19

How to Select An Appropriate Statistical Tool?

Good Statistical Design

Good statistical design is a pivotal factor in animal research. However, replication, randomization and blinding, which are key components of good statistical design, are less often used in animal research (Kilkenny *et al.*, 2009). Hess (2011) reviewed statistical design given in 100 articles on animal experiments published in Cancer Research in 2010. In 14 of the 100 articles, the number of animals used per group was not reported. In none of the 100 articles the method used to determine the number of animals per group was reported. Among the 74 articles in which randomization seemed feasible, only 21 reported that they had randomly allocated animals to treatment groups. None of these articles described how the randomization was carried out. Selection of appropriate statistical tools is very crucial in the analysis of data obtained from toxicological and pharmacological studies. Selection of a non-appropriate statistical tool during the design of a study or using a different statistical tool from that mentioned in the study plan with improper justification may lead to misinterpretation of the data (Kobayashi *et al.*, 2011).

Decision Trees

Several attempts have been made to standardize statistical methodologies for the analysis of data obtained from the toxicological and pharmacological studies. One of the methodologies proposed by several authors is the tree-type algorithms (Gad and Weil, 1986; Healey, 1997; Hamada *et al.*, 1998; Gad, 2006). The tree-type algorithms are called as decision trees, which are graphical representation of decisions involved in the choice of the statistical procedure (Howell, 2008). The decision tree-diagram is an excellent tool for determining the optimum course of action in situations offering several alternatives with uncertain outcomes. The first tree-type algorithm for toxicity studies reported in Japan by Yamazaki *et al.* (1981) is given in Figure 19.1.

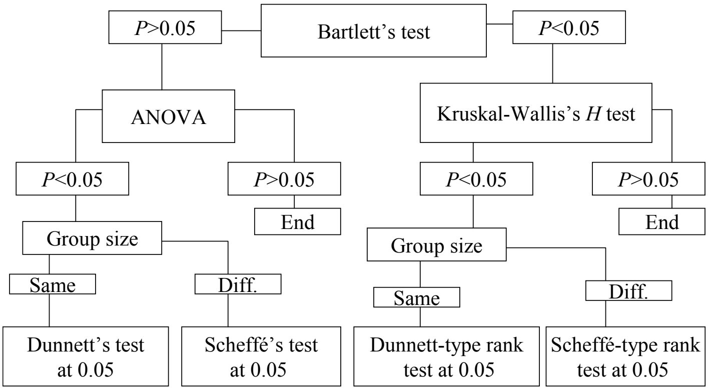


Figure 19.1. The first tree-type algorithm for toxicity studies reported in Japan

This tree-type algorithm was criticized by Kobayashi *et al.* (1995), who identified three major weaknesses which included: selection of a parametric or non-parametric test is based on the highly sensitive Bartlett's homogeneity test; test for normality is not covered in this algorithm; and outliers and dose-dependency are not evaluated.

Hamada *et al.* (1998) proposed a tree-type algorithm for the analysis of quantitative data, which is given in Figure 19.2.

Kobayashi *et al.* (2000) proposed a simple tree-type algorithm for the analysis of quantitative data obtained from toxicological experiments involving more than 2 groups (Figure 19.3).

Sakaki *et al.* (2000) proposed a tree-type algorithm for the analysis of quantitative data, particularly body weight, hematology, and organ weight data, obtained from repeated dose administration studies. This tree-type algorithm does not recommend homogeneity and normality tests; the data are directly analysed by Williams's test (Figure 19.4).

Gad and Weil (1986) proposed a flow chart covering most of the situations that can be encountered in toxicology and pharmacology (Figure 19.5).

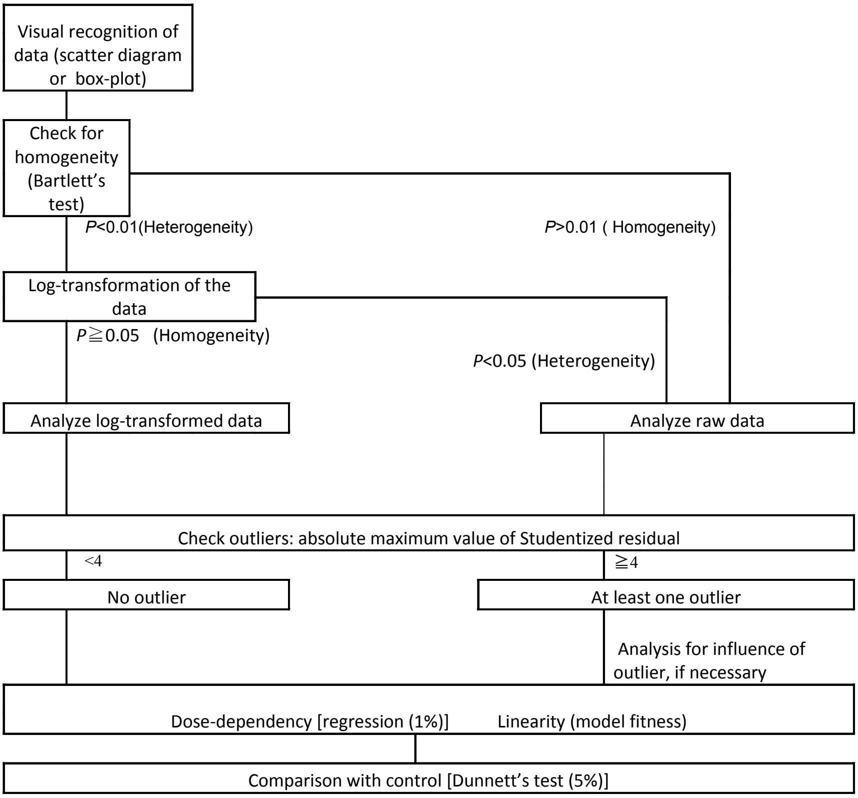

Visual recognition of data (scatter diagram or box-plot)
Check for homogeneity (Bartlett's test)
P<0.01(Heterogeneity)
P>0.01 (Homogeneity)
Log-transformation of the data
P≧0.05 (Homogeneity)
P<0.05 (Heterogeneity)
Analyze log-transformed data
Analyze raw data
Check outliers: absolute maximum value of Studentized residual
<4
≧4
No outlier
At least one outlier
Analysis for influence of outlier, if necessary
Dose-dependency [regression (1%)]
Linearity (model fitness)
Comparison with control [Dunnett's test (5%)]

Figure 19.2. Tree-type algorithm for the analysis of quantitative data proposed by Hamada *et al.* (1998)

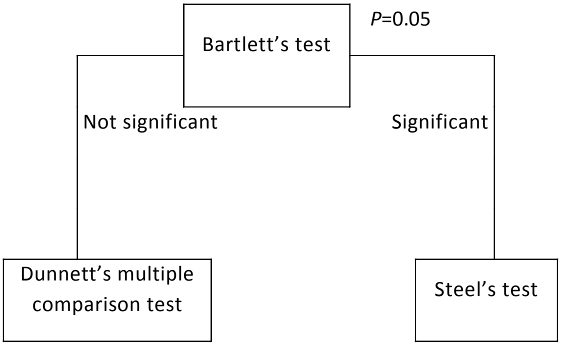


Figure 19.3. The tree-type algorithm for the analysis of toxicological data proposed by Kobayashi *et al.* (2000)

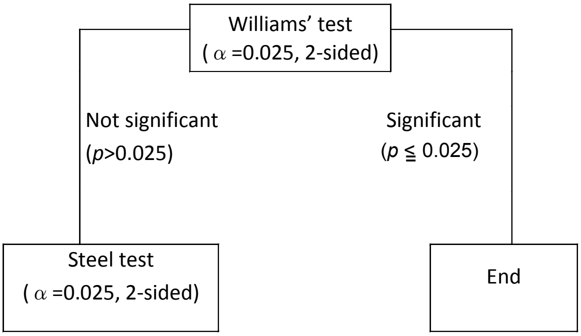


Figure 19.4. The tree-type algorithm for the analysis of quantitative data obtained from repeated dose administration studies proposed by Sakaki *et al.* (2000)

Statistical Procedures Used by National Toxicology Program (NTP), USA

The statistical procedures used in the analysis of data of 2-year toxicity/ carcinogenesis studies presented in the Technical Reports of the NTP are given below:

a. Survival Analyses

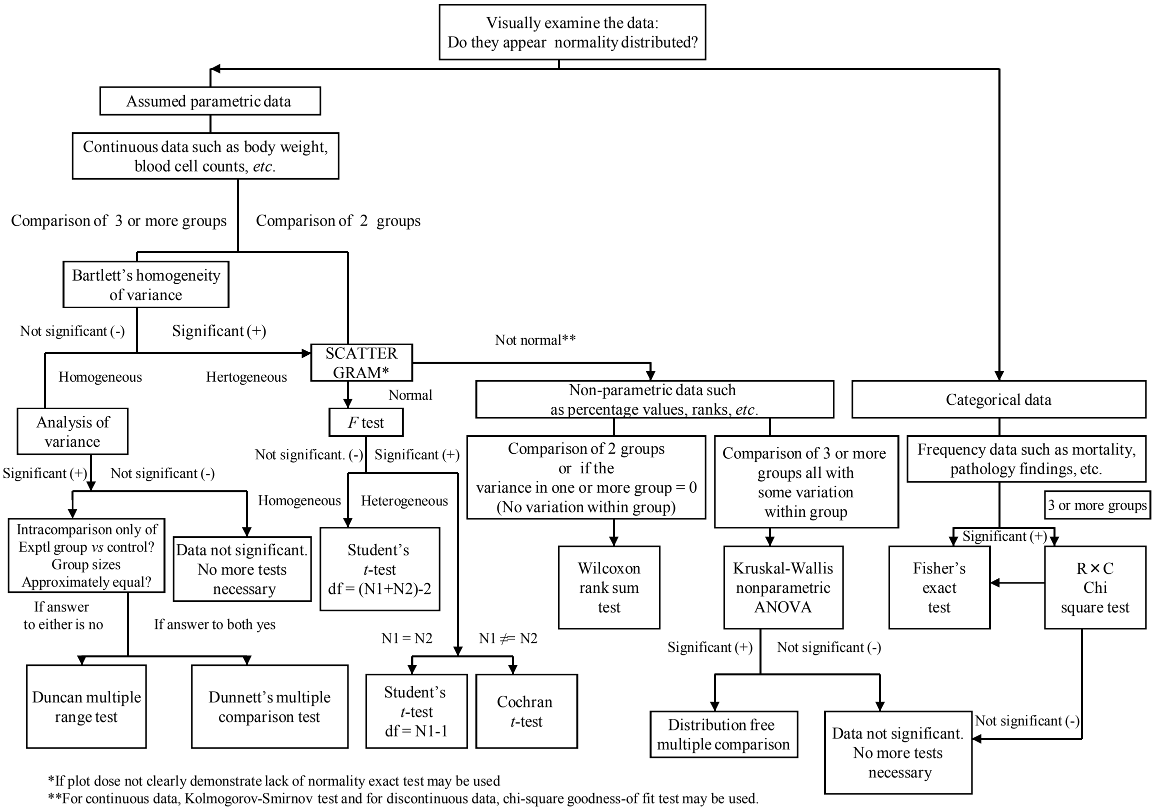
The product-limit procedure of Kaplan and Meier (1958) is used to estimate the probability of survival. Animals found dead due to causes other than natural causes are censored from the survival analyses, while animals dying from natural causes are not censored. Dose-related effects on survival is calculated using Cox's method (Cox, 1972) (for testing two groups for equality) and Tarone's (1975) life table test (to identify dose-related trends). The *P* values are two-sided.

b. Analysis of neoplasm and non-neoplastic lesion incidences

The Poly-*k* test (Bailer and Portier, 1988; Portier and Bailer, 1989; Piegorsch and Bailer, 1997) is used to assess neoplasm and non-neoplastic lesion prevalence. Tests of significance include pair-wise comparisons of each exposed group with controls and a test for an overall exposure-related trend. Continuity-corrected Poly-3 tests are used in the analysis of lesion incidence. The *P* values are one-sided.

c. Analysis of continuous variables

Organ and body weight data is analyzed with the parametric multiple comparison procedures of Dunnett (1955) and Williams (1971, 1972). Hematology, clinical chemistry, urinalysis, urine concentrating ability,


Visually examine the data:
Do they appear normality distributed?
Assumed parametric data
Continuous data such as body weight, blood cell counts, *etc*.
Comparison of 3 or more groups
Comparison of 2 groups
Bartlett's homogeneity of variance
Not significant (-)
Significant (+)
SCATTER GRAM*
Not normal**
Homogeneous
Hertogeneous
Normal
Analysis of variance
F test
Significant (+)
Not significant (-)
Not significant. (-)
Significant (+)
Homogeneous
Heterogeneous
Intracomparison only of Exptl group *vs* control? Group sizes Approximately equal?
Data not significant. No more tests necessary
Student's *t*-test df = (N1+N2)-2
If answer to either is no
If answer to both yes
N1 = N2
N1 ≠ N2
Duncan multiple range test
Dunnett's multiple comparison test
Student's *t*-test df = N1-1
Cochran *t*-test
Non-parametric data such as percentage values, ranks, *etc*.
Comparison of 2 groups or if the variance in one or more group = 0 (No variation within group)
Comparison of 3 or more groups all with some variation within group
Wilcoxon rank sum test
Kruskal-Wallis nonparametric ANOVA
Significant (+)
Not significant (-)
Distribution free multiple comparison
Data not significant. No more tests necessary
Categorical data
Frequency data such as mortality, pathology findings, etc.
3 or more groups
Significant (+)
Fisher's exact test
R × C Chi square test
Not significant (-)

*If plot dose not clearly demonstrate lack of normality exact test may be used

**For continuous data, Kolmogorov-Smirnov test and for discontinuous data, chi-square goodness-of fit test may be used.

Figure 19.5 Flow chart proposed by Gad and Weil (1986)

cardiopulmonary, cell proliferation, tissue concentrations, spermatid, and epididymal spermatozoal data are analyzed using the non-parametric multiple comparison methods of Shirley (1977), as modified by Williams (1986) and Dunn (1964). Jonckheere's test (Jonckheere, 1954) is used to assess the significance of the dose-related trends and to determine whether a trend-sensitive test (Williams' or Shirley's test) is more appropriate for pair-wise comparisons than a test that does not assume a monotonic dose-related trend (Dunnett's or Dunn's test).

Average severity values are analyzed for significance with the Mann-Whitney *U* test (Hollander and Wolfe, 1973). Vaginal cytology data are transformed to arcsine values and then the treatment effects are investigated by applying a multivariate analysis of variance (Morrison, 1976).

Immunological data is initially tested for homogeneity using Bartlett's test. For data that is determined to be homogeneous, one-way analysis of variance (ANOVA) is conducted. If the ANOVA is significant at $P < 0.05$, Dunnett's multiple range *t*-test is used for multiple treatment-control comparisons. If the data is not homogeneous, the Kruskal-Wallis test or the Wilcoxon rank sum test is used to compare treatment groups with controls groups. The level of statistical significance is set at $P < 0.05$ and $P < 0.01$.

Values are routinely presented as mean ± standard error.

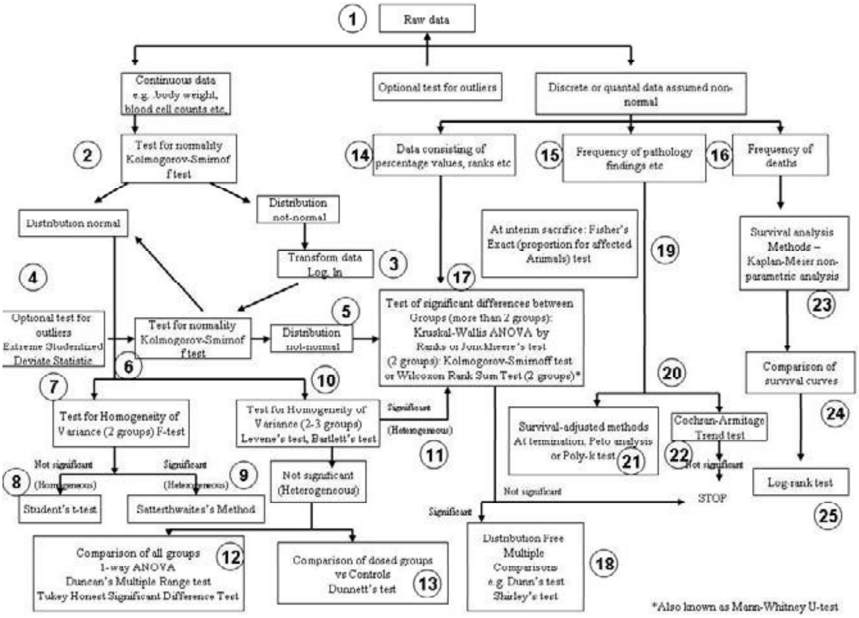
Decision Tree Produced by OECD

OECD produced a decision tree for analyzing data in long-term toxicology studies by summarizing common statistical procedures (OECD, 2010). This decision tree, more or less similar to an approach used by the US National Toxicology Program, is given in Figure 19.6.

A detailed description on this decision tree is given in the guideline (OECD, 2010) by providing explanation on each circled number given in the Figure.

Decision tree has also been used *in vitro* assays and pharmacological experiments. Decision-tree approaches were proposed for the analysis of the chromosome aberration assay (Kim *et al.*, 2000; Hothorn, 2002) and for evaluating drug-specific effects of quantitative pharmaco-EEG (Dago *et al.,* 1994).

Though the decision trees are used in the statistical analysis of data of various toxicological studies (Krores *et al.*, 2004), critics point out that, 'although there are efficiency gains in the application of flow charts, there


1
Raw data
Continuous data e.g. body weight, blood cell counts etc.
Optional test for outliers
Discrete or quantal data assumed non-normal
2
Test for normality Kolmogorov-Smirnof f test
Distribution normal
Distribution not-normal
Transform data Log. ln
3
4
Optional test for outliers Extreme Studentized Deviate Statistic
Test for normality Kolmogorov-Smirnof f test
5
Distribution not-normal
6
7
Test for Homogeneity of Variance (2 groups) F-test
10
Test for Homogeneity of Variance (2-3 groups) Levene's test, Bartlett's test
Not significant (Homogeneous)
Significant (Heterogeneous)
8
Student's t-test
9
Satterthwaites's Method
Not significant (Heterogeneous)
Significant (Heterogeneous)
11
Comparison of all groups 1-way ANOVA Duncan's Multiple Range test Tukey Honest Significant Difference Test
12
Comparison of dosed groups vs Controls Dunnett's test
13
14
Data consisting of percentage values, ranks etc
15
Frequency of pathology findings etc
16
Frequency of deaths
At interim sacrifice: Fisher's Exact (proportion for affected Animals) test
19
17
Test of significant differences between Groups (more than 2 groups): Kruskal-Wallis ANOVA by Ranks or Jonckheere's test (2 groups): Kolmogorov-Smirnoff test or Wilcoxon Rank Sum Test (2 groups)*
Survival analysis Methods – Kaplan-Meier non-parametric analysis
23
20
Comparison of survival curves
24
Survival-adjusted methods At termination, Peto analysis or Poly-k test
21
Cochran-Armitage Trend test
22
Not significant
Not significant
STOP
Significant
Log-rank test
25
Distribution Free Multiple Comparisons e.g Dunn's test Shirley's test
18

*Also known as Mann-Whitney U-test

Figure 19.6. Decision tree produced by OECD for the analysis of data in long-term toxicology studies (OECD, 2010)

is a 'deskilling' of the task, an over-emphasis on significance testing for decision making, and vulnerability to artefactual results'. There is also the methodological problem with a multiple testing procedure where one hypothesis test is used to select another test which can complicate quantifying the true probability values associated with various comparisons (OECD, 2010).

Incongruence in Selection of a Statistical Tool

Nomura (1994) compared the tree-type algorithms used at the contract research laboratories in Japan and other countries. He observed that the countries developed their own tree-type algorithms. Kobayashi *et al*. (2011) compared the statistical tools used for analysing the data of repeated dose toxicity studies with rodents conducted in 45 countries, with that of Japan. The study revealed there was no congruence among the countries in the use of statistical tools for analysing the data obtained from the above studies. For example, to analyse the data obtained from repeated dose toxicity studies with rodents, Scheffé's multiple range and Dunnett type (joint type Dunnett) tests are commonly used in Japan, but in other countries use of these statistical tools is not so common. In most of the countries, the data are generally not tested for normality. The authors observed that out of 127 studies examined, data of only 6 studies were analysed for both homogeneity of variance and normal distribution.

The decision trees mentioned above are developed based on the classical statistical principles sidelining biological principles. For example a sensitive Bartlett's test for examining homogeneity of variance may not be suitable in most of the animal studies. The below mentioned decision trees or flow charts are developed providing due consideration to biological principles:

Selection of a Statistical Tool—Suggested Decision Tress or Flow Charts

1. Selection of a statistical tool when the data show a normal or non-normal distribution (Kobayashi *et al.*, 2008).

Situation 1 (Number of Groups, 2)

When the data of each group show a normal distribution by Shapiro-Wilk's *W* test, then the *F*-test is applied. If the *F*-test is insignificant, the data are analysed using Student's *t*-test and if it is significant, Aspin-Welch's *t*-test is used to analyse the data.

When the data of any group show a non-normal distribution by Shapiro-Wilk's *W* test, they are subjected to Mann-Whitney's *U* test (Rank sum test).

Flow chart of situation 1 is given in Figure 19.7.

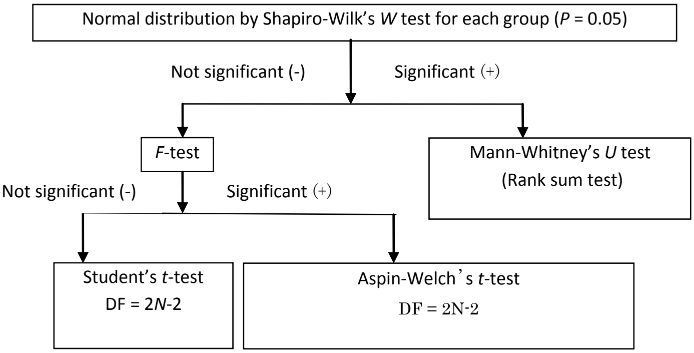


Figure 19.7. Flow chart for selecting the statistical tool when the data show a normal or non-normal distribution (Situation 1, Number of group = 2)

Situation 2 (Number of Groups, ≥3)

When each group shows a normal distribution by Shapiro-Wilk's *W* test, the Dunnett's multiple comparison test is used. When control group or all groups do not show a normal distribution, non-parametric Steel's test (Dunnett's separate type test) is used. When normal distribution is not observed by one or two treatment groups, they are excluded from the analysis and the remaining groups are analyzed by Dunnett's multiple comparison test. The clinical relevance of the excluded groups is assessed in the light of other observations.

Flow chart of situation 2 is given in Figure 19.8.

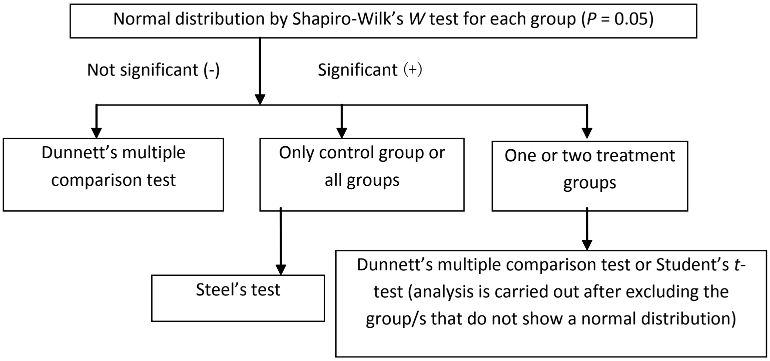

Normal distribution by Shapiro-Wilk's W test for each group (P = 0.05)
Not significant (-)
Significant (+)
Dunnett's multiple comparison test
Only control group or all groups
One or two treatment groups
Steel's test
Dunnett's multiple comparison test or Student's t-test (analysis is carried out after excluding the group/s that do not show a normal distribution)

Figure 19.8. Flow chart for selecting the statistical tool when the data show a normal or non-normal distribution (Situation 2, Number of group ≥ 3)

2. Analysis of qualitative data of urinalyses and pathological findings.

Analysis of qualitative data of urinalysis and pathological findings presented in 2×2 and 4×4 Tables is given in Table 19.1.

Statistical Tools Suggested for the Analysis of Toxicology Data

The suggested statistical tools for the analysis of parametric and non-parametric data are given in Table 19.2 and for the comparison of two and multi-groups are given in Table 19.3.

Use of Statistics in Toxicology-Limitations

There are limitations in the use of statistics in toxicology. According to Gad and Weil (1986), the limitations are: 1. statistics cannot make poor data better; 2. statistical significance may not imply biological significance; 3. an effect that may have biological significance may not be statistically significant; 4. the lack of statistical significance does not prove safety. Statistical analysis cannot rescue poor data resulting from a flawed design or

Table 19.1. Analysis of qualitative data of urinalyses and pathological findings (Kobayashi, 2010)

Incidence

2×2 Table	
Control: Observed (+)	Control: None (–)
Treatment: Observed (+)	Treatment: None (–)
(1) Chi square test	
(2) Fisher's test	
Note: Small numerical values (0–5) are not suitable for Chi square analysis in the four-values data set (Control: +, – and Treatment: +, –). Fisher's test (one-sided) is suitable for the data with small numerical values.	

4×4 Table, Grades and number of findings with the grades in Groups				
Group	No finding (–)	Slight (+)	Moderate (++)	Marked (+++)
Control	10	1	0	0
Low	4	3	2	1
Mid	1	4	3	2
High	0	3	4	3
Note 1: If Chi square analysis by 4×4 Table shows a significant difference, Control Group *vs* Low dose Group, Control Group *vs* Mid dose Group and Control Group *vs* High dose Group are analysed by 2×4 Table by division. Note 2: If the number of animals in a group is ≥5, use of Mann-Whitney's *U* test is preferred. Note 3: Cochran-Armitage trend test is the preferred tool for examining dose-related pattern.				

Table 19.2. Parametric and non-parametric statistical tools for the analysis of data obtained from toxicology studies

Group settings	Parametric test	Non-parametric test
Only two groups	Student, ◎Aspin-Welch, Cochran-Cox *t*-tests	Mann-Whitney *U* test, Wilcoxon test
Three or more group	ANOVA	Kruskal-Wallis rank sum test
	◎Dunnett's multiple comparison test, General, multiple comparison test	Nonparametric type Dunnett's rank sum test
		◎Steel's test
	Tukey's multiple range test (the size of the group is the same)	Nonparametric type Tukey's rank sum test
	Tukey-Kramer's multiple range test (the size of the group is different)	◎Steel-Dwass' test
	◎Duncan's multiple range test	Nonparametric type Duncan's rank sum test
	Scheffé's multiple comparison test	Nonparamteric type Scheffé's rank sum test
	◎Williams's *t*-test (analyzes the difference of the mean values between each treated group and control, when the mean value of the treated groups changes in one direction.)	Shirley-Williams's test
	—	Jonckheere's trend test

◎Tests recommended.

Table 19.3. Statistical tools suggested for the comparison of two and multi-groups

Group setting	Comparison	Analysis
Only two groups	Only one time	Aspin-Welch's *t*-test
Control(x_0), Low dose(x_1), Mid-dose(x_2), High dose(x_3)	Analysis of difference of the chisel between control group and each dose group (the analysis frequency is three times)	Dunnett's multiple comparison test; Williams's *t*-test (assumption: data possess a dose-dependency)
Control, Drug A, Drug B, Drug C or Group A, Group B, Group C, Group D	Analysis of difference between control group and each drug or group (total number of comparisons made is three)	Dunnett's multiple comparison test
	Comparison of all pairs (total number of comparisons made is six)	Tukey's multiple range test; Duncan's multiple range test
Control(x_0), Low dose(x_1), Mid dose(x_2), High dose(x_3), Reference drug (R_1)	Analysis of difference between control group and Reference drug followed by comparison of control group with each dose group.	Dunnett's test or Williams's *t*-test. Examine if there is a significant difference between x_0 and R_1 by *t*-test; if there is a significance, then compare the control with x_1, x_2 and x_3, excluding R_1 using the tests of Dunnett or Williams.

a poorly conducted study. An appropriate data analysis will follow directly from a correct experimental design (including the selection of statistical methods to be applied) and implementation (OECD, 2010). According to Altman and Bland (1994), 'failing to reject the hypothesis often leads to the conclusion of evidence in favour of safety, simply because absence of evidence is not evidence of absence'.

References

Altman, D. and Bland, M. (1994): Regression towards the mean. British Med. J., 308, 1499.

Bailer, A.J., and Portier, C.J. (1988): Effects of treatment-induced mortality and tumor-induced mortality on tests for carcinogenicity in small samples. Biometrics, 44, 417–431.

Cox, D.R. (1972): Regression models and life-tables. J.R. Stat. Soc., B34, 187–220.

Dago, K.T., Luthringer, R., Lengellé, R., Rinaudo, G. and Macher, J.P. (1994): Statistical Decision Tree: A Tool for Studying Pharmaco-EEG Effects of CNS-Active Drugs. Neuropsychobiol., 29, 91–96.

Dunn, O.J. (1964): Multiple comparisons using rank sums. Technometrics, 6, 241–252.

Dunnett, C.W. (1955): A multiple comparison procedure for comparing several treatments with a control. J. Am. Stat. Assoc., 50, 1096–1121.

Gad, S. (2006): Statistics and Experimental Design for Toxicologists and Pharmacologists. 4th Edition, Taylor and Francis, Boca Raton, FL, USA.

Gad, S. and Weil, C.W. (1986): Statistics and Experimental Design for Toxicologists. The Telford Press Inc., New Jersey, U.S.A.

Hamada, C., Yoshino, K., Matsumoto, K., Nomura, M. and Yoshimura, I. (1998): Tree-type algorithm for statistical analysis in chronic toxicity studies. J. Toxicol. Sci., 23(3), 173–181.

Healey, G.F. (1997): How to achieve standardization of statistical methods in toxicology. Drug Inf. J., 31–32, 327–334.

Hess, K.R. (2011): Statistical design considerations in animal studies published recently in cancer research. Cancer Res., 15, 71(2), 625.

Hollander, M. and Wolfe, D.A. (1973): Nonparametric Statistical Methods, John Wiley and Sons, New York, USA.

Hothorn, L.A. (2002): Selected biostatistical aspects of the validation of *in vitro* toxicological assays. ATLA, 30 (Supp. 2), 93–98.

Howell, D.C. (2008): Fundamental Statistics for the Behavioral Sciences. 6th Edition, Thomson Wadsworth, Belmont, USA.

Jonckheere, A.R. (1954): A distribution-free *k*-sample test against ordered alternatives. Biometricka, 41, 133–145.

Kaplan, E.L. and Meier, P. (1958): Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc., 53, 457–481.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, M.F.W., Cuthill, I.C., Fry, D. Hutton, J. and Altman, D.J. (2009): Survey of the quality of experimental design, statistical analysis and reporting of research using animals. PLoSOne, 4(11), 1–11.

Kim, B.S., Zhao, B., Kim, H.J. and Cho, M. (2000): The statistical analysis of the *in vitro* chromosome aberration assay using Chinese hamster ovary cells. Mutation Res., 469, 243–252.

Kobayashi, K. (2010): Trend of statistics used for toxicity studies. Yakuji-niposha, Tokyo, Japan.

Kobayashi, K., Kanamori, M., Ohori, K. and Takeuchi, H. (2000): A new decision tree method for statistical analysis of quantitative data obtained in toxicity studies in rodents. San Ei Shi., 42(4), 125–129.

Kobayashi, K., Pillai, K.S., Guhatakurta, S., Cherian, K.M. and Ohnishi, M. (2011): Statistical tools for analysing the data obtained from repeated dose toxicity studies with rodents: A comparison of the statistical tools used in Japan with that of used in other countries. J. Environ. Biol., 32(1), 11–16.

Kobayashi, K., Pillai, K.S., Suzuki, M. and Wang, J. (2008): Do we need to examine the quantitative data obtained from toxicity studies for both normality and homogeneity of variance? J. Environ. Biol., 29(1), 47–52.

Kobayashi, K., Watanabe, K. and Inoue, H. (1995): Questioning the usefulness of the non-parametric analysis of quantitative data by transformation into ranked data in toxicity studies. J. Toxicol. Sci., 20(1), 47–53.

Krores, R., Renwick, A.G., Cheeseman, M., Kleiner, J., Piersma, A., Schilter, B., Schlatter, J., van Schothorst, F., Vos, J.G. and Wurtzen, G. (2004): Structure-based thresholds of toxicological concern (TTC): Guidance for application to substances present at low levels in the diet. Food Chem. Toxicol., 42(1), 65–83.

Morrison, D.F. (1976): Multivariate Statistical Methods, 2nd Edition, McGraw-Hill Book Co., New York, USA.

Nomura, M. (1994): International comparison of statistical analysis methods for toxicological study. Jap. Soc. Biopharm. Statistics, 40, 1–36.

NTP. National Toxicology Program, USA. http://ntp.niehs.nih.gov/go/10007

OECD (2010): Organisation for Economic Cooperation and Development. OECD Draft Guidance Document N° 116 on the Design and Conduct of Chronic Toxicity and carcinogenicity Studies, Supporting TG 451, 452, 453. OECD, Paris, France.

Piegorsch, W.W. and Bailer, A.J. (1997): Statistics for Environmental Biology and Toxicology, Section 6.3.2., Chapman and Hall, London.

Portier, C.J. and Bailer, A.J. (1989) : Testing for increased carcinogenicity using a survival-adjusted quantal response test. Fund. Appl. Toxicol., 12, 731–737.

Sakaki, H., Igarashi, T., Ikeda, Y., Mizoguchi, K., Omichi, T., Kadota, M., Kawada, T., Takizawa, O., Tsukamoto, Y., Terai, K., Tozuka, A., Hirata, J., Handa, H., Mizuma, Z., Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., 71–81.

Shirley, E. (1977): A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. Biometrics, 33, 386–389.

Tarone, R.E. (1975): Tests for trend in life table analysis. Biometrika, 62, 679–682.

Williams, D.A. (1971): A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics, 27, 103–117.

Williams, D.A. (1972): The comparison of several dose levels with a zero dose control. Biometrics, 28, 519–531.

Williams, D.A. (1986): A note on Shirley's nonparametric test for comparing several dose levels with a zero-dose control. Biometrics, 42, 182–186.

Yamazaki, M., Noguchi, Y., Tanda, M. and Shintani, S. (1981): Statistical method appropriate for general toxicological studies in rats. J. Takeda Res. Lab., 40 (3), 163–187.