Statistical analysis for toxicity studies

Chikuma Hamada*

1 Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science, 6–3-1 Niijuku, Katsushika-ku, Tokyo 125–8585, Japan

Abstract: Generally, multiple statistical analysis methods can be applied for certain kinds of data, and conclusion could differ, depending on the selected statistical method. Therefore, it is necessary to fully understand the performance of each statistical method and to examine which method is appropriate to use and to standardize statistical methods for toxicity studies to be carried out routinely. Several viewpoints for selecting appropriate statistical methods are discussed in this review paper. According to the distribution form, i.e., whether a distribution has a bell shape without outliers or not, either a parametric or a nonparametric approach should be selected. The nonparametric approach is also available for categorical data. Depending on the design and purpose of a study, several forms of statistical analysis are available. Assuming dose dependency, comparisons with a control are conducted by Williams test (nonparametric: Shirley-Williams test). When a dose dependent relationship is not expected, comparisons with the control are conducted by Dunnett test (nonparametric: Steel test). All possible pairwise comparisons among groups are conducted by Tukey test (nonparametric: Steel-Dwass test). If we are interested in several specific comparisons among groups, the Bonferroni-adjusted Student’s t-test (nonparametric: the Bonferroni-adjusted Wilcoxon test) can be used. (DOI: 10.1293/tox.2017-0050; J Toxicol Pathol 2018; 31: 15–22)

Key words: decision tree, parametric method, nonparametric method and multiple comparison

Introduction

First, the purpose of toxicity studies and the strategy of statistical analysis are described briefly.

Toxicity studies are conducted to ascertain the toxicity of chemical substances, and the information that toxicologists are required to obtain, depends on the application field of the chemical substance and the purpose of toxicity studies. According to the purpose of the studies, the animal species, sample size, administration route, dose, period, evaluation measurements, study design, and other details of the studies are determined. In addition, an appropriate statistical analysis method for the obtained data should also be selected. There are various types of toxicity studies, and specific data is obtained in each study. Basically, in toxicity studies, certain numbers of animals, tissues, or cell populations (groups) are exposed to different dose of a test substance, and differences among groups, dose response relationships, and time trend are investigated. Since the study subjects are organisms, the data from toxicity studies show variation due to biological mechanisms. Generally, multiple statistical analysis methods can be applied for certain kinds of data, and the conclusion could differ, depending on the selected statistical method. Therefore, it is necessary to fully understand the performance of each statistical method to examine which method is appropriate to use, and to standardize statistical methods for toxicity studies to be carried out routinely.

Through the interpretation of individual results in toxicity studies, a comprehensive judgment is made by adding the knowledge of toxicology to the results of statistical analysis. The main topic of this review, however, concerns statistical analysis. Statistical methods widely used in toxicity studies and how to select an appropriate method are described herein.

Efforts at an international level have been made to standardize procedures and methods in many areas of toxicology. In the statistical area, the first steps have been taken. It is obvious that biostatistics plays a basic role in adequate toxicity assessment. Biostatistical evaluations in repeated toxicity studies should support the positive or negative findings of the study and support the quantitative assessment of toxic effects.

Because of limitation with respect to space, the details required for actual statistical computations are not presented in this article, so readers should refer to standard statistical textbooks and appropriate statistical software for the contents of individual statistical methods and actual computation methods.
Analysis of Quantitative Data

In repeated dose toxicity studies (general toxicity studies), blood laboratory tests, biochemical examinations, urinalysis, organ weight measurement, etc., are evaluated. Many of these tests provide metric quantitative data taking continuous values.

Statistical methods commonly used in toxicity studies are shown in Table 1.

Parametric and nonparametric methods

Statistical methods are roughly classified into parametric (para) and nonparametric (nonpara). In the strict sense, the former requires that the distribution form follows a symmetric normal distribution. The latter does not need a distribution form assumption, and it is a valid method for data other than normal distribution. Assuming a normal distribution, Student’s t-test is a method for evaluating the difference in mean values between two groups parametrically. For a right skewed distribution, such as liver function measurements and serum lipids, application of Student’s t-test requires the variable transformation, such as logarithmic transformation, to be performed in advance in order to make the distribution closer to a normal distribution. In contrast, the Wilcoxon test is a typical nonparametric method for comparing the distributions of two groups. Since the data in the two groups are converted into ranks to calculate the sum of the ranks for each group, this method is also referred to as the rank-sum test. The Wilcoxon test is also called the Mann-Whitney U test.

Since both methods conduct essentially the identical evaluation, they are sometimes described as the Mann-Whitney U test as an alias. This method is also applicable to ordered categorical data other than quantitative data, as in the case of pathological findings, that is, ordered categorical data (−, ±, +, ++, and +++), in which the response variable is represented by a grade. Since ordered categorical data are not metric data, they cannot be evaluated by their mean values and standard deviations; however, because there is a natural order relationship between categorical levels, it is possible to rank them. Even if there is data below the detecting measurement limit, we can apply the Wilcoxon test by giving them the lowest ranking. Basically, nonparametric statistical methods corresponding to parametric ones are available, including methods for multiple comparisons. For example, the Kruskal-Wallis test extends the two-sample nonparametric Wilcoxon test to comparisons of more than two groups, and it corresponds to the nonparametric version of parametric one-way ANOVA (analysis of variance), which compares the means among more than two groups, assuming a normal distribution and homogeneity of variance.

Nonparametric methods have a great advantage in that they can be applied irrespective of the distribution form. Even if they are applied when the distribution form is closer to a normal distribution, it is known that they can demonstrate considerably good performance (high statistical power). A common disadvantage in nonparametric approaches is that when the sample size per group is small (less than 7), there is extreme performance deterioration. In toxicity studies using large animals (such as dogs or primates), the sample size per group is often less than 5, so use of a nonparametric approach is not recommendable. In addition, there is a detailed option in the nonparametric approaches, and it is necessary to pay attention to the fact that the result will differ slightly depending on the statistical software and such things as the ranking methods, separate ranking (ranking of each pair comparison) or joint ranking (overall ranking among all groups), and the presence or absence of a continuous correction for approximating the discrete distribution to the theoretical continuous normal distribution.

A preliminary test, such as Bartlett test, can be used to evaluate the hypothesis of homogeneity of variance, and if it is significant, a nonparametric comparison is selected: if it is not significant, a parametric statistical method is selected. This is the common practice with respect to how to choose between a parametric or nonparametric method; however, it has a logical contradiction because both approaches require the assumption on an equal distribution when comparing two groups. Assuming an equal distribution, assumption of equal variance is held, both approaches require an assumption concerning homogeneity of variance.

The choice of parametric or nonparametric methods should be based on the distribution forms of the data. With respect to evaluating the distribution form of data, the distribution form should be explored using a boxplot or a scatter plot, etc., the deviation from the normal distribution should be demonstrated by calculating summary statistics, such as skewness and kurtosis, and it is useful to apply a statistical test for the assumption of a normal distribution.

Multiplicity problems of statistical tests

The usual design in toxicity studies consists of multiple groups (4–5 including a control group), and differences between the groups are often assessed in various combinations of comparison. When the Student’s t-test is repeatedly applied for these comparisons, according to the multiplicity problems for statistical tests, the type I error (false positive error) increases. Here the multiplicity problem and multiple comparison are illustrated using a repeated toxicity study in rats (red blood cell count data).

Table 2 shows that administration of the drug induced a trend of erythrocyte count decline. In toxicity studies, we compare a control group and each dose group in order to investigate the dose that exhibits change compared with the control group. For this purpose, the results of applying the unpaired Student’s t-test at the two-tailed 5% significance level are shown in Table 3.

In comparisons between the control and middle-dose group and the control and high-dose group, the absolute t values are greater than the critical value, 2.101 (corresponding to a p value of less than 0.05), and the results are significant at the 5% level.

In the comparison between the control group and the
When several comparisons among groups are carried out in an experiment, in order to deal with the problem of multiplicity, the entire set of comparisons is regarded as an analysis, and the probability of incorrectly identifying significance throughout the comparisons is controlled to below the nominal significance level; this statistical method is referred to as multiple comparisons. There are several methods for multiple comparisons depending on the combination of groups being compared. Hereinafter, typical multiple comparisons, Bonferroni, Dunnett, Tukey, and Williams test, are discussed.

Bonferroni test: To counter the multiplicity problem, Bonferroni test, when performed k times, determines conservatively the significance level at alpha/k. The p value that is not adjusted, as comparing the significance level at alpha/k is equivalent to comparing the adjusted p value(k×p value) with the significance level alpha, so in Bonferroni test, a multiplicity-adjusted p value is generated by multiplying the raw p value by k. In experiment with four groups, k equals 6 for Tukey type comparisons, and k equals 3 for Dunnett type comparisons. Although Bonferroni test can flexibly cope with various comparisons, assuming independence of comparisons, when there is a correlation among actual comparisons, Bonferroni test is less likely to provide significant results (larger p value), compared with Dunnett or Tukey test. Table 4 shows the results of application of Bonferroni test to RBC data. Since three comparisons are conducted, each p value for Bonferroni test is three times that obtained for the Student’s t-test shown in Table 3; the p value for Bonferroni test is six times the non-adjusted p value for all combinations of pairwise comparisons among the four groups.

Dunnett test: Dunnett test is used for multiple comparisons when comparing a certain reference group with...
Table 4. Bonferroni Test (at Two-sided 5%) Applied to the RBC Data

| Comparison   | Difference in mean value | t value | p value |
|--------------|--------------------------|---------|---------|
| C-L          | -14.1                    | -1.367  | 0.570   |
| C-M          | -34.5                    | -2.305  | 0.099   |
| C-H          | -33.0                    | -2.387  | 0.064   |

Critical value: 2.531. All absolute t values are less than 2.531 and not significant. C: Control group, L: Low-dose group, M: Middle-dose group, H: High-dose group.

Table 5. Dunnett Test (at Two-sided 5%) Applied to the RBC Data

| Comparison   | Difference in mean value | t value | p value |
|--------------|--------------------------|---------|---------|
| C-L          | -14.1                    | -1.010  | 0.622   |
| C-M          | -34.5                    | -2.472  | 0.048   |
| C-H          | -33.0                    | -2.364  | 0.061   |

Critical value: 2.452. *: t values are different from Table 3, because of pooling variance among four groups in Dunnett test. *: Significant in Dunnett test (two-sided 5%). C: Control group, L: Low-dose group, M: Middle-dose group, H: High-dose group.

all other groups (a total of 3 times in the example shown in Table 2)8. Because Dunnett test calculates critical values considering multiplicity, with comparisons (three in the example in Table 2) conducted simultaneously, the critical value is large, compared with that in the Student’s t-test, and it is less likely to be significant. The results of Dunnett test are shown in Table 5 at the two-sided 5% significance level. The critical value of Dunnett test (DF, degree of freedom = 36) in this example is 2.452, and the comparison between the control and middle-dose group is barely significant. The p value for Dunnett test is larger than that for the Student’s t-test and smaller than that for Bonferroni test.

Tukey test: Tukey test conducts all possible pairwise comparisons among four groups (six in this example)9. Due to the increased types of comparisons of interest compared with Dunnett test, individual comparisons in Tukey test are less likely to be significant than those in Dunnett test. The critical value of Tukey test at the two-sided 5% significance level in this example is 2.694. The results of Tukey test are shown in Table 6. None of the comparisons are significant.

Williams test: In the results of Dunnett test shown in Table 5, the comparisons between the control and low-dose groups and between the control and high-dose groups are not significant at the 5% level (p value greater than 0.05). In contrast, the result of the comparison between the control and middle-dose group is subtle, but it is significant at the 0.05 level. In such a case, however, interpretation of the results is difficult. Although these results are from an actual toxicity study and it is biologically plausible for a dose dependent change to occur if the chemical substance has the effect of decreasing the red blood cell count, in these data, only the intermediate middle-dose group showed a significant decrease in the number of erythrocytes compared with the control group. Exploring the details of the data, the mean value is 891.5 in the middle-dose group and 893.0 in the high-dose group, so it can be seen that the value is almost unchanged. In this case, it is natural to interpret this as meaning that the effect is almost the same in the middle-dose group and high-dose group, rather than considering that only the middle-dose group showed decrease. In fact, Dunnett test, often indicates such a significant difference in only an intermediate-dose group, so it is difficult to interpret the results. Basically, assuming a monotonic dose response relationship, if there is a significant difference at a certain dose, above the dose, all doses are considered to result in significant differences; because Dunnett test does not assume monotonicity between dose and response, it is not always possible to obtain such a plausible result. In contrast, Williams test assumes a monotonic dose response relationship, if a middle-dose results in a significant change, all higher doses are considered to result in significant changes, so there is an advantage with respect to the consistency of results among the dose groups20. Dunnett test is a method for detecting whether there is a difference in a group compared with the control group; in contrast, Williams test is considered to find the dose at which a change begins to be shown compared with the control.

Case 1: control group, drug A group, drug B group, drug C group

Dunnett test, which takes into account the multiplicity in the context of case 1) shown above, was originally developed to identify a drug effect compared with a control group. Here, drug A, B, and C groups have no natural order relationship. However, in practice, Dunnett test is also frequently used in the context of case 2) shown below.

Case 2: control group, low-dose group, middle-dose group, high-dose group

The difference from the context of case 1) is that there is a natural order relationship among the dose groups (low-dose group < middle-dose group < high-dose group). If a monotonic relationship is biologically plausible between dose and response, the reaction should increase or decrease with increasing dose. In fact, because Dunnett test does not utilize information about the sequential relationship between such groups, in the context of case 2, it may provide results that are difficult to interpret. In contrast, due to the necessity for ordering relationships among groups, Williams test cannot be used in the context of case 1), and if the monotonic dose response in the context of case 2) is satisfied, Williams test has higher power than Dunnett test.

The mean values of the groups in the Table 7 are 926.0, 911.9, 891.5, and 893.0, and it can be seen that they almost monotonically decrease, although the mean values are reversed at the high and middle doses.

Since monotonicity is assumed for the dose response in Williams test, this reversal is regarded as occurring by
chance, and there is actually no difference in population mean value between the middle-dose and the high-dose. An estimate of the effect at the high-dose can be calculated as the aggregate mean value of the mean values for the middle-dose (891.5) and high-dose (893.0), which corresponds to 892.25 in this example.

In this manner, if the assumption concerning monotonicity among groups is not held, Williams test calculates the weighted mean between adjacent groups. After conducting such an operation and holding monotonicity among groups, comparison with the control group is continued from higher dose sequentially, until any significant difference disappears. Multiple comparisons accompanying such a hierarchical procedure are called step-down methods. Based on the assumption of monotonicity, unless the dose response is significant at higher doses, the dose response at lower doses will not be significant and application of the step-down procedure is reasonable.

Table 7 shows the results of applying Williams test at the lower significance level of 2.5%, when applied to these data, becomes significant above the middle-dose group.

Williams test is necessary to determine whether a monotonically decreasing (lower) or monotonically increasing (upper) dose response before conducting statistical test. In this sense, Williams test is considered to be an essentially one-sided oriented method. If a two-sided test is desired, the significance level should be split into lower and upper hypotheses (half and half) as in the case of the Bonferroni test. In addition, since Williams test is a step-down procedure, the lower level hypothesis is only evaluated, when the upper level hypothesis is significant. Therefore, the p value of Williams test cannot be interpreted as the p value obtained from simultaneously conducted statistical tests, as in the case of Dunnett test.

Although Dunnett test has been used as the standard analysis methods for comparing the dose groups with a control group in repeated dose toxicity studies, if a monotonic relationship can be assumed, Williams test has considerably higher power than Dunnett test and is recommendable.

Table 6. Tukey Test (at Two-sided 5%) Applied to the RBC Data

| Comparison | Difference in mean value | t value | p value |
|------------|--------------------------|---------|---------|
| C-L        | -14.1                    | -1.010  | 0.745   |
| C-M        | -34.5                    | -2.472  | 0.082   |
| C-H        | -33.0                    | -2.364  | 0.103   |
| L-M        | -20.4                    | -1.462  | 0.471   |
| L-H        | -18.9                    | -1.354  | 0.536   |
| M-H        | 1.5                      | 0.108   | 1.000   |

Critical value: 2.694. All absolute t values are less than 2.694 and not significant. C: Control group, L: Low-dose group, M: Middle-dose group, H: High-dose group.

Table 7. Williams Test (at Lower 2.5%) Applied to the RBC Data

|                   | Control group | Low-dose group | Middle-dose group | High-dose group |
|-------------------|---------------|----------------|-------------------|-----------------|
| Mean value        | 926.0         | 911.9          | 891.5 *           | 893.0 *         |
| Aggregate mean    | 926.0         | 911.9          | 892.25            | 892.25 *        |

*: Significant in Williams test. #: Aggregate mean assuming monotonicity (892.25 = (893.0 + 891.5) / 2).

Table 8. Comparison of Dunnett Test and Williams Test

| Analysis method                        | Dunnett test                                      | Williams test                                    |
|----------------------------------------|---------------------------------------------------|-------------------------------------------------|
| Assumption on dose response            | Non-monotonic                                     | Monotonic                                        |
| Alternative hypothesis                  | One-sided and two-sided                           | Essentially one-sided                            |
| Hypothesis                             | Whether there is a difference between the control group and drugs | From which dose is different to the control group |
| Context                                | Control, drug A, drug B, drug C                    | Control, low-dose, middle-dose, high-dose        |
| alpha error                            | Below alpha considering multiplicity               | Below alpha considering multiplicity             |
| Procedure                              | One step procedure                                | Step-down procedure                              |
| Monotonic cases                        | Low power                                         | High power                                       |
| Non-monotonic cases                    | Easy to interpret                                 | Difficult to interpret                           |

Table 6. Tukey Test (at Two-sided 5%) Applied to the RBC Data

| Table 6. Tukey Test (at Two-sided 5%) Applied to the RBC Data | Table 7. Williams Test (at Lower 2.5%) Applied to the RBC Data | Table 8. Comparison of Dunnett Test and Williams Test | Table 9. Alpha errors for Dunnett Test | Table 10. Summary of Dunnett Test and Williams Test |
|---------------------------------------------------------------|---------------------------------------------------------------|-----------------------------------------------------|--------------------------------------|------------------------------------------------|
| Comparison                                                   | Difference in mean value                                      | t value                                             | p value                              | Analysis method                        |
|---------------------------------------------------------------|---------------------------------------------------------------|-----------------------------------------------------|--------------------------------------|--------------------------------------|
| C-L                                                         | -14.1                                                       | -1.010                                             | 0.745                                | Dunnett test |                                                        |
| C-M                                                         | -34.5                                                       | -2.472                                             | 0.082                                |                                                        |
| C-H                                                         | -33.0                                                       | -2.364                                             | 0.103                                |                                                        |
| L-M                                                         | -20.4                                                       | -1.462                                             | 0.471                                |                                                        |
| L-H                                                         | -18.9                                                       | -1.354                                             | 0.536                                |                                                        |
| M-H                                                         | 1.5                                                         | 0.108                                              | 1.000                                |                                                        |

Critical value: 2.694. All absolute t values are less than 2.694 and not significant. C: Control group, L: Low-dose group, M: Middle-dose group, H: High-dose group.

Table 7. Williams Test (at Lower 2.5%) Applied to the RBC Data

|                   | Control group | Low-dose group | Middle-dose group | High-dose group |
|-------------------|---------------|----------------|-------------------|-----------------|
| Mean value        | 926.0         | 911.9          | 891.5 *           | 893.0 *         |
| Aggregate mean    | 926.0         | 911.9          | 892.25            | 892.25 *        |

*: Significant in Williams test. #: Aggregate mean assuming monotonicity (892.25 = (893.0 + 891.5) / 2).
systems for toxicity studies until recently. When the GLP systems were introduced, selective use of Dunnett test and Scheffe test became the standard algorithm because multiple comparisons without problems in unbalanced cases, besides Scheffe, were widely unknown, Scheffe test makes it possible to control the alpha error to lower than the nominal significance level, and it satisfies the prerequisites for multiple comparisons in this sense, however, it is well known that Scheffe test is too conservative to detect true significance.

An example of the actual significance level for a comparison between a control group and individual groups using Scheffe test is shown in Table 9.

As the number of groups increases, it can be confirmed that the actual significance level gradually decreases compared with the nominal significance level. For example, at five groups, when the t value exceeds 2.442, Dunnett test becomes significant at the two-sided 5% significance level, but Scheffe test does not become significant if it does not exceed 3.080. Compared with the nominal significance level of 5%, the actual significance level of 0.8% is much lower. In this way, Scheffe test has a conservative alpha error, and as a result, it provides a larger probability of beta error, which misses the true toxicological response. Therefore, we should not apply Scheffe test in toxicity studies.

As already stated, Dunnett test and Tukey test, they originally assumed that the sample sizes among groups are equal, and they were subsequently extended to cases in which the sample sizes are not equal among groups. Exact calculation methods or precise approximation methods are known and several statistical software are available for unbalanced data. In addition, even though sample sizes among groups are different, the difference less than 20% does not provide a big problem even if the critical value is used, when sample sizes are equal.

Although many statistical packages can conduct multiple comparisons using Duncan method, this method does not satisfy the requirement for multiple comparisons (the probability of alpha errors in the overall comparisons can be kept below the nominal level); therefore, Duncan test should not be used. The total size of the alpha error is much higher than the nominal level; as in the case of repeated t tests, and ignoring the multiplicity, performance of Duncan test is almost the same as that of repeated t test. For this reason, although it is likely to provide more significant results than other multiple comparison methods, this is because the alpha error cannot be controlled and Duncan test does not satisfy the requirement for multiple comparisons. Although the Student-Newman-Keuls method, which is similar to Duncan test, is also often used in Europe and the USA, this method inflates the alpha error, compared with the nominal level and is not recommended to be used.

**ANOVA and multiple comparison**

When several factors consist of variation in experimental data, analysis of variance is a generic statistical method for decomposing the overall variation into the individual elements. For example, one-way ANOVA decomposes the total sum of squares into the between-group variation and within-group variation. Assuming the homogeneity of variances and normality, the magnitude of the variation of each factor is shown by an analysis of variance table, the F test, comparing its magnitude with error variation. Analysis of variance is often abbreviated as ANOVA (Analysis of Variance). Classical statistical textbooks may have positioned multiple comparisons as an ad hoc analysis after one-way analysis of variance, and ANOVA is traditionally conducted before multiple comparisons. When a significant difference is only observed by ANOVA, comparisons among groups is conducted in toxicity studies. In principle, however, multiple comparisons are designed to be applied independently from one-way analysis of variance, as each can independently control the significance level. If the results are judged to be significant only when there is a significant difference in both ANOVA and multiple comparisons, the overall significance level is smaller than the significance level of the individual method, and the judgment becomes conservative.
Table 10 shows the overall significance levels for different sample sizes and the different number of groups, when a combined analysis with ANOVA and Dunnett test at the 5% level, respectively, is applied. The actual significance level is reduced to 3–4%.

**Decision Tree for Statistical Analysis in Toxicity Studies**

Since the performance of each statistical method depends on the distribution form and we usually do not have exact information about this, selection of one method in preference to another can be very difficult. Therefore, it is
convenient to choose a suitable method according to the data obtained. Figure 1 shows a decision tree for statistical analysis in toxicity studies[13–15].

The first step is visual expression of the data using a scatter plot or a stratified boxplot. A scatter plot and stratified boxplot for RBC data are shown in Fig. 2 and Fig. 3, respectively. The viewpoints of visualization of data are existence of outliers, normality of the distribution and homogeneity of variance. The stratified boxplot for the RBC data shows two outliers in the control group and middle-dose group.

An increase of variance with an increasing mean value of the response variable is frequently observed in repeated toxicity studies. We also often find outliers that could be related to any compound effect, could be from an abnormality in the animal, or might not be directly related to the compound, e.g., wrong operations in the experiment. These phenomena make statistical analysis more difficult. When heterogeneity of variance is observed, variable transformation should be conducted. It is well known that log transformation often improves heterogeneity of variance.

According to the distribution form, i.e., whether the distribution has a bell shape without outliers or not, either a parametric approach or nonparametric approach should be selected. The nonparametric approaches are also available for categorical data (e.g., graded histopathological data).

Depending on the design and purpose of a study, several statistical analysis are available. Assuming dose dependency, comparisons with a control are conducted by Williams test (nonparametric: Shirley-Williams test). When a dose dependent reaction is not expected, comparisons with a control are conducted by Dunnett test (nonparametric: Steel test). All possible pairwise comparisons among groups are conducted by Tukey test (nonparametric: Steel-Dwass test). If we are interested in specific several comparisons among groups, Bonferroni-adjusted Student’s t-test (nonparametric: Bonferroni-adjusted Wilcoxon test) can be used.

Disclosure of Potential Conflicts of Interest: The author declares that he has no conflict of interest.

Reference

1. Hothorn AL, Lin KK, Hamada C, and Rebel W. Recommendation for biostatistics of repeated toxicity studies. Drug Inf J. 31: 327–334. 1997. [CrossRef]
2. Uesaka H. Multiple comparison. In: Handbook of Statistics in Medicine. Tango T and Miyahara H (eds). Asakura Publishing, Tokyo, 77–120. 1995. (in Japanese).
3. Hamada C. Statistics for Presentation and Writing Articles. Shinko Trading. Tokyo, 2012. (in Japanese).
4. SAS website: https://www.sas.com/ja-jp/home.html
5. SPSS website: https://www-01.ibm.com/software/jp/marketplace/spss/
6. R website: http://www.statistics.co.jp/reference/software_R/free_software-R.html
7. Dunnett CW. A multiple comparisons procedure for comparing several treatments with a control. J Am Stat Assoc. 50: 1096–1121. 1955. [CrossRef]
8. Dunnett CW. Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. J Am Stat Assoc. 75: 789–795. 1980. [CrossRef]
9. Tukey JW. Comparing individual means in the analysis of variance. Biometrics. 5: 99–114. 1949. [Medline] [CrossRef]
10. Williams DA. The comparison of several dose levels with a zero dose control. Biometrics. 28: 519–531. 1972. [Medline] [CrossRef]
11. Yamazaki M, Noguchi Y, Tanda M, and Shintani S. Statistical methods appropriate for general toxicological studies in rats. Algorithms for multiple comparisons of treatment groups with control. J Takeda Res Lab. 40: 163–187. 1981; (in Japanese).
12. Duncan DB. T tests and intervals for comparisons suggested by the data. Biometrics. 31: 339–359. 1975. [CrossRef]
13. Hamada C, Yoshino K, Abe I, Matsumoto K, Nomura M, and Yoshimura I. A study on the consistency between statistical evaluation and toxicological judgment. Drug Inf J. 31: 413–421. 1997.
14. Hamada C, Yoshino K, Matsumoto K, Nomura M, and Yoshimura I. Tree-type algorithm for statistical analysis in chronic toxicity studies. J Toxicol Sci. 23: 173–181. 1998. [Medline] [CrossRef]
15. Hamada C, Yoshino K, Abe I, Matsumoto K, Nomura M, and Yoshimura I. Detection of an outlier and evaluation of its influence in chronic toxicity studies. Drug Inf J. 32: 201–212. 1998. [CrossRef]

In memory of the late Prof. Chikuma Hamada.

Dr. Chikuma Hamada passed away on December 21, 2017, at the age of 52. Dr. Hamada was a foremost authority on toxicological statistics providing an important contribution to the field of toxicology. He was a valued and respected professor and he will be greatly missed. We are saddened by his death and our thoughts and condolences go to his family at this difficult time. We pray that his soul may rest in peace.