



Thailand Statistician
July 2009; 7(2) : 201-209
www.statassoc.or.th
Contributed paper

Confidence Intervals for the Differences between the Means Using Jackknife Method after ANOVA for Data with Outliers

Jeerapa Sappakitkamjorn and Sa-aat Niwitpong*

Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand.

* **Author for correspondence**; e-mail: snw@kmutnb.ac.th

Received: 8 February 2009

Accepted: 14 July 2009

Abstract

This study is a continuing investigation into the performance of confidence intervals for the differences between pairs of means based on the jackknife method for multiple comparisons after the ANOVA F-test. The main focus was to examine the jackknife based confidence intervals in detecting the differences between the means in multiple comparisons tests, especially when datasets contain outliers. A simulation with 5,000 runs each under various situations was conducted to investigate the performance of the jackknife based intervals. As equal sample sizes were considered in this study, we compared the performance of the jackknife based intervals with those obtained by the Tukey's HSD (Honesty Significant Difference), a method that allows the comparison of all pairs of means, in terms of coverage probability and expected length. The results from a simulation study show that the jackknife based intervals enable better detection of differences between the means when outliers are present in the datasets. The intervals constructed by the jackknife method not only have shorter lengths but also offer control over the probability of a Type I error.

Keywords : ANOVA, confidence Interval, jackknife method, multiple comparisons, outliers.

1. Introduction

In the presence of outliers, one should not simply remove them from a dataset unless there is a proof that they are false values. Outliers are often true values that are important since they contain valuable information [1]. However, taking outliers into account, particularly in the one-way ANOVA, may cause unequal variances and non-normality. As a result, assumptions to be satisfied are violated when ANOVA is applied. In order to protect against the effects of these violations, one should apply balanced designs where sample sizes are equal.

As reviewed in [2], when a balanced design is applied, the ANOVA F-test is relatively resistant to violations of assumptions. Similarly, the multiple comparisons tests after the rejection of the null hypothesis in the one-way ANOVA will be resistant to those violations. There are several powerful multiple comparison procedures available, such as Tukey's HSD, Scheffé's and Bonferroni methods, see e.g. Lapin [3]. The Tukey's HSD is used in this study because it allows the comparison of all pairs of means when used with equal sample sizes.

According to our previous studies [4,5], the confidence intervals constructed by the jackknife method perform well in detecting true differences between the means in multiple comparisons tests. However, outliers were not considered in those studies. Therefore, it is of interest to further investigate the accuracy of the jackknife method for making paired comparisons in the presence of outliers.

The outline of this paper is as follows: in Section 2, a formula for a $(1 - \alpha)100\%$ confidence interval for the difference between two means based on the jackknife method is given. In section 3, details of the simulation study are presented. The results and conclusion are given in Section 3 and Section 4 respectively.

2. The Jackknife based confidence interval

Following our preliminary studies that proposed a confidence interval and an adjusted confidence interval based on the jackknife method [4,5], the $(1 - \alpha)100\%$ confidence interval for the difference between two means, when all sample sizes are equals, is given by

$$(\bar{x}_{i(JK)} - \bar{x}_{j(JK)}) \pm 4t_{\alpha/2, N-r} \sqrt{MSE_{(JK)} \left(\frac{2}{n} \right)} \quad (1)$$

$$\text{where } MSE_{(JK)} = \frac{1}{r} \sum_{h=1}^r Var_{(JK)}(\bar{x}_h) \quad (2)$$

$$Var_{(JK)}(\bar{x}_h) = \frac{(n-1)}{n} \sum_{k=1}^n \left(\bar{x}_{h(k)} - \bar{x}_{h(.)} \right)^2 \quad (3)$$

$$\bar{x}_{h(.)} = \frac{1}{n} \sum_{k=1}^n \bar{x}_{h(k)} \quad ; k = 1, 2, \dots, n \quad (4)$$

Here $MSE_{(JK)}$ is the jackknife pooled error variance (see [6] for more details), $\bar{x}_{h(k)}$ is the mean of sample h obtained by eliminating the k^{th} observation, where $\bar{x}_{i(JK)}$ and $\bar{x}_{j(JK)}$ are jackknife means of samples i and j . N and r denote total number of observations and number of groups respectively. A constant 4 in equation (1) is an adjusting constant used to improve the jackknife based interval proposed by Sappakitkamjorn et al. [4] in order to attain the nominal confidence level of $(1 - \alpha)100\%$. The confidence interval for the difference between two means by the Tukey method can be found in various textbooks (e.g., Lapin [3]). Refer to them for details.

3. Simulation study

In order to assess the accuracy of the jackknife based confidence intervals in comparison with those of the widely used Tukey's HSD pairwise comparison procedure, a simulation study was conducted with 5,000 repetitions for each case under consideration. The focus in the simulation study was on the case of the presence of outliers in datasets where mild and extreme outliers were considered. A box-and-whisker plot or boxplot was used to identify which observations would be considered as mild or extreme outliers. That is an observation would be labeled as a mild outlier if its value lies either in the range of $(Q_3 + 1.5 * IQR, Q_3 + 3 * IQR)$ or $(Q_1 - 3 * IQR, Q_1 - 1.5 * IQR)$. The observation would be labeled as an extreme outlier if its value lies beyond $Q_3 + 3 * IQR$ or $Q_1 - 3 * IQR$, where Q_1 , Q_3 and IQR are the first quartile, third quartile and interquartile range respectively.

For simplicity, the case of 3 population means was considered. We generated 3 samples of equal sizes ($n_1 = n_2 = n_3 = n = \{10, 20, 30, 40, 50\}$) from normal populations with parameters $\mu_1 = \mu_2 = 10$, $\mu_3 = 15$ and equal error variances ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2 = \{1.25, 2.50, 3.75\}$). The values of σ were chosen in such a way that produces a coefficient of variation of 0.1, 0.2, 0.3. Let λ and K denote the percentage of outliers and the number of outliers respectively. In this study, λ was chosen as 0.1, 0.2 and 0.3.

As discussed by Wellmann and Gather [7], although contamination models are frequently used to generate outliers in univariate statistical data, the fraction of outliers produced by the models is not only low but also the models do not produce sufficient extreme outliers. Therefore, contamination models were not used to generate the datasets in this study. To generate a sample of size n with K outliers, $n - K$ regular observations were generated from normal distribution with the predefined parameters and then the data were used to calculate the values of Q_1, Q_3 and IQR . Next, mild outliers were randomly taken from either $[Q_3 + 1.55 * IQR, Q_3 + 3 * IQR]$ or

$[Q_1 - 3 * IQR, Q_1 - 1.55 * IQR]$ while extreme outliers were taken from $[Q_3 + 3.5 * IQR, Q_3 + 7 * IQR]$ or $[Q_1 - 7 * IQR, Q_1 - 3.5 * IQR]$.

To assure that a generated dataset contains the required number of outliers, a box-plot is used to examine if a dataset contains K outliers before proceeding to the next step. For example, box-plots of simulated data with mild and extreme outliers where $n = 30$, $\mu = 15$, $\sigma = 1.25$ and the percentage of outliers are 0.1, 0.2, 0.3 are shown in Figures 1 and 2 respectively.

After having obtained the 3 datasets of equal size n with K mild or extreme outliers described previously, the ANOVA F-test was applied to examine the differences among sample means. After the rejection of the null hypothesis in the ANOVA F-test, 95% confidence intervals for all possible pairs of the different means were obtained via the Tukey and jackknife methods. Then coverage probabilities and expected lengths of confidence intervals for each pair were evaluated.

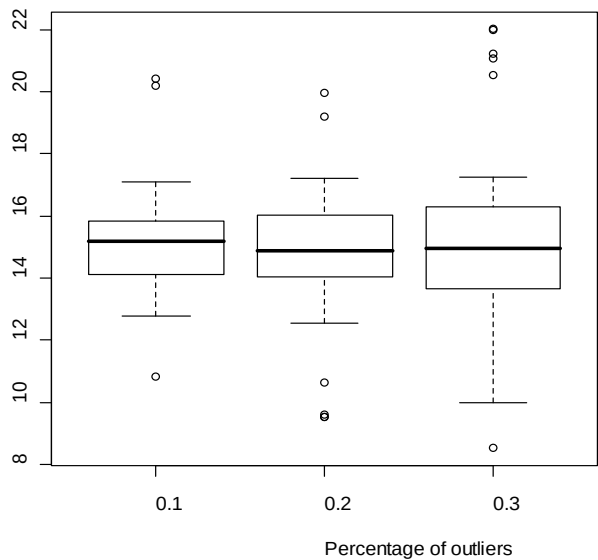


Figure 1. Box-plot of simulated data with mild outliers where $n = 30$, $\mu = 15$, $\sigma = 1.25$ and the percentage of outliers are 0.1, 0.2, 0.3.

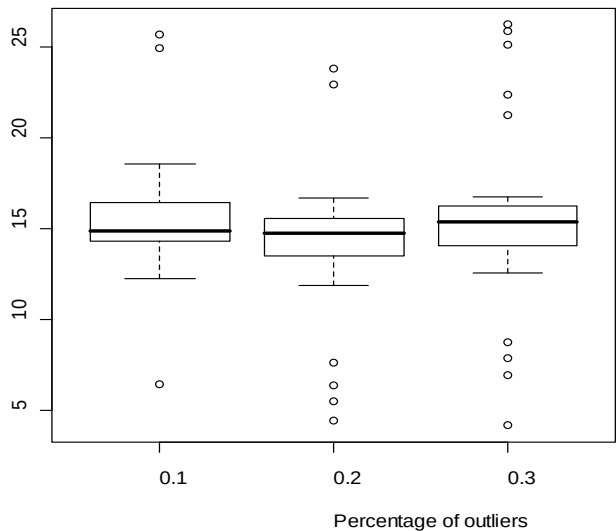


Figure 2. Box-plot of simulated data with extreme outliers where $n = 30$, $\mu = 15$, $\sigma = 1.25$ and the percentage of outliers are 0.1, 0.2, 0.3.

4. Results

The results from our investigation are partly summarized in Table 1. The table shows the coverage probabilities and expected lengths of nominal 95% confidence intervals for three pairs of the different means obtained by the Tukey and jackknife methods. Each pair respectively represents the difference between μ_1 vs. μ_2 denoted as Pair 1, μ_1 vs. μ_3 denoted as Pair 2 and μ_2 vs. μ_3 denoted as Pair 3. It can be seen from Table 1, the jackknife method provides either equal or higher coverage probabilities for Pair 2 where $\mu_1 \neq \mu_3$ and Pair 3 where $\mu_2 \neq \mu_3$. On the contrary, the jackknife method gives lower coverage probabilities than the Tukey method for Pair 1 where $\mu_1 = \mu_2$. Note that the coverage probabilities for both methods tend to be smaller when outliers become extreme and variances of datasets are high. At a given variance the coverage probabilities decrease upon an increase in the percentage of outliers. By comparing the expected lengths between the two methods as shown in the last two columns, it is clear that the jackknife based intervals are shorter than those obtained by the Tukey method for all cases under consideration. Note that the expected lengths for all three pairs for a given method are equal.

5. Conclusion

The study was carried out to investigate the performance of the jackknife based intervals for detecting the differences between the means after the ANOVA F-test in the presence of outliers. As seen in Table 1, the results suggest that both methods are robust to mild outliers with the jackknife method performing better than the Tukey method particularly for Pair 2 and Pair 3 where the population means are truly different. For Pair 1 where the population means are equal, however, the jackknife method performs adversely. Both methods offer control over the probability of a Type I error or attain coverage probabilities higher than the nominal level of 0.95 when outliers are mild and variances of datasets are low. The jackknife method is not only more accurate but also gives shorter intervals.

Overall, the power of the jackknife method, meaning its probability of detecting true differences between the means competes well with the Tukey method. Nevertheless, when outliers are extreme and variances of datasets are high, the power of two methods can be relatively low.

Further research is to examine the confidence intervals for these paired means using ANOVA with non-normally distributed data particularly when outliers do not occur

randomly. In addition, the other multiple comparison procedures, such as Scheffé's and Bonferroni methods could be taken into account.

Table 1. Coverage probabilities and expected lengths of nominal 95% confidence intervals for the difference of means by the Tukey (TK) and Jackknife (JK) methods.

<i>n</i>	Degree of outliers	% of outliers	σ	Coverage probability						Expected length	
				Pair 1		Pair 2		Pair 3		TK	JK
				TK	JK	TK	JK	TK	JK		
10	Mild	0.1	1.25	0.986	0.967	0.996	0.999	0.999	0.999	3.975	3.454
			2.50	0.852	0.832	0.722	0.800	0.736	0.789	6.650	5.778
			3.75	0.516	0.506	0.371	0.439	0.364	0.419	6.001	5.214
		0.2	1.25	0.966	0.936	0.974	0.988	0.985	0.988	4.689	4.074
			2.50	0.700	0.668	0.569	0.632	0.543	0.624	6.754	5.869
			3.75	0.386	0.359	0.269	0.316	0.266	0.313	5.441	4.728
		0.3	1.25	0.950	0.924	0.943	0.963	0.946	0.968	5.327	4.629
			2.50	0.596	0.563	0.459	0.532	0.459	0.515	6.288	5.464
			3.75	0.342	0.318	0.226	0.265	0.250	0.292	5.716	4.967
	Extreme	0.1	1.25	0.923	0.903	0.846	0.889	0.840	0.878	6.060	5.265
			2.50	0.468	0.460	0.316	0.383	0.326	0.369	5.791	5.032
			3.75	0.244	0.238	0.152	0.182	0.143	0.174	4.463	3.878
		0.2	1.25	0.759	0.732	0.619	0.693	0.638	0.697	6.533	5.677
			2.50	0.288	0.269	0.197	0.235	0.186	0.222	4.826	4.193
			3.75	0.142	0.137	0.076	0.098	0.088	0.105	3.787	3.291
		0.3	1.25	0.594	0.577	0.463	0.521	0.457	0.517	5.973	5.190
			2.50	0.184	0.172	0.119	0.146	0.130	0.153	3.849	3.345
			3.75	0.100	0.094	0.058	0.078	0.064	0.079	3.523	3.061
20	Mild	0.1	1.25	0.987	0.961	1	1	1	1	2.811	2.470
			2.50	0.970	0.937	0.941	0.959	0.948	0.970	5.563	4.889
			3.75	0.801	0.775	0.653	0.717	0.652	0.728	6.781	5.960
		0.2	1.25	0.974	0.949	1	1	1	1	3.434	3.018
			2.50	0.911	0.889	0.827	0.875	0.851	0.891	6.363	5.592
			3.75	0.615	0.597	0.472	0.536	0.464	0.529	6.504	5.716
		0.3	1.25	0.972	0.950	0.997	0.998	0.998	0.999	4.011	3.526
			2.50	0.805	0.776	0.693	0.759	0.687	0.756	6.653	5.847
			3.75	0.499	0.483	0.359	0.420	0.366	0.426	6.197	5.446
	Extreme	0.1	1.25	0.981	0.959	0.987	0.992	0.985	0.992	4.832	4.247
			2.50	0.689	0.670	0.532	0.604	0.533	0.609	6.614	5.813
			3.75	0.397	0.374	0.281	0.325	0.240	0.301	5.717	5.024
		0.2	1.25	0.942	0.923	0.881	0.917	0.889	0.915	6.157	5.411
			2.50	0.443	0.422	0.323	0.363	0.311	0.353	5.821	5.116
			3.75	0.211	0.200	0.136	0.158	0.129	0.154	4.217	3.706
		0.3	1.25	0.825	0.809	0.739	0.798	0.719	0.772	6.547	5.754
			2.50	0.322	0.304	0.213	0.250	0.202	0.246	4.939	4.340
			3.75	0.152	0.139	0.105	0.125	0.084	0.108	3.884	3.413

Table 1. (continued)

n	Degree of outliers	% of outliers	σ	Coverage probability						Expected length	
				Pair 1		Pair 2		Pair 3			
				TK	JK	TK	JK	TK	JK	TK	JK
30	Mild	0.1	1.25	0.976	0.950	1	1	1	1	2.294	2.023
			2.50	0.975	0.945	0.992	0.995	0.992	0.994	4.570	4.030
			3.75	0.918	0.893	0.835	0.878	0.807	0.869	6.483	5.718
		0.2	1.25	0.970	0.948	1	1	1	1	2.829	2.495
			2.50	0.955	0.927	0.939	0.963	0.950	0.966	5.651	4.984
			3.75	0.776	0.755	0.646	0.708	0.655	0.715	6.806	6.003
		0.3	1.25	0.964	0.948	1	1	1	1	3.307	2.917
			2.50	0.918	0.895	0.852	0.883	0.877	0.903	6.163	5.435
			3.75	0.638	0.612	0.487	0.560	0.517	0.577	6.604	5.824
	Extreme	0.1	1.25	0.978	0.958	0.999	1	1	1	3.971	3.502
			2.50	0.849	0.824	0.734	0.783	0.726	0.784	6.775	5.976
			3.75	0.504	0.491	0.356	0.413	0.358	0.411	6.033	5.321
		0.2	1.25	0.972	0.943	0.967	0.975	0.970	0.983	5.319	4.691
			2.50	0.623	0.605	0.471	0.542	0.477	0.545	6.807	6.003
			3.75	0.328	0.316	0.213	0.254	0.219	0.260	5.322	4.694
		0.3	1.25	0.933	0.920	0.885	0.915	0.874	0.917	6.112	5.390
			2.50	0.462	0.446	0.319	0.372	0.326	0.371	6.012	5.302
			3.75	0.209	0.193	0.130	0.153	0.131	0.163	4.261	3.758
40	Mild	0.1	1.25	0.980	0.964	1	1	1	1	1.986	1.755
			2.50	0.973	0.946	1	1	0.999	1	3.976	3.512
			3.75	0.959	0.935	0.923	0.951	0.927	0.947	5.826	5.147
		0.2	1.25	0.977	0.952	1	1	1	1	2.493	2.202
			2.50	0.970	0.945	0.986	0.992	0.992	0.997	4.969	4.390
			3.75	0.877	0.851	0.762	0.814	0.773	0.832	6.639	5.865
		0.3	1.25	0.973	0.940	1	1	1	1	2.866	2.532
			2.50	0.956	0.940	0.935	0.954	0.938	0.959	5.618	4.963
			3.75	0.749	0.724	0.612	0.670	0.632	0.688	6.652	5.877
	Extreme	0.1	1.25	0.980	0.964	1	1	1	1	3.461	3.057
			2.50	0.925	0.900	0.845	0.895	0.837	0.889	6.496	5.739
			3.75	0.634	0.612	0.463	0.525	0.485	0.547	6.760	5.972
		0.2	1.25	0.977	0.954	0.995	0.997	0.994	0.998	4.650	4.108
			2.50	0.698	0.681	0.545	0.612	0.552	0.615	6.636	5.863
			3.75	0.390	0.372	0.249	0.307	0.259	0.305	5.572	4.922
		0.3	1.25	0.960	0.940	0.938	0.955	0.948	0.964	5.520	4.876
			2.50	0.568	0.552	0.387	0.465	0.419	0.481	6.442	5.691
			3.75	0.256	0.246	0.172	0.198	0.158	0.192	4.547	4.017
50	Mild	0.1	1.25	0.978	0.953	1	1	1	1	1.776	1.570
			2.50	0.979	0.951	1	1	1	1	3.569	3.156
			3.75	0.965	0.936	0.974	0.984	0.976	0.988	5.302	4.689
		0.2	1.25	0.976	0.946	1	1	1	1	2.220	1.963
			2.50	0.978	0.953	0.994	0.996	0.997	0.999	4.433	3.920
			3.75	0.925	0.898	0.854	0.894	0.854	0.894	6.294	5.566
		0.3	1.25	0.974	0.950	1	1	1	1	2.587	2.288
			2.50	0.957	0.935	0.968	0.982	0.980	0.986	5.126	4.533
			3.75	0.835	0.815	0.750	0.792	0.710	0.766	6.620	5.854
	Extreme	0.1	1.25	0.980	0.956	1	1	1	1	3.088	2.731
			2.50	0.960	0.935	0.908	0.936	0.922	0.945	6.053	5.353
			3.75	0.746	0.719	0.588	0.661	0.603	0.664	7.012	6.201
		0.2	1.25	0.965	0.951	0.998	0.999	1	1	4.186	3.702
			2.50	0.824	0.809	0.693	0.755	0.674	0.743	7.080	6.261
			3.75	0.488	0.466	0.333	0.378	0.339	0.390	6.262	5.537
		0.3	1.25	0.973	0.952	0.980	0.987	0.977	0.986	5.032	4.450
			2.50	0.637	0.617	0.486	0.546	0.475	0.544	6.513	5.759
			3.75	0.319	0.304	0.198	0.244	0.213	0.254	5.027	4.445

Acknowledgements

This research was supported by the 2007 general researcher fund from King Mongkut's University of Technology North Bangkok. The authors would like to thank Dr. Gareth Clayton for his editorial comments and anonymous referees for valuable suggestions that led to an improvement of the paper.

References

- [1] Goupy, J. Outliers and Experimental designs. *Chemometrics and Intelligent Laboratory Systems*, 1996; 35: 145-156.
- [2] Kao, L.S. and Green, C. E. Analysis of Variance: Is there a difference in Means and What Does It Mean? *Journal of Surgical Research*, 2008; 144: 158-170.
- [3] Lapin, L.L. Probability and Statistics for Modern Engineering. 2nd ed. Belmont: Duxbury Press, 1990.
- [4] Sappakitkamjorn, J, Niwitpong, S. and Nangsue, N. Confidence Intervals for the Different Means Following ANOVA F-Test Using Jackknife. Proceedings of the 26th CASI, May 17-19. Kerry, Ireland, 2006; 150-151.
- [5] Sappakitkamjorn, J. and Niwitpong, S. Adjusted confidence intervals based on jackknife method for the differences between means. Proceedings of Statistics 2007 – Statistics under one umbrella, March 27-30. Bielefeld, Germany, 2007; 265.
- [6] Variyath, A.M. and Abraham, B. Analysis of Performance Measures in Experimental Designs Using Jackknife. *Journal of Quality Technology*, 2005; 37: 91-100.
- [7] Wellmann, J., and Gather, U. A Note on Contamination Models and Outliers. *Communications in Statistics A*, 1999; 28: 1-9.