The Japanese Society of Toxicology

Fundamental Toxicological Sciences

URL : http://www.fundtoxicolsci.org/index_e.html

Original Article

Testing normality for quantitative values obtained from repeated dose administration toxicity studies – Fraught with challenges

Katsumi Kobayashi[1], Kalathil Sadasivan Pillai[2], Lakshmi Narashimhan Ramana[3] and Mebin Wilson Thomas[4]

[1]Ex-Faculty of Policy Studies, Keio University Shonan Fujisawa Campus (SFC), Part-time Lecturer, Endo, Fujisawa-shi, Kanagawa, 252-0816, Japan
[2]PNB Vesper Life Science. Kochi, Kerala, India
[3]Department of Chemistry and Biochemistry, Jain University, Bengaluru, Karnataka, India
[4]Department of Forensic Science, School of Sciences, Jain University, Bengaluru, Karnataka, India

(Received February 8, 2025; Accepted February 13, 2025)

ABSTRACT — In repeated dose administration toxicity studies, which are regulatory requirements for the safety evaluation of drugs, pesticides, etc., the analysis of between-group differences is typically carried out using either parametric or nonparametric statistical methods. The choice of the method depends on whether the data distributions within the groups are normal or not, and or are homogeneous. In theory, testing for normality is important because many parametric tests, such as the *t*-test and ANOVA, assume that the data within each group follows a normal distribution. However, in repeated dose administration toxicity studies, the data are not always explicitly tested for normality. One reason for this is that there is no universally accepted threshold for deciding whether data is "normal enough" for parametric tests. Another reason is the power of normality tests varies depending on the sample size. In repeated dose administration studies, the number of animals in each group is often small (5–20). In such cases, normality tests may not provide meaningful results due to low statistical power, and the decision to use a parametric or nonparametric test often relies on other considerations, such as variance. It is a common practice to use equal-variance tests (for example Bartlett's test, Levene's test, etc.) to determine whether parametric or nonparametric methods should be used for analyzing data. In repeated dose administration toxicity studies for assessing normality the Shapiro-Wilk's *W* (Shapiro-Wilk) test is recommended.

Key words: Normal distribution, Normality, Repeated dose administration toxicity studies, Power of normality test, Shapiro-Wilk's *W* test

INTRODUCTION

In repeated dose administration toxicity studies, before comparing groups (eg., treatment vs. control), there is a need to confirm whether the data follow a normal distribution. This is important because many statistical tests, especially parametric ones, assume data are normally distributed (Hamada, 2018). Statistical decision trees in the literature recommend testing the data for normality as the first approach for selecting a parametric or non-parametric test for subsequent statistical analysis (Hothorn, 2014; OECD, 2010). The other assumption of parametric tests is equal variance across the groups (Hazra and Gogtay, 2016). However, in real-world applications, in these studies, the assumption of normality is often not tested before performing statistical analyses. The reason for not testing

Correspondence: Katsumi Kobayashi (E-mail: student-kobayashi@beige.plala.or.jp)

normality is that the Type I error and power of the F-statistic are not altered by moderate violation of normality (Blanca *et al.*, 2023). However, if the violation of normality is serious, not checking normality could result in inappropriate statistical methods, which could mislead interpretations in safety evaluation studies that are critical for public health. This paper discusses the merits/demerits of examining data for normality and recommends Shapiro-Wilk's *W* test for normality, if normality testing is deemed necessary.

MATERIALS AND METHODS

This paper used data from long-term repeated dose administration toxicity studies with rodents conducted at Anpyo Center Inc., Shizuoka, Japan, NIHS (National Institute of Health Sciences, Japan) database, and several publications. SAS JMP software version 5.1 and Excel 2008 were used for the statistical analysis.

RESULTS

A note on normal distribution

Normal distribution is often referred to as a bell-shaped curve or Gaussian distribution. This distribution is characterized by its symmetric shape, where most of the data points are clustered around the mean, and the frequency of extreme values decreases symmetrically from the mean. In repeated dose administration toxicity studies, body weight, feed intake, red blood cell count, etc of rats usually follow a normal distribution. On the other hand, Kobayashi *et al*. (2011) observed that in carcinogenicity/chronic toxicity test using rats, cholinesterase, white blood cell count, platelet count, urine osmolality, uterine weight, lactate dehydrogenase, methemoglobin, triglyceride, urine volume, neutrophil fraction, creatine phosphokinase, and γ-glutamyltranspeptidas, aspartate aminotransferase, and alanine aminotransferase do not follow a normal distribution. Kobayashi (2005) observed that the data that do not show a normal distribution usually have a high coefficient of variation (CV) greater than 20%.

Necessity of performing normality tests in repeated dose administration toxicity studies

Textbooks generally recommend performing a normality test before applying parametric statistical tests (Mishra *et al*., 2019) such as the Student's *t*-test, and multiple comparison/range tests (eg., Dunnett's, Tukey's, Duncan's, and Scheffé's tests) (Rochon *et al*., 2012). These tests rely on the assumption that the data are normally distributed. Usually, the data are not examined for normality but examined for homogeneity of variance (homoscedasticity). If the test for homogeneity of variance passes, parametric tests are opted for the analysis of the data.

Amano *et al*. (1999) stated that most *t*-tests and *F*-tests rely on the assumption that the population being studied is normally distributed. Ichikawa (1986) mentioned that before applying the *F*-test to compare two groups, it is necessary to assume that the data follows a normal distribution. In this context, Ichikawa refers to the normal distribution as a precursor to applying the *Z*-test. According to the Graduate School of Pharmaceutical Sciences (2017) (REF MISSING), parametric methods assume that the population distribution follows a normal distribution. This is crucial because commonly used statistical methods, such as *t*-tests or ANOVA, compare population means and require normality to ensure that the comparisons are valid (Kashy *et al*., 2009). Katabami *et al*. (1977) stated that when using the *t*-test to compare the means of two groups, it is essential for both groups' data to follow a normal distribution. This is in line with the general assumption for *t*-tests, where the distribution of the sample means is expected to be normal. Nagata and Yoshida (1997) stated that in the context of two-sample *t*-tests when comparing treated *vs* non-treated groups, the data for both groups are assumed to follow a normal distribution. This again aligns with the standard assumptions for applying the two-sample *t*-test. The authors further stated that several multiple comparison methods, such as Tukey's, Dunnett's, Williams's, and Scheffé's, are also based on the normality assumption. Murata and Yano (2002) discussed the distinction in statistical analysis methods for the analysis of two groups, when data are normally distributed versus when they are not. They specified that if the data are normally distributed, then Student's *t*-test (for equal variances) and Welch's *t*-test (for unequal variances) can be used. However, if the data are not normally distributed, non-parametric methods such as the Wilcoxon signed-rank test, Wilcoxon rank-sum test, or Median test are used instead.

Many statisticians have stated that the *t*-test and various multiple comparison and range tests for parametric tests using sums of squares, variances, or standard deviations require a normality test before an equal variances test. However, no specific test for normality has been described.

Normality tests used in the US National Toxicology Program (NTP) technical report (TR)

An "empirical judgment" based on observations from the data is used for the evaluation of the normality of quantitative values (continuous variables) obtained

from the NTP TR long-term 2-year carcinogenic toxicity studies (TR 514/2004, TR 581/2014, TR 582/2015, TR 589/2016, TR 594/2017). In NTP studies, the body weight distributions of rodents are typically assumed to be approximately normal (Taylor-LaPole *et al.*, 2022). On the other hand, some of the hematological, biochemical, and urinalysis test results may show skewed distributions. Similarly, several data on hematology, clinical chemistry, urinalysis, cell proliferation, and sperm cells are usually not normally distributed. Before proceeding with statistical analysis, the NTP checks the data for outliers to maintain the validity and reliability of the study results.

Normality tests used in repeated-dose administration toxicity studies in Japan

In Japan, the normality test is generally omitted for 28-day repeated-dose administration toxicity test, combined repeated-dose toxicity test, and reproductive toxicity screening test for existing chemicals published by the NIHS, and the toxicity tests for pesticides and packaging materials published by the Food and Agricultural Materials Inspection Center and Food Safety Commission of Japan of Cabinet Office. In Japan, the focus is more on homogeneity of variance, than normality, and in most studies, the homogeneity of variance is determined by employing Bartlett's test.

Evaluation of normality in repeated dose administration studies-A brief literature review

Most of the literature did not provide information on whether the data were tested for normality before performing an ANOVA (Bemidinezhad *et al*., 2023; Benrahou *et al.*, 2022; Alelign *et al.,* 2020; Koriem *et al*., 2019; Traesel *et al.*, 2014). A few studies mentioned the tests used to assess the assumptions of normality and homogeneity of variances before conducting ANOVA (Wu *et al*., 2021; de Azevedo Mello *et al.*, 2020; Rodríguez-Lara *et al*., 2019; Aldana *et al*., 2005).

Reasons for preferring a homogeneity of variance test to normality

The reason normality tests are often not conducted before examining group differences in quantitative values in repeated dose administration toxicity studies can be attributed to - (1) Lack of clear guidelines: While several normality tests such as the Kolmogorov-Smirnov (Kolmogorov, 1933; Smirnov, 1948) Lilliefors (Peat and Barton, 2005), D'Agostino (D'Agostino, 1971), Anderson-Darling (Oztuna *et al*., 2006), Shapiro-Wilk (Shapiro and Wilk, 1965), and chi-square goodness-of-fit tests (Zar, 2010), are available, there is no widely adopted consensus or guideline for which test should be used, (2) Small Sample Sizes: In repeated dose administration toxicity studies with rats, the sample sizes for each group are typically small—usually 5-10 animals/sex/group for 28- and 90- day duration studies. With such small sample sizes, many normality tests lose power and may not provide reliable conclusions. For example, the Shapiro-Wilk test is commonly used for smaller sample sizes and is typically assumed to show normality when the group size is small, even though the underlying data may not strictly adhere to a normal distribution, (3) Non-normality in one group: When a single group fails the normality test, choosing an appropriate multiple comparison test can be challenging, (4) Non-normality across multiple groups: When the control and high-dose groups (or any other groups) exhibit non-normality, this creates additional complexity when comparing groups. In this situation, researchers often choose non-parametric tests, like Kruskal-Wallis test for multiple comparisons or the Mann-Whitney *U* test for pairwise comparisons, (5) Appropriate use of non-parametric tests: In cases where the assumption of normality is not met, non-parametric tests mentioned-above are commonly used. These tests do not require the assumption of normality and are more appropriate when data are skewed. However, applying these tests can be less powerful than parametric methods, especially when normality is not severely violated, (6) Challenges of testing group differences in small sample sizes: With small sample sizes, it is not always possible to confidently determine normality or to test for differences between groups in a statistically powerful way. In such studies, researchers may often rely on visual assessments (such as histograms or Q-Q plots). (7) When data distribution is uncertain from visual assessments, formal goodness-of-fit tests can be used (Hazra and Gogtay, 2016). However, the assumption of the homogeneity of variance (homoscedasticity) should be met for the validity of the ANOVA test (Azizi *et al.*, 2022).

For approximating normality and homogeneity of variance, continuous data may be transformed to log or square root. The analysis of such transformed data may result in erroneous conclusions if normality and equal variances assumptions are seriously violated (Shockley and Kissling, 2018).

Outline of the normality test methods

A brief description of the commonly used normality tests in repeated dose administration toxicity studies is given below:

(1) Shapiro-Wilk's Test: This test is one of the most widely used tests for normality, especially when dealing

with smaller sample sizes. It tests the null hypothesis that the data comes from a normal distribution. A significant result (*P*-value less than the chosen alpha, typically 0.05) suggests that the data do not follow a normal distribution. The SAS JMP recommends using this test when the sample size (N) is less than 2000. The weakness of this test is that in large sample sizes, it may detect minor deviations from normality that are not practically significant.

(2) D'Agostino test: This widely used test performs with similar power to Shapiro-Wilk's Test (Ehsanian *et al.*, 2024; Le Boedec, 2016).

(3) Anderson-Darling test: This test is also recommended for normality (Henderson, 2006).

(4) Kolmogorov-Smirnov's Test (K-S Test): This is yet another commonly used test for normality. The K-S test compares the observed cumulative distribution function (CDF) of data with the CDF of a normal distribution. It is generally used when comparing an empirical distribution with a normal distribution. This test can be used for larger sample sizes (N>2000). The K-S test is less powerful than the Shapiro-Wilk test for small sample sizes, and its power decreases as the sample size increases.

(5) Lilliefors's Test: Lilliefors's test is a variation of the Kolmogorov-Smirnov test designed specifically for normality testing when the mean and variance are unknown *a priori*, which is often the case in real-world studies. It is useful for small sample sizes. Like the K-S test, it can have reduced power in large sample sizes.

(6) Chi-Square Distribution Method: This method involves dividing the data into intervals and testing the frequencies of observations within those intervals against the expected frequencies under the normality assumption. The chi-square test is applied to the difference between the observed and expected counts. This method can be used as an alternative to formal normality tests, particularly when normality is assessed visually using histograms. The weakness of this test is that the results can be sensitive to the number of intervals (degrees of freedom), and the choice of interval boundaries can influence the outcome. It also works best with larger sample sizes and may not be reliable for small sample sizes or if the intervals are poorly chosen.

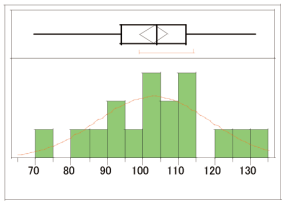
(7) Visual Judgment Based on Histogram: Visual inspection of histograms is a simple but common way to assess normality. By looking at the shape of the histogram, one can make a subjective judgment about whether the data approximate a normal distribution (bell-shaped curve). In toxicology, this can serve as a quick check or be used in conjunction with statistical tests. The weakness of this method is that it is subjective and prone to bias. Small deviations from normality may be missed or overemphasized based on how the histogram is presented or interpreted.

The Shapiro-Wilk test is often the preferred choice to other normality tests for testing normality in repeated dose administration toxicity studies, where smaller sample sizes are involved. The K-S test is versatile but is less powerful for small sample sizes. Lilliefors's test is a variant of the K-S test and is not as powerful as the Shapiro-Wilk test for small samples. Though the Shapiro-Wilk test is widely used for testing normality in repeated dose administration toxicity studies, this test has certain weaknesses (see below):

Shapiro-Wilk test does not hold normality as the number of animals increases

Shapiro-Wilk test was applied to examine the normality of body weight distribution of rats in four different group sizes (Group sizes 1, 2. 3, and 4). In Group size 1, n=17. In Group size 2, the body weights of rats of Group size 1 were repeated 2 times, in Group size 3 repeated 3 times, and in Group size 4 repeated 4 times. The results are shown in Table 1.

Table 1. Power of Shapiro-Wilk test for different group sizes of rats.

Number of rats	Histograms	Mean	Coefficient of variation (%)	Shapiro-Wilk test		Kolmogorov-Smirnov's test	
				W	*P*	*D*	*P*
17 (Group size 1)		103	15.5	0.98727	0.9891 (NS)	0.15669	> 0.2 (NS)
34 (Group size 2)			15.3	0.96874	0.5017 (NS)	0.12969	> 0.2 (NS)
51(Group size 3)			15.2	0.95988	0.1486 (NS)	0.12071	> 0.2 (NS)
68 (Group size 4)	Number = 68		15.2	0.95486	**0.0383** (S)	0.11623	0.1162 (NS)

NS; not significant (normal distribution). S; significant (non-normal distribution).

Table 2. Change in the normality of body weight of male F344 rats in control groups in 52-week studies.

Calculated value	Study No.									
	1	2	3	4	5	6	7	8	9	10
Mean ± S.D.	355 ± 20	396 ± 26	344 ± 24	351 ± 21	361 ± 22	384 ± 20	355 ± 18	358 ± 18	371 ± 29	358 ± 16
N	59	70	50	60	49	69	50	50	50	49
Coefficient of variation (%)	5.6	6.6	7.0	6.0	6.1	5.2	5.1	5.0	7.8	4.5
W	0.9587	0.9721	0.9800	0.7928	0.9743	0.9780	0.9760	0.9767	0.9778	0.9787
P (Prob < *W*)	0.0912	0.3140	0.7299	0.3964	0.5262	0.5499	0.5809	0.6089	0.6496	0.6895
P-value of cumulative number of animals	0.1069									
		0.1218								
			0.0144(S)							
				0.0070(S)						
					0.0144(S)					
						0.0460(S)				
							0.0141(S)			
								0.0365(S)		
									0.0153(S)	

S; significant (not normally distributed)

Table 3. Change in normality of platelet count in 7 tests at 104 weeks in male F344 rats used as the control groups.

Calculated value	Study No.						
	1	2	3	4	5	6	7
Mean ± S.D.	611 ± 136	648 ± 137	647 ± 104	797 ± 194	679 ± 125	724 ± 115	733 ± 150
N	41	38	40	37	40	38	41
Coefficient of variation (%)	22	21	16	24	18	16	20
W	0.7876	0.8487	0.9172	0.8690	0.8731	0.8366	0.8501
P (Prob < *W*)	< 0.0001(S)	< 0.0001(S)	0.0069(S)	0.0003(S)	0.0002(S)	< 0.0001(S)	< 0.0001(S)
P-value of cumulative number of animals	0.0000(S)						
		< 0.0001(S)					
			0.0000(S)				
				< 0.0001			
					< 0.0001(S)		
						0.0000(S)	

S: significant (not normally distributed).

The body weight distribution shown in the histogram is identical, irrespective of group sizes. The Shapiro-Wilk value decreases, as the sample size increases, showing a significance when n=68. The power for the detection of non-normality increases with the increase in the sample size. The Kolmogorov-Smirnov test showed normality at all the Group sizes.

Table 2 shows the results of the test of normality (Shapiro-Wilk test) for the body weight data (control groups) obtained from 2-year combined chronic toxicity/carcinogenicity studies of pesticides (10 studies). Body weight normally distributed for each of the 10 groups, loses normality with the increase in the number of animals in the group. As the number of animals (sample size) in each group increases, the Shapiro-Wilk test begins to detect non-normality. This implies that, with larger sample sizes, the test becomes more sensitive to even small deviations from normality.

Table 3 shows the results of the Shapiro-Wilk test for normality for platelet counts at 104 weeks in male F344 rats. The platelet counts are usually not normally distributed. The platelet counts that did not show a normal distribution in the control groups did not show a normal distribution when the number of samples in the groups increased.

Shapiro-Wilk test is a popular test for normality, but it does not allow for adjusting the number of groups, making it less flexible in some contexts, particularly for large samples. It might become less reliable as the sample size increases, potentially leading to a rejection of normality even if the data are approximately normal. The K-S test, while less powerful than the Shapiro-Wilk test, is recommended for larger samples. The K-S test has lower power in detecting deviations from normality, meaning it might

not detect subtle departures from normality in smaller datasets, but it works reasonably well for large samples due to its flexibility. The goodness-of-fit test using the chi-squared distribution involves grouping data into classes (or bins) and comparing the observed frequencies with expected frequencies under the assumption of normality. A critical point here is that the selection of class intervals (binning) significantly impacts the results.

The Shapiro-Wilk test is best suited for smaller samples, the K-S test is better for larger samples but with lower sensitivity, and the chi-squared test requires careful selection of class intervals. However, the Shapiro-Wilk test has limitations, particularly in detecting deviations from normality when the sample size is substantially small (Rochon *et al.,* 2012). The choice among these tests is context-dependent, and the researcher should tailor their approach based on the sample size, data distribution, and analytical goals.

DISCUSSION

Testing for normality in continuous data is a crucial step in deciding the statistical methods for data analysis (Mishra *et al.*, 2019). According to Neville *et al*. (2006), parametric tests can cause erroneous results unless a test for normality has been conducted beforehand. Currently, no universal tests are available to examine the normality of continuous variables in repeated dose administration toxicity studies. In these studies, the number of animals in each dose group typically ranges from 5 to 10. For large sample sizes ($n > 30$), normality is typically ensured by the Central Limit Theorem, and for small sample sizes ($n < 30$), normality must be specifically assessed (Sullivan *et al.*, 2016).

Normality can be assessed by visual examination (Ghasemi and Zahediasl, 2012). But, visual assessment of normality becomes more difficult when the sample size is between 5 and 20. Moreover, if normality is not observed in one of the groups, it complicates the selection of tests for comparing between-group differences. This makes reliance on normality tests less practical in these cases. Across the world, most scientists instead of focusing on normality tests, equality of variances across groups is assessed for selecting parametric or non-parametric statistical analysis. This is because variance homogeneity (equal variances) is a key assumption in widely used parametric tests like *t*-test and ANOVA. To conclude, while the use of normality tests is critical in the analysis of data from repeated dose administration toxicity studies, the reality is that these tests are not consistently applied. This gap between theory and practice can have significant implications for the validity and reliability of the results.

Conflict of interest---- The authors declare that there is no conflict of interest.

REFERENCES

Aldana, L., Bacardi, D., Merino, N., Cosme, K., Porras, D., Carreras, I., Alí, A., Suárez, J., Ariel Vázquez, A. and Cruz, Y. (2005): Safety evaluation of granulocyte colony-stimulation factor obtained at CIGB. Biotecnol. Apl., **22**, 50-53.

Alelign, T., Chalchisa, D., Fekadu, N., Solomon, D., Sisay, T., Debella, A. and Petros, B. (2020): Evaluation of acute and subacute toxicity of selected traditional antiurolithiatic medicinal plant extracts in Wistar albino rats. Toxicol. Rep., **7**, 1356-1365.

Amano, S., Nomura, K. and Yokohama, M. (1999): Keiryo-seibutsugaku, pp. 20, Sankyo Shuppan Co., Ltd. Tokyo.

Azizi, F., Ghasemi, R. and Ardalan, M. (2022): Two common mistakes in applying ANOVA test: Guide for biological researchers. Available at SSRN: https://ssrn.com/abstract=4221619 or http://dx.doi.org/.

Bemidinezhad, A., Zojaji, S.A., Taraz Jamshidi, S., Mohammadi, M., Alavi, M.S. and Ghorbani, A. (2023): Evaluation of acute, subacute, and subchronic toxicity of a hepatoprotective herbal formulation. Toxicol. Rep., **11**, 452-459.

Benrahou, K., Mrabti, H.N., Assaggaf, H.M., Mortada, S., Salhi, N., Rouas, L., El Bacha, R., Dami, A., Masrar, A., Alshahrani, M.M., Awadh, A.A., Bouyahya, A., Goh, K.W., Ming, L.C., Cherrah, Y. and Faouzi, M.E. (2022): Acute and subacute toxicity studies of Erodium guttatum extracts by oral administration in rodents. Toxins (Basel), **14**, 735.

Blanca, M.J., Arnau, J., García-Castro, F.J., Alarcón, R. and Bono, R. (2023): Non-normal data in repeated measures ANOVA: impact on Type I error and power. Psicothema, **35**, 21-29.

D'Agostino, R.B. (1971): An omnibus test of normality for moderate and large sample sizes. Biometrika, **58**, 341-348.

de Azevedo Mello, F., Magalhaes Silva, B.B., Barreiro, E.B., Franco, I.B., Nogueira, I.M., Nahas Chagas, P.H., Santos Parizi, J.L., Pereira, D.R., Rossi, R.C. and Nai, G.A. (2020): Evaluation of genotoxicity after acute and chronic exposure to 2,4-dichlorophenoxyacetic acid herbicide (2,4-D) in rodents using machine learning algorithms. J. Toxicol. Sci., **45**, 737-750.

Ehsanian, R., Buttner, J.A., Rivers, W.E., Nagpal, A., Patel, J., Zheng, P., McCormick, Z. and Schneider, B.J. (2024): Normality analysis of numeric rating scale scores in patients with chronic axial spine pain before and after medial branch blocks: a multicenter study. Pain Med., **25**, 591-599.

Ghasemi, A. and Zahediasl, S. (2012): Normality tests for statistical analysis: a guide for non-statisticians. Int. J. Endocrinol. Metab., **10**, 486-489.

Hamada, C. (2018): Statistical analysis for toxicity studies. J. Toxicol. Pathol., **31**, 15-22.

Hazra, A. and Gogtay, N. (2016): Biostatistics series module 3: Comparing groups: Numerical variables. Indian J. Dermatol., **61**, 251-260.

Henderson, A.R. (2006): Testing experimental data for univariate normality. Clin. Chim. Acta, **366**, 112-129.

Hothorn, L.A. (2014): Statistical evaluation of toxicological bioassays – a review. Toxicol. Res., **3**, 418-432.

Ichikawa, K. (1986): Tokeigaku-Shuchukouza, pp. 40, Nihon-sotsugo-

Kyoiku-center, Tokyo.

Kashy, D.A., Donnellan, M.B., Ackerman, R.A. and Russell, D.W. (2009): Reporting and interpreting research in PSPB: practices, principles, and pragmatics. Pers. Soc. Psychol. Bull., **35**, 1131-1142.

Katabami, J., Kanamori, M. and Honda, Y. (1977): Keito-Kangogaku-Koza Kiso, **4**, 108. Igaku-Shoin Ltd, Tokyo.

Kobayashi, K. (2005): Analysis of quantitative data obtained from toxicity studies showing non-normal distribution. J. Toxicol. Sci., **30**, 127-134.

Kobayashi, K., Sakuratani, Y., Abe, T., Yamazaki, K., Nishikawa, S., Yamada, J., Hirose, A., Kamata, E. and Hayashi, M. (2011): Influence of coefficient of variation in determining significant difference of quantitative values obtained from 28-day repeated-dose toxicity studies in rats. J. Toxicol. Sci., **36**, 63-71.

Kolmogorov, A. (1933): Sulla determinazione empirica di una legge di distribuzione. Attuari, **4**, 83-91.

Koriem, K.M., Arbid, M.S. and El-Attar, M.A. (2019): Acute and subacute toxicity of *Ammi visnaga* on rats. Interdiscip. Toxicol., **12**, 26-35.

Le Boedec, K. (2016): Sensitivity and specificity of normality tests and consequences on reference interval accuracy at small sample size: a computer-simulation study. Vet. Clin. Pathol., **45**, 648-656.

Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C. and Keshri, A. (2019): Descriptive statistics and normality tests for statistical data. Ann. Card. Anaesth., **22**, 67-72.

Murata, K. and Yano, E. (2002): Igaku-Tokei, pp 34, Nankodo Co., Ltd, Tokyo. NIHS: JECDB (Japan Existing Chemical Database): National Institute of Health Sciences (nihs.go.jp) Accessed on August 27, 2023.

NTP Technical Report: https://ntp.niehs.nih.gov/publications/reports/tr Accessed on January 27, 2024.

Nagata, Y. and Yoshida, M. (1997): Tokeiteki-Tajuhikaku-no-Kouza, pp. 20 & 33, Scientist Press Co., Ltd, Tokyo.

Neville, J.A., Lang, W. and Fleischer, A.B. Jr. (2006): Errors in the Archives of Dermatology and the Journal of the American Academy of Dermatology from January through December 2003. Arch. Dermatol., **142**, 737-740.

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

Oztuna, D., Elhan, A.H. and Tuccar, E. (2006): Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. Turk. J. Med. Sci., **36**, 171-176.

Peat, J. and Barton, B. (2005): Medical Statistics: A guide to data analysis and critical appraisal. Blackwell Publishing, Carlton Australia.

Rochon, J., Gondan, M. and Kieser, M. (2012): To test or not to test: preliminary assessment of normality when comparing two independent samples. BMC Med. Res. Methodol., **12**, 81.

Rodríguez-Lara, A., Mesa, M.D., Aragón-Vela, J., Casuso, R.A., Vázquez, C.C., Zúñiga, J.M. and Huertas, J.R. (2019): Acute/subacute and sub-chronic oral toxicity of a hidroxytyrosol-rich virgin olive oil extract. Nutrients, **11**, 2133.

Shapiro, S.S. and Wilk, M.D. (1965): An analysis of variance test for normality (complete samples). Biometrica, **52**, 591-611.

Shockley, K.R. and Kissling, G.E. (2018): Statistical guidance for reviewers of toxicologic pathology. Toxicol. Pathol., **46**, 647-652.

Smirnov, N. (1948): Table for estimating the goodness of fit of empirical distributions. Ann. Math. Stat., **19**, 279-281.

Sullivan, L. M, Weinberg, J and Keaney, J. F. Jr. (2016): Common statistical pitfalls in basic science Research. J. Am. Heart Assoc., **29**, e004142.

Taylor-LaPole, A.M., Cunny, H.C. and Shockley, K.R. (2022): Statistical analysis of rodent body weight data is robust to departures from normality in historical national toxicology program studies Dated 1980-2013. J. Young Investig., **25**, 1-12.

Traesel, G.K., de Souza, J.C., de Barros, A.L., Souza, M.A., Schmitz, W.O., Muzzi, R.M., Oesterreich, S.A. and Arena, A.C. (2014): Acute and subacute (28 days) oral toxicity assessment of the oil extracted from Acrocomia aculeata pulp in rats. Food Chem. Toxicol., **74**, 320-325.

Wu, J., Zhang, Y., Lv, Z., Yu, P. and Shi, W. (2021): Safety evaluation of Aloe vera soft capsule in acute, subacute toxicity and genotoxicity study. PLoS One, **16**, e0249356.

Zar, J.H. (2010): Biostatistical analysis 5th Ed. Pearson Education. Inc. New Jersey. https://bayesmath.com/wp-content/uploads/2021/05/Jerrold-H.-Zar-Biostatistical-Analysis-5th-Edition-Prentice-Hall-2009.pdf