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

Kevin Nyamai

Computational statistics

Université de Neuchâtel

Under supervision of
Prof. Alina Matei

Aim

Original Paper: (Thavatchai, et al, 2004) Link

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 for homogeneity of variance in three situations:

- When the data distributions are normal and non-normal
- The sample sizes were equal and unequal
- The sample variances were equal and unequal

Overview- Background (1/4)

- The tests are primarily used in checking homogeneity of variance between groups. This key in tests like ***anova*** that is preferred in checking sample differences in central tendency and whether they reflect parent populations:

- *Anova* assumes normality and homogenity of variance.
  If the group sizes are vastly unequal and homogeneity of variance is violated, then the F statistic will be biased when large sample variances are associated with small group sizes, implying that the level of significance will be underestimated leading false rejection of the null hypothesis (Tjen-Sien Lim 1996).

Overview-Bartlett's (2/4)

- **Bartlett's test** (Bartlett, 1937) involves 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)

- **Hypothesis:**

  H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

  H_a: $\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j)

- Test Statistic:

$$\chi^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right) - \frac{1}{N-k}\right)}$$

Overview-Cochran's (3/4)

- **Cochran's test** (Cochran 1947) is computationally simpler than Bartlett's test and also affected by non-normality (Phil, 1999).:

- **Hypothesis**:

  H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

  H_a: $\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j)

- **Test Statistic**: $C = \dfrac{Largest\ S^2}{\sum S^2}$

Overview-Levene's (4/4)

- **Levene's test** (Levene 1960)) is used to test if k samples are from equal variance populations. It is less sensitive than the Bartlett test to departures from normality
- **Hypothesis**:

  H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

  H_a: $\sigma_i^2 \neq \sigma_j^2$ for at least one pair (i,j)

- **Test Statistic**: $$W = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_{i\cdot} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i\cdot})^2},$$

.

Methodology-Robustness(1/5)

Two criteria are used to detect the appropriate statistics under violation are:

Robustness (τ_hat): is the capability to control *Type I error*. (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.

$$\boldsymbol{\tau\_hat} = \frac{\textit{the number if } H_0 \textit{ rejected when Ho is true}}{\text{(the number of replications, 10,000 times)}}$$

α = the nominal level of significance or the theoretical alpha

Cochran Limit Evaluation (*Peechawanich*, 1992) :

- 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)

Methodology-Power(2/5)

Two criteria are used to detect the appropriate statistics under violation are:

- **Power of the test, (1 – β_hat) :** the probability of rejecting a null hypothesis when it is false and therefore should be rejected. (Where β is the probability of Type II error)

- $$\boldsymbol{\beta\_hat} = \frac{the\ number\ if\ H_0\ failed\ to\ reject\ when\ H_0\ is\ false}{\text{(the number of replications,10,000 times)}}$$

- $$\boldsymbol{Power} = 1 - \beta\_\text{hat}$$

Methodology- Computations (3/5)

- The **computations** were done in **R software** using :
  - *stats* Package for Bartlett's Test
  - *car* Package for Levene's test
  - *GAD* Package for Cochran's Test

- **Sample sizes** 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

- **Sample Variances** in each group of three populations were in the ratios 1:1:1 (under H0); 1:1:2; and 1:2:4 (under H1) (Box, 1953)

- The populations were generated in three groups with the same distribution as ***Normal***- Mean=10, Variance=26 (N N N) and ***Chi-square*** (C C C) degrees freedom=2

- Number of **Simulations** was **10,000**

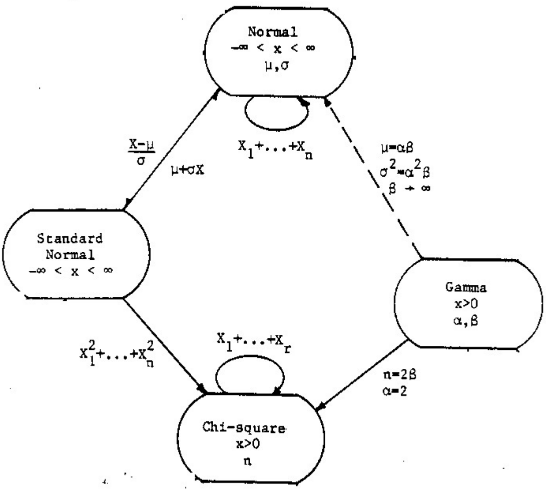
Methodology- Deviations from Paper (4/5)

- **Number of Simulations:** The paper used 1000. I used 10,000 for stability of results

- **Chi- Square Degrees of Freedom :** The paper was not clear on the number of degrees of freedom, I used *df=2* under null hypothesis.

- **Choice of Distribution.** The paper generetad samples from the normal, chi-square and gamma distributions. The normal and chi-sqaure distributions were used for simplicity as well as avoiding direct relationship

Methodology- Deviations from Paper (5/5)



Results- Empirical Type I Error (1/2)

Nominal Level of Significance	Sample Sizes, n_i	Normal Distribution				Chi-square Distribution		
		B	L	C		B	L	C
0.01	15, 15, 15	0.0091*	0.0038	0.0108*		0.2017	0.0075*	0.1721
	30, 30, 30	0.0105*	0.0069	0.0105*		0.2493	0.0100*	0.2061
	45, 45, 45	0.0114*	0.009*	0.0098*		0.2652	0.0092*	0.2193
	15, 30, 45	0.0107*	0.0056	0.0107*		0.2353	0.0084*	0.2083
0.05	15, 15, 15	0.0517*	0.0294	0.0510		0.3682	0.0382	0.311
	30, 30, 30	0.0484*	0.0410	0.0484		0.4024	0.0521*	0.3543
	45, 45, 45	0.0469*	0.0417	0.0487		0.4215	0.0494*	0.3712
	15, 30, 45	0.0495*	0.0368	0.0494		0.3952	0.0451*	0.3446

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

Results- Power (2/2)

Nominal Level of Significance	Ration of σ^2,S	Sample Sizes, n_i	Normal Distribution				Chi-square Distribution			
			B	**L**	**C**	**Order of Power (Largest to Smallest)**	**B**	**L**	**C**	**Order of Power (Largest to Smallest)**
0.01	1:1:2	15, 15, 15	0.5572	0.3266	0.6365	C→B→L	0.3108	0.0525	0.2623	B→C→L
		30, 30, 30	0.9299	0.8225	0.9475	C→B→L	0.4062	0.1546	0.3988	B→C→L
		45, 45, 45	0.9899	0.9940	0.9940	C→B=L	0.5787	0.2665	0.2193	B→C→L
		15, 30, 45	0.9393	0.8351	0.4937	B→L→C	0.4842	0.1583	0.2548	B→C→L
	1:2:4	15, 15, 15	0.9796	0.7715	0.8985	B→C→L	0.8843	0.5417	0.7788	B→C→L
		30, 30, 30	1.0000	0.9985	0.9957	B→L→C	0.9848	0.9484	0.9723	B→C→L
		45, 45, 45	1.0000	1.0000	1.0000	B=L=C	0.9985	0.9977	0.9970	B→L→C
		15, 30, 45	0.9980	0.9925	0.6328	B→L→C	0.9801	0.9253	0.7536	B→L→C
0.05	1:1:2	15, 15, 15	0.7575	0.5857	0.8090	C→B→L	0.4834	0.1702	0.4208	B→C→L
		30, 30, 30	0.9472	0.9350	0.9830	C→B→L	0.6263	0.3341	0.5636	B→C→L
		45, 45, 45	0.9972	0.9929	0.9987	C→B→L	0.7308	0.4901	0.6749	B→C→L
		15, 30, 45	0.9867	0.9593	0.7171	B→L→C	0.6590	0.3663	0.4322	B→C→L
	1:2:4	15, 15, 15	0.9959	0.9476	0.9561	B→C→L	0.9249	0.7922	0.8800	B→L→C
		30, 30, 30	1.0000	1.0000	0.9995	B=L→C	0.9949	0.9881	0.9899	B→L→C
		45, 45, 45	1.0000	1.0000	1.0000	B=L=C	0.9996	0.9996	0.9992	B=L→C
		15, 30, 45	1.0000	0.9998	0.8341	B→L→C	0.9922	0.9859	0.9172	B→L→C

Conclusion

- 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. Consist with the power but not robustness. Power wise, performs well even under non-normal cases. The paper doesn't distinguish the two aspects for preference.

- When data are non-normally distributed, Levene's test is a good choice for equal and unequal variances based on robustness and not power.

- Based on power, If only one of the three variances is larger, the data are normally distributed, and are of equal sample sizes, Cochran's test would be recommended. This however in contrary to the paper that gives the same assertion but under unequal sample sizes

- According to (ISixSigma, 2001)
  - Test is efficient/powerful if it performs optimally under ideal circumstances.
  - Test is robust if it performs well under less-than-ideal circumstances.

  A trade off between two under most circumstances, robust tests are more often efficient/powerful than efficient tests are robust.

References

1 Bartlett, M. S. "Properties of Sufficiency and Statistical Tests." Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences (The Royal Society Publishing), May 1937: 268-282.

2 BOX, G. E. P. "Non-normality and Tests of Variances." Biometrika 40, no. 3-4 (December 1953): 318-335.

3 D.C, Montgomery. Design and Analysis of Experiments. 4th Edition. New York: John Wiley and Sons, 1997.

4 Hartley, H. O. "The Maximum F-Ratio as a Short-Cut Test for Heterogeneity of Variance." Oxford Journals (Oxford University Press on behalf of Biometrika Trust), December 1950: 308-312.

5 Levene, H. "Robust tests for equality of variances." Contributions to Probability and Statistic (Stanford University Press), 1960: 278-292.

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

7 Thavatchai, Vorapongsathorn,Taejaroenkul Sineenart,Chukiat Viwatwongkasem. "A comparison of type I error and power of Bartlett's test, Levene's test and Cochran's test under violation of assumptions." Journal of Science and Technology, 2004: 537-547.

8 Tjen-Sien Lim, Wei-Yin Loh,. "A comparison of tests of equality of variances." Computational Statistics & Data Analysis, 22, no. 3 (1996): 287-301.

9 W.G.COCHRAN. "Some consequences when the assumptions for the analysis of variance are not satisfied." Biometrics (International Biometric Society) 3, no. 1 (March 1947): 22-38.