Pergamon

Computers ind. Engng Vol. 31, No. 1/2, pp. 265 – 268, 1996

0360- 8352/96 $15.00 + 0.00

S0360-8352(96) 00127-1

Robustness of the ANOVA and Tukey-Kramer Statistical Tests

Wade C. Driscoll

Department of Industrial and Systems Engineering, Youngstown State University
Youngstown, OH 44555

ABSTRACT

Analysis of variance (ANOVA) allows one to compare the means $\mu_1, \mu_2, \ldots, \mu_m$ of m populations. For a randomized block single-factor design of experiments with m treatments, the analyst uses ANOVA to test the null hypothesis $H_0 : \{\mu_1 = \mu_2 = \ldots = \mu_m\}$. If H_0 is rejected, the Tukey-Kramer procedure may be used to determine which pairs of means have statistically significant differences. Unfortunately, ANOVA and the Tukey-Kramer procedures both depend upon the use of data from normal populations. Departure from normality may cause errors. These errors cannot be determined analytically; however, they can be estimated via simulation. This paper reports on the use of computer simulation to estimate the errors that attend the use of ANOVA with Tukey-Kramer to test for differences of population means.

KEYWORDS

ANOVA; Robustness; Simulation; Tukey-Kramer

INTRODUCTION

The analysis of variance (ANOVA) and Tukey-Kramer procedures* apply to the analysis of industrial engineering and other data. For example, suppose that an industrial engineer has data available on a quality characteristic of interest for each of m similar machines on each of n days. The machines are known to have been in statistical quality control. He/she observes variations in quality characteristics from one machine to another, and wishes to determine if any machines have a significantly different quality characteristic. One may use a two-sample t test to test for statistically significant differences between the means of any single pair of machines. However, the Type I error—α, the probability of rejecting a good null hypothesis—becomes confounded** if multiple comparisons are made between several pairs of machines. Both the widely-known ANOVA and the lesser-known Tukey-Kramer procedures simultaneously test for equality of all means. In contrast to ANOVA, the Tukey-Kramer procedure also identifies which means are significantly different from one another.

Unfortunately, the theoretical developments of both procedures are based on the assumption that the data are random samples from normal probability distributions. In practice it is seldom known if this is true. Furthermore, data may not be available to support the performance of a Chi-Squared test for goodness of fit, and it is frequently not cost-effective to gather more data. Consequently, the analyst must often use procedures rooted in the assumption of normality on data which may not be normally distributed.

This paper reports the results of a series of simulation experiments which were performed to investigate the errors that attend using these statistical procedures on nonnormal data. The rest of the paper is organized as follows. The next section provides a brief overview of the framework within which the statistical procedures are to be applied, and introduces the notation required to describe the investigation that was conducted. That is followed by two sections which describe the series of simulation experiments that were performed, and report the results of the experiments. The final section summarizes the results and draws conclusions from them.

---

* Many engineering statistics textbooks, both classical [Bowker and Leiberman (1972); Brownlee (1965)] and contemporary [Barnes (1994); Walpole and Myers (1993)], describe these procedures.

** For example, Barnes (1994) has an example illustrating the propagation of the Type I error for multiple comparisons of treatment means.

STATISTICAL PROCEDURES INVESTIGATED

The analyst has a set of n data points X_{ij} $(j = 1, \ldots, n)$ available for each* of m treatments or conditions $(i = 1, \ldots, m)$. E.g., for the example cited earlier, X_{ij} would denote the value of the quality characteristic of interest for machine i on day j. The theoretical underpinnings of both the ANOVA and Tukey-Kramer procedures are met when the data from the i^{th} treatment are independent random samples from a $N(\mu_i, \sigma)$ population, where $N(\mu_i, \sigma)$ denotes a normal (or, Gaussian) population with mean μ_i and standard deviation σ. The data may be summarized in a table such as:

Factor	Replication 1	Replication 2	...	Replication n	Sample Average
Treatment 1	X_{11}	X_{12}	...	X_{1n}	$\bar{X}_{1\bullet}$
Treatment 2	X_{21}	X_{22}	...	X_{2n}	$\bar{X}_{2\bullet}$
...	...	...	...	...	...
Treatment m	X_{m1}	X_{m2}	...	X_{mn}	$\bar{X}_{m\bullet}$

The equation $\bar{X}_{i\bullet} = \sum_{j=1}^{n} X_{ij}/n$ defines the i^{th} sample average. One computes the overall average from $\bar{X}_{\bullet\bullet} = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}/(mn)$. The sum of squares SS_{WITHIN}, computed from $SS_{WITHIN} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_{i\bullet}\right)^2$, can be divided by its number of degrees of freedom $m(n-1)$ to form a pooled estimate $S_P^2 = SS_{WITHIN}/(mn-m)$ of the common variance σ^2 of the random variables. Similarly, $SS_{BETWEEN} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right)^2 = n\sum_{i=1}^{m}\left(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}\right)^2$ has $m-1$ degrees of freedom. If all treatment means are equal $(\mu_1 = \mu_2 = \ldots = \mu_m = \mu)$, then $S_B^2 = SS_{BETWEEN}/(m-1)$ is also an estimate of the common variance.

ANOVA requires the comparison of the two estimates of the variance. One compares the statistic $F(m-1, mn-m) = S_B^2/S_P^2$, which has an F distribution with $m-1$ and $m(n-1)$ degrees of freedom if each $X_{ij} = N(\mu, \sigma)$, to table values of probabilities for the F distribution. If an unreasonably large value of F is obtained, the null hypothesis H_0: $\{\mu_1 = \mu_2 = \ldots = \mu_m = \mu\}$ is rejected. Note that H_0 can be rejected because (i) the X_{ij}s are not normally distributed, or (ii) their variances are unequal, or (iii) H_0 is not true.

The analysis of variance reveals whether or not to accept the null hypothesis that the treatment means are equal. Unfortunately, it does not indicate which means are different when the null hypothesis is rejected. The Tukey-Kramer procedure overcomes this deficiency by simultaneously testing the $m(m-1)/2$ null hypotheses that $\mu_{i_1} = \mu_{i_2}$ for all $i_1 = 1, \ldots, m$, and $i_2 = 1, \ldots, m$, where $i_1 \neq i_2$. In applying the Tukey-Kramer procedure for a given pair i_1 and i_2 of treatments, one computes bounds for the $100(1-\alpha)\%$ confidence interval for the difference $\mu_{i_1} - \mu_{i_2}$ in treatment means from the relation

$$\bar{X}_{i_1\bullet} - \bar{X}_{i_2\bullet} \pm R_\alpha(m, mn-m)\frac{S_P}{\sqrt{n}}, \text{ where}$$

$R_\alpha(m, mn-m)$ denotes a value from tables of probabilities for the Studentized Range distribution. If the confidence interval brackets zero, the null hypothesis H_0: $\{\mu_{i_1} = \mu_{i_2}\}$ is accepted. If both ends of the confidence interval are negative (positive), one concludes that $\mu_{i_1} < \mu_{i_2}$ $(\mu_{i_1} > \mu_{i_2})$.

The Tukey-Kramer procedure appears to be more useful than ANOVA in the sense that it specifies which treatment means have a statistically significant difference. However, the statistics that must be determined for the Tukey-Kramer procedure require the computation of many statistics that are used in ANOVA. Consequently, analysts using the Tukey-Kramer procedure often employ it as a supplement to ANOVA rather than as a substitute for it.

The theoretical development of both procedures requires that the X_{ij}s be independent random variables from normal distributions**. The following section describes a research methodology which was employed to estimate the errors that result when this assumption is not met.

ESTIMATING ERRORS FROM NONNORMALITY

A given design of experiments specifies the number of treatments m, the number of replications n per treatment, and the level of significance α. A trial uses a set of data X_{ij} $(i = 1, \ldots, m$ and $j = 1, \ldots, n)$. From those X_{ij}s, one can compute the data's attendant statistics for the ANOVA and Tukey-Kramer

* Equal sample sizes are used. Kramer (1956) and Hayter (1984) report that the Tukey-Kramer procedure applied to the case having unequal sample sizes yields conservative results.

** Brownlee (1965) states that, with experience, one "...ceases to worry unduly about the normality assumption in most situations." Similarly, Walpole and Myers (1993) state that "slight departures from normality result in minor deviations from the ideal for the standard parametric tests." The results reported herein allow one to quantify these statements for selected designs of experiments and departures from normality.

procedures, and determine whether to accept or reject their corresponding null hypotheses. For a given common distribution for all of the X_{ij}s, one can estimate the Type I error for a procedure by using simulation as follows:

(a) Perform steps (i)-(ii) for the index $k = 1, \ldots, K$.

(i) Fill the X_{ij} matrix with independent random variables from the common probability distribution.

(ii) Compute the statistic(s) for the appropriate statistical procedure for which the Type I error is to be estimated. Define $B_k = \begin{cases} 0 & \text{if the null hypothesis is accepted} \\ 1 & \text{if the null hypothesis is rejected.} \end{cases}$

(b) Estimate the corresponding Type I error for that particular design of experiments and distribution for the X_{ij}s by $\sum_{k=1}^{K} B_k/K$.

Thus simulation can be used to estimate the Type I error that results for a given design of experiments and a given common probability distribution for the X_{ij}s.

The family of gamma distributions serves as a convenient vehicle to use in investigating the errors that result from nonnormality. The random variable T_κ will be said to have a gamma distribution with index κ if its probability density function is given by

$$f_{T_\kappa}(t) = \begin{cases} 0 & t < 0 \\ \frac{\lambda(\lambda t)^{\kappa-1} e^{-\lambda t}}{(\kappa-1)!} & t \geq 0. \end{cases}$$

The extreme $\kappa = 1$ yields the distribution in the gamma family that is the least like the normal distribution; the gamma distribution approaches the normal as κ increases. Experiments were performed for gamma distributions having $\lambda = 1$ and $\kappa \in \mathcal{K} = \{1, 2, 3, 4, 5, 7, 10, \infty\}$. A simulation model yielded results for a total of 64 combinations of designs of experiments. The number of treatments m took on values in the set $\mathcal{M} = \{2, 3, 4, 5, 7, 10, 16, 21\}$. The number n of replications per treatment took on values in the set $\mathcal{N} = \{2, 3, 4, 5, 7, 10, 15, 20\}$. The sets $\mathcal{M}$ and $\mathcal{N}$ were selected in consideration of the goal of investigating a broad range of designs of experiments.

The overall series of simulation experiments performed may be represented in pseudocode as

```
for all m ∈ M do
   for all n ∈ N do
      for all κ ∈ K do
         for k = 1 to K do
            • perform steps (i) and ii) above to determine B_k for trial k
            • estimate the Type I error for trial k using step (b) above
         end loop on k
         • Average the results from the K trials to form an estimate
           of the Type I Error given m, n and κ
      end loop on distributions (κ)
   end loop on number of replications (n)
end loop on number of treatments (m)
```

A nominal level of significance of $\alpha = 0.05$ was used for all experiments. The sample size of $K = 2{,}500$ results in a 95% confidence interval for the values of the Type I error having a spread of $\pm 1.96\sqrt{\frac{.05(1-.05)}{2{,}500}} = \pm 0.009 \approx \pm 0.010$. The results from running the series of experiments described in the above pseudocode are described next.

RESULTS

One series of results aids in model verification. The gamma distribution with parameter $\kappa = \infty$ (the normal distribution) should yield estimates of the Type I error that in the long run average to the nominal value of $\alpha = 0.05$ that was used. This proved to be the case. Furthermore, the ANOVA and Tukey-Kramer procedures should have identical results for $m = 2$ treatments. This also proved to be true. In addition, the values of the F statistic generated for a given design of experiments when $\kappa = \infty$ passed Chi-Squared tests for goodness of fit to the theoretical $F(m-1, mn-m)$ distribution. Similarly, the values of the t statistic generated for a given design of experiments passed Chi-Squared tests for goodness of fit to the theoretical $t(mn-m)$ distribution. These favorable experimental results helped to establish the validity of the simulation model.

The experimental methodology resulted in a three-dimensional array of estimates of Type I errors. One addresses a specific element in the array by specifying the number of treatments m, the number of replications n per treatment, and the distribution index κ being considered. The effect of the *distribution* (indexed by $\kappa \in K$) being considered may be discerned by considering the data appearing in Table 1.

κ	ANOVA Type I Error	T-K Type I Error
1	0.047	0.046
2	0.049	0.048
3	0.049	0.049
4	0.049	0.049
5	0.050	0.049
7	0.050	0.050
10	0.051	0.051
∞	0.051	0.051

Table 1

Number of Treatments m	ANOVA Type I Error	T-K Type I Error
2	0.047	0.047
3	0.049	0.048
4	0.048	0.047
5	0.049	0.047
7	0.049	0.049
10	0.050	0.049
16	0.052	0.053
21	0.053	0.053

Table 2

Sample Size n	ANOVA Type I Error	T-K Type I Error
2	0.052	0.052
3	0.048	0.048
4	0.048	0.048
5	0.047	0.048
7	0.047	0.048
10	0.050	0.050
15	0.051	0.051
20	0.051	0.050

Table 3

Table 1 displays the Type I errors aggregated over number of treatments $m \in M$ and replications per treatment $n \in N$ for the ANOVA and Tukey-Kramer procedures. As one might anticipate, the difference between the Type I error and its nominal value of $\alpha = .05$ tends to decrease as κ increases. The results for $\kappa = 7$, $\kappa = 10$ and $\kappa = \infty$ (the normal distribution) have no statistically significant differences. A linear regression straight-line fit for the data over $\kappa = 1, 2, 3, 4, 5, 7$ and 10 yields a difference between the nominal and observed Type I Errors of

$$\texttt{Difference in Type I Error} = \begin{cases} 0.0024 - 0.00036\kappa & \text{for ANOVA, } \kappa \in K, \kappa \neq \infty \\ 0.0032 - 0.00046\kappa & \text{for Tukey-Kramer, } \kappa \in K, \kappa \neq \infty \end{cases}$$

The data appearing in Table 2 allows one to analyze the effect of the *number of treatments* (indexed by m). Table 2 displays the Type I errors aggregated over distributions ($\kappa \in K$) and number of replications ($n \in N$) for the ANOVA and Tukey-Kramer procedures. A linear regression straight-line fit for the data over $m \in M$ yields a difference between the nominal and observed Type I Errors of

$$\texttt{Difference in Type I Error} = \begin{cases} 0.0028 - 0.00028m & \text{for ANOVA, } m \in M \\ 0.0039 - 0.00036m & \text{for Tukey-Kramer, } m \in M \end{cases}$$

Finally, the effect of the *number of replications* (indexed by $n \in N$) may be discerned by considering the data appearing in Table 3. It displays the Type I errors aggregated over distribution ($\kappa \in K$) and number of treatments ($m \in M$) for the ANOVA and Tukey-Kramer procedures. As the reader would surmise from Table 3, a straight-line fit to its data fails to capture the its dynamics.

SUMMARY

The effects of departures from normality on the ANOVA and Tukey-Kramer procedures has been investigated via simulation. The results support the contention that both procedures are 'robust' in the sense that they yield insignificant differences from the nominal Type I errors when used on a wide range of nonnormal populations and designs of experiments. The largest differences occurred for the exponential distribution with a small number of treatments, where a discrepancy of roughly 0.3% results. This difference diminishes for larger numbers of treatments and a larger index κ for the gamma distribution. Fortunately, an analyst's reputation will in all likelihood remain untarnished if he or she thinks that the level of significance is 0.050 when in fact the true level of significance is 0.047! Thus the results from this series of simulation experiments support the use of these statistical procedures, even if the normality assumption cannot be verified, under a wide variety of designs of experiments.

REFERENCES

Barnes, J. Wesley (1994) *Statistical Analysis for Engineers and Scientists. A Computer-Based Approach.* McGraw Hill, New York, 226.

Bowker, Albert H. and Gerald J. Lieberman (1972). *Engineering Statistics.* Prentice Hall, New Jersey.

Brownlee, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering.* John Wiley and Sons, New York, 241

Hayter, Anthony J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *The Annals of Statistics*, **12**, 1, 61-75.

Kramer, Clyde Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, **12**, 307-310.

Walpole, Ronald E. and Raymond H. Myers (1993). *Probability and Statistics for Engineers and Scientists.* Prentice Hall, New Jersey, 624.