7

Tests for Significant Differences

Null Hypothesis

The main objective of conducting an animal experiment is to know whether the treatment with a test item causes any effect compared to the control group. The comparison between the treatment group/s and the control group is made using various statistical tools. The selection of an appropriate statistical tool is based on certain assumptions. Before we go further, we need to understand a hypothesis called 'null hypothesis'.

In the statistical context, a hypothesis is a statement about a distribution (example, normal distribution), or its underlying parameter (example, mean value, μ) or a statement about the relationship between probability distribution (example, there is no statistical difference between the treated and the control groups) or its parameter ($\mu_1 = \mu_2$) (Le, 2009). Why is it called as 'null hypothesis'? Let us try to understand 'null hypothesis' using the explanation proposed by Yoshida (1980). No pharmaceutical company will venture in developing a new drug, A1, if it is not superior to the drug currently in use, A2. In a statistical analysis, we first hypothesize that drugs A1 and A2 have the same therapeutic value. That is, we hypothesize A1 = A2, which is contrary to our assumption A1 > A2. When the experimental data fail to show A1 = A2, we judge that A1 differs from A2 and reject the hypothesis. Thus, in a statistical test, we first hypothesize A1 = A2 in contrast to our assumption A1 > A2, and then show that it is not true (A1 $\neq$ A2). The original hypothesis A1 = A2, which is desirably rejected, is called the null hypothesis. In most of the statistical books null hypothesis is notated as:

$H_0=\mu_1=\mu_2$, and the alternate hypothesis is notated as:

$H_1=\mu_1\neq\mu_2$, where μ_1 and μ_2 are the mean values of two groups.

Generally, a statistical process starts from the null hypothesis, which assumes no difference between the control group and the treated group or among the groups, and if a significant difference is detected at 5% significant level ($P<0.05$), the null hypothesis is rejected.

Significant Level, Type I and Type II Errors

In the publications of pharmacological and toxicological experiments one would have come across authors using $P<0.05$, usually as footnotes of the Tables to denote a significant difference. *P* stands for probability. In order to detect a significant difference we have to challenge the null hypothesis. When $P<0.05$, the null hypothesis is rejected. It means there is only a 5% chance of rejecting null hypothesis, when it is true. We are not supposed to reject null hypothesis, when it is true, if we reject it, it is called as Type I error. In a pharmacological experiment, if you reject the null hypothesis, when actually it is true, *i.e.*, $H_0=\mu_1=\mu_2$ (there is no difference between the treated and control groups), you would report that the drug that you tested had an effect, causing a Type I error. Hence, this Type I error is also called as 'false positive'. Experimental design in pharmacology should be proper so that misleading claims concerning the effectiveness of a treatment (Type I error) are not made (Spina, 2007). Type II error is opposite to Type I error, also called as 'false negative', occurs when you falsely accept the null hypothesis.

Why at 5% Significant Level?

In statistical analysis, the smallest probability for rejecting a null hypothesis, when it is true, is considered as 5% (Madsen, 2011). The same is used in most of the pharmacological and toxicological studies, where a significant difference between the treated and the control groups is judged at 5% probability level. Why the statisticians look upon 5% probability as the cut-off point for assessing a significant difference? Let us try to explain it with an example: A tennis player loses several matches against an opponent of supposedly equal skill level. How many losses will be required for the player to regard the opponent as a better player than him? It is not odd for a player to lose three consecutive games to his opponent with equal ability, but the fourth consecutive loss leads the player to believe that his opponent to be a better player. After losing five consecutive games, the

player may abandon the null hypothesis (null hypothesis in this case is that the player and his opponent have equal skill level) and consider that his opponent is a better player than him. If the player and his opponent have equal ability, the probability of losing the game once by the player is 1/2, but the probabilities of losing four and five games consecutively by the player are $(1/2)^4 = 6.3\%$ and $(1/2)^5 = 3.2\%$, respectively. The mid-point of these probabilities is about 5% [(6.3+3.2)/2=4.8%)].

The five percent significant level which implies 1 mistake in 20 observations (1/20) is normally unavoidable in biological experiments and has been used for more than half a century in bioassays including toxicity tests (Dunnett, 1955; Kornegay *et al.*, 1961). Hence, according to Bailey (1995), the five percent significant level can be generally used for flagging a significant difference. Conventionally, a *P* value of <0.05 indicates statistical significance (Doll and Carney, 2005).

However, strictly adhering to a 5% significant level to delineate a significant difference has been questioned by few statisticians. Fisher (1955) recommended a 5% significant level based on a single hypothesis, H_0. Neyman and Pearson (1928, 1936) proposed a decision process which seeks to confirm or reject *a priori* hypothesis and rejected Fisher's idea that only the null hypothesis needs to be tested. Statisticians posed questions against Fisher's 5% probability level; the question was 'what should be the smallest *P* value that warrants rejection of the null hypothesis?' In later years, Fisher (1971) stated that the *Q* value can be significant at a 'higher standard, if *P* is 1%' and at a 'lower standard if *P* is 5%'. It again states, though indirectly, that a significant difference can be obtained only when the *P* is between 1 and 5%. (Note: *Q* value is the 'false discovery rate' analogue of *P*).

Many statisticians do not favor strictly characterizing the result of a statistical analysis into a positive or negative finding on the basis of a *P* value. They suggest, when reporting the results of significance tests, precise *P* values (example, $P<0.049$ or $P<0.051$) should be reported rather than referring to specific critical values. Interpretation of the results of a statistical analysis should not be made solely on the basis of null hypothesis. The hypothesis testing has been challenged and there has been suggestion to report confidence intervals rather than *P* (Krantz, 1999). According to Gelman and Stern (2006) 'dichotomization into significant and non-significant results encourages the dismissal of observed differences in favor of the usually less interesting null hypothesis of no difference'. In the case of experiments conducted in pharmacology and toxicology, biological

relevance of the results also should be considered for interpreting the data. Declaring a result non-significant does not mean that the effect is not biologically relevant; it only means that there is not sufficient evidence to reject the null hypothesis. In a nutshell, statistical analysis should not override the experience of the experimenter in interpreting the results of the experiments.

How to Express *P*?

The published articles express the *P* in two ways: $P < 0.05$ or $P \leq 0.05$. The question is how the *P* should be expressed—$P < 0.05$ or $P \leq 0.05$? Though, technically, it may be better to express $P \leq 0.05$, $P < 0.05$ also conveys similar information on statistical significance. We conducted a small investigation on the expression of *P* in toxicological/pharmacological articles published in few journals. In most of the journals investigated, we observed that $P \leq 0.05$ and $P < 0.05$ were used at similar frequencies. In the toxicological/pharmacological experiments conducted in Japan, $P < 0.05$ tended to be used slightly more frequently than $P \leq 0.05$. In the technological report of the National Toxicology Program of NIH, USA, $P < 0.05$ is more widely used.

One-sided and Two-sided Tests

Generally, it has been stated that a one-sided test is used in the following cases: 1) the difference, large or small is questioned and 2) the inter-group difference (plus or minus) is known in advance. On the other hand, a two-sided test is used in the following cases: 1) only the presence or absence of an inter-group difference is questioned and 2) it is not certain whether the inter-group difference is plus (positive) or minus (negative). The detection rate of a significant difference differs depending on the selection of a one-sided or a two-sided test. Let us work out an interesting example: A customer went to a grocery shop to buy a loaf of bread. The weight of a loaf of bread printed on the bread wrappers was 450 g. On a hunch, the customer purchased one loaf of bread from the shop daily for seven days and weighed the loaves. The weights were 444, 434, 450, 430, 458, 446 and 422 g. He informed the grocer that the weight printed on the bread wrapper did not match with the actual weight of the bread. The grocer offered to analyse the data provided by the customer using a two-sided test. The calculated *t* value (2.14) was less than the value of *t*-distribution Table

(2.447), hence the null hypothesis was not rejected (Note: Normally we analyse the data using a statistical formula to obtain a 'calculated value'. Then, we compare this 'calculated value' with the value (critical value) given in the appropriate statistical Table. If the calculated value is greater than the Table value (critical value), we consider the null hypothesis is rejected. In this particular example we have analysed the data using a *t*-test and got a *t* value. This *t* value was compared with the value given in the *t* Table. You shall learn about various statistical tools and their applications in later chapters). Not-rejection of the null hypothesis means there is no statistical significant difference among the weights of seven loaves of the bread that the customer purchased. The customer was not convinced with the result of the two-sided test provided by the grocer. The customer decided to analyse the data using a one-sided test, with the assumption that the weight of the loaf of the bread is less than 450 g. When the customer analysed the data using the one-sided test, he found that the calculated *t* value (2.14) was greater than the value of *t*-distribution Table (1.943). Therefore, "Null hypothesis" is rejected, which means that there is a statistical significant difference among the weights of seven loaves of the bread that he purchased.

Which Test to Use: One-sided or Two-sided?

It is interesting to note that scientists have different views in choosing between one-sided test and two-sided test. Kobayashi *et al*. (2008) examined whether a one-sided or a two-sided test was used in the analysis of the data obtained from 122 numbers of 28-day repeated dose administration studies in rats. The studies were conducted as per Chemical Substances Control Law, Japan (CSCL, 1986) or OECD test guideline (OECD, 2008). Out of 122 studies examined, quantitative data of 22 studies were analysed by the one-sided test, 87 studies were analysed by two-sided test, whereas there was no mention about whether the one-sided or two-sided test was used in 13 studies. With regard to qualitative data, in 34 and 22 studies the data were analysed by the one-sided and two-sided tests, respectively, whereas there was no mention about whether the one-sided or two-sided test was used in 66 studies.

Drewitt *et al*. (1993) used a two-sided *t*-test for preliminary studies and one-sided test for the main studies. Shertzer and Sainsbury (1991) used a one-sided *t*-test for the detection of a significant difference between two groups. Yoshimura and Ohashi (1992) recommended the one-sided test because the results of a toxicity study are evaluated by the presence or absence of an increase in the mean value of the treated groups in comparison

with the control group. Shirley (1997) used a two-sided test for Student's *t*-test and Cochran's *t*-test, and if a significant difference is observed in the ANOVA, used the one-side test in Dunnett's multiple comparison test. Dunnett (1955) recommended the use of a two-sided test to determine simultaneously the upper and lower limits to the difference between the control group and each treated group and a one-sided test to determine either the upper or lower limit to the difference between the control group and each treated group. Gad and Weil (1988) explained the significant difference between the control and treated groups in body weight by using the two-sided test. Sakuma (1977) suggested to select either a one- or a two-sided test referring to the reports of similar studies. Nakamura (1986) stated that selection of one- or two-sided test may depend on the objective of the study, and he suggested that the statistical significance of the data should not be foreseen. Kobayashi (1997) recommended a one-sided test for the analysis of data obtained from toxicological studies.

A significant difference is more apt to be observed in a one-sided test than in a two-sided test. According to a survey, the detectability of a significant difference by the two-sided test was 71–95% of that by a one-sided test in the Dunnett's *t*-test (Table 7.1) (Kobayashi, 1997).

Table 7.1. Difference in number of significant differences ($P < 0.05$) by one- and two-sided test by Dunnett's *t*-test in a chronic toxicity and carcinogenicity study

Items	No. of statistical analyses	Dunnett's *t*-test	
		One-sided	Two-sided
Body weight (b.w.)	528	223	212 (95)
Feed consumption	832	235	189 (80)
Hematology	352	123	105 (85)
Blood chemistry	576	215	181 (84)
Urinalysis	64	7	5 (71)
Organ weight	224	47	42 (89)
Organ weight/b.w.	224	82	67 (81)
Total	2800	932	801 (86)

Note: Values in parentheses show the percent significant difference by two-sided test with regard to one-sided test.

Overall significant difference shown by the two-sided test is 86% of the one-sided test. The reason for this is that one-sided test requires less strength of evidence than the two-sided test for a significant difference. It is likely that an item shown as insignificant by a two-tailed test can be significant by a one-sided test. One-sided test should never be used to make a conventionally non-significant difference significant (Bland and

Bland, 1994). Therefore, it is important to decide to use a one-sided or a two-sided test before the data collection (Rosner, 2010).

The rejection limit value at 5% probability level of *t*-test and Dunnett's multiple comparison test was excerpted and is shown in Table 7.2. The rejection limit value of the one-sided test does not become 1/2 the value of the Table of the two-sided test, but it becomes 78% of two-sided test in *t*-test, and it becomes 85% of two-sided test at four groups setting in Dunnett's multiple comparison test.

Table 7.2. Rejection limits of *t*-test and Dunnett's multiple comparison test with one- and two-sided (Yoshimura, 1987)

DF	Rejection limit at 5% level			
	t-Table		Dunnett's Table[a]	
	Two-sided	One-sided	Two-sided	One-sided
1	12.706	6.314	–	–
2	4.303	2.920	–	–
3	3.182	2.353	3.867	2.912
4	2.776	2.132	3.310	2.598
5	2.571	2.015	3.030	2.433
6	2.447	1.943	2.863	2.332
7	2.365	1.895	2.752	2.264
8	2.306	1.860	2.673	2.215
9	2.262	1.833	2.614	2.178
10	2.228	1.812	2.268	2.149
•	•	•	•	•
•	•	•	•	•
21	2.080	1.721	2.370	2.021
22	2.074	1.717	2.363	2.016
•	•	•	•	•
•	•	•	•	•
31	2.040	1.696	2.317	1.986
32	2.037	1.694	2.314	1.984
•	•	•	•	•
•	•	•	•	•
41	2.020	1.683	2.291	1.969
42	2.018	1.682	2.289	1.968
•	•	•	•	•
•	•	•	•	•
60	2.000	1.671	2.265	1.952
120	1.980	1.657	2.238	1.934
240	1.970	1.651	–	–
∽	1.960	1.645	2.212	1.916
Rate[b]	1:0.78		1:0.85	

[a]Four groups setting.

[b]Value when total of two-sided is assumed to be one.

The decision to use a one-sided or a two-sided test should be made carefully, as it has an impact on sample size calculation. Minimum sample size required for one-sided test is less, because it focuses on only tail of the probability distribution (Moye and Tita, 2002). The decision to use a one-sided or a two-sided test also has an impact on assessment of study results by regulatory authorities (Freedman, 2008). When you carry out initial pharmacological or toxicological tests with an unknown molecule, it would be appropriate to use a two-sided test. In subsequent tests, for confirming the findings of the initial tests, one-sided test may be used.

References

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

Bland, J.M. and Bland, D.G. (1994): Statistics notes: One and two sided tests of significance. BMJ, 309, 248.

CSCL (1986): Chemical Substance Control Law. http://www.safe.nite.go.jp/ kasinn/ genkou/kasinhou02.html.

Doll, H. and Carney, S. (2005): Statistical approaches to uncertainty: p values and confidence intervals unpacked. Evid. Based Med., 10, 133–134.

Drewitt, P.N., Butterworth, C.D., Springall, C.D. and Moorhouse, S.R. (1993): Plasma levels of aluminum after tea ingestion in healthy volunteers. Food Chem. Toxic., 31, 19–23.

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

Fisher, R.A. (1955): Statistical methods and scientific induction. J. Royal Stat. Soc. B., 17, 69–78.

Fisher, R.A. (1971): The Design of Experiments. 9th Edition. Hafner Press, New York, USA.

Freedman, L. (2008): An analysis of the controversy over classical one-sided tests. Clin. Trials, 5(6), 635–640.

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

Gelman, A. and Stern, H. (2006): The difference between "significant" and "not significant" is not itself statistically significant. Am. Stat., 60(4), 328–331.

Kobayashi, K. (1997): A comparison of one- and two-sided tests for judging significant differences in quantitative data obtained in toxicological bioassay of laboratory animals. J. Occup Health, 39, 29–35.

Kobayashi, K., Pillai, K.S., Sakuratani, Y., Abe, T., Kamata, E. and Hayashi, M. (2008): Evaluation of statistical tools used in short-term repeated dose administration toxicity studies with rodents. J. Toxicol. Sci., 33(1), 97–104.

Kornegay, E.T., Clawson, A.J., Smith, F.H. and Barrick, E.R. (1961): Influence of protein source on toxicity of gossypol in swine ration. J. Anim. Sci., 20, 597–602.

Le, C.T. (2009). Health and Numbers-A Problems-Based Introduction to Biostatistics, 3rd Edition, John Wiley & Sons Inc., New Jersey, USA.

Krantz, D.H. (1999): The null hypothesis testing controversy in psychology. J. Am. Stat. Assoc., 94, 1372–1381.

Madsen, B. (2011): Statistics for Non-Statisticians. Springer-Verlag, Berlin, Germany.

Moye, L.A. and Tita, A.T.N. (2002): Defending the rationale for the two-tailed test in clinical research. Circulation, 150, 3062–3065.

Nakamura, G. (1986): Practice, Statistical Analyses. Kaiumeisha, Tokyo, Japan.

Neyman, J. and Pearson, E.S. (1928): On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. Biometrika, 20A, 263–94.

Neyman, J. and Pearson, E.S. (1936): Sufficient statistics and uniformly most powerful tests of statistical hypotheses. Stat. Res. Mem., 1, 113–137.

OECD (2008): Organization for Economic Cooperation and Development. OECD Guidelines for the Testing of Chemicals. Repeated Dose 28-Day Oral Toxicity Study in Rodents., No. 407. OECD, Geneva, France.

Rosner, B. (2010): Fundamentals of Biostatistics. 7th Edition. Brooks/cole, Cengage Learning, Boston, USA.

Sakuma, A. (1977): Statistical Methods in Pharmacometrics I. 56, Tokyodaigaku Shupankai, Tokyo, Japan.

Shertzer, H.G. and Sainsbury, M. (1991): Chemoprotective and hepatic enzyme induction properties of indol and indenoindol antioxidants in rats, Food Chem. Toxic., 29, 391–400.

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

Spina, D. (2007): Statistics in Pharmacology. Br J. Pharmacol., 152(3), 291–293.

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

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

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

8

t-Tests

Student's *t*-Test—History

The history of statistical significance tests dates back 17th century. Perhaps the earliest statistical analysis published was by John Arbuthnot on London birth rates with regards to gender in 1710 (Hacking, 1965). One of the most popular significance tests is the Student's *t*-test, which has wide scientific applications (Papana and Ishwaran, 2006). The Student's *t*-test is a parametric test for comparing two groups. Readers may be interested to know why it is called as Student's *t*-test. 'Student' was the pseudonym of W.S. Gossett (1876–1937) (Box, 1987). He worked as a chemist at the Guinness brewery, Ireland. He chose this pseudonym because his company did not allow its scientists to publish confidential data (Raju, 2005). His company regarded use of statistics in quality control as a trade secret. In an article published in Biometrika, Gossett described a procedure to assess population means by using small samples under the pseudonym, "Student" (Student, 1908).

t-Test for One Group

The temperature of an animal room was set at 22°C. The temperature of the room measured everyday at 9.00 am for seven days is given in Table 8.1. The temperature measured was not the same as the temperature set (22°C) in any of these days. Let us examine whether the temperature measured during the seven days is statistically similar to the temperature set (22°C).

Table 8.1. Temperature of the animal room

Day	1	2	3	4	5	6	7
Temperature (°C)	22.3	22.6	22.4	22.4	22.6	22.5	22.4

N = 7; Mean = 22.46; SD = 0.1134; SE =0.0429

$$tcal = \frac{22.46 - 22.0}{0.0429} = 10.723$$

The *t*-distribution Table value (Table 8.2.) at 0.05 probability, for 6 (7–1) degrees of freedom is 2.447 (two-sided). Since calculated value (10.723) is greater than the Table value (2.447), it is considered that the temperature measured in the animal room during the seven days differed from the temperature set (22°C).

Table 8.2. *t*-distribution Table (Yoshimura, 1987)

DF\2α*	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF\α**	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	**2.447**	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408

DF, Degrees of freedom; *One-sided; **Two-sided

t-Test for Two Groups

The use of a repeated *t*-test for comparison of three or more groups might cause the error of the first kind (Type I error). Three kinds of *t*-tests are commonly used (Figure 8.1). Depending on the variance ratio (*F*) and the number of samples in the group, a *t*-test is selected.

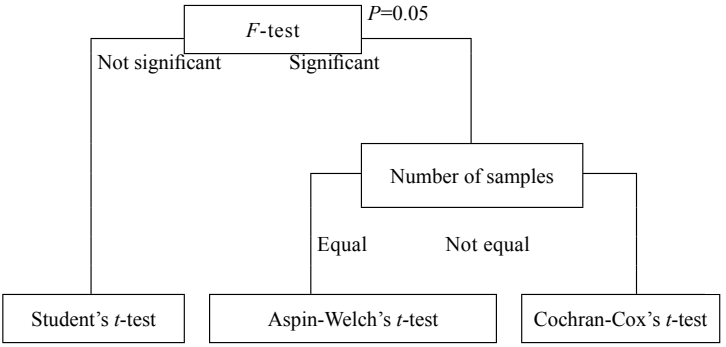

Figure 8.1. Selection of a *t*-test

F-value is the variance ratio. It is calculated by dividing the larger variance by the smaller variance. If the calculated *F*-value is smaller than the value given in *F*-distribution Table at 5% probability level, the two groups are regarded to have the same distribution and the data are analysed

using Student's *t*-test. On the contrary, if the calculated *F*-value is greater than the value given in *F*-distribution Table at 5% probability level, the two groups are regarded to have different distributions and the data are analysed using either Aspin-Welch's *t*-test (if the number of samples in the two groups is equal) or Cochran-Cox's *t*-test (if the number of samples in the two groups is not equal). Cochran-Cox's test has a low power to detect a significant difference. This may be the reason why Aspin-Welch's *t*-test is often used regardless of the number of samples in the two groups.

Student's t-test

The height of male and female students in a class room is given in Table 8.3. We would like to examine whether the male and female students have similar heights.

Table 8.3. Height (cm) of male and female students

Male (Group 1)	Female (Group 2)
170	160
168	154
170	162
169	160
179	151
162	159
172	148
169	159
169	150
179	162

Statistics

Estimates	Male (Group 1)	Female (Group 2)
N	10	10
Sum	1707	1565
Mean	170.7	156.5
SD	5.0783	5.2546
Variance	25.79	27.61
Sum of squares	232.10	248.50

Let us examine the distribution of the data of males and females by calculating *F*-value:

$$F_9^9 = \frac{27.6}{25.8} = 1.07$$

Note: F_9^9—The superscript and subscript to *F* indicate the degrees of freedom of the numerator and denominator, respectively.

Compare the calculated *F*-value with the *F*-distribution Table value (Table 8.4). *F*-distribution Table value is the value, where the degrees of freedom of numerator and denominator intercept.

(Note: The *F*-distribution is named after Sir Ronald A. Fisher (1890–1962), who is known to be the father of modern statistics (Kennedy, 2003). *F*-test is a ratio of the sample variances. However, *F*-test is not suitable for the data showing a non-normal distribution.).

Table 8.4. *F*-distribution values at 5% probability level (Yoshimura, 1987)

$N_1 \diagdown N_2$	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	30
7	5.59	4.73	4.34	4.12	3.97	3.86	3.78	3.72	3.67	3.63	3.57	3.52	3.49	3.46	3.44	3.37
8	5.31	4.45	4.06	3.83	3.68	3.58	3.50	3.43	3.38	3.34	3.28	3.23	3.20	3.17	3.15	3.07
9	5.11	4.25	3.96	3.63	3.48	3.37	3.29	3.22	**3.17**	3.13	3.07	3.02	2.98	2.96	2.93	2.86
10	4.96	4.10	3.70	3.47	3.32	3.21	3.13	3.07	3.02	2.97	2.91	2.86	2.82	2.79	2.77	2.69
11	4.84	3.98	3.58	3.35	3.20	3.09	3.01	2.94	2.89	2.85	2.78	2.73	2.70	2.67	2.64	2.57

N_1 = Degrees of freedom of numerator, N_2 = Degrees of freedom of denominator

The calculated *F* value (1.07) is less than the Table value (3.17). Hence, F_9^9 is not considered significant, indicating that the variances of both the groups having a similar distribution. Therefore, as given in Figure 8.1, the data can now be analysed using Student's *t*-test.

The *t* value is calculated using the equation,

$$tcal = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{SS_1 + SS_2}} \times \sqrt{\frac{N_1 \times N_2}{N_1 + N_2}\left(N_1 + N_2 - 2\right)}$$

Where,

$\overline{X}_1$ = Mean of Group 1; $\overline{X}_2$ = Mean of Group 2; SS_1 = Sum of squares of Group 1; SS_2= Sum of squares of Group 2; N_1 = Degrees of freedom of Group 1; N_2 = Degrees of freedom of Group 2.

$$tcal = \frac{\left|170.7 - 156.5\right|}{\sqrt{232.1 + 248.5}} \times \sqrt{\frac{10 \times 10}{10 + 10}(10 + 10 - 2)} = 6.145$$

Compare the calculated *t* value with the *t*-test critical value given in Table 8.5.

Table 8.5. *t*-test critical values (Yoshimura, 1987)

P= 2α	0.20	0.10	0.05	0.02	0.01	0.002	0.001
P= α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
DF							
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	**1.734**	2.101	2.552	2.878	2.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.850

Note: α=one-sided, 2α=two-sided.

The *t*-test critical value at 5% probability level for N_1+N_2-2 (10+10–2=18) degrees of freedom is 1.734. Since calculated *t*-value (6.145) is greater that the *t*-test critical value, it is considered that the height of male and female students is different.

Aspin-Welch's t-test

This test is used to compare the means of two groups having different distributions, but number of samples (observations) is the same.

A study was conducted in volunteers to find the effect of high fat content. Diet containing high fat content was given to 10 individuals (Group 1). Concurrently, normal diet was given to another 10 individuals for comparison (Group 2). At the end of the 7 days treatment, alanine aminotransferase (ALT) activity was measured in the individuals of both the Groups. The ALT determined in the individuals is given in Table 8.6.

Table 8.6. Alanine aminotransferase activity (IU/l) of individuals

Diet containing high fat content (Group1)	Normal diet (Group 2)
42	30
60	34
26	35
48	32
56	36
31	41
30	42
80	28
79	71
93	35

Statistics

Estimates	Diet containing high fat content (Group 1)	Normal diet (Group 2)
N	10	10
Sum	545	384
Mean	55	38
SD	23.4011	12.2493
Variance (Sx^2)	548	150

F-ratio =

$$F_9^9 = \frac{548}{150} = 3.65$$

Compare the calculated *F*-value with the Table value (Table 8.4). The derived *F* value (3.65) is greater than the Table value (3.17). Hence, F_9^9 is considered significant, indicating that the variances of both the groups are distributed differently. According to Figure 8.1, Aspin-Welch's *t*-test is the appropriate statistical tool for the analysis of this data. The *t* is calculated using the following formula:

$$tcal = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{Sx_1}{N_1} + \frac{Sx_2}{N_2}}}$$

Where,

$\overline{X}_1$ = Mean of Group 1; $\overline{X}_2$ = Mean of Group 2; Sx_1= Variance of Group 1; Sx_2= Variance of Group 2; N_1= Degrees of freedom of Group 1; N_2 = Degrees of freedom of Group 2.

$$tcal = \frac{\left|55 - 38\right|}{\sqrt{\frac{548 + 150}{10}}} = 2.03$$

Unlike Student's *t*-test, where the degrees of freedom is N_1+N_2-2, degrees of freedom needs to be calculated for Aspin-Welch's *t*-test. The degrees of freedom for Aspin-Welch's *t*-test is calculated as given below:

$$N = \frac{1}{\frac{C^2}{N_1 - 1} + \frac{(1-C)^2}{N_2 - 1}}$$

Where,

$$C = \frac{\frac{Sx_1^{\,2}}{N_1}}{\frac{Sx_1^{\,2}}{N_1} + \frac{Sx_2^{\,2}}{N_2}}$$

$$C = \frac{54.8}{54.8 + 15.0} = 0.79$$

$$N = \frac{1}{\frac{0.79^2}{9} + \frac{(1-0.79)^2}{9}} = 13.5$$

Compare the derived *t* value with the *t*-test critical value given in Table 8.7 at 5% probability level for fourteen degrees of freedom (14 degrees of freedom is obtained by rounding up the calculated *N*, 13.5). Since the calculated *t*-value, 2.03 is greater than the *t*-test critical value given in the Table 8.7 (1.761), it can be stated that there is a difference in ALT between the high fat diet-treated and normal diet treated-individuals.

Table 8.7. *t*-test critical values (Yoshimura, 1987)

2α	0.20	0.10	0.05	0.02	0.01	0.002	0.001
α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
DF=14	1.345	**1.761**	2.145	2.624	2.977	3.787	4.140

α=one-sided, 2α=two-sided.

Cochran-Cox's t-test

Cochran-Cox's *t*-test is used to compare the means of two samples having different distributions and different number of observations. We shall modify the data given in Table 8.6 and analyse it using Cochran-Cox's *t*-test. The values modified are given in Table 8.8. We have not made any change in the ALT values of Group 1. But, the values of Group 2 are changed and only nine individuals of this group are used for the analysis.

Table 8.8. Alanine aminotransferase activity (IU/l) of individuals

Diet containing high fat content (Group1)	Normal diet (Group 2)
42	57
60	45
26	55
48	46
56	26
31	33
30	41
80	35
79	43
93	-

Statistics

Estimates	Diet containing high fat content (Group 1)	Normal diet (Group 2)
N	10	9
Sum	545	381
Mean	55	42
SD	23.4011	10.0374
Variance (Sx^2)	548	101

F-ratio =

$$F_8^9 = \frac{548}{101} = 5.43$$

Compare the derived *F*-value with the Table value (Table 8.4). The calculated *F*-value (5.43) is greater than the Table value (3.38). Hence, F_8^9 is considered significant, indicating that the variances of both the groups are distributed differently. According to Figure 8.1, Cochran-Cox's *t*-test is the appropriate statistical tool for the analysis of the data given in Table 8.8.

In Cochran-Cox's *t*-test, we need to calculate two *t* values (*t* calculated and *t*' calculated).

$$tcal = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{\frac{Sx_1^{\,2}}{N_1} + \frac{Sx_2^{\,2}}{N_2}}}$$

$$tcal = \frac{\left| 55 - 42 \right|}{\sqrt{\frac{548}{10} + \frac{101}{9}}} = 1.21$$

$$t'cal = \frac{\frac{t_1 \times Sx^2{}_1}{N_1} + \frac{t_2 \times Sx^2{}_2}{N_2}}{\frac{Sx^2{}_1}{N_1} + \frac{Sx^2{}_2}{N_2}}$$

$$t'cal = \frac{\frac{1.833 \times 548}{10} + \frac{1.860 \times 101}{9}}{\frac{548}{10} + \frac{101}{9}} = 1.83$$

Since the *t* calculated (*tcal* = 1.21) is smaller than the *t*' calculated (t'cal=1.83), it is concluded from the analysis that there is no significant difference in ALT between the high fat diet-treated and normal diet treated-individuals.

Paired *t*-Test

Let us assume one needs to test an antidiabetic drug in diabetic rats. One way to do is to measure the blood sugar before and after treatment with the drug and calculate the respective mean values, and compare the mean values using an appropriate *t*-test (select the appropriate *t*-test as per Figure 8.1). Another way is to analyse the data using paired *t*-test.

Blood sugar values of individual rats before and after the drug treatment is given Table 8.9.

Table 8.9. Blood sugar values (mg/dl) of individual rats

Rat Number	1	2	3	4	5	Mean	Variance	SD	SE
Before treatment	274	287	277	259	237	-	-	-	-
After treatment	165	142	215	209	198	-	-	-	-
Difference between before and after treatments	109	145	62	50	39	81	1992	44.6	19.9

$$tcal = \frac{Mean}{SE}$$

$$tcal = \frac{81}{19.9} = 4.07$$

Compare the calculated *t*-value with the *t*-test critical value given in Table 8.10 at 5% probability level for N-1 degrees of freedom. N is number of pairs, hence N–1=4. Since the calculated *t* value, 4.07 is greater than the *t*-test critical value given in the Table 8.10 (2.132), it can be stated that treatment with the drug significantly decreased the blood sugar in rats.

Table 8.10. *t*-test critical values (Yoshimura, 1987)

2α	0.20	0.10	0.05	0.02	0.01	0.002	0.001
α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
DF=4	1.533	**2.132**	2.776	3.747	4.604	7.173	8.610

α=one-sided, 2α=two-sided.

A Note of Caution

It is well known that with Student's two-independent-sample *t*-test, the actual level of significance can be well above or below the nominal level, confidence intervals can have inaccurate probability coverage, and power can be low relative to other methods.

In Student's two-independent-sample *t*-test, the variance heterogeneity can distort rates of Type I error (Kaselman *et al*., 2004). Therefore, when the variance of the two populations is different, Student's *t*-test is not suitable (Ruxton, 2006). When the number of the groups is more than two, multiple comparison with Student's *t*-test can cause Type I error.

References

Box, J.F. (1987): Guinness, Gosset, Fisher, and Small Samples. Stat. Sci., 2(1), 45–52.

Hacking, I. (1965): Logic of Statistical Inference. Cambridge University Press, New York.

Kaselman, H.J., Othman, A.R., Wilcox, R.R. and Fradette, K. (2004): The new and improved two-sample *t*-test. Psych. Sci., 15(1), 47–51.

Kennedy, P. (2003): A Guide to Econometrics. 5th Edition. MIT Press, UK.

Papana, A. and Ishwaran, H. (2006): CART variance stabilization and regularization for high-throughput genomic data. Bioinformatics, 22(18), 2254–2261.

Raju, T.N.K. (2005): William Sealy Gosset and William A. Silverman: Two "Students" of Science. Pediatrics, 116(3), 732–735.

Ruxton, G. (2006): The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney U test. Behavioral Ecol., 17(4), 688–690.

Student (1908): The Probable Error of a Mean. Biometrika, 6(1), 1–25.

Yoshimura, I. (1987): Statistical Analysis of Toxicity and Drug Efficacy Data. Scientist Inc., Tokyo, Japan.

9
Correlation Analysis

Correlation and Association

Correlations are relationships between two or more variables or sets of variables (Cohen and Cohen, 1983). In statistics there is a distinction between an association and a correlation, though these terms are often used interchangably. Two variables are associated if one of them provides information about the likely value of the other. If the association between two variables is linear, there is a correlation. Therefore, strictly speaking, "non-linear correlation" is an incorrect terminology, a better term is "non-linear association".

Statisticians' definition of correlation is that it is 'a parameter of the bivariate normal distribution'. The variables are random variables when one variable is not depended on the other. In statistics, correlation is referred to as coefficient of correlation (Paler-Calmorin and Calmorin-Piedad, 2008). The correlation coefficient is denoted by the letter r which might have originated from the letter, r of the word, relation. A number between –1 and +1 is used to 'quantify' the correlation of the variables (Glantz, 2005). The closer the absolute value of r to 1 or –1, the higher the degree of correlation. When one variable increases as the other variable increases, it is called a 'positive correlation', and when one variable decreases as the other variable increases, it is called a 'negative correlation'. When $r = -1$, there is a 100% negative correlation, when $r = +1$, there is a 100% positive correlation and when $r = 0$, there is a 100% no correlation. But, if $r = 0.5$, it does not mean that there is a 50% correlation. Therefore, r does not indicate the percent of correlation (Gurumani, 2005).

Pearson's Product Moment Correlation Coefficient

A commonly used measure of correlation is Pearson's product moment correlation coefficient. Correlation coefficient is a standardised covariance (Field, 2009; Berkman and Reise, 2011). Covariance is a measure of joint variances of two variables; the deviation of each variable is computed and multiplied. Since there are two variables, there are two standard deviations. Multiply these standard deviations and divide joint variances by it.

Standardised covariance, $r = \dfrac{1}{n-1}\dfrac{\sum(x-\bar{x})(y-\bar{y})}{(s_x)(s_y)}$, where

s_x and s_y are the standard deviations of variable x and variable y, respectively. Above equation can be rewritten as follows:

$$r = \frac{1}{n-1}\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\left(\dfrac{\sum(x-\bar{x})^2}{n-1}\right)\left(\dfrac{\sum(y-\bar{y})^2}{n-1}\right)}}$$

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

The above equation was formulated by Karl Perason, hence called Pearson's correlation coefficient.

Let us compute correlation coefficient, r for the variables x and y given in Table 9.1.

Table 9.1. Calculation of correlation coefficient

x	y	x^2	y^2	xy
1	93	1	8649	93
2	87	4	7569	174
3	76	9	5776	228
4	70	16	4900	280
5	62	25	3844	310
6	45	36	2025	270
7	40	49	1600	280
8	32	64	1024	256
9	25	81	625	225
10	10	100	100	100
Σx= 55	Σy=540	Σx^2=385	Σy^2=36112	Σxy= 2216

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

$$\sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \times \sum y}{n} = 2216 - \frac{55 \times 540}{10} = -754$$

$$\sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 385 - \frac{(55)^2}{10} = 82.5$$

$$\sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n} = 36112 - \frac{(540)^2}{10} = 6952$$

$$r = \frac{-754}{\sqrt{82.5 \times 6952}} = \frac{-754}{757.32} = -0.996$$

Significance of *r*

When the sample size is not too large, the significance of a correlation coefficient can be tested using a *t*-test:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.996\sqrt{10-2}}{\sqrt{1-(-0.996)^2}} = \frac{-2.8171}{0.0894} = -31.51$$

Above is Students *t*-test with n–2 degrees of freedom.

Alternatively, significance of a correlation coefficient can be tested as given below, which involves no calculation procedure:

Compare the correlation coefficient, *r* with the value given in correlation coefficient table (Table 9.2) for eight degrees of freedom. The computed correlation coefficient, *r* (–0.996) is less than the correlation coefficient Table value (–0.765) at 1% probability level. Hence the correlation coefficient is considered to be significant. The negative sign of the correlation coefficient indicates that the variables *x* and *y* are negatively correlated. Had the *r* been 0.996 (positively correlated), we would have compared it with 0.765 (without a minus sign). In this case, in order to consider the *r* to be significant, it has to be greater than 0.765.

Table 9.2. Correlation coefficient Table (Shibata, 1970)

DF	5%	1%	DF	5%	1%	DF	5%	1%
1	0.997	1.000	17	0.456	0.575	45	0.288	0.372
2	0.950	0.990	18	0.444	0.561	50	0.273	0.354
3	0.878	0.959	19	0.433	0.549	60	0.250	0.325
4	0.811	0.917	20	0.423	0.537	70	0.232	0.302
5	0.754	0.874	21	0.413	0.526	80	0.217	0.283
6	0.707	0.834	22	0.404	0.515	90	0.205	0.267
7	0.666	0.798	23	0.396	0.505	100	0.195	0.254
8	0.632	0.765	24	0.388	0.496	125	0.174	0.228
9	0.602	0.735	25	0.381	0.487	150	0.159	0.208
10	0.576	0.708	26	0.374	0.478	200	0.138	0.181
11	0.553	0.684	27	0.367	0.470	300	0.113	0.148
12	0.532	0.661	28	0.361	0.463	400	0.098	0.128
13	0.514	0.641	29	0.355	0.456	500	0.088	0.115
14	0.497	0.623	30	0.349	0.449	1000	0.062	0.081
15	0.482	0.606	35	0.325	0.418			
16	0.468	0.590	40	0.304	0.393			

Confidence Interval of Correlation Coefficient

A confidence interval of correlation coefficient, r can be determined by using a transformation of r to a quantity z, which has an approximately normal distribution. This transformed z is calculated using the equation:

$$Z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right]$$

For the example given in Table 9.1, $r = 0.996$. The transformed Z is:

$$Z = \frac{1}{2}\ln\left[\frac{1+(-0.996)}{1-(-0.996)}\right] = \frac{1}{2}\ln\left[\frac{0.004}{1.996}\right] = -3.1063$$

Now, we need to calculate an estimate called Error of Estimate:

Error of Estimate = $1/\sqrt{n-3} = 1/\sqrt{10-3} = 0.3780$

Using the Error of Estimate we can calculate Z_1 and Z_2 with 95% confidence level:

$$Z_1 = -3.1063 - (1.96\text{x}0.3780) = -3.8472$$
$$Z_2 = -3.1063 + (1.96\text{x}0.3780) = -2.3654$$

Next step is to transform the Z_1 and Z_2 back to original scale. Confidence interval of *r* is:

$$\frac{e^{2z_1}-1}{e^{2z_1}+1} \text{ to } \frac{e^{2z_2}-1}{e^{2z_2}+1} =$$

$$\frac{e^{2\times-3.8472}-1}{e^{2\times-3.8472}+1} \text{ to } \frac{e^{2\times-2.3654}-1}{e^{2\times-2.3654}+1} =$$

$$\frac{0.9995}{1.0005} \text{to} \frac{0.9912}{1.0088}$$

Confidence interval of *r* is calculated as 0.983–0.999.

Coefficient of Determination

The coefficient of determination is the square of *r* (R^2; coefficient of determination is usually denoted by the capital letter R^2), which expresses the strength of the relationship between the *x* and *y* variables (McDonald, 2009). This is reviewed in Chapter 10, in greater detail.

Rank Correlation

When the variables are not linearly associated, Pearson's product moment correlation analysis does not work well. In this situation the association is transformed into linear by ranking the variables. Rank correlation is a nonparametric alternative to the linear correlation coefficient (Ruby, 2008). There are several rank correlation analyses available, amongst them, Spearman's rank correlation is more commonly employed (Hassard, 1991).

Spearman's Rank Correlation

As stated, in Spearman correlation analysis, the variables are converted to ranks. Spearman rank correlation analysis is also used, when there are two measurement variables and one "hidden" nominal variable. If you measure body weight and body surface area of rats with the rat identification number, the identification number of the rat is the nominal variable. The major advantages of Spearman's rank correlation are that it is not affected by the distribution of the population and it can be applied to small samples (Gauthier, 2001).

Canonical Correlation

Canonical correlation analysis developed by Hotelling (1936), is the study of the linear relationships between two sets of variables, and is considered as a fundamental statistical tool (Bulut *et al*., 2010). It is the multivariate extension of correlation analysis and it measures the interrelationships among sets of multiple dependent variables and multiple independent variables (Green, 1978). Canonical correlation simultaneously predicts multiple dependent variables from multiple independent variables. It is a very useful tool in pharmacology and toxicology (Kelder, 1982; Hu *et al*., 2003; Tanaka, 2010), where interrelationships between several dependent and independent variables need to be assessed.

An elaborative discussion on canonical correlation is beyond the scope of this book. Several books are available that cover the subject in depth (Green and Carroll, 1978; Das and Sen, 1994).

Misuse of Correlation Analysis

There are a several situations in which the correlation coefficient can be misinterpreted. Fifteen errors related to correlation and regression were identified in articles published in three leading medical journals in the year, 1997 (Porter, 1999). Perhaps the most important error committed in these articles was, not presenting confidence intervals of correlation coefficient (this error could be seen in many of the scientific articles, even today). Another error in interpreting the correlation coefficient is, the failure to consider that there may be a third variable related to both of the variables being investigated, which is responsible for the apparent correlation. Often the correlation coefficient fails to detect the existence of a nonlinear association between two variables (Bewick *et al*., 2003).

A high correlation coefficient (for example, $r = > 0.997$) is not always a useful indicator of linearity in method validation; other statistical tests like Lack-of-fit and Mendel's fitting test may be used for evaluating the linearity (Loco *et al.*, 2002).

A correlation coefficient will have limited use as a stand-alone quantity without reference to the number of observations, the pattern of the data and the slope of the regression line (Sonnergaard, 2006). It is recommended to plot the variables and understand the pattern of the data before interpreting the correlation analysis.

References

Berkman, E.T and Reise, S.P. (2011): A Conceptual Guide to Statistics Using SPSS. SAGE Publications Inc., California, USA.

Bewick, V., Cheek, L. and Ball, J. (2003): Statistics review 7: Correlation and regression. Critical Care, 7, 451–459.

Bulut, M., Gultepe, N., Mendes, M., Guroy, D. and Palaz, M. (2010): According to Canonical correlation, the evaluation of bluefish (*Pomatomus saltatrix*) blood chemistry. J. Animal Vet. Adv., 9(4), 666–670.

Cohen, J. and Cohen, P. (1983): Multiple Regression/Correlation for the Behavioral Sciences. 2nd Edition. Erlbaum Associates, Hillsdale, New Jersey, USA.

Das, S. and Sen, P.K. (1994): Restricted canonical correlations. Linear algebra and its applications, 210, 29–47.

Field, A. (2009): Discovering Statistics Using SPSS. 3rd Edition. SAGE Publications Ltd., London, UK.

Gauthier, T.D. (2001): Detecting trends using Spearman's rank correlation coefficient. Exp. Forensics, 2, 359–362.

Glantz, S.A. (2005): Primer of Biostatistics. Mc Graw-Hill Companies Inc., USA.

Green, P.E. (1978): Analyzing Multivariate Data. Holt, Rinehart & Winston, Illinois, USA.

Green, P.E., and Carroll, J.D. (1978): Mathematical Tools for Applied Multivariate Analysis. Academic Press, New York, USA.

Gurumani, N. (2005): An Introduction to Biostatistics. 2nd Edition. MJP Publishers, Chennai, India.

Hassard, T.H. (1991): Understanding Biostatistics. Mosby-Year Book Inc., St. Loius, Missouri, USA.

Hotelling, H. (1936): Relations between two sets of variates. Biometrika, 28, 321–377.

Hu, Q.N., Liang, Y.Z., Peng, X.L., Hong, Y. and Zhu, L. (2003): Application of orthogonal block variables and canonical correlation analysis in modeling pharmacological activity of alkaloids from plant medicines. J. Data Sci., 1, 405–423.

Kelder, J. (1982): Prediction of the Bobon clinical profile of neuroleptics from animal pharmacological data. Psychopharmacol., 77(2), 140–145.

Loco. J.V., Elskens, M., Croux, C. and Beernaert, H. (2002): Linearity of calibration curves: use and misuse of the correlation coefficient. Accred. Qual. Assur., 7, 281–285.

McDonald, J.H. (2009): Handbook of Biological Statistics. 2nd Edition. Sparky House Publishing Baltimore, Maryland, USA.

Paler-Calmorin, L. and Calmorin-Piedad, M.L.P. (2008): Nursing Biostatistics with Computer. Rex Printing Co. Inc., Florentino St., Quezon City, Philippines.

Porter, A.M.W. (1999): Misuse of correlation and regression in three medical journals. J. Royal Soc. Med., 92, 123–128.

Ruby, J. (2008): Elementary Statistics. Thompson Brooks. Cole, Belmont, USA.

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

Sonnergaard, J.M. (2006): On the misinterpretation of the correlation coefficient in pharmaceutical sciences. Int. J. Pharm., 321(1-2), 12–17.

Tanaka, T. (2010): Biological factors influencing exploratory behavior in laboratory mice, *Mus musculus*. Mammal Study, 35(2), 139–144.