10
Regression Analysis

History

The origin of the term 'regression' in statistics has an interesting history. Francis Galton (1822–1911) had deep interest in heredity, biometrics and eugenics (Crow, 1993). He found that sons of tall men to be shorter than their fathers. He called this phenomenon regression towards the mean, and thus the term 'regression' originated (Dupont, 2002).

Unlike correlation, where there is no 'dependence relationship', there are dependent and independent variables in regression analysis. In regression analysis, y is assumed to be a random variable and x is assumed to be a fixed variable. The underlying assumption of regression analysis is that the dependent variable follows a normal distribution and scatter about the regression line.

In animal experiments regression analysis is used to evaluate cause (variable x) and effect (variable y) relationships; for example in a repeated dose administration study, the rate of decrease in body weight (y) as the exposure period (x) increases can be determined using regression analysis.

Linear Regression Analysis

The regression equation is:

$y = a + bx,$ where y = Dependent variable, x = Independent variable, a = Intercept and b = slope.

The intercept represents the estimated average value of y when x equals zero and the slope represents the estimated average change in y when x increases/decreases by one unit. Slope and intercept are derived using the least-square method.

If the underlying assumptions of the least-square model are not met, the regression slope and intercept may be incorrect. Two factors that cause incorrect regression coefficients are: (i) imprecision in the measurement of the independent (x) variable and (ii) inclusion of outliers in the data analysis (Cornbleet and Gochman, 1979). Outliers have profound effect on the slope (Farnsworth, 1990; Glaister, 2005).

The slope, b is calculated using the formula:

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

The intercept a can be calculated from the equation:

$$\bar{y} = a + b\ \bar{x}$$

Let us work out an example for calculating b and a. Body weight of babies measured in different months is given in Table 10.1. Month is the independent variable (x) and the body weight is the dependent variable (y).

Table 10.1. Body weight of babies measured in different months

Age (Month) (x)	Body weight (kg) (y)	x^2	y^2	xy
1	3.8	1	14.44	3.8
2	4.2	4	17.64	8.4
3	4.8	9	23.04	14.4
5	5.7	25	32.49	28.5
6	6.4	36	40.96	38.4
7	6.9	49	47.61	48.3
8	7.1	64	50.41	56.8
9	7.8	81	60.84	70.2
10	8.6	100	73.96	86
12	10.4	144	108.16	124.8
$\Sigma x = 63$ $\bar{X} = 6.3$	$\Sigma y = 65.7$ $\bar{y} = 6.57$	$\Sigma x^2 = 513$	$\Sigma y^2 = 469.55$	$\Sigma xy = 479.6$

We shall calculate the slope, b first:

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \times \sum y}{n} = 479.6 - \frac{63 \times 65.7}{10} = 65.69$$

$$\sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 513 - \frac{(63)^2}{10} = 116.1$$

$$\sum (y-\bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} == 469.55 - \frac{(65.7)^2}{10} = 37.90$$

$$b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{65.69}{116.1} = 0.5658$$

Once the slope, *b* is calculated, it is easy to calculate the intercept, *a*:

$$\bar{y} = a + b\,\bar{x}$$

$$6.57 = a + 0.5658 \times 6.3$$

$$a = 6.57 - (0.5658 \times 6.3) = 3.005$$

Regression equation:

$$y = a + bx$$

$$y = 3.005 + 0.5658\,x$$

Significance of regression line can be determined by ANOVA (Table 10.2).

We wish to test the hypothesis:

H_0: b = 0 *vs* H_1: b ≠ 0, where b is the slope.

Table 10.2. Significance of regression line by ANOVA

Source of variation	Degrees of freedom	SS	Mean SS	*F*
Total SS for $y = \sum y^2 - \frac{(\sum y)^2}{n}$	9	37.90	4.21	-
Reduction due to regression (Residual SS) = $\frac{\left[\sum xy - \frac{\sum x \times \sum y}{N}\right]^2}{\sum \left[x-\bar{x}\right]^2}$	1	37.17	37.17	407
Error	8	0.73	0.0913	-

SS—Sum of squares

Since the calculated *F*-value is greater than the *F*-Table value (Table 10.3), the null hypothesis is rejected and the alternative hypothesis (H_1: b ≠ 0) is accepted. This means the slope of the regression line is significantly different from 0, which implies that there is a significant relationship between age and body weight of the babies.

Table 10.3. *F*-distribution values at 0.1% probability level (Yoshimura, 1987)

$N_1 \backslash N_2$	1	2	3	4	5	6	7	8	9	10
8	25.42	18.49	15.83	14.39	13.49	12.86	12.40	12.05	11.77	11.54

The test of significance is based on the assumption that the distribution of the deviation from the regression line (residual values) of all the values of dependent variable, *y* is the same for all the independent varable, *x*. The residue of each observation is given by the difference between the observed value and the fitted value of the regression line (Chan, 2004). Let us understand the terminology the residue of *y,* by plotting the data given in the Table 10.1. Figure 10.1 is the body weight *vs* age plot.

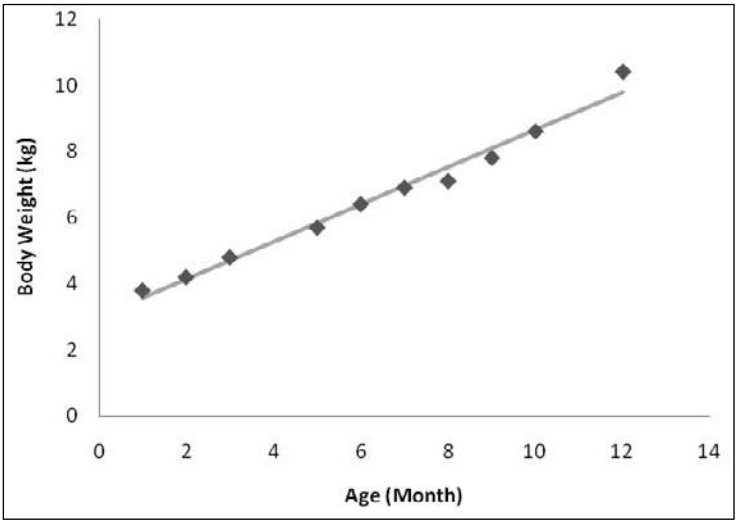

Figure 10.1. Body weight of babies measured in different months

Solid squares are the actual values. The line passing through the actual values is the regression line. For each value of *x* variable, the predicted *y* value is computed using the regression equation, $y' = 3.005 + 0.5658\ x$ (predicted *y* is denoted as *y*' in order to differentiate it from the actual *y*). Thus, *y*' is derived for each *x*, and the predicted *y*'s are joined together to obtain the regression line. By closely observing the plot, one can find that all the actual values do not fall on the regression line, though they are very close to the regression line. Linear regression line is called a 'best fit line', since it best fits the data points. The "best" fit line minimizes the squared vertical distances between the actual values and the line. An estimate of the squared vertical distances between the actual values and the line

(in other words, variation of the actual values from the predicted values) can easily be arrived at (*vide* Table 10.4). You would have noticed that this estimate is the sum of squares for error component given in the ANOVA Table (Table 10.2).

Table 10.4. Calculation of variation of the actual *y* values from the predicted *y*' values

Age (Month) (x)	Body weight (kg) (y)	y' ($y' = 3.005 + 0.5658\,x$)	$y - y'$	$(y - y')^2$
1	3.8	3.5708	0.2292	0.052533
2	4.2	4.1366	0.0634	0.00402
3	4.8	4.7024	0.0976	0.009526
5	5.7	5.834	–0.134	0.017956
6	6.4	6.3998	0.0002	0.00000004
7	6.9	6.9656	–0.0656	0.004303
8	7.1	7.5314	–0.4314	0.186106
9	7.8	8.0972	–0.2972	0.088328
10	8.6	8.663	–0.063	0.003969
12	10.4	9.7946	0.6054	0.366509
-	-	-	-	$\sum(y - y')^2 =$ 0.733249

Confidence Limits for Slope

95% confidence limits for the slope (b) can be derived by using the formula:

$b \pm t_{0.05.n-2}$ SE (b), where b is the slope (0.5658); $t_{0.05.n-2}$ is the critical value for t at 5 % probability level for n–2 degrees of freedom (2.306);

SE (b) is the standard error of $b = \sqrt{\dfrac{ErrorMeanSS}{\sum\left[x - \bar{x}\right]^2}} = \sqrt{\dfrac{0.0913}{116.1}} = 0.0280$

95% confidence limits for the slope (b) = 0.5658 ± (2.306 x 0.0280) = 0.5658 ±0.0646.

The significance of slope can be tested using the t-test, when the number of samples is smaller than about 30 (Bailey, 1995):

$t_{0.05.n-2} = \dfrac{b - \beta}{s / \sqrt{\sum\left[x - \bar{x}\right]^2}}$ where $t_{0.05.n-2}$ is the critical value for t at 5% probability level for n–2 degrees of freedom; b is the slope (b=0.5658);

β is the hypothetical value ($\beta = 0$) (we are testing whether the observed b value is different from the hypothetical value); s is the square root of error mean sum of squares

$$s = \sqrt{0.0913} = 0.3022 \;\; ; \;\; \sum\left[x - \bar{x}\right]^2 = 116.1.$$

$$t_{0.05.\text{n-2}} = \frac{0.5658 - 0}{0.3022 / \sqrt{116.1}} = \frac{0.5658}{0.0280} = 20.17$$

The derived t value (20.17) is greater than the Table t-value (2.228) at 5% probability level and 10 degrees of freedom; hence the slope is significant.

Comparison of Two Regression Coefficients

The regression coefficient, b measures how much the dependent variable, y changes (increases or decreases), for each unit change in the independent variable, x. The slopes of two similar studies can be compared using the formula:

$$\text{d} = \frac{b_1 - b_2}{\sqrt{\left[\frac{{s_1}^2}{\sum\left[X_1 - \overline{X_1}\right]^2} + \frac{{s_2}^2}{\sum\left[X_2 - \overline{X_2}\right]^2}\right]}}$$

Suffix 1 refers to independent variable x_1, and 2 independent variable x_2. Since d is normally distributed, the difference between b_1 and b_2 can be examined for statistical significance using t-test:

$$t = \frac{b_1 - b_2}{s\sqrt{\left[\frac{1}{\sum\left[X_1 - \overline{X_1}\right]^2} + \frac{1}{\sum\left[X_2 - \overline{X_2}\right]^2}\right]}}, \text{ where}$$

$$s = \sqrt{\frac{(n_1 - 2){s_1}^2 + (n_2 - 2){s_2}^2}{n_1 + n_2 - 4}}$$

The calculated t value is compared with the Table t-value at $n_1 + n_2 - 4$ degrees of freedom.

R^2

R^2 is interpreted as the proportion of total variability of the outcome that is accounted by the model (Vittinghoff *et al*., 2005). In other words, it is the proportion of the variation in the *y* variable that is "explained" by the variation in the *x* variable. R^2 is called as the 'coefficient of determination'. R^2 can vary from 0 to1. An R^2 close to 1 indicates that the actual *y* values fall almost right on the regression line. An R^2 close to 0 indicates that there is little or no relationship between *x* and *y*.

Multiple Linear Regression Analysis

In most situations, the dependent variable is associated with more than one independent variable. For example, the body weight of rats measured in a repeated dose administration study is associated with several independent variables like, age, sex and feed consumption of the animals. Multiple regression analysis is a very useful tool for finding out which independent variable/s has/have genuine relationship with the dependent variable. Multiple linear regression model is an extension of the simple linear regression model (Ambrosius, 2007).

The regression equation for two independent variables is:

$y = a + b_1x_1 + b_2x_2$, where y = Dependent variable, x_1 and x_2 are the independent variables, a = Intercept and b_1 = Slope of x_1 and b_2 = Slope of x_2.

We shall examine the steps involved in calculating multiple linear regression coefficient:

$$\sum\left(x_1-\overline{x_1}\right)^2 = A$$

$$\sum\left(x_1-\overline{x_1}\right)\left(x_2-\overline{x_2}\right) = B$$

$$\sum\left(x_2-\overline{x_2}\right)^2 = C$$

$$\sum\left(x_1-\overline{x_1}\right)\left(y-\overline{y}\right) = D$$

$$\sum\left(x_2-\overline{x_2}\right)\left(y-\overline{y}\right) = E$$

$$\sum\left(y-\overline{y}\right)^2 = F$$

$$b_1 = \frac{CD - BE}{AC - B^2}$$

$$b_2 = \frac{AE - BD}{AC - B^2}$$

Once the slopes are derived, *a* can be calculated using the formula:

$$y = a + b_1\bar{x}_1 + b_2\bar{x}_2$$

Multiple correlation coefficient can be computed using the formula:

$$R = \frac{\Sigma yy'}{\sqrt{\sum y^2 y'^2}}, \text{ where}$$

R = Multiple correlation coefficient; y = Actual value; $\bar{y}$'= Predicted y (calculated using the regression equation, $y = a + b_1\bar{x}_1 + b_2\bar{x}_2$;

$$\Sigma yy' = \left[\sum yy' - \frac{\sum y \times \sum y'}{n}\right]$$

$$\sum y^2 = \sum y^2 - \frac{(\sum y)^2}{n}; \quad \sum y'^2 = \sum y'^2 - \frac{(\sum y')^2}{n}$$

Significance of the multiple regression equation can be checked by ANOVA (Table 10.5).

Polynomial Regression

Linear regression does not hold good, when the data of your dependent variable follows a curved line, rather than a straight line. Transforming the *y* or *x* or both the variables to their logarithms, reciprocals, square roots etc., may straighten certain curves, but not all. Another way to solve this issue is to use a curvilinear regression equation. Polynomial regression equation is an example of curvilinear regression equation, which is used to predict toxicological variables (Vogt, 1989). Given the complexity of the calculations in polynomial regression analysis, it is not being included in the coverage of this book. The purpose of touching upon polynomial

Table 10.5. Significance of multiple regression equation by ANOVA

Source of variation	Degrees of freedom	SS	Mean SS
Total SS for *Y*	n–1	$\sum y^2 - \frac{(\sum y)^2}{n}$	-
Reduction due to regression (Residual SS)	k	$\sum y'^2 - \frac{(\sum y')^2}{n}$	$\sum Y'^2 - \frac{(\sum Y')^2}{n}$ /k
Error	n–k–1	$\left[\sum yy' - \frac{\sum y \times \sum y'}{n}\right]^2$	$\left[\sum yy' - \frac{\sum y \times \sum y'}{n}\right]^2$ /n–k–1

k is the number of independent variables.
F value is calculated by dividing Reduction due to regression (Residual SS) with error.

regression analysis, is to create awareness that before carrying out linear regression analysis one should ensure that the trend of the association between the two variables is linear.

Misuse of Regression Analysis

Use of a regression equation is considered to be inappropriate for estimating an independent variable, rather than a dependent variable (Williams, 1983). It is important to understand the nature of the data before choosing a regression model. This can be easily done by plotting the data, which will help understanding the nature of the data and selecting appropriate regression model. One should not fit a straight line using a linear regression equation for a 'non-linear data'.

References

Ambrosius, W.T. (2007): Topics in Biostatistics. Humana Press Inc., New Jersey, USA.

Bailey, N.T.J. (1995): Statistical Methods in Biology. Cambridge University Press, Cambridge, UK.

Chan, Y.H. (2004): Biostatistics 201: Linear regression analysis. Singapore Med. J., 45 (2), 55–61.

Cornbleet, P.J. and Gochman, N. (1979): Incorrect least-squares regression coefficients in method-comparison analysis. Clin. Chem., 25, 432–438.

Crow, J.F. (1993): Francis Galton: Count and measure, measure and count. Genetics, 135, 1–4.

DuPont, W.D. (2002): Statistical Modeling for Biomedical Researchers. Cambridge Univ. Press, Cambridge, U.K.

Farnsworth, D.L. (1990): The effect of a single point on correlation and slope. Internat. J. Math. Math. Sci., 13(4), 799–806.

Glaister, P. (2005): Robust linear regression using Theil's method. J. Chem. Educ., 82(10), 1472–1473.

Vittinghoff, E., Glidden, D.N., Shiboski, S.C. and McCulloch, C.E. (2005): Statistics for Biology and Health. Springer Science+Business Media, Inc., New York, USA.

Vogt, N.B. (1989): Polynomial principal component regression: An approach to analysis and interpretation of complex mixture relationships in multivariate environmental data. Chemometrics Intelligent Lab Systems, 7(1-2), 119–130.

Williams, G.P. (1983): Improper use of regression equations in earth sciences. Geology, 11(4), 195–197.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Inc., Tokyo, Japan.

11

Multivariate Analysis

Analysis of More than Two Groups

Student's *t*-test is used to test the equality of the means from two different populations (Rothmann, 2005). Use of Student's *t*-test for comparing more than two groups can cause Type I error. This can be better understood from the example below:

Absolute weight of the liver of female mice in a 13-week repeated dose administration study is given in Table 11.1.

Table 11.1. Liver weight (g) of female mice in a 13-week repeated dose administration study

Group	N	Mean ± SD	Tukey's multiple range test			Repeated comparison with Student's *t*-test		
			A	B	C	A	B	C
A	10	1.083±0.057	-	-	-	-	-	-
B	10	1.098±0.077	NS	-	-	NS	-	-
C	10	1.154±0.050	NS	NS	-	S	NS	-
D	10	1.273±0.062	S	S	S	S	S	S

NS—Not significant; S—Significant ($P<0.05$).

Repeated analysis by Student's *t*-test revealed a significant difference between Groups A and C. Actual increase in liver weight in Group C compared to Group A is only 6.6%. In this case, the significant difference between Groups A and C detected by repeated comparison with the *t*-test is caused by Type I error. When the groups were compared using Tukey's multiple range test, no significant difference was observed between Groups A and C (Tukey's multiple range test is the ideal test in this situation, since the number of groups to be compared is more than two).

There are several methods available for multiple comparison of means, but most of them have often been misused (Gill, 1990). An appropriate tool for analyzing more than two groups is analysis of variance (Wallenstein *et al*., 1980). One advantage of ANOVA (Analysis of Variance is abbreviated as ANOVA) is that it is easy to execute (Muir *et al*., 2006) and it has great utility and flexibility (Armstrong *et al.*, 2000). Like Student's *t*-test, for carrying out ANOVA, it is a prerequisite that homogeneity of variance prevails across all the groups (Moder, 2007) and the data has normal distribution. However, normality is rarely tested in ANOVA, because, a slight departure from normality does not affect the conclusion drawn from the analysis (Norman and Streiner, 2008).

ANOVA is also an excellent tool for analysing data obtained from factorial experiments. In a factorial experiment, there can be several factors at several levels. For example, to test a drug against hypercholesterolemia in rats, we may use a standard drug for comparison. The test drug and the standard drug are called factors. We may test these drugs at different dose levels. Depending upon the number and levels of factors, an ANOVA can be one-way, two-way or multi-way.

One-way ANOVA

One-way ANOVA is used to find if the given factor has significant effect on the expected outcome of the experiment. Jaundice index (x) of a newborn baby measured in weeks 36, 38 and 40 is presented in Table 11.2. We want to examine if the factor (week) has any significant effect on the jaundice index.

Table 11.2. Jaundice index (x) of newborn baby

Week					
36 (Group 1)		38 (Group 2)		40 (Group 3)	
x_1	13	x_{11}	9	x_{21}	5
x_2	6	x_{12}	11	x_{22}	5
x_3	11	x_{13}	11	x_{23}	4
x_4	12	x_{14}	10	x_{24}	7
x_5	14	x_{15}	7	x_{25}	7
x_6	10	x_{16}	7	x_{26}	3
x_7	9	x_{17}	5	x_{27}	3
x_8	11	x_{18}	8	x_{28}	4
x_9	11	x_{19}	7	x_{29}	5
x_{10}	10	x_{20}	10	x_{30}	3

Statistics

Estimates	Week		
	36 (Group 1)	38 (Group 2)	40 (Group 3)
N	10	10	10
Mean ± SD	10.7 ± 2.2	8.5 ± 2.0	4.6 ± 1.5
Sum	107	85	46
Grand sum		238	

Total sum of squares = $(x_1^{\,2} + x_2^{\,2} \cdots\cdots x_{29}^{\,2} + x_{30}^{\,2}) - \dfrac{(\sum x)^2}{\sum N} =$

$$= (13^2 + 6^2 \cdots\cdots 5^2 + 3^2) - \frac{(238)^2}{30} = 291.9$$

Sum of squares of among the groups

$$= \frac{107^2}{10} + \frac{85^2}{10} + \frac{46^2}{10} - \frac{(238)^2}{30} = 190.9$$

Total sum of squares for error = Total sum of squares—Sum of squares of among the groups
= 291.9–190.9 = 101

We have all the estimates required for constructing the ANOVA Table. See Table 11.3 given below:

Table 11.3. ANOVA Table

Source of variation	SS	DF	Variance (MS)	*F*-value	*P*
Total	291.9	29	-	-	-
Groups	190.9	2	95.5	25.5	$P<0.001$
Error	101	27	3.74		

SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.
Note: There are 30 observations, hence the DF for SS total is 30–1 = 29; Total number of groups are three, hence the DF for SS groups is 3–1 = 2; DF for error SS = DF for SS total—DF for groups SS (29–2 = 27).

$$F_{27}^{2} calc = \frac{95.5}{3.74} = 25.5$$

Compare the derived *F* value with the value given in the *F* distribution Table (Table 11.4):

Table 11.4. *F*-distribution values at 0.1% probability level (Yoshimura, 1987)

$N_1 \backslash N_2$	1	2	3	4	5	6	7	8	9	10
27	13.613	**9.019**	7.272	6.326	5.726	5.308	4.998	4.759	4.568	4.412

N_1—DF associated with the numerator (in this example, the DF associated with 95.5); N_2—DF associated with the denominator (in this example, the DF associated with 3.74). Since the calculated *F*-value is greater than the Table value, it is considered that the jaundice index of the newborn baby is significantly different among the weeks at 0.1% probability level.

post hoc Comparison

ANOVA indicates that the jaundice index of the newborn baby is significantly different among the groups. The question is, which group is different from the other group or groups? Are all the groups are different from each other? The possible comparisons that we can make in this particular example are:

Group 1 *vs* Group 2

Group 1 *vs* Group 3

Group 2 *vs* Group 3

There are several tests available in the literature for *post hoc* comparison. Few tests that are commonly used in pharmacology and toxicology are explained below:

Dunnett's multiple comparison test

Dunnett's multiple comparison test (Dunnett, 1955) is a widely used approach for comparing all groups with the control (Cheung and Holland, 1991).

To compare the Jaundice indices of weeks 36 and 38 with that of week 40 (*i.e.*, Group 1 *vs* Group 3 and Group 2 *vs* Group 3), Dunnett's multiple comparison test is the most appropriate statistical tool. Here, we are considering Group 3 as some sort of 'standard' or 'control'. Dunnett's multiple comparison test should not be used for other comparison, such as, comparison between Group 1 and Group 2.

Comparison between Group 1 and Group 3:

$$= \frac{10.7 - 4.6}{\sqrt{3.74} \times \sqrt{\frac{2}{10}}} = \frac{6.1}{0.8648} = 7.05 \qquad \therefore p < 0.001$$

Comparison between Group 2 and Group 3:

$$= \frac{8.5 - 4.6}{\sqrt{3.74} \times \sqrt{\frac{2}{10}}} = \frac{3.9}{0.8648} = 4.51 \qquad \therefore p < 0.001$$

The calculated values (7.05 and 4.51) are greater than the Dunnett's *t*-test critical value given in Table 11.5. Dunnett's *t*-test critical value at 3 (numerator)/27 (denominator) degrees of freedom is 3.674

Hence, it is considered that Jaundice indices of weeks 36 and 38 are different from that of week 40.

Table 11.5. Dunnett's *t*-test critical values (one-sided test at 0.1% probability level) (Yoshimura, 1987)

DF	2	3	4	5	6	7	8
27	3.422	**3.674**	3.821	3.922	3.999	4.061	4.114

Note: One-sided *t*-test is more appropriate in this example as it is an established fact that the jaundice index decreases in newborn babies as their age increases.

The power of the Dunnett's test decreases as the number of groups increases. This could be better understood from the data given in Table 11.6.

Table 11.6. Change in the power of the Dunnett's test when the number of groups increases

Data and tests	Control	Low dose	Mid dose	High dose	Top dose
Hemoglobin level (g/dl) of B6C3F1 male mice at Week 78	13.9, 14.3	14.0, 13.3	14.0, 13.8	14.1, 13.9	14.2, 14.2
	13.7, 13.8	15.0, 13.8	13.7, 13.8	14.3, 14.0	14.7, 13.9
	14.0, 14.3	14.1, 13.3	13.5, 14.1	14.2, 14.1	14.3, 13.7
	13.9, 13.7	14.1, 13.9	14.2, 13.8	14.3, 14.4	14.3, 14.4
	13.9, 13.5	13.8, 13.4	14.1, 14.0	14.4, 14.4	14.0, 14.3
N	10	10	10	10	10
Mean ± SD	13.9 ± 0.3	13.9 ± 0.4	13.9 ± 0.2	14.2 ± 0.2	14.2 ± 0.3
Rejection value in Dunnett's Table at 0.05 (two-sided)	2.45				
Statistical result		NS	NS	S	
Rejection value in Dunnett's Table at 0.05 (two-sided)	2.53				
Statistical result		NS	NS	NS	NS

NS-Not significant; S-Significant ($P < 0.05$)

In the four-group setting (control, low dose, mid dose and high dose), the high dose group showed a significant difference from the control group, whereas in the in the five-group setting (control, low dose, mid dose, high dose and top dose), no significant difference was seen in the high dose group compared to the control group, indicating a decrease in the power of Dunnett's test to detect a significant difference as the number of groups increases.

Tukey's multiple range test (Yoshida, 1980)

Tukey's multiple range test, also known as Tukey range test, Tukey's honest significance test (Tukey's HST) or the Tukey–Kramer test (Mathews, 2005), is used to compare all possible pairs of means.

This is exemplified by reviewing the example given in Table 11.2.

The variance of the error is 3.74 (Table 11.3).

$$S\bar{x} = \sqrt{\frac{3.74}{10}} = 0.6116$$

Find the Q (critical) value from the Table of Tukey (Table 11.7). In this example, Q at 5% probability level is 2.8882 [Number of groups = 2 ; Degrees of freedom for error = 30. Actual degrees of freedom of error is 27 (Table 11.3); since this value is not given in Table 11.7, the value 30 is considered].

Table 11.7. Tukey's critical value at 5% probability level (Yoshida, 1980)

Degrees of freedom for error	Number of Groups						
	2	3	4	5	6	8	10
24	2.9188	3.5317	3.9013	4.1663	4.3727	4.6838	4.9152
30	**2.8882**	3.4864	3.8454	4.1021	4.3015	4.6014	4.8241

Next step is the calculation of significant difference D. It is the product of $S\bar{x}$ and Q ($S\bar{x}$×Q):

$$D = S\bar{x} \times Q = 0.6116 \times 2.8882 = 1.7664$$

If the difference between any two means is greater than D, the difference is considered significant.

The difference between the means is given in Table 11.8. All means are different from each other.

Table 11.8. Jaundice index of newborn baby-Difference between mean values

Estimates	Week		
	36 (Group 1)	**38 (Group 2)**	**40 (Group 3)**
Mean	10.7	8.5	4.6
Difference of Means	Group 1 and Group 2	2.2[a]	Significant (P<0.05)
	Group 1 and Group 3	6.1[b]	Significant (P<0.05)
	Group 2 and Group 3	3.9[c]	Significant (P<0.05)

Note: The superscripts of the mean values can be explained as—"Values bearing similar superscripts are statistically the same". Since the superscripts of the mean values are different, it can be stated that each mean value is different from the other.

Williams's test

Most of the regulatory guidelines prescribe that the repeated-dose administration studies with rodents should be conducted with a minimum of three levels of doses (low, mid and high doses) and a control group (OECD, 1995). The high dose is chosen with the aim to induce toxicity but not death or severe suffering (OECD, 1998; EPA, 2000), whereas the low dose is chosen with the assumption that animals exposed to this dose level will not show any effect of the treatment compared to the control group (Kobayashi *et al*., 2010). However, these guidelines do not state how to determine the mid dose. It only indicates that this dose is required to examine dose dependency. According to Gupta (2007), the mid dose selection should consider threshold in toxic response and mechanism of toxicity. Choosing the mid dose is as important as choosing the high and low doses in repeated dose administration studies, since mid dose plays a determining role in establishing the dose dependency. It is not uncommon to encounter situations where mid dose alone shows an insignificant difference compared to the control group, whereas low and high doses show a significant difference. In this situation the data are examined for a dose-related trend. Williams' test is generally carried out to test dose-related trend (Bretz, 2006).

For the data that show a dose-related trend and a significant difference by Dunnett's test (Dunnett, 1955), the interpretation of the data analysis can be done in a straight forward manner. In a four group-setting repeated dose administration study, seven different situations can be expected (Table 11.9). Interpretation is relatively easier in situations 1–3, whereas it is difficult in situations 4–7, where further investigation on dose-related trend is required.

Table 11.9. Significant difference shown by the treatment groups by Dunnett's test—Possible situations

Test Group	●: Significant difference, ○: No significant difference from the control group						
	Situation 1	Situation 2	Situation 3	Situation 4	Situation 5	Situation 6	Situation 7
Control	○	○	○	○	○	○	○
Low dose	●	○	○	●	●	○	●
Mid dose	●	●	○	○	●	●	○
High dose	●	●	●	○	○	○	●
Investigation	Not required	Not required	Not required	Required	Required	Required	Required
Visual dose-related trends	Yes	Yes	Yes	No	No	No	No

Absolute kidney weight of rats from a repeated dose administration study is given in Table 11.10. These data were analysed using Dunnett's and Williams' tests. Dunnett's test showed a significant difference in low and high dose groups, whereas Williams' test showed a significant difference in all the groups.

Table 11.10. Absolute kidney weights of rats

Absolute kidney weights	Dose group			
	Control	Low	Mid	High
Individual data, (g)	2.558	3.269	3.116	2.706
	2.789	3.428	2.791	3.293
	2.764	3.083	2.981	3.535
	2.707	3.532	3.337	3.387
	2.793	3.546	2.432	3.064
	3.041	2.677	2.934	3.102
	3.000	2.822	3.388	3.279
	-	3.656	2.911	-
	-	3.271	2.798	-
	-	3.348	3.208	-
	-	3.031	2.876	-
	-	3.742	2.703	-
Number of animal	7	12	12	7
Mean ± Standard deviation	2.807±0.167	3.284±0.329	2.956±0.273	3.195±0.269
Bartlett's homogeneity test	$P = 0.4130$ (No heterogeneity)			
Dunnett's test		$P = 0.0026$*	$P = 0.5190$	$P = 0.0332$*
Mean value used for Williams' test	2.807	3.284	3.120	3.195
Williams' test		$P<0.05$*	$P<0.05$*	$P<0.05$*
Jonckheere's trend test	No significant difference			

*Significantly different from control group.

Use of Williams' test is not recommended when the number of animals in the groups is different (Williams, 1972) and extremely less (Williams, 1971; 1972). But, Sakaki *et al.* (2000) stated that Williams' test can be used even if number of the animals in a group differs about 2 times compared to other group/s.

Williams' test analyzes the difference of the mean values between each treated group and the control, like Dunnett's test, when the mean value of the treated groups changes in one direction. The example given in Table 11.11 does not show a dose-dependence as the mid dose showed an insignificant liver weight compared to control (by Dunnett's test). When the data were analysed by Williams' test, significance in the liver weight is observed in the mid dose group. The reason for this may be better explained by elucidating the calculation procedure of Williams' test as given below:

Table 11.11. Liver weight of rats in a 4-week repeated dose administration study

Group	Liver weight (g), N=5, (Sum)	Mean ± SD (% change with respect to control)	Results of Dunnett's test	Mean for Williams' test (% change with respect to control)	Results of Williams' test
Control	10.7, 11.5, 11.6, 12.0, 11.0 (56.8)	11.36 ± 0.51 (100)		11.36 (100)	
Low dose	11.6, 12.3, 12.5, 12.3, 12.7 (61.4)	12.28 ± 0.41 (108.1)	$P<0.05$	12.28 (108.1)	$P<0.05$
Mid dose	11.2, 11.5, 11.6, 11.5, 11.5 (57.3)	11.46 ± 0.15 (100.9)	Not significant	11.87 (104.5)	$P<0.05$
High dose	12.2, 12.5, 12.0, 11.9, 13.0 (61.6)	12.32 ± 0.44 (108.5)	$P<0.05$	12.32 (108.5)	$P<0.05$

Calculation procedure of Williams' test:

(1) Control *vs* High dose

$$\frac{61.4+57.3+61.6}{5+5+5}=12.02$$

(Note: Numerator—sums of low dose + mid dose + high dose; denominator—number of observations of low dose + mid dose + high dose).

$$\frac{57.3+61.6}{5+5}=11.89$$

(Note: Numerator—sums of mid dose + high dose; denominator—number of observations of mid dose + high dose).

$\frac{61.6}{5} = 12.32 \leftarrow$This largest value is used for the calculation of *t* value.

(Note: Numerator—sum of high dose; denominator—number of observations of high dose).

We have all estimates for calculating the *t* value, except the mean SS of error variance. Let us analyse the data using ANOVA:

Liver weight of rats in a 4-week repeated dose administration study

Statistics

Estimates	Liver weight (g)			
	Control	Low dose	Mid dose	High dose
N	5	5	5	5
Mean ± SD	11.36 ± 0.51	12.28 ±0.41	11.46±0.15	12.32 ± 0.44
Sum	56.8	61.4	57.3	61.6
Grand sum	237.1			

Total sum of squares

$$= (x_1^{\,2} + x_2^{\,2} \cdots\cdots x_{29}^{\,2} + x_{30}^{\,2}) - \frac{(\sum x)^2}{\sum N} =$$

$$= (10.7^2 + 11.5^2 \cdots\cdots 11.9^2 + 13^2) - \frac{(237.1)^2}{20} = 6.6095$$

Sum of squares of among the groups

$$= \frac{56.8^2}{5} + \frac{61.4^2}{5} + \frac{57.3^2}{5} + \frac{61.6^2}{5} - \frac{(237.1)^2}{20} = 3.9895$$

Total sum of squares for error = Total sum of squares—Sum of squares of among the groups

$= 6.6095 - 3.9895 = 2.62$

The ANOVA Table constructed is given below (Table 11.12).

Table 11.12. ANOVA Table

Source of variation	SS	DF	MS	*F* value	*P*
Total	6.6095	19	-	-	-
Groups	3.9895	3	1.32983	8.12	$P<0.001$
Error	2.62	16	0.16375		

SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.

Mean SS for error is 0.16375. Now we have all the required estimates for calculating *t*:

$$t = \frac{11.36 - 12.32}{\sqrt{0.16375\left(\frac{1}{5} + \frac{1}{5}\right)}} = 3.751$$

t-value is significant at 5% level (Table 11.13, Number of groups-4; DF-16).

(2) Control *vs* Mid dose

$\frac{61.4 + 57.3}{5 + 5} = 11.87$←This largest value is used for the calculation of *t*-value.

(Note: Numerator—sums of low dose + mid dose; denominator—number of observations of low dose + mid dose).

$$\frac{57.3}{5} = 11.46$$

(Note: Numerator- sum of mid dose; denominator- number of observations of mid dose).

$$t = \frac{11.36 - 11.87}{\sqrt{0.16375\left(\frac{1}{5} + \frac{1}{5}\right)}} = 1.993$$

t value is significant at 5% level (Table 11.13, Number of groups-3; DF-16).

(3) Control *vs* Low dose

$$\frac{61.4}{5} = 12.28$$

(Note: Numerator- sum of low dose; denominator- number of observations of low dose).

$$t = \frac{11.36 - 12.28}{\sqrt{0.16375\left(\frac{1}{5} + \frac{1}{5}\right)}} = 3.595$$

t-value is significant at 5% level (Table 11.13, Number of groups-2; DF-16).

Table 11.13. Williams' Table

DF	Number of groups							
	2	3	4	5	6	7	8	9
15	1.753	1.839	1.868	1.882	1.891	1.896	1.900	1.903
16	**1.746**	**1.831**	**1.860**	1.873	1.882	1.887	1.891	1.893
17	1.740	1.824	1.852	1.866	1.874	1.879	1.883	1.885

The reason for Williams' test showing a significant difference in the weight of the liver of the mid dose group, when compared with the control group, is that the test used 11.87 as the mean value of the mid dose group for the comparison instead of the actual value (11.46).

Williams' test is a useful statistical tool in toxicology as it provides information on evidence of toxicity and also the dose level that causes the toxicity (Shirley, 1977). Williams' test is similar to Dunnett, Tukey and Duncan multiple comparison (range) tests as it uses the error variance of the ANOVA (Nagata and Yoshida, 1997) in the calculation procedure. Williams' test is a closed procedure. If no significant difference between control group and highest dose group is seen, all the other treated groups are considered to have no significant difference compared to the control group and no further analysis is carried out. If there is a significant difference in the highest dose group, then the next highest dose level is examined for the significant difference from the control. If this dose group does not show a significant difference, no further analysis is carried out. But if it shows a significant difference, the next highest dose level is examined for the significant difference from the control group. Thus all the dose groups are sequentially examined.

Williams' test is effective in monotonic and non-monotonic dose-response relationships (Dmitrienko *et al.*, 2007). Since estimated mean values are used in the calculation procedure of Williams' test, it is likely that this test might show a dose-related trend, where it actually does not exist. It also may be noted in this context that, according to Gad and Weil (1988) dose-related trend is necessarily not evident in all the parameters.

Duncan's multiple range test (Shibata, 1970)

Duncan's multiple range test is generally used for comparison of more than 2 groups, when the number of observations of the groups is different. We shall work on the example given in Table 11.2. The data is slightly modified by changing the number of observations of Groups 1 and 2. The changed data are given in Table 11.14.

Table 11.14. Jaundice index of newborn baby. Reproduced from Table 11.2. Number of observations of Groups 1 and 2 was changed

Week		
36 (Group 1)	38 (Group 2)	40 (Group 3)
13	9	5
6	11	5
11	11	4
12	10	7
14	7	7
10	7	3
9	5	3
	8	4
		5
		3

Statistics

Estimates	Week		
	36 (Group 1)	38 (Group 2)	40 (Group 3)
N	7	8	10
Mean ± SD	10.7±2.7	8.5±2.1	4.6±1.5
Sum	75	68	46
Grand sum		189	

Calculation steps:

Total sum of squares =

$$(x_1^{\,2} + x_2^{\,2} \cdots\cdots x_{24}^{\,2} + x_{25}^{\,2}) - \frac{(\sum x)^2}{\sum N} =$$

$$(13^2 + 6^2 \cdots\cdots 5^2 + 3^2) - \frac{(189)^2}{25} = 260.2$$

Sum of squares of among the groups

$$= \frac{75^2}{7} + \frac{68^2}{8} + \frac{46^2}{10} - \frac{(189)^2}{25} = 164.3$$

Total sum of squares for error = Total sum of squares—Sum of squares among the groups

= 260.2 – 164.3 = 95.9

Let us construct the ANOVA Table (Table 11.15).

Table 11.15. ANOVA Table

Source of variation	SS	DF	MS	*F* value	*P*
Total	260.2	24	-	-	-
Groups	164.3	2	82.2	19.1	$P<0.001$
Error	95.9	22	4.3		

SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.

Note: There are 25 observations, hence the DF for total SS is 25–1 = 24; Total number of groups are three, hence the DF for SS groups is 3–1 = 2; DF for error SS = DF for total SS—DF for SS among groups (24 – 2 = 22).

$$F_{22}^{2}calc = \frac{82.2}{4.3} = 19.1$$

Compare the derived *F* values with that of the value given in the *F* Distribution Table (Table 11.16.)

Table 11.16. *F*-distribution values at 0.1% probability level (Yoshimura, 1987)

$N_1 \backslash N_2$	1	2	3	4	5	6	7	8	9	10
22	14.380	**9.612**	7.796	6.814	6.191	5.758	5.438	5.190	4.993	4.832

N_1—DF for the numerator; N_2–DF for the denominator.

Since the derived *F*-value is greater than the Table value, it is considered that the jaundice index of the newborn baby is significantly different among the weeks at 0.1% probability level.

Let us carry out *post hoc* comparison using Duncan's multiple range test. The first step is calculation of 'least significant range', *Rp*:

R_p = Sm×Q, where

$$Sm = \sqrt{\frac{\text{MS for error variance}}{\sum N / \text{Number of groups}}}$$

Q = Critical value from Duncan's table

$$Sm = \sqrt{\frac{4.3}{25/3}} = 0.72$$

Note: 4.3 is variance of error (see Table 11.15); Total number of observation = 25; Total number of groups = 3).

Critical Q values are obtained from Duncan's Table (Table 11.17). Q values at 22 degrees of freedom (Degrees of freedom of the error component; see Table 10.15) for 2 and 3 Groups are 2.93 and 3.08, respectively.

Table 11.17. Duncan's critical values at 5% probability level (Shibata, 1970)

DF	Group								
	2	3	4	5	6	7	8	9	10
22	**2.93**	**3.08**	3.17	3.24	3.29	3.32	3.35	3.37	3.39

$$R_{2(0..05)} = 0.72 \times 2.93 = 2.11$$

$$R_{3(0..05)} = 0.72 \times 3.08 = 2.22$$

Arrange the mean values orderly:

Group 1 (Week 36) = 10.7

Group 2 (Week 38) = 8.5

Group 3 (Week 40) = 4.6

Let us compare the largest sample means range, *i.e.,* 10.7 and 4.6. The difference between these two mean values is 6.1, which is greater than the 'least significant range', R_3. Hence, the difference between these two mean values (Group 1 and Group 3) is considered significant. Let us compare the next set of mean values, 10.7 and 8.5. The difference between these two mean values is 2.2, which is greater than the 'least significant range', R_2. Hence the difference between the mean values of Group 1 and Group 2 is also considered significant.

Scheffé's multiple comparison test (Scheffe, 1953)

We shall use the data given in Table 11.14 for demonstrating Scheffé's multiple comparison test.

Statistics

Estimates	Week		
	36 (Group 1)	38 (Group 2)	40 (Group 3)
N	7	8	10
Mean ± SD	10.7±2.7	8.5±2.1	4.6±1.5
Sum	75	68	46
Grand sum		189	

Comparisons:

Group 3 *vs* Group 2

$$F = \frac{(4.6-8.5)^2}{(3-1)\times 4.3\times(\frac{1}{10}+\frac{1}{8})} = 7.86 \qquad \therefore p < 0.05$$

Group 3 *vs* Group 1

$$F = \frac{(4.6-10.7)^2}{(3-1)\times 4.3\times(\frac{1}{10}+\frac{1}{7})} = 17.82 \qquad \therefore p < 0.05$$

Group 2 *vs* Group 1

$$F = \frac{(8.5-10.7)^2}{(3-1)\times 4.3\times(\frac{1}{8}+\frac{1}{7})} = 2.10 \qquad \therefore p > 0.05(NS)$$

Note: 4.3 is the variance of error (*vide* Table 11.15).

These derived *F*-values are compared with the values given in *F* distribution Table (Table 11.18) given below:

Table 11.18. *F*-distribution values at 5% probability level (Yoshimura, 1987)

$N_1 \backslash N_2$	1	2	3	4	5	6	7	8	9	10
22	4.301	**3.443**	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297

N_1-DF for the numerator; N_2-DF for the denominator.

All the derived *F* values, except the one computed for the comparison between Group 2 and Group 1, are significant at 5% probability level.

The Scheffé's multiple comparison test is used for all-pair comparisons, like the Duncan's multiple comparison test. However, the power to detect a significant difference is low with the Scheffé's multiple comparison test compared to that of the Duncan's multiple comparison test (*vide* Table 11.19).

Duncan's multiple comparison test showed a significant difference in the mid dose and high dose groups, whereas the Scheffé's multiple comparison test did not show a significant difference in these groups, indicating it's low power to detect a significant difference. Therefore, use of Scheffé's multiple comparison test should be done with little caution in the safety evaluation studies with animals.

Table 11.19. Comparison of the power to detect a significant difference between Scheffés and Duncan's multiple comparison tests. LDH activity (U/l) of F344 female rats at week 78 in a repeated dose administration study is given.

Estimates	Control	Low dose	Mid dose	High dose
-	168, 188, 181, 250, 122, 89, 125, 135, 211, 204	112, 168, 175, 241, 218, 49, 49, 76, 66, 30	69, 86, 145, 244, 135, 46, 105, 40, 53, 73	43, 59, 73, 99, 129, 181, 49, 69
N	10	10	10	8
Mean ± SD	167 ± 49	118 ± 76	100 ± 62	88 ± 47
In % of control	-	71	60	53
ANOVA	$P < 0.05$			
Duncan's test		N.S.	S	S
Scheffé's test		N.S.	N.S.	N.S.

N.S.—Not significant ($P > 0.05$); S—Significant ($P < 0.05$)

Two-way ANOVA

It is an extension of one-way ANOVA. The difference in 2-way ANOVA is that it has 2 independent factors. The data is arranged in tabular fashion in such a way that the column represents one factor and the row, the other factor (Belle *et al*., 2004).

An example is provided to illustrate the computations required in two-way ANOVA (Kibune and Sakuma, 1999).The diameter of the head of the three human embryos was measured by four observers. Each observer measured the diameter of three embryos. The data is arranged in a tabular fashion as given in Table 11.20. We are interested to know: 1. Among the observers, is there any difference in the diameter of embryos measured 2. Among the embryos, is there any difference in the diameter of embryos measured and 3. Is there any simultaneous influence of observer and embryo in the diameter measured (interaction)

Calculation steps:

1) Correction factor (CF)
   $=(\text{Grand sum})^2/N = 558.1^2/36=8652.1$
2) Total sum of squares
   $= (14.3^2+14.0^2+.....+12.9^2+13.8^2)\text{-CF}=8979.7–8652.1=\mathbf{327.6}$
3) Sum of squares of among the observers
   $=1/9\ (141.0^2+137.6^2+138.2^2+141.3^2)\text{-CF}=8653.2–8652.1=\mathbf{1.199}$

4) Sum of squares of among the embryos
=1/12 (167.9^2+236.3^2+153.9^2)-CF=8976.1–8652.1=**324**
5) Embryo × Observer (Interaction)
=1/3 (43.1^2+59.4^2+38.5^2+41.0^2+58.9^2+37.7^2+41.4^2+58.8^2
+38.0^2+42.4^2+59.2^2+39.7^2)-CF=8977.8–8652.1=**325.7**
Sum of squares of interaction is calculated as given below:
325.7–1.199–324= **0.501**. The DF for interaction is (3–1) (4–1)=6.
6) Sum of squares of error
327.6–1.199–324–0.501=**1.9**. The DF for error is 35–2–3–6=24

Table 11.20. Diameter of three human embryos (cm) measured by four observers

	Observer 1	Observer 2	Observer 3	Observer 4	Sum
Embryo 1	14.3	13.6	13.9	13.8	167.9 (109)
	14.0	13.6	13.7	14.7	
	14.8	13.8	13.8	13.9	
Sum	43.1	41.0	41.4	42.4	
Embryo 2	19.7	19.8	19.5	19.8	236.3 (154)
	19.9	19.3	19.8	19.6	
	19.8	19.8	19.5	19.8	
Sum	59.4	58.9	58.8	59.2	
Embryo 3	13.0	12.4	12.8	13.0	153.9 (100)
	12.6	12.8	12.7	12.9	
	12.9	12.5	12.5	13.8	
Sum	38.5	37.7	38.0	39.7	
Total sum	141.0 (99.8)	137.6 (97.4)	138.2 (97.8)	141.3 (100)	558.1

Let us construct the ANOVA Table (Table 11.21):

Table 11.21. Two-way layout ANOVA

Source of variation	SS	DF	MS	*F* value	*P*
Embryo*	324	2	162	2051	$P<0.001$
Observer*	1.199	3	0.399	5.05	$P<0.01$
Embryo×Observer**	0.501	6	0.084	1.06	NS
Error	1.9	24	0.079		
Total sum	327.6	35			

*Main effects **Interaction, SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.

The computed *F* values are compared with the *F* distribution values given in *F* distribution Table (Table 11.22). For the comparison of all the sources of variation (embryo, observer and embryo × observer interaction), the denominator remains the same (DF of error, which is 24), but the numerator differs. The *F* values should be compared with *F* distribution values at 2/24 (numerator/denominator) for embryo, 3/24 for observer and 6/24 for embryo × observer interaction.

Table 11.22. *F* distribution values at 1% probability level (Yoshimura, 1987)

$N_1 \backslash N_2$	1	2	3	4	5	6	7	8	9	10
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168

N_1- DF for the numerator; N_2 –DF for the denominator.

Discussion:

1. Embryo: The *F*-value is greater than the Table *F*-value (2051>5.614); hence there is a significant difference among embryos.
2. Observer: The *F*-value is greater than the Table *F*-value (5.05>4.718); hence there is a significant difference among observers.
3. The embryo × observer interaction: The *F*-value is less than the Table *F* value (1.06<3.667); hence embryo×observer interaction is not significant.

Since the interaction is not significant, the ANOVA Table can be reconstructed excluding interaction as a source of variation. The SS of interaction is added to the SS of error and the DF of the interaction is added to the DF of error. The Table thus reconstructed after excluding interaction as a source of variation is given below (Table 11.23):

Table 11.23. ANOVA Table excluding the interaction

Source of variation	SS	DF	MS	*F* value	*P*
Embryo	324	2	162	2024	*P*<0.001
Observer	1.199	3	0.399	4.99	*P*<0.001
Error	2.40	30	0.080		
Total sum	327.6	35	-		

SS-Sum of squares; DF-Degrees of freedom; MS-Mean sum of squares.

Dunnett's Multiple Comparison Test and Student's *t* Test—A Comparison

In pharmacological and toxicological experiments the number of groups usually employed is more than two. If the data obtained from such studies are anlaysed by Student's *t*-test (picking up any two groups and analyzing by Student's *t*-test), it may cause Type I error.

We analysed data obtained from several repeated dose administration studies in rats using Dunnett's multiple comparison test and Student's *t*-test to know to what extent repeated analysis by Student's *t*-test shows a Type I error. Our finding is given in Table 11.24.

Table 11.24. Analysis of data obtained from repeated dose administration studies in rats by Dunnett's multiple comparison test and Student's *t*-test

Item	Number of analyses	Dunnett's multiple comparison test	Student's *t*-test[a]
Body weight	528	223	246 (10)
Feed consumption	832	235	349 (49)
Hematology	352	123	159 (29)
Blood chemistry	576	215	272 (27)
Urinalysis	64	7	11 (57)
Organ weight	224	47	80 (70)
Organ weight/ body weight ratio	224	82	104 (27)
Total	2800	932	1221 (31)

[a]Values given in parentheses are percent increase compared to Dunnett's multiple comparison test.

The number of items showing a significant difference by Student's *t*-test increased, compared to those showing a significant difference by Dunnett's multiple comparison test. Overall, there was an increase by 31% in the items, when they were analysed by Student's *t*-test. This increase is due to the Type I error. Yoshimura and Tsubaki (1993) suggested that to assess the toxicity, Dunnett's multiple comparison test is the appropriate statistical approach; on the contrary, from a consumer point of view, Student's *t*-test, may be more appropriate.

References

Armstrong, R.A., Slave, S.V. and Eperjesi, F. (2000): An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry. Ophthalmic Physiol. Opt., 20(3), 235–241.

Belle, G.V., Fisher, L.D., Heagerty, P.J. and Lumley, T. (2004): Biostatistics—A Method for Health Sciences. John Wiley & Sons, Inc., New Jersey, USA.

Bretz, F. (2006): An extension of the Williams' trend test to general unbalanced linear models. Comp. Stat. Data Anal., 50(7), 1735–1748.

Cheung, S.H. and Holland, B. (1991): Extension of Dunnett's multiple comparison procedure to the case of several groups. Biometrics, 47, 21–32.

Dmitrienko, A., Chaung-Stein, C. and D'Agostino, R. (2007): Pharmaceutical Statistics. Using SAS—A Practical Guide. SAS Institute, NC, USA.

Dunnett, C.W. (1955): A multiple comparison procedure for comparing several treatments with a control. Am. Stat. Assoc., 50, 1096–1211.

EPA (2000): United States Environmental Protection Agency. Health Effects Test Guidelines, OPPTS 870.3050, Repeated Dose 28—Day Oral Toxicity Study in Rodents, EPA 712–C–00–366 2000. EPA, USA.

Gad, S. and Weil, C.S. (1988): Statistics and Experimental Design for Toxicologists. Telford Press, New Jersey, USA.

Gill, L. (!990): Uses and abuses of statistical methods in research in parasitology. Vet. Parasitol., 36(3-4), 189–209.

Gupta, R.C. (2007): Veterinary Toxicology—Basic and Clinical Principles. Academic Press, New York, USA.

Kibune, Y. and Sakuma, A. (1999): Practical Statistics for Medical Research, Scientist Press, Tokyo, Japan.

Kobayashi, K., Pillai, K.S., Michael, M., Cherian, KM., Ohnishi, M. (2010): Determining NOEL/NOAEL in repeated-dose toxicity studies, when the low dose group shows significant difference in quantitative data. Lab. Anim. Res., 26(2),133–137.

Mathews, P.G. (2005): Design of Experiments with MINITAB. American Society for Quality, Milwaukee, USA.

Moder, K. (2007): How to keep the Type I error rate in ANOVA if variances are heteroscedastic. Aust. J. Stat., 6(3), 179–188.

Muir, W.M., Romero-Severson, J., Rider Jr., S.D., Simons, A. and Ogas, J. (2006): Application of one sided *t*-tests and a generalized experiment-wise error rate to high-density oligonucleotide microarray experiments: An example using Arabidopsis. J. Data Sci., 4, 323–341.

Nagata, Y. and Yoshida, M. (1997): Tokeiteki-tajuhikakuho-no-kiso. Scientist Co. Ltd, Tokyo, Japan.

Norman, G.R. and Streiner, D.L. (2008): Biostatistics—The Bare Essentials. 3rd Edition. BC Decker Inc., Ontario, Canada.

OECD (1995): Organization for Economic Cooperation and Development. OECD Guidelines for Testing of Chemicals. Repeated Dose 28-Day Oral Toxicity Study in Rodents. No. 407, OECD, France.

OECD (1998): Organization for Economic Co-operation and Development. OECD Guideline for the Testing of Chemicals. Repeated Dose 90 Day Oral Toxicity Study in Rodents, No. 408. OECD, France.

Rothmann, M. (2005): Type I error probabilities based on design-stage strategies with applications to noninferiority trials. J. Biopharm. Stat., 15(1), 109–127.

Sakaki, H., Igarashi, S., Ikeda, T., Imamizo, K., Omichi, T., Kadota, M., Kawaguchi, T., Takizawa, T., Tsukamoto, O., Terai, K., Tozuka, K., Hirata, J., Handa, J., Mizuma, H., Murakami, M., Yamada, M. and Yokouchi, H. (2000): Statistical method appropriate for general toxicological studies in rats. J. Toxicol. Sci., 25, 71–98.

Scheffé, H. (1953): A method for judging all contrasts in the analysis of variance. Biometrica, 40, 87–104.

Shibata, K. (1970): Biostatistics.Tokyo University of Agriculture, Tokyo, Japan.

Shirley, E. (1997): A non-parametric equivalent of Williams' test for contrasting increasing dose levels of a treatment. Biometrics, 33(2), 386–389.

Wallenstein , S., Zucker, C.L. and Fleiss, J.L. (1980): Some statistical methods useful in circulation research. Circ. Res., 47, 1–9.

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 zero dose control. Biometrics, 28, 519–531.

Yoshida, M. (1980): Design of Experiments for Animal Husbandry, Yokendo Press, Tokyo, Japan.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Inc., Tokyo, Japan.

Yoshimura, I and Tsubaki, H. (1993): Multiple comparison test for more than three dosed groups settings (debate). Japanese Society for Biopharmaceutical Statistics, August 7, 1993, Sohyo Kaikan, Tokyo, Japan.

12

Non-Parametric Tests

Non-parametric and Parametric Tests—Assumptions

Statistical methods are based on certain assumptions. For applying parametric statistical tools, the assumptions made are that data follow a normal distribution pattern and are homogeneous. In many situations, the data obtained from animal studies contradict these assumptions, and are not suitable to be analysed with the parametric statistical methods. Non-parametric tests do not require the assumption of normality or the assumption of homogeneity of variance. Hence, these tests are referred to as distribution-free tests. Non-parametric tests usually compare medians rather than means, therefore influence of one or two outliers in the data is annulled. We shall deal with some of the most commonly used non-parametric tests in toxicology/pharmacology.

Sign Tests

Perhaps, the sign test is the oldest distribution-free test which can be used either in the one-sample or in the paired sample contexts (Sawilowsky, 2005). Sign test is probably the simplest of all the non-parametric methods (Whitley and Ball, 2002; Crawley, 2005). The null hypothesis of the sign test is that given a pair of measurements (xi, yi), then xi and yi are equally likely to be larger than each other (Surhone *et al*., 2010). Though the sign test is rarely used in toxicology, it can be used in certain pharmacological *in vivo* experiments to evaluate whether a treatment is superior to the other. The sign test may be used in clinical trials to know whether either of the two treatments that are provided to study subjects is favored over the other (Nietert and Dooley, 2011).

The calculation procedure of sign test for small sample size ($n \leqq 25$) is different from that of large sample size ($n > 25$):

Calculation procedure of sign test for small sample size

A study was conducted to evaluate the hypoglycemic effect of an herbal preparation in rats. Hyperglycemia was induced in rats by administering streptozotozin. Following the administration of streptozotozin, the blood sugar was measured in individual rats to confirm hyperglycemia. Then the hyperglycemic rats were given the herbal preparation daily for 14 consecutive days. On day 15, again blood sugar was measured in these rats. The blood sugar measured in hyperglycemic rats before and after the administration of the herbal preparation is given in Table 12.1.

Table 12.1. Blood sugar level (mg/dl) in hyperglycemic rats

Rat No.	1	2	3	4	5	6	7	8
Blood sugar level before administration of herbal preparation (X_a)	236	223	211	229	205	245	243	231
Blood sugar level after administration of herbal preparation (X_b)	155	156	172	198	209	181	231	231
Difference (X_b- X_a)	–81	–67	–39	–31	+4	–64	–12	0
Sign	- (–1)	- (–1)	- (–1)	- (–1)	+ (+1)	- (–1)	- (–1)	± (0)

$$p = {}_7C_1\left(\frac{1}{2}\right)^6\frac{1}{2} + {}_7C_0\left(\frac{1}{2}\right)^7$$

$$= \left({}_7C_1 + {}_7C_0\right)\left(\frac{1}{2}\right)^7$$

$$= 0.0546 + 0.0078 = 0.0624$$

Note: ${}_nC_r = \frac{n!}{r!(n-r)!}$; Rat No. 8, which did not show any change in the blood sugar is not included in the analysis.

Since *P*=0.0624 is >0.05, it is considered that the decrease in blood sugar in rats administered with herbal preparation is insignificant.

Calculation procedure of sign test for large sample size

The effect of two analgesics, drugs A and B was evaluated five times by 32 doctors and their findings are given in Table 12.2. The objective of the

Table 12.2. Analgesic effect of drugs A and B evaluated by 32 doctors

Doctor No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Drug A (X_a)	4	5	4	5	5	3	5	4	3	4	5	3	3	4	5	4	3	4	4	5	4	5	4	3	3	5	2	1	4	5	3	5
Drug B (X_b)	2	3	3	4	2	3	2	5	3	5	3	4	2	4	3	5	4	3	5	3	4	3	5	4	4	2	4	3	2	4	2	2
Sign (X_b- X_a)	-	-	-	-	-	±	-	+	±	+	-	+	-	±	-	+	+	-	+	-	±	-	+	+	+	-	+	+	-	-	-	-

study was to know whether the analgesic effect of drugs A and B is similar or different.

The pairs, which showed a difference of 0 (± sign) are excluded from the calculation procedure. In this example four pairs showed a difference of 0 (± sign). Therefore, number (n) of data becomes 32–4=28. Number of + sign, which indicates that the effect of drug B is better than drug A, is 11. *Z* is obtained from the equation given below:

$$z = \frac{(r + 0.5) - \mu_r}{\sigma_r} = \frac{11.5 - 14}{2.65} = -0.94$$

$$\text{Mean } \mu_r = \frac{28}{2} = 14 \qquad \sigma_r(SD) = \frac{\sqrt{28}}{2} = 2.65$$

r = Total number of + sign = 11

The $p(|z| > 0.94 = 0.9) = 0.36812$ from normal distribution Table (Table 12.3) is greater than 0.05 (two-sided test). Therefore, it can be concluded that both the drugs have similar effect.

Table 12.3. Normal distribution table (Yoshimura, 1987)

%	Two-sided *P*	Upper *P*
Z	2α	α
0.8	0.423711	0.211855
0.9	**0.368120**	0.184060
1.0	0.317311	0.158655

Signed Rank Sum Tests

The major disadvantage of the sign test is that it considers only the direction of difference between pairs of observations, not the size of the difference (Mc Donald, 2009). Ranking the observations and then carrying out the statistical analysis can solve this issue. Signed rank sum test is more powerful than the sign test (Elston and Johnson, 1994).

Wilcoxon Rank-Sum test (Wilcoxon, 1945)

The Wilcoxon rank-sum test is one of the most commonly used non-parametric procedures (Le, 2003). This is the non-parametric analogue to the paired *t*-test. The null hypothesis of Wilcoxon rank-sum test is that the median difference between pairs of observations is zero.

The performance of six classes of two schools expressed in average scores is given in Table 12.4. We shall analyse this data using Wilcoxon rank-sum test.

Table 12.4. Average scores of six classes of two schools

School	Average score					
School A	79.5	85.5	83.5	93.5	91.5	77.5
School B	95.5	87.5	89.5	98.0	97.5	81.5

Step 1: Combine the scores of both the schools and arrange them from the smallest to the largest. Then assign a rank from 1 to 12 to the scores as given in Table 12.5. (Note: if there are tied observations, assign average rank to each of them).

Table 12.5. Ranks assigned to the combined scores of two schools

Scores arranged from smallest to largest	Rank
77.5	1
79.5	2
81.5	3
83.5	4
85.5	5
87.5	6
89.5	7
91.5	8
93.5	9
95.5	10
97.5	11
98	12

Step 2: Arrange the rank corresponding to the original scores as given in Table 12.6 and calculate the sum of the ranks.

Table 12.6. Ranks arranged to the original scores

School	Ranks						Sum of rank
School A	2	5	4	9	8	1	29
School B	10	6	7	12	11	3	49

Calculation Procedure:

The number of samples (classes) in each group = 6

Sum of rank of School B, R_2=10+6+7+12+11+3=49

$$V = \frac{\left[\begin{array}{l}(2-6.5)^2+(5-6.5)^2+(4-6.5)^2+(9-6.5)^2+(8-6.5)^2+(1-6.5)^2 \\ +(10-6.5)^2+(6-6.5)^2+(7-6.5)^2+(12-6.5)^2+(11-6.5)^2+(3-6.5)^2\end{array}\right]\times 6\times 6}{12\times 11}$$

$$= 39$$

Where,

$$6.5 = \frac{29+49}{12}$$

12 = Sum of number of samples (classes) of School A and School B

11 = (Sum of number of samples (classes) of School A and School B) – 1

Let us calculate *T*

$$T = \frac{49 - 6\times\frac{13}{2}}{\sqrt{39}} = 1.601$$

Where,

13 = (Sum of number of samples (classes) of School A and School B) + 1

2 = Constant

Calculated *T* value (*T*=1.601) is smaller than the *U(α)* = 1.644854 at *P*= 0.05 (see Table 12.7). Hence, it is considered that there is no significant difference in scores between the schools.

Table 12.7. Standard normal distribution Table (Yoshimura, 1987)

Two tailed *P*	Upper *P*	% point
2α	α	U(α)
0.05000	0.025000	1.959964
0.06000	0.030000	1.880791
0.07000	0.035000	1.811911
0.08000	0.040000	1.750686
0.09000	0.045000	1.695398
0.10000	0.050000	**1.644854**

Fisher's exact test

Fisher's exact test is used in the analysis of contingency tables with small sample sizes (Fisher, 1922; 1954). It is similar to χ^2 test, since both Fisher's exact test and χ^2 test deal with nominal variables. In Fisher's exact test, it is assumed that the value of the first unit sampled has no effect on the value of the second unit. It is interesting to learn how the Fisher's exact test was originated. Dr Muriel Bristol of Rothamsted Research Station, UK claimed that she could tell whether milk or tea had been added first to a cup of tea. Fisher designed an experiment to verify the claim of Dr Muriel

Bristol. Eight cups of tea were made. In four cups, milk was added first and in the other four cups tea was added first. Thus, the column totals were fixed. Dr. Bristol was asked to identify the four to 'tea first', and the four to 'milk first' cups. Thus, the row totals were also fixed in advance. Fisher proceeded to analyse the resulting 2 × 2 table, thus giving birth to Fisher's exact test (Clarke, 1991; Ludbrook, 2008).

Manual analysis of data using Fisher's exact test is beyond the scope of this book, hence not covered. The power to detect a significant difference is more with Fisher's exact test than the χ^2 test as seen in Table 12.8.

Table 12.8. Power to detect a significant difference—Comparison between χ^2 test and Fisher's exact test

Incidence of pathological lesions (Control *vs* dosed group)	*P*-value	
	Chi-square test*	Fisher's test (α)
0/5 *vs* 1/5	1.00000	0.50000
0/5 *vs* 2/5	0.42920	0.22222
0/5 *vs* 3/5	0.16755	0.08333
0/5 *vs* 4/5	0.05281	0.02381
0/5 *vs* 5/5	0.01141	0.00397
1/5 *vs* 2/5	1.00000	0.50000
1/5 *vs* 3/5	0.51861	0.26190
1/5 *vs* 4/5	0.20590	0.10317
1/5 *vs* 5/5	0.05281	0.02381
2/5 *vs* 3/5	1.00000	0.50000
2/5 *vs* 4/5	0.51861	0.26190
2/5 *vs* 5/5	0.16755	0.08333

*Yetes's correction (Note on Yetes's correction: χ^2 slightly overestimates the 'difference between expected and observed' results. This overestimation can be corrected by decreasing the 'difference between expected and observed' by 0.5).

McKinney *et al*. (1989) reviewed the use of Fisher's exact test in 71 articles published between 1983 and 1987 in six medical journals. Nearly 60% of articles did not specify use of a one- or two-sided test. The authors concluded that the use of Fisher's exact test without specification as a one- or two-sided version may misrepresent the statistical significance of data.

Mann-Whitney's U test

Mann-Whitney's *U* test, a test equivalent of Student's *t*-test for comparing two groups, was independently developed by Mann and Whitney (1947) and Wilcoxan (1945). The calculation procedure of Mann-Whitney's *U* test is very much similar to Wilcoxan signed rank sum test. For understanding Mann-Whitney's *U* test in a detailed manner, let us analyse the data given in Table 12.9. Our objective of the analysis is to find whether there is a significant difference in hemoglobin content between Group A and Group B.

Table 12.9. Hemoglobin content (g/dl) in two experimental groups of rats following the administration of a drug at 10 mg/kg b.w. (Group A) and at 20 mg/kg b.w. (Group B)

Group A	9.3	6.4	10.8	5.6
Group B	5.9	9.7	9.9	6.7

Let us pool the data and arrange them from the smallest to the largest, ignoring the Group to which they belong and rank them. Then, tag them with the identity of the Group to which they belong (Table 12.10).

Table 12.10. Ranking the data

Pooled data	5.6	5.9	6.4	6.7	9.3	9.7	9.9	10.8
Ranked data	1	2	3	4	5	6	7	8
Tagged data with respective group	A	B	A	B	A	B	B	A

(Note for tied observations: Assign mean score for the tied observations. For example, if the value of ranks 2nd and 3rd is 5.9, give each value a rank of 2.5).

Let n_a = Number of observations in Group A, n_b = Number of observations in Group B, T_a = Rank sum for Group A, T_b = Rank sum for Group B:

$T_a = 1+3+5+8 = 17$

$T_b = 2+4+6+7 = 19$

Let us calculate U_1 and U_2:

$$U_1 = T_a - \frac{n_a(n_a+1)}{2} = 17 - \frac{4(4+1)}{2} = 7$$

$$U_2 = T_b - \frac{n_b(n_b+1)}{2} = 19 - \frac{4(4+1)}{2} = 9$$

The smallest value 7 is the *U* value.

The smallest *U* value, 7 is compared with the Mann-Whitney *U* Table value at n_1=4 and n_2=4. Relevant part of the *U* Table is reproduced in Table 12.11.

Table 12.11. Mann-Whitney *U* Table

n_1	n_2	Two-sided		One-sided	
		α=0.05	α=0.01	α=0.05	α=0.01
2	2	---	---	---	---
3	3	---	---	---	---
4	4	0	---	1	---
5	5	2	0	4	1
6	6	5	2	7	3

Since the computed *U* value is greater than the values given in the Mann-Whitney *U* Table, it is not significant at 5% level by two-sided and one-sided tests (at 5 % significant level the *U* Table values are 0 and 1 for two-sided and one-sided tests, respectively).

When the size of either of the groups exceeds 20, the significance of *U* can be tested using the *Z* statistic:

$$Z = \frac{U - n_1 n_2 / 2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

Z score for normal distribution is shown in Appendix 3.

(A note on *Z* statistic: *Z* is designated to a standard normal variate. It is computed by subtracting the measured value from the population mean, then dividing by the population SD(σ). A standard normal variate has a normal distribution with mean 0 and variance 1. The total area under a normal distribution curve is unity (or 100%). The notation, Prð(–1 < z < 1) = 0.6826, indicates that about 68% of the area is contained within ± 1 SD).

Mann-Whitney's *U* test works well in the analysis of data obtained from toxicity studies, where the number of animals in each group is 27 or less. By Mann-Whitney's *U* test, a significant difference (one-sided test) can be detected even with three animals in each group. Therefore, this test can be used in experiments with dogs, where each group usually consists of three animals/sex. This test seems to be extensively used for analyzing urinalyses data and pathological findings in repeated dose administration studies in rodents.

The power to detect a significant difference is more with Mann-Whiney's *U* test than the Fisher's test. Analysis of pathological findings of a repeated dose administration study by Mann-Whitney's *U* and Fisher's tests is given in Table 12.12.

Table 12.12. Analysis of pathological findings of a repeated dose administration study by Mann-Whitney's *U* and Fisher's tests

Groups	Lesions grades and number of animals with lesions grade				Mann-Whitney's *U* test	Lesions grades and number of animals with lesions grade		Fisher's test
	-	±	+	++	*P*=0.0032 (One-sided)	-	> ±	*P*=0.0238 (One-sided)
Control	4	1	0	0		4	1	
High dose	0	0	3	2		0	5	

The computed *P* value for Mann-Whitney's *U* test (*P*=0.0032) is considerably less than that of the Fisher's test (*P* = 0.0238), indicating that the power to detect a significant difference is more with Mann-Whitney's *U* test than the Fisher's test.

The power of the Mann-Whitney's *U* test decreases when the groups to be compared have the same order of rank. There is a possibility in having the same order of rank, when the number of digits after decimal of the raw data is truncated. This can be better understood from the data given in Table 12.13.

Table 12.13. Change in the pattern of significant difference detection as the number of digits after decimal of the raw data decreases. Absolute liver weight (g) of male rats from a 28-day repeated dose administration study is given in the Table.

Number of digits after decimal	Items	Groups		P
		Control (N = 6)	High dose (N = 6)	Mann-Whitney's *U* test
3	Raw data	10.391, 11.442, 13.653, 10.224, 10.783, 10.414	13.194, 11.444, 13.701, 11.572, 12.683, 12.661	< 0.05
	Mean ± SD	11.151 ± 1.301	12.543 ± 0.889	
	Mean rank	4.3	8.6	
2	Raw data	10.39, 11.44, 13.65, 10.22, 10.78, 10.41	13.19, 11.44, 13.70, 11.57, 12.68, 12.66	Not significant
	Mean ± SD	11.15 ± 1.30	12.54 ± 0.89	
	Mean rank	4.4	8.5	
1	Raw data	10.4, 11.4, **13.7**, 10.2, 10.8, 10.4	13.2, 11.4, **13.7**, 11.6, 12.7, 12.7	Not significant
	Mean ± SD	11.2 ± 1.3	12.6 ± 0.9	
	Mean rank	4.5	8.5	

The high dose group is significantly different from the control group as per Mann-Whitney's *U* test, when the data of both the groups have three digits after decimal and no data from the control group is repeated in the high dose group and *vice versa*. When the number of digits after the decimal of the data was truncated to two decimals, the value 11.44 was repeated in both the groups, resulting in an insignificant difference between the control and high dose groups. When the number of digits after the decimal of the data was restricted to one decimal, the values 11.4 and 13.7 were repeated in both the groups, resulting in an insignificant difference between the control and high dose groups.

There are two methods for calculating the Mann-Whitney's *U* test. When the number of observations in each group is small (N= <27), the Mann-Whitney's *U* test can be calculated by using a ready reckoner (*http://aoki2.si.gunma-u.ac.jp/lecture/Average/u-tab.html*). When the number of observations in each group is large (N= >27), it is calculated using the *Z* distribution Table method. Table 12.14 demonstrates the analysis of a simulated data with a strong dose-related pattern by Mann-Whitney's *U* test using the *Z* distribution Table method. Table 12.15 demonstrates the analysis of a simulated data with strong dose-related pattern by Mann-Whitney's *U* test using the ready reckoner.

Table 12.14. Power of Mann-Whitney's *U* test for three and four samples with a strong dose-related pattern (calculated by using *Z* distribution Table)

Number of samples	Group	Raw data (ranked)	Mean rank	*Z* value	*P* value	
					Two-sided	One-sided
3	Control	1, 2, 3	2	1.96	0.04953	0.02500
	Dose	4, 5, 6	5			
4	Control	1, 2, 3, 4	2.5	2.30	0.0209	0.010
	Dose	5, 6, 7, 8	6.5			

Table 12.15. Power of Mann-Whitney's *U* test for three and four samples with a strong dose-related pattern (calculated by using the ready reckoner—*http://aoki2.si.gunma-u.ac.jp/lecture/Average/u-tab.html*)

Number of samples	Group	Raw data (ranked)	Mean rank	*U* value	*P* value	
					Two-sided	One-sided
3	Control	1, 2, 3	2	0.0	Not significant	$P<0.05$.
	Dose	4, 5, 6	5			
4	Control	1, 2, 3, 4	2.5	0.0	$P=0.05$	$P<0.05$.
	Dose	5, 6, 7, 8	6.5			

The Tables 12.14 and 12.15 indicate that there is not much difference in *P* values between *Z* distribution Table and ready reckoner methods, when the number of samples is as small as 3 to 4. However, we recommend a ready reckoner when the number of observations in each group is small (N= <27) and a *Z* distribution Table when the number of observations in each group is large (N= >27).

Kruskal-Wallis Nonparametric ANOVA by Ranks
(Kruskal and Wallis, 1952)

The Kruskal–Wallis test is identical to one-way ANOVA with the data replaced by their ranks. It has also been stated that this test is an extension of the two-group Mann-Whitney's *U* (Wilcoxon rank) test (Mc Kight and Najab, 2010). It assumes that the observations in each group come from populations with the same shape of distribution, so if different groups have different shapes (for example, one is skewed to the right and another is skewed to the left or they have different variances), the Kruskal–Wallis test may give inaccurate results (Fagerland and Sandvik, 2009).

Calculation Procedure:

The data is ranked and the sum of the ranks is calculated. Then the test statistic, *H*, is calculated (hence this test is also called as Kruskal-Wallis *H* test). *H* is approximately chi-square distributed. Kruskal-Wallis test is not suitable if the sample size is small, say less than 5.

The formula for the calculation of chi-square value is given below (Equation 1):

$$X^2 = \frac{12 \times \left(\frac{r_1^{\,2}}{N_1} + \frac{r_2^{\,2}}{N_2} + \cdots\cdots + \frac{r_a^{\,2}}{N_a} \right)}{N(N-1)} - 3(N+1)$$

If the groups have data with same ranks, the chi-square value is calculated as given below (Equation 2):

$$X^2 = \frac{(N-1)S}{\left\{ \left(r_{11} - \frac{N+1}{2} \right)^2 + \cdots\cdots + \left(r_{ana} - \frac{N+1}{2} \right)^2 \right\}}$$

$$S = \frac{\left(\frac{r_1 - N_1(N+1)}{2} \right)^2}{N_1} + \frac{\left(\frac{r_2 - N_2(N+1)}{2} \right)^2}{N_2} + \cdots\cdots + \frac{\left(\frac{r_a - N_a(N+1)}{2} \right)^2}{N_a}$$

If the derived chi-square value is larger than the chi distribution Table value, then it indicates a significant difference.

Let us work out an example. Lymphocyte count determined in four groups in a clinical study is given in Table 12.16.

Table 12.16. Lymphocyte counts (%) determined in a clinical study

	Group A	Group B	Group C	Group D
	40.6	31.9	32.7	30.6
	38.0	36.8	31.3	35.9
	41.1	32.4	32.9	29.6
	52.7	34.8	31.9	29.2
	48.8	43.1	**28.5**	**28.5**
	41.1	39.0	31.2	30.8
	39.9	33.6	33.1	30.5
	43.1	34.3	34.1	29.4
	32.7	34.0	31.2	30.8
	30.1	33.8	31.7	32.0
Mean	40.8	35.4	31.9	30.7
N	10	10	10	10

Number group = 4; Total number of samples = 40.

Combine the lymphocytes counts of all the four groups, and arrange them from the smallest to the largest. Then assign a rank from 1 to 40 to them as given in Table 12.17. (Note: we have done a similar exercise while working out the example of scores for performance of six classes of two schools for explaining Wilcoxon rank-sum test; *vide* Tables 12.4 and 12.5).

Table 12.17. Ranks assigned to the lymphocyte counts (%) of four groups

	Group A	Group B	Group C	Group D
	34	15.5	19.5	8
	31	30	13	29
	35.5	18	21	5
	40	28	15.5	3
	39	37.5	**1.5**	**1.5**
	35.5	32	11.5	9.5
	33	23	22	7
	37.5	27	26	4
	19.5	25	11.5	9.5
	6	24	14	17
Mean rank	31.1	26	15.55	9.35

Equation 2 (page 117) is used to calculate the chi-square value.

Let us calculate r_1, r_2, r_3 and r_4:

$$r_1 = 34+31+\cdots\cdots\cdots+19.5+6 = 311$$

$$r_2 = 15.5+30+\cdots\cdots\cdots+25+24 = 260$$

$$r_3 = 19.5+13+\cdots\cdots\cdots+11.5+14 = 155.5$$

$$r_4 = 8+29+\cdots\cdots\cdots+9.5+17 = 93.5$$

S is calculated as 2914.35 (see below):

$$S = \frac{\left(311-\frac{10\times 41}{2}\right)^2}{10} + \frac{\left(260-\frac{10\times 41}{2}\right)^2}{10} + \frac{\left(155.5-\frac{10\times 41}{2}\right)^2}{10} + \frac{\left(93.5-\frac{10\times 41}{2}\right)^2}{10} = 2914.35$$

X^2 is calculated as 21.3 (see below):

$$X^2 = \frac{(40-1)\times 2914.35}{\left(34-\frac{(40+1)}{2}\right)^2 + \left(31-\frac{(40+1)}{2}\right)^2 + \cdots\cdots + \left(9.5-\frac{(40+1)}{2}\right)^2 + \left(17-\frac{(40+1)}{2}\right)^2}$$

$$= \frac{113659.7}{5326.5} = 21.3$$

The computed X^2 value is compared with the X^2 Table value (Table 12.18) at 4–1=3 degrees freedom. Since the computed X^2 value (21.3) is greater than the X^2 Table value (16.266), it is considered that there is a significant difference in lymphocyte counts among the groups (P<0.001).

Table 12.18. Chi square Table (Yoshimura, 1987)

DF\α	0.1	0.05	0.01	0.001
1	2.706	3.841	6.635	10.828
2	4.605	5.991	9.210	13.816
3	6.251	7.815	11.345	**16.266**
4	7.779	9.488	13.277	18.467
5	9.236	11.070	15.086	20.515

Comparison of Group Means

Wilcoxon Rank-Sum test or Kruskal-Wallis test provides the information, whether a significant difference exists among the group means. If these

tests reveal a significant difference, it does not indicate that every group means are significantly different from each other. One of the robust tests used to find out which group means are significantly different from each other is the Dunn's multiple comparison test. Dunn's multiple comparison test can be used to find the difference of 3 or more groups (Israel, 2008).

Dunn's multiple comparison test for more than three groups (Gad and Weil, 1986; Hollander and Wolf, 1973)

Let us review the example given in Table12.17. The mean rank values are reproduced in Table 12.19.

Table 12.19. Mean rank of lymphocyte (%)

	Group A	Group B	Group C	Group D	
Mean rank	31.1	26	15.6	9.4	
N	10	10	10	10	Sum=40

Calculation procedure

Group A** vs **Group B:

Difference of mean rank: 31.1–26=5.1

The Probability value:

$$\left[\frac{0.05}{4(3)}\right] = Z_{0.00417} = 2.63\sqrt{\frac{(40)(41)}{12}} \times \sqrt{\frac{1}{10}} + \sqrt{\frac{1}{10}} = 13.7$$

Group A vs Group C:

Difference of mean rank: 31.1–15.6=15.5

The Probability value:

$$\left[\frac{0.05}{4(3)}\right] = Z_{0.00417} = 2.63\sqrt{\frac{(40)(41)}{12}} \times \sqrt{\frac{1}{10}} + \sqrt{\frac{1}{10}} = 13.7$$

Group A vs Group D:

Difference of mean rank: 31.1-9.4=21.7

The Probability value:

$$\left[\frac{0.05}{4(3)}\right] = Z_{0.00417} = 2.63\sqrt{\frac{(40)(41)}{12}} \times \sqrt{\frac{1}{10}} + \sqrt{\frac{1}{10}} = 13.7$$

4 (3) = Number of group × Number of group – 1; The value 2.63 is obtained from Table 12.20 **(**the value, 0.00417 can be rounded to 0.0042. This value lies between 0.0043 and 0.0041 of *Z* value. In this case, 0.0043 was considered. The *Z* value corresponding to 0.0043 is 2.63).

The numerator (40) is total number of samples, (41) is total number of sample + 1; The denominator 12 is a constant, whereas 10 is number of samples in the groups.

Table 12.20. *Z* score for normal distribution (Gad and Weil, 1986)

Z	Proportional parts									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.6	0.0047	0.0045	0.0044	**0.0043**	**0.0041**	0.0040	0.0039	0.0038	0.0037	0.0036

The difference between the two mean scores is compared with the Probability (critical) value (13.7). If the difference between the two mean scores is greater than the Probability (critical) value, then the difference is considered significant (see below given Table 12.21).

Table 12.21. Significant difference between the groups

Analysis	Difference	Critical value	P
Group A *vs* Group B	31.1–26=**5.1**	13.7	Not significant (P>0.05)
Group A *vs* Group C	31.1–15.6=**15.5**		Significant (P<0.05)
Group A *vs* Group D	31.1–9.4=**21.7**		Significant (P<0.05)

Steel's multiple comparison test for more than three groups
(Steel, 1961)

The power of Steel's test is higher than the other multiple comparison tests. Usually the number of groups employed is four (three treatment groups + one control group) in most of the animal studies. For a parameter which shows a strong dose-related pattern, a significant difference can be detected by Steels's test, even if the number of animals in a group is as low as four (Yoshimura and Ohashi, 1992; Inaba, 1994). Let us work out an example (Table 12.22).

Calculation procedure:

Control group *vs* Low dose group

1) Sum of rank of low dose group, R_2=5+6+7+8=26

Table 12.22. Quantitative data from a toxicity study

Group	Control	Low dose	Mid dose	High dose
	1	5	9	13
	2	6	10	14
	3	7	11	15
	4	8	12	16
Mean rank	2.5	6.5	10.5	14.5

Note: Ranked values are given.

2) Calculation of SS(S_2) and Variance (V_2)
$S_2 = (1–4.5)^2 + (2–4.5)^2 + (3–4.5)^2 + (4–4.5)^2 + (5–4.5)^2 + (6–4.5)^2 + (7–4.5)^2 + (8–4.5)^2 = 42$, where

4.5 = Sum of number of samples of control group and number of samples of low dose group + 1 divided by number of groups [(4+4+1)/2]= 4.5).

$$V_2 = \frac{4 \times 42}{4 \times 8 \times 7} = 0.75 \text{, where}$$

4×42 = Number of sample in control group × S_2value, 42; 4×8×7 = Number of sample in low dose × Sum of number of samples of control and low dose groups ×Sum of number of samples of control and low dose groups – 1.

3) Calculation of t_2

$$t_2 = \frac{\frac{26}{4} - \frac{4+4+1}{2}}{\sqrt{0.75}} = \frac{2}{0.866} = 2.309 \text{, where}$$

26/4 =R_2/4 (4=Number of sample in low dose), (4+4+1)/2 =(Number of samples in control group + Number of samples in low dose group + 1)/2; 0.75 = V_2.

4) Calculated t_2 value, 2.309 is compared with the critical value given in Table 12.23. As the size of each group is similar, the critical value becomes (∞, 4) =2.062.

5) Since computed t_2 value, 2.309 is greater than the Table value,2.062, it is considered that the low dose group is significantly different from the control.

Table 12.23. Dunnett's *t* test critical values, one-sided at 0.05 probability level (Yoshimura, 1987)

Number of group	2	3	4	5	6	7	8
∞	1.645	1.916	**2.062**	2.160	2.234	2.292	2.340

Using the calculation procedure mentioned above for comparing control group *vs* low dose group, comparison between other groups (control group *vs* mid dose group and control group *vs* high dose group) can be made.

Rank Sum Tests—Some Points

An interesting example of a rank sum test analysis is given in Table 12.24. Creatinine value of F344 rats on week 52 in a repeated dose administration study is given in the Table.

Table 12.24. Creatinine value (mg/dl) of F344 rats at 52 weeks after dosing

Group	Individual value (20 animals/group)	Mean ± SD
Control	0.70 0.68 0.70 0.74 0.60 0.65 0.65 0.72 0.63 0.78 0.67 0.64 0.63 0.66 0.88 0.73 0.57 0.79 0.78 0.65	0.69 ± 0.07
Low dose	0.72 0.64 0.66 0.66 0.88 0.68 dead 0.51 0.65 0.63 0.79 0.60 0.69 0.68 0.62 0.57 dead 0.66 0.59 0.54	0.65 ± 0.09
Middle dose	0.56 0.59 0.66 0.68 0.57 0.67 0.70 0.83 0.86 0.68 0.60 0.68 0.57 0.67 0.53 0.57 0.64 0.61 0.86 0.67	0.66 ± 0.10
High dose	0.51 0.59 0.49 0.60 0.58 0.62 0.51 0.57 0.60 2.96 0.56 0.65 0.71 0.55 0.54 0.41 0.52 0.62 0.59 0.59	0.69 ± 0.54**

**Significantly different from control by rank sum test ($P<0.01$).

Bartlett's test for homogeneity of variance showed a significant difference, therefore Dunnett type rank test was used for the analysis of the data. The Dunnett type rank test revealed a significant difference between the high dose group and the control group ($P<0.01$), though the mean values of these groups are the same (0.69). Close examination of the individual values of the high dose group revealed that one of the values among them (2.96) is extremely high compared with the other values. If a number slightly higher than 0.88, which is the next highest value among the high dose and control groups, replaces 2.96 of the high dose group, the mean value of this group becomes lower than that in the control group, but the rank is not changed, *i.e.,* the result of the rank sum test will not be changed. Thus, the significant difference between the control group and high dose group detected by the rank sum test is understandable, though the mean values of these groups are the same.

Another important point in rank sum test analysis is that one should know the minimum number of animals required in each group to detect a significant difference. Table 12.25 shows the minimum number of animals required in four-group and five-group settings to detect a significant difference.

Table 12.25. Minimum number of animals in four-group and five-group settings necessary to show a significant difference

Test	Four groups	Five groups
Scheffé type	22	40
Hollander-Wolfe*	19	30
Tukey type	18	32
Dunnett type	15	26
Wilcoxon	8	12
Steel type	4	6
Mann-Whitney *U***	3	

*Dunn's test. **Test for 2 group alone.

The power also depends on the number of treatment groups, which implies that inclusion of further non-significant treatment group/s can result in overlooking significant effects (Hothorn, 1990).

As mentioned earlier, the power to detect a significant difference is high with Steel's test. A comparison of the power to detect a significant difference between Dunnett type rank test and Steel's test is given in Table 12.26.

Table 12.26. Comparison of the power to detect a significant difference between Dunnett type rank test and Steel's test

Parameter analysed and tests	Control (N=5)	Low dose (N=5)	Mid dose (N=5)	High dose (N=4)	Top dose (N=4)
Urine volume (ml)	2.4, 2.8, 2.4, 2.4, 2.4	43, 45, 40, 41, 46	62, 48, 68, 52, 55	73, 72, 102, 104	52, 97, 99, 103
Mean ± SD	2.5 ± 0.18	43 ± 2.55	57 ± 8.0	87.8 ± 17.6	87.8 ± 24
Bartlett's homogeneity test	P = 0.0001				
Kruskal-Wallis's test	P = 0.0006				
Dunnett type rank test		NS	S	S	S
Steel's test		S	S	S	S

NS-Not significant (P>0.05); S-Significant (P<0.05)

The low dose group was not significantly different, when analysed using Dunnett type rank test, whereas, this dose group was significantly different, when analysed using Steel's test.

Most of the pharmacologists and toxicologists express their concern about use of non-parametric tests like rank sum tests, because of their low sensitivity in detecting a significant difference. However, some

biostatisticians are of the opinion that the rank sum tests are more useful for analyzing the biological data than the parametric tests.

References

Clarke, S.C. (1991): Invited commentary on R. A. Fisher. Am. J. Epidemiol., 134(12), 1371–1374.

Crawley, M.J. (2005): Statistics: An Introduction Using R. John Wiley and Sons Ltd., Chichester, UK.

Elston, R.C. and Johnson, W.D. (1994): Essentials of Biostatistics. F.A. Davis & Co., Philadelphia, USA.

Fagerland, M.W. and Sandvik, L. (2009): The Wilcoxon-Mann-Whitney test under scrutiny. Statist. Med., 28, 1487–1497.

Fisher, R.A. (1922): On the interpretation of χ^2 from contingency tables, and the calculation of *P*. J. Royal Stat. Soc., 85(1), 87–94.

Fisher, R.A. (1954): Statistical Methods for Research Workers. Oliver and Boyd, London, UK.

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

Hollander, M. and Wolf, D.A. (1973): Non-Parametric Statistical Methods.John Wiley, New York, USA.

Hothorn, L. (1990): Biometrische Analyse spezieller Untersuchungen der regulatorischen Toxikologie. In: Aktuelle Probleme der Tbxikologie, Vol. 5 Grundlagen der Statistik fuer Toxikologen (M. Horn and L. Hothorn, Eds.) Verlag Gesundheit Gmbh, Berlin, Germany.

Inaba, T. (1994): Problem of multiple comparison method used to evaluate medicine of enzyme inhibitor X1, Japanese Society for Biopharmaceutical Statistic, 40, 33–36.

Israel, D. (2008): Data Analysis in Business Research-A Step by Step Non-Parametric Approach. SAGE Publications India Pvt. Ltd., New Delhi, India.

Kruskal, W.H. and Wallis, A.W (1952): Use of ranks in one criterion analysis of variance. J. Am. Stat. Assoc., 47(260), 583–621.

Le, C.T. (2003): Introductory Biostatistics. John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

Ludbrook, J. (2008): Analysis of 2 × 2 tables of frequencies: Matching test to experimental design. Int. J. Epidemiol., 37(6), 1430–1435.

Mann, H.B. and Whitney, D.R. (1947): On a test of whether one of 2 random variables is stochastically larger than the other. Ann. Math. Stat., 18, 50–60.

Mc Donald, J.H. 2009: Handbook of Biological Statistics, 2nd Edition. Sparky House Publishing, Baltimore, USA.

Mc Kight, P.E. and Najab, J. (2010): Kruskal-Wallis Test. In: Corsini Encyclopedia of Psychology. Editors, Weiner, I.B. and Craighead, W.E., Wiley Online Library, DOI: 10.1002/9780470479216.

Mc Kinney, W.P., Young, M.J., Hartz, A. and Lee, M.B. (1989): The inexact use of Fisher's exact test in six major medical journals. JAMA, 16, 261(23), 3430–3433.

Nietert, P.J. and Dooley, M.J. (2011): The power of the sign test given uncertainty in the proportion of tied observations, 32(1), 147–150.

Sawilowsky, S. (2005): Encyclopedia of Statistics in Behavioral Science. Wiley Online Library, DOI: 10.1002/0470013192.bsa615.

Steel, R.G.D. (1961): Some rank sum multiple comparison tests. Biometrics, 17(4), 539–552.

Surhone, L.M., Timpledon, M.T. and Marseken, S.F. (2010): Sign Test. VDM Verlag Dr Mueller AG&Co., KG, Germany.

Whitley, E. and Ball, J. (2002): Statistics review 6: Nonparametric methods, Crit. Care, 6(6), 509–513.

Wilcoxan, F. (1945): Individual comparisons by ranking methods. Biometrics Bull., 1(6), 80–83.

Yoshimura, I. (1987): Statistical Analysis of Toxicological Data. Scientist Press, Tokyo, Japan.

Yoshimura, I. and Ohashi, Y. (1992): Statistical Analysis for Toxicology Data. Chijin-Shokan, Tokyo, Japan.