption may not result in the loss of power of Levene's test. When normal assumption, homogeneity of variance assumption and equality of sample sizes are violated, Cochran's test still is a good choice that gives the highest power. The result is the same for different empirical alpha.

2.3 Chi-square Distribution

When sample sizes are equal, but homogeneity of variances assumption and normal assumption are violated, Bartlett's test seems to be

a good statistical test which has the highest power in all equal sample sizes. When normal assumption, homogeneity of variances assumption and equality of sample sizes are violated, Cochran's test is a good choice that gives the highest power when one variance is larger and sample sizes are equal. That is the same as in Gartside (1972). However, when unequal variances ratio (1:2:4) is met under non-normality and equal sample sizes, Bartlett's test is a good choice since it gives the highest power.

Implication of the Study

When data are normally distributed, Bartlett's test is a good choice for testing homogeneity of variances since it is not affected by sample sizes. When data are non-normally distributed, Levene's test is a good choice for small equal sample sizes and equal variances. If one variance is larger, the data are non-normally distributed, and are of unequal sample sizes, Cochran's test would be recommended, since it still gives high power.

References

Bodhisuwan, W. 1991. A comparison of the test statistics for homogeneity of variances. [M.S.Thesis in Statistics]. Bangkok: Faculty of Graduate Studies, Chulalongkorn University.

Box, G.E.P. 1953. Nonnormality and tests on variances. Biometrika, 40: 318-35.

Brown, M.B and Forsythe, A.B. 1974. Robust tests for the equality of variances. JASA, June; 69: 364-7.

Cochran, W.G. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. Biometics, 3: 22-38

Filliben, J.J and Heckert, A. 2000a. Bartlett's test. [Online] 2000;[4]. Available from: URL:http://www.nist.gov/itl/div898/handbook/eda/section3/eda357.htm Accessed 2000 July 12.

Filliben, J.J. and Heckert, A. 2000b. Levene's test for equality of variances. [Online]2000; Available from: http://www.nist.gov/itl/div898/handbook/eda/section3/eda 358.htm [Accessed 2000 July 12].

Gartside, P.S. 1972. A study of methods for comparing several variances. JASA, June; 67(338): 342-6.

IsixSigma. 2001. Small sample size calculation [Online]. Available from: http://www.sixsigma.com: library/conent/c000709smsample.asp [Accessed 2001 May 18].

Karlen, D. 1995. Part III: Monte carlo methods. Random number generation. [Online]. Ottawa: Available from: http://www.physics.carleton.ca/courses/75.502/slides/monte21 /p015.html. [Accessed 2002 August 21].

Kingman, A and Zion, G. 1994. Some power considerations when deciding to use transformations. Stat Med, 13: 769-83.

Kyle, S. 2001. Special distributions. The gamma distribution. [Online]. Huntsville:1997, updated 27 Sept. 2001. Available from: http://www. mathuah. edu/stat/special/special3.html [Accessed 2001 May 29].

Montgomery, D.C. 1997. Design and analysis of experiments. 4th ed. New York: John Wiley & Sons.

Neter, J, Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. 1996. Applied linear regression model. 3rd ed. Illinois: Irwin.

Peechawanich, V. 1992. Probability theory and the applications. Bangkok: Prakyprueg.

Phil, E. 1999. Checking assumptions. Education 230 B/C Linear Statistical Models. [Online]. Graduate School of Education & Information Studies; 4 June 99. Available from: http://www.gseis.ucla.edu/courses/ed230bc1/cnotes1 [Accessed 2000 July 13].

Srisunsanee, J. 1998. Effects of failure to meet assumptions of homogeneity of variance and normal on Type I error in one-way analysis of variance. [M.S. Thesis in statistics]. Bangkok: Faculty of Graduate Studies, Kasetsart University.

Winer, B.J. 1974. Statistical principles in experimental design. 2nd ed. New York: McGraw-Hill.