Thailand Statistician
January 2010; 8(1) : 81-92
www.statassoc.or.th
Contributed paper

Confidence intervals for the variance and the ratio of two variances of non-normal distributions with missing data

Pawat Paksaranuwat and Sa-aat Niwitpong*
Department of Applied Statistics, Faculty of Applied Science,
King Monkut's University of Technology North Bangkok, Bangkok 10800, Thailand.
* **Author for correspondence**; e- mail: snw@kmutnb.ac.th

Received: 15 October 2009
Accepted: 11 January 2010

Abstract

This paper compares confidence intervals for the variance and the ratio of two variances when the population distributions are non-normal and item nonresponse is occurring. The data after random hot deck imputation used to define the confidence interval. The confidence intervals considered are the classical confidence intervals in text books and the adaptive confidence interval based on the Bonett confidence intervals. Our simulation study shows that the use of the adaptive confidence intervals for variance and ratio of two variances when the underlying distributions are generally skewed and unknown and missing data occur give better coverage probabilities. Therefore their use is recommended.

Keywords: confidence interval, coverage probability, item nonresponse, kurtosis, prior information, random hot deck imputation.

1. Introduction

Calculating confidence interval for the variance and the ratio of two variances is an important problem in manufacturing and quality management. Generally when sample distributions are normal the classical confidence intervals by χ^2 and F statistics in text books are used for the variance and the ratio of two variances respectively. However in this paper we are interested in non-normal distributions. Bonett [1] showed that when

the normal assumption is violated performance of the classical confidence interval for variance σ^2 is not acceptable. With nominal level 0.95 its coverage probability drops below 0.60 in some situations. For this case Bonett [1] proposed a confidence interval using a normal approximation to $\ln s^2$ and the kurtosis of the distribution γ_4. Simulation results showed that the Bonett confidence interval has better coverage than the classical confidence interval but its coverage probability is still less than nominal level for some non-normal distributions. The Bonett confidence interval has been adjusted by Niwitpong and Kirdwichai [2]. The adjusted confidence interval has coverage probability closes to the nominal level than the Bonett confidence interval. However that study results made with the assumption that the data is complete or not missing.

Item nonresponse for certain questions is a general missing data problem in sample surveys. Imputation methods are usually used for item nonresponse. Kalton and Brick [3] concluded that the advantages of using imputed data that can be used for internally consistent standard analysis or multivariate analysis. Qin et al.[4] proposed confidence intervals for marginal parameters such as mean distribution function or quantile under imputation for item nonresponse but they did not discuss about the important parameters such as the variance or the ratio of two variances. This paper proposes adaptive confidence intervals for the variance and the ratio of two variances for non-normal distributions with missing data. In this study we assumed data is missing completly at random (MCAR) and used random hot deck imputation to fill in missing data. We compared four confidence intervals: the classical and adaptive confidence intervals for the variance and for the ratio of two variances with missing data.

2. Confidence Intervals Considered

Let $\{x_i, \delta_i\}$ with $i = 1,...,n$ be random samples of incomplete data from population $\{X, \delta\}$ where $\delta_i = 0$ if x_i is missing, and $\delta_i = 1$ otherwise. Let $r = \sum_{i=1}^{n} \delta_i$ and $m = n - r$. We denote s_r to be the set of respondents with respect to x. Let x_i^* be the imputed values for the missing data with respect to x. Random hot deck imputation uses a simple random sample with replacement size m from s_r to fill in the missing data, i.e., $x_i^* = x_j$ for some $j \in s_r$. Let $x_{I,i} = \delta_i x_i + (1 - \delta_i) x_i^*$ be the complete data after imputation.

2.1 Classical Interval for Variance with Missing Data

Let s_I^2 be the sample variance for x_I. The Classical $100(1-\alpha)\%$ confidence interval for variance σ^2 with missing data is:

$$(n-1)s_I^2 / \chi^2_{1-\alpha/2,n-1} < \sigma^2 < (n-1)s_I^2 / \chi^2_{\alpha/2,n-1} \quad (1)$$

where $\chi^2_{p,k}$ is the $1-p$ quantile of the χ^2 distribution with degrees of freedom k.

2.2 Adaptive Interval for Variance with Missing Data

For adaptive confidence intervals for a single variance σ^2 with missing data we used complete data after random hot deck imputation to calculate an adjusted statistic t confidence interval of Niwitpong and Kirdwichai [2]. The adaptive $100(1-\alpha)\%$ confidence interval for variance σ^2 with missing data is defined by

$$\exp\left\{\ln\left(Rs_I^2\right) \pm t_{\alpha/2,n-1} se_I\right\} \quad (2)$$

where $t_{p,k}$ is the $1-p$ quantile of the t distribution with degrees of freedom k,

$$se_I = R\left[\left\{\hat{\gamma}^*_{4I}(n-3)/n\right\}/(n-1)\right]^{1/2}, \quad (3)$$

$R = n/\left(n - t_{\alpha/2,n-1}\right)$ is a small sample adjustment, $\hat{\gamma}^*_{4I} = \left(n_0\tilde{\gamma}_4 + n\gamma'_{4I}\right)/\left(n_0+n\right)$, $\tilde{\gamma}_4$ is a prior estimate of γ_4 obtained from a larger sample of size n_0 (we used $\hat{\gamma}^*_{4I} = \gamma'_{4I}$ if $\tilde{\gamma}_4$ is not available), and

$$\gamma'_{4I} = \frac{n\sum_{i=1}^{n}\left(x_{I,i} - med_I\right)^4}{\left(\sum_{i=1}^{n}\left(x_{I,i} - \overline{x}_I\right)^2\right)^2} \quad (4)$$

where med_I is median for x_I.

2.3 Classical Interval for Ratio of Two Variances with Missing Data

Let $x_{1I,i}$ and $x_{2I,j}$ with $i=1,...,n_1$ and $j=1,...,n_2$ be complete data after random hot deck imputation from populations X_1 and X_2 with variances σ_1^2 and σ_2^2 respectively. Let s_{1I}^2 and s_{2I}^2 be sample variances for x_{1I} and x_{2I} respectively. The Classical $100(1-\alpha)\%$ confidence interval for variance σ_1^2/σ_2^2 with missing data is:

$$\frac{s_{1I}^2}{s_{2I}^2} f_{\alpha/2,n2-1,n1-1} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_{1I}^2}{s_{2I}^2} f_{1-\alpha/2,n2-1,n1-1} \quad (5)$$

where f_{p,k_1,k_2} is the $1-p$ quantile of the F distribution with degrees of freedom k_1 and k_2 respectively.

2.4 Adaptive Interval for Ratio of Two Variances with Missing Data

Bonett [1] used a normal approximation of $\ln s^2$ to propose a confidence interval for a single variance from non-normal distributions. For adaptive confidence intervals for the ratio of two variances σ_1^2/σ_2^2 with nonnormal distributions and missing data we used a normal approximation of $\ln\left(s_1^2/s_2^2\right)$. The adaptive $100(1-\alpha)\%$ confidence interval for ratio of two variances σ_1^2/σ_2^2 with missing data is defined by

$$exp\left\{\ln\left(\frac{R_1 s_{I1}^2}{R_2 s_{I2}^2}\right) \pm z_{\alpha/2}\sqrt{se_{I1}^2 + se_{I2}^2}\right\} \quad (6)$$

where $R_k = n_k/(n_k - z_{\alpha/2})$ and se_{Ik}^2 is defined by (3) in subsection 2.2 respect to X_1 and X_2 with $k=1,2$.

3. Simulation Results

This section provides simulation studies for the coverage probabilities of confidence intervals as proposed in section 2 and the ratio of expected lengths of the adaptive and classical confidence intervals RE. The nominal level is 95%. If $RE<1$

that mean the adaptive confidence intervals are shorter than the classical confidences intervals. The results, based on 10,000 simulations, are written in R program.

For confidence intervals for variances the distributions we considers are normal, uniform, t, exponential and chi-squared. Sample sizes $n = 10, 25, 50, 100$ are used. We defined probability of item response is p (if $p = 1$ the data is not missing). Table 1 and 2 show that the adaptive confidence interval (2) has coverage probability closer to nominal level 0.95 than the classical confidence interval (1) for all distributions considered except the uniform distribution. They also show that the coverage probability of the adaptive confidence interval will be closer to the nominal level 0.95 when a prior estimate of γ_4 for large samples is used. Although the length of the classical confidence interval is lower than the adaptive confidence interval in most cases but their coverage probabilities are not acceptable.

Table 1: Estimated coverage probabilities and ratio of expected lengths of (2) and (1).

	n	p	(1)	(2)	RE
Normal(0,1)	10	0.9	0.9312	0.9782	1.6562
		0.8	0.9022	0.9641	1.6988
	25	0.9	0.9264	0.9724	1.3486
		0.8	0.8998	0.9580	1.3585
	50	0.9	0.9263	0.9727	1.2753
		0.8	0.9004	0.9554	1.2786
	100	0.9	0.9232	0.9698	1.2456
		0.8	0.9012	0.9545	1.2477
Uniform(0,1)	10	0.9	0.9840	0.9924	1.5019
		0.8	0.9657	0.9848	1.6030
	25	0.9	0.9904	0.9947	1.1446
		0.8	0.9842	0.9919	1.1744
	50	0.9	0.9924	0.9946	1.0415
		0.8	0.9852	0.9893	1.0537
	100	0.9	0.9943	0.9931	0.9946
		0.8	0.9884	0.9870	1.0009

	n	p	(1)	(2)	RE
t (5)	10	0.9	0.8456	0.9502	2.0639
		0.8	0.8093	0.9295	2.0535
	25	0.9	0.8109	0.9390	1.7742
		0.8	0.7685	0.9077	1.7538
	50	0.9	0.7808	0.9286	1.7349
		0.8	0.7431	0.9037	1.7370
	100	0.9	0.7516	0.9283	1.7569
		0.8	0.7209	0.9030	1.7414
Exponential(3)	10	0.9	0.7368	0.9070	3.0377
		0.8	0.7010	0.8745	2.9323
	25	0.9	0.6926	0.9121	2.4954
		0.8	0.6488	0.8850	2.4386
	50	0.9	0.6740	0.9305	2.4157
		0.8	0.6360	0.9048	2.3892
	100	0.9	0.6495	0.9466	2.4097
		0.8	0.6121	0.9177	2.3696
χ^2 (1)	10	0.9	0.6117	0.8640	3.9111
		0.8	0.5789	0.8229	3.7028
	25	0.9	0.5642	0.8881	3.3129
		0.8	0.5308	0.8576	3.1685
	50	0.9	0.5524	0.9122	3.1876
		0.8	0.5075	0.8824	3.0815
	100	0.9	0.5347	0.9295	3.1307
		0.8	0.4948	0.9090	3.0797

Table 2: Estimated coverage probabilities and ratio of expected lengths of (2) and (1) when prior kurtosis information is available.

	n_0	$\tilde{\gamma}_4$	n	p	(1)	(2)	RE
Uniform(0,1)	200	1.8	25	0.9	0.9908	0.9928	1.0523
				0.8	0.9844	0.9884	1.0541
			100	0.9	0.9944	0.9933	0.9755
				0.8	0.9871	0.9861	0.9768
	500	1.8	25	0.9	0.9917	0.9940	1.0635
				0.8	0.9815	0.9863	1.0664
			100	0.9	0.9940	0.9929	0.9819
				0.8	0.9880	0.9867	0.9842
t(5)	200	6.1	25	0.9	0.8077	0.9777	2.1678
				0.8	0.7651	0.9692	2.1660
			100	0.9	0.7899	0.9681	1.8387
				0.8	0.7175	0.9361	1.8278
	500	7.0	25	0.9	0.8055	0.9844	2.4026
				0.8	0.7699	0.9766	2.4003
			100	0.9	0.7574	0.9623	1.9537
				0.8	0.7153	0.9511	1.9550
Exponential(3)	200	7.9	25	0.9	0.6905	0.9676	2.6541
				0.8	0.6527	0.9509	2.6457
			100	0.9	0.6600	0.9528	2.2245
				0.8	0.6077	0.9282	2.2179
	500	8.5	25	0.9	0.6935	0.9713	2.7933
				0.8	0.6556	0.9574	2.7904
			100	0.9	0.6547	0.9509	2.2477
				0.8	0.6101	0.9342	2.2419
χ^2 (1)	200	12.2	25	0.9	0.5613	0.9534	3.7144
				0.8	0.5320	0.9388	3.7048
			100	0.9	0.5270	0.9355	2.8647
				0.8	0.4979	0.9123	2.8419
	500	13.6	25	0.9	0.5692	0.9641	4.0843
				0.8	0.5344	0.9529	4.0767
			100	0.9	0.5303	0.9483	2.9329
				0.8	0.4926	0.9233	2.9268

For the uniform distribution the adaptive confidence interval has coverage probability and expected length close to the classical confidence interval when a prior estimate of γ_4 for large samples is used. Therefore, the use of the adaptive confidence interval for variance when the distribution is unknown is recommended.

For confidence intervals for the ratio of two variances we consider cases when two populations X_1 and X_2 have the same and different distributions. Sample sizes $n_1 = n_2 = 20$ and $n_1 = 100, n_2 = 50$ are considered. We define $p_1, p_2, n_{10}, n_{20}, \tilde{\gamma}_{10}$ and $\tilde{\gamma}_{20}$ as probability of item response of sample unit, prior sample sizes and kurtosis information with respect to X_1 and X_2 respectively. Tables 3 and 4 show that when the two distributions are normal the adaptive confidence interval for the ratio of two variances (6) is wider than the classical confidence interval (5) if no data are missing ($p_1 = p_2 = 1.0$) but when there is missing data the adaptive confidence interval will have coverage probability closer to nominal level 0.95 than the classical confidence interval. When either or both distributions are non-normal, except the uniform distribution, the adaptive confidence interval has coverage probability closer to nominal level 0.95 than the classical confidence interval whether there is missing data or not. They also show that the coverage probability of adaptive confidence interval will be closer to the nominal level 0.95 when a prior estimate of kurtosis for large samples n_{10} and n_{20} are used.

When both distributions are uniform distributions the adaptive confidence interval has coverage probability close to the classical confidence interval but in large sample sizes $n_1 = 100, n_2 = 50$ the adaptive confidence interval are shorter than the classical confidence interval.

4. Conclusions

In this paper we compare the efficiency of the classical confidence intervals and adaptive confidence intervals for the variance and the ratio of two variances when there is missing data and the samples come form non-normal distributions. The simulation results show that the adaptive confidence interval with imputed data by random hot deck method has higher efficiency than the classical confidence interval based on the χ^2 statistic which coincides with results of Niwitpong and Kirdwichai [2] for complete data.

For two populations when two data are missing the adaptive confidence interval based on Bonett confidence interval have higher efficiency than the classical confidence

interval base on the F statistic interval in all distributions considered except the uniform distribution. They also show that the adaptive confidence interval has coverage probability close to the classical confidence interval and is shorter with large sample sizes. Therefore, the use of the adaptive confidence intervals when the underlying distributions are generally skewed and unknown and missing data occur give better coverage probabilities.

Table 3: Estimated coverage probabilities and ratio of expected lengths of (6) and (5).

			(5)	(6)	RE
Normal(0,1)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.9499	0.9769	1.3082
		$n_1=100, n_2=50$	0.9499	0.9840	1.1980
	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.9105	0.9553	1.3335
		$n_1=100, n_2=50$	0.9114	0.9608	1.2000
Uniform(0,1)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.9946	0.9970	1.0906
		$n_1=100, n_2=50$	0.9959	0.9969	0.9591
	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.9823	0.9898	1.1480
		$n_1=100, n_2=50$	0.9900	0.9928	0.9722
t(5)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.8399	0.9510	1.6783
		$n_1=100, n_2=50$	0.7884	0.9544	1.6026
	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.7927	0.9190	1.6435
		$n_1=100, n_2=50$	0.7579	0.9288	1.5671
Exponential(3)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.7402	0.9498	2.5037
		$n_1=100, n_2=50$	0.6941	0.9646	2.2872
	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.6892	0.9095	2.3773
		$n_1=100, n_2=50$	0.6447	0.9390	2.2416
χ^2 (1)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.6155	0.9222	3.4927
		$n_1=100, n_2=50$	0.5689	0.9545	3.0580
	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.5657	0.8845	3.2191
		$n_1=100, n_2=50$	0.5266	0.9282	2.9790
X_1 is Normal(0,1) X_2 is χ^2 (1)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.7263	0.9181	2.1040
		$n_1=100, n_2=50$	0.6466	0.9395	2.2977
	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.6887	0.8931	2.0711
		$n_1=100, n_2=50$	0.6071	0.9131	2.2576

			(5)	(6)	RE
X_1 is Exponential(3)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.6657	0.9302	2.8912
X_2 is χ^2 (1)		$n_1=100, n_2=50$	0.6028	0.9500	2.7131
	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.6167	0.8994	2.7690
		$n_1=100, n_2=50$	0.5631	0.9258	2.6638

Table 4: Estimated coverage probabilities and ratio of expected lengths of (6) and (5) when prior kurtosis information is available.

			(5)	(6)	RE
Uniform(0,1)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.9944	0.9946	1.0052
$n_{10}=200, \tilde{\gamma}_{10}=1.8$		$n_1=100, n_2=50$	0.9954	0.9956	0.9302
$n_{20}=500, \tilde{\gamma}_{20}=1.8$	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.9833	0.9843	1.0095
		$n_1=100, n_2=50$	0.9915	0.9900	0.9328
t(5)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.8305	0.9871	2.5970
$n_{10}=200, \tilde{\gamma}_{10}=6.1$		$n_1=100, n_2=50$	0.7910	0.9750	1.9366
$n_{20}=500, \tilde{\gamma}_{20}=7.0$	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.7908	0.9776	2.5938
		$n_1=100, n_2=50$	0.7493	0.9600	1.9369
Exponential(3)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.7283	0.9796	3.2690
$n_{10}=200, \tilde{\gamma}_{10}=7.9$		$n_1=100, n_2=50$	0.7055	0.9694	2.2961
$n_{20}=500, \tilde{\gamma}_{20}=8.5$	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.6839	0.9624	3.2567
		$n_1=100, n_2=50$	0.6528	0.9503	2.2895
χ^2 (1)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.6046	0.9745	5.4144
$n_{10}=200, \tilde{\gamma}_{10}=12.2$		$n_1=100, n_2=50$	0.5693	0.9614	3.1693
$n_{20}=500, \tilde{\gamma}_{20}=13.6$	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.5550	0.9563	5.3876
		$n_1=100, n_2=50$	0.5281	0.9424	3.1477
X_1 is Exponential(3)	$p_1=p_2=1.0$	$n_1=n_2=20$	0.6591	0.9762	4.3645
X_2 is χ^2 (1)		$n_1=100, n_2=50$	0.6057	0.9630	2.8938
$n_{10}=200, \tilde{\gamma}_{10}=7.9$	$p_1=0.8, p_2=0.9$	$n_1=n_2=20$	0.6300	0.9613	4.3491
$n_{20}=500, \tilde{\gamma}_{20}=13.6$		$n_1=100, n_2=50$	0.5707	0.9457	2.8821

5. Numerical Example

Suppose a manufacturer takes random sample size $n_1 = n_2 = 10$ items of his product and competitor product respectively. But the performance values of the competitor have missing in sampling survey. Therefore random hot deck imputation is used to handle. The resulting performance values are generated from χ^2 (1) and χ^2 (2) respectively :

Manufacturer's Product	Competitor's Product	imputed data
2.13	0.91	0.91
1.57	0.99	0.99
0.28	0.90	0.90
0.38	3.62	3.62
0.56	0.28	0.28
0.01	-	0.43
1.24	4.38	4.38
0.02	0.43	0.43
0.01	1.81	1.81
1.61	2.23	2.23

The sample variances and the small-sample adjustments are $s_1^2 = 0.62,$ $s_2^2 = 2.01$ and $R_1 = R_2 = \frac{10}{10 - z_{0.025}} = 1.244$. The pooled estimate of $\mathrm{var}\left(\ln s_1^2\right)$ and $\mathrm{var}\left(\ln s_2^2\right)$ are $se_1^2 = 0.66^2$ and $se_2^2 = 0.84^2$. From equation (6) the 95% adaptive confidence interval for ratio of variances of performance values from two products is

$$exp\left\{\ln\left(\frac{0.62}{2.01}\right) \pm 1.96\sqrt{0.66^2 + 0.84^2}\right\} = \left[0.038, 2.513\right]$$

Acknowledgements

The authors thank the anonymous referees for their comments that resulted in an improved present of the paper.

References

[1] Bonett, D., Approximate confidence interval for standard deviation of nonnormal distributions. *Comput Statist Data Anal.*, 2006;50:775-782.

[2] Niwitpong, S., and Kirdwichai, P., Adjusted Bonett Confidence Interval for Standard Deviation of Non-normal Distributions. *Thail Stat.*, 2008;6:1-6.

[3] Kalton, G., and Brick, J.M., Handing missing data in survey research. *Stat Meth Med Res.*, 1996;5:215-238.

[4] Qin, Y., Rao, J.N.K. and Ren, Q., Confidence intervals for marginal parameters under imputation for item nonresponse. *J Stat Pl Inf.*, 2008;99:1232-1259.