Biometrical Letters
Vol. 47 (2010), No. 2, 107-118

The impact of distributional shape on the power of randomization tests for two independent groups: a simulation study using small balanced samples

Fernando Branco[1], T. Azinheira Oliveira[2], Amílcar Oliveira[2]

[1]ULHT – Universidade Lusófona de Humanidades e Tecnologias, Avenida do Campo Grande, 376, 1749-024 Lisboa, Portugal, branco.fernando@netcabo.pt
[2]Departamento de Ciências e Tecnologia, Universidade Aberta and CEAUL-Center of Statistics and Applications, University of Lisbon, Rua Fernão Lopes, 9, 2º Dir. 1000-132 Lisboa, Portugal, aoliveira@univ-ab.pt, toliveir@univ-ab.pt

SUMMARY

The importance of randomization tests is very well known in experimental research, particularly in biometry. The aim of the present research is to evaluate the impact of distributional shape on the power of the randomization test for difference between the means of two independent groups (with n_1=n_2=16). To manipulate shape in terms of asymmetry and kurtosis, we used *g*-and-*h* distributions. We evaluated the power of the randomization test, and also the power of the Student-t test, as a comparison standard, with data simulated from 12 *g*-and-*h* distributions for seven values of effect size. For each condition, we generated 20 000 samples, and for each one the power of randomization tests was estimated using 1000 permutations. We set the value of Type I error probability at 0.05. The results show gains in power for both tests with increasing skewness and/or kurtosis, with a slight advantage for the randomization tests over the Student-t test.

Key words: Randomization tests, statistical power, *g*-and-*h* distributions.

1. Introduction

Randomization tests are significance tests based on the random assignment of experimental units to treatments in order to test hypotheses about treatment effects. Their validity is based on a random-assignment model, whereas the validity of classical tests, e.g. Student-t test, is based on a random-sampling model. Given the widespread use of non-random samples in experimental research, namely in the behavioural and social sciences, as well in biometry, randomization tests are not only a way of avoiding distributional assumptions, but they allow us to drop the most implausible assumption of typical experimental research: random sampling from a specified population.

The randomization idea stems from Fisher (1935), but it was Pitman (1937a, 1937b, 1938) who first presents a type of significance tests, “which may be applied to samples from any population”, based on random assignment alone. These tests were further developed by Kempthorne (1952, 1955), Hinkelmann and Kempthorne (1994), Edgington (1964, 1966, 1969a, 1969b, 1995) and recently Edgington and Onghena (2007).

With the advent of computers, interest in these tests has shifted from theoretical considerations – the validation of classical methods – to practical applicability. Even with moderate sample sizes, there may be so many data permutations that it would not be feasible to generate them all. Contributions from Dwass (1957) and Chung and Fraser (1958) provided the possibility of using only a subset of all possible data permutations, thus rendering this computer-intensive technique practical. Some research applications can be found in Manly (1997) and Edgington and Onghena (2007).

When analysing data from an experiment, where the experimental units are randomly assigned to treatments, if we use a test statistic, like t or F, the distinction between a randomization and a classical test is the way of calculating the significance. In the case of a randomization test, the significance is calculated by a procedure in which the data are repeatedly permuted, and the significance thus obtained is exact, conditional on the data. With this procedure, the researcher can calculate the significance of any statistical test, even of one whose sampling distribution has not yet been analytically derived. Thus to analyse the data, the researcher is free to choose the test that is most likely to be sensitive to the type of treatment effect that is expected.

When the assumptions for using classical tests are met, the classical and randomization tests are equivalent in terms of statistical power.

The concept of statistical power, the probability of rejecting a false null hypothesis, dates back to the work of Neyman and Pearson. In a series of papers

(Neyman, Pearson 1928a, 1928b, 1933), these authors stated that the choice of test must take into consideration not only the hypothesis, but also the alternatives against which it is being tested, introducing the distinction between errors of the first and second kind and proposing the likelihood-ratio criterion as a general method of test construction.

The Neyman-Pearson theory of statistical inference is mainstream in mathematical statistics (see e.g. Lehmann 1986; Mood, Graybill, Boes 1974) and also in the social and behavioural sciences (see, e.g., Hays 1994; Marascuilo, Serlin 1988; Winer, Brown, Michels 1991). However, in these sciences power analyses were neglected, and we must credit Cohen (1962) for introducing the notion of statistical power to behavioural scientists. The handbook on power analysis, by Cohen (1969), updated in 1988, allowed researchers planning an experiment to determine the sample size needed to detect a given population effect size, taking into account the two types of errors.

As stated above, the classical and randomization tests are equivalent in terms of power, when the assumptions for using classical tests are met. However, in empirical research, the data seldom are well behaved, frequently presenting a non-normal shape.

To study distributional shape, Tukey (1977) introduced the *g*-and-*h* distributions. The investigation of their properties was extended by Hoaglin (1983, 1985), Martinez and Iglewicz (1984), Badrinath and Chatterjee (1988 e 1991), Mills (1995), Dutta and Babbel (2002), Field and Genton (2006), and Headrick, Kowalchuk and Sheng (2008).

Tukey presented this family of distributions by the following transformation of a standard normal random variable Z:

$$Y_{g,h}(Z) = \left(\frac{e^{gZ} - 1}{g} \right) e^{hZ^2/2},$$

where the parameters g and h represent the degrees of skewness and kurtosis respectively.

When *h*=0, the *g*-and-*h* distribution reduces to the first term of the right-hand side of the above expression and is known as the *g* distribution. When g=0, the *g*-and-*h* distribution reduces to the second term of the right-hand side of the above expression, multiplied by Z, and is known as the *h* distribution.

To see graphically how the *g*-and-*h* distribution takes different shapes depending on values of the parameters *g* and *h*, refer to Figure 1 in the Method section, where we plot the graphs of the density functions for several combinations of *g* and *h*.

The *g*-and-*h* family of non-normal distributions are often used in Monte Carlo or statistical modelling studies. Since these distributions are merely a transformation of the standard normal distribution, they provide useful probability functions for the generation of random numbers in the course of a Monte Carlo simulation.

The aim of the present research is to evaluate the impact of distributional shape on the power of the randomization test for the difference between the means of two independent groups (with $n_1 = n_2 = 16$). To manipulate shape in terms of asymmetry and kurtosis, we simulate data from *g*-and-*h* distributions. As a comparison standard, we also evaluate, for the same distributions, the power of the Student-t test.

2. Method

We evaluated the power of the randomization test, and also the power of the Student-t test, for the difference between the means of two independent groups, with $n_1 = n_2 = 16$, with data simulated from 12 *g*-and-*h* distributions and seven effect sizes (-0.8, -0.5, -0.2, 0, 0.2, 0.5 and 0.8).

We chose these values for the effect size (the difference between the population means in population standard deviation units), using Cohen (1988) conventional figures for small, medium and large effect sizes in the behavioural sciences. We simulated data from 12 *g*-and-*h* distributions, with *g* values of 0, 0.4 and 0.8, *h* values of 0, 0.1, 0.2 and 0.3 and with the combinations of those values. In Table 1 we list these *g*-and-*h* distributions, presenting their means and standard deviations.

Table 1. Means and standard deviations for the 12 *g*-and-*h* simulated distributions

Distribution		*g*	*h*	μ	σ
1	Gaussian	0	0	0	1
2	skewed	0.4	0	0.21	1.128
3		0.8	0	0.47	1.630
4	kurtotic	0	0.1	0	1.182
5		0	0.2	0	1.467
6		0	0.3	0	1.988
7	skewed and kurtotic	0.4	0.1	0.24	1.381
8		0.4	0.2	0.29	1.816
9		0.4	0.3	0.36	2.758
10		0.8	0.1	0.56	2.207
11		0.8	0.2	0.69	3.420
12		0.8	0.3	0.87	7.165

In Figure 1 we present the graphs of the density functions for these 12 distributions.

To simulate data from these distributions, we have written programs in R (R Development Core Team 2008), version 2.7.1. For each distribution we generated 20 000 samples, and for each one and each effect size we estimated the power of the Student-t and randomization tests. For the latter test we used as a test statistic the sum of the values of the first group, which is an equivalent test statistic to the difference between means. To estimate the significance of the randomization test for each sample we used 999 permutations, plus the one observed. For values for the number of samples and the number of permutations, we followed the guidelines in Westfall and Young (1993).

We set the value of Type I error probability at 0.05; the power of a test was obtained as the proportion of samples in which the significance was equal to or smaller than that value. As the power of the randomization test was estimated using 1000 of the 601 080 390 possible permutations, we computed a 99%

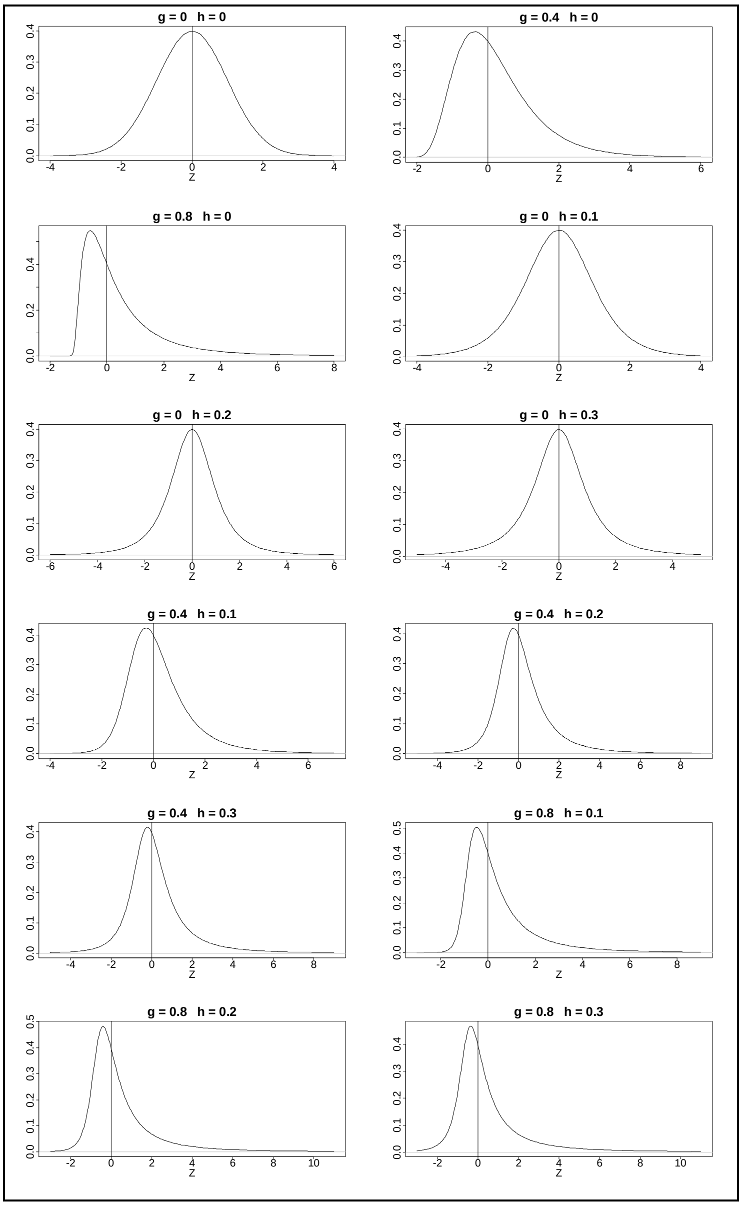

Figure 1. Graphs of the density functions for the 12 simulated *g*-and-*h* distributions

confidence interval for each power value. Whenever we compared the power values of this test with those of the Student test, we used the information provided by these confidence intervals.

3. Results

In Table 2 we present the results for the power of the two tests. To begin with we will analyse the results, comparing the randomization and Student-t tests. Then, for the randomization tests, we will describe the differences in power between skewed and/or kurtotic distributions (conditions 2 to 12) and the Gaussian (condition 1).

Randomization test vs. Student-t test

Comparing these two tests in terms of power, we can say that they are close to identical, with a small advantage for the randomization test. This advantage is only statistically significant in respect of the last two distributions ($g = 0.8$, $h = 0.2$ and $g = 0.8$, $h = 0.3$), for two-tailed tests and small and medium effect sizes: the gains in power range from 0.03 to 0.06.

As regards Type I errors, the two tests adequately controlled this type of error: For an effect size of zero, no value exceeded the nominal level for more than 0.004. But for some conditions, with increasing skewness or/and kurtosis, the Student-t test was unduly conservative, presenting values of power below the nominal level.

Randomization tests: Gaussian vs. *g*-and-*h* distributions

For the randomization tests we present, in Table 3, descriptive statistics (minimum, maximum and mean) for the differences in power between the *g*-and-*h* distributions and the Gaussian:

Analysing Table 3, we can see that all comparisons show gains in power for lower-, upper- and two-tailed tests. The increase in the gains is connected with increases in skewness, in kurtosis and in skewness combined with kurtosis.

Table 2. Power of the Student-t and randomization tests for two independent samples (n_1=n_2=16)

		Student-t test			Randomization test		
	Effect size	Lower-tailed	Upper-tailed	Two-tailed	Lower-tailed	Upper-tailed	Two-tailed
g = 0, h = 0	-0.8	0.714	0.000	0.592	0.713	0.000	0.590
(Gaussian)	-0.5	0.394	0.001	0.277	0.391	0.001	0.275
	-0.2	0.137	0.015	0.087	0.136	0.015	0.087
	0	0.049	0.050	0.050	0.049	0.051	0.050
	0.2	0.013	0.137	0.083	0.013	0.137	0.083
	0.5	0.001	0.395	0.277	0.002	0.394	0.277
	0.8	0.000	0.712	0.590	0.000	0.711	0.587
g = 0.4, h = 0	-0.8	0.725	0.000	0.608	0.724	0.000	0.611
	-0.5	0.414	0.001	0.296	0.414	0.001	0.300
	-0.2	0.144	0.012	0.088	0.144	0.012	0.091
	0	0.050	0.049	0.048	0.050	0.050	0.049
	0.2	0.013	0.144	0.087	0.013	0.143	0.090
	0.5	0.001	0.412	0.296	0.001	0.412	0.300
	0.8	0.000	0.724	0.611	0.000	0.725	0.613
g = 0.8, h = 0	-0.8	0.770	0.000	0.681	0.773	0.000	0.693
	-0.5	0.489	0.000	0.374	0.494	0.000	0.393
	-0.2	0.168	0.008	0.100	0.174	0.009	0.113
	0	0.047	0.047	0.041	0.051	0.049	0.051
	0.2	0.008	0.163	0.098	0.009	0.170	0.111
	0.5	0.000	0.488	0.371	0.000	0.495	0.391
	0.8	0.000	0.769	0.679	0.000	0.772	0.692
g = 0, h = 0.1	-0.8	0.725	0.000	0.611	0.724	0.000	0.611
	-0.5	0.414	0.001	0.296	0.414	0.001	0.300
	-0.2	0.139	0.013	0.086	0.140	0.013	0.087
	0	0.052	0.049	0.048	0.053	0.049	0.049
	0.2	0.013	0.140	0.088	0.012	0.141	0.089
	0.5	0.001	0.422	0.298	0.001	0.421	0.303
	0.8	0.000	0.722	0.603	0.000	0.720	0.606
g = 0, h = 0.2	-0.8	0.751	0.000	0.650	0.751	0.000	0.656
	-0.5	0.450	0.001	0.331	0.451	0.001	0.341
	-0.2	0.149	0.011	0.090	0.152	0.011	0.095
	0	0.052	0.048	0.045	0.054	0.049	0.049
	0.2	0.010	0.150	0.092	0.010	0.153	0.098
	0.5	0.001	0.460	0.335	0.001	0.461	0.344
	0.8	0.000	0.748	0.642	0.000	0.750	0.650
g = 0, h =0.3	-0.8	0.806	0.000	0.728	0.810	0.000	0.742
	-0.5	0.529	0.000	0.412	0.537	0.000	0.433
	-0.2	0.172	0.006	0.103	0.179	0.007	0.116
	0	0.050	0.046	0.042	0.053	0.049	0.050
	0.2	0.007	0.174	0.106	0.007	0.179	0.119
	0.5	0.000	0.539	0.420	0.000	0.546	0.442
	0.8	0.000	0.804	0.725	0.000	0.807	0.740

g = 0.4, h = 0.1	-0.8	0.752	0.000	0.652	0.753	0.000	0.660
	-0.5	0.451	0.001	0.333	0.454	0.001	0.344
	-0.2	0.157	0.011	0.092	0.159	0.012	0.099
	0	0.049	0.047	0.045	0.051	0.048	0.051
	0.2	0.010	0.151	0.089	0.010	0.153	0.095
	0.5	0.001	0.456	0.333	0.001	0.458	0.343
	0.8	0.000	0.754	0.651	0.000	0.754	0.660
g = 0.4, h = 0.2	-0.8	0.802	0.000	0.720	0.807	0.000	0.734
	-0.5	0.519	0.000	0.402	0.527	0.000	0.423
	-0.2	0.176	0.007	0.106	0.182	0.008	0.117
	0	0.048	0.045	0.041	0.051	0.048	0.050
	0.2	0.008	0.173	0.100	0.008	0.179	0.112
	0.5	0.000	0.523	0.404	0.000	0.531	0.423
	0.8	0.000	0.800	0.723	0.000	0.804	0.737
g = 0.4, h = 0.3	-0.8	0.878	0.000	0.829	0.885	0.000	0.847
	-0.5	0.664	0.000	0.563	0.675	0.000	0.592
	-0.2	0.234	0.003	0.149	0.250	0.004	0.172
	0	0.046	0.044	0.038	0.051	0.049	0.051
	0.2	0.003	0.235	0.146	0.004	0.247	0.169
	0.5	0.000	0.670	0.568	0.000	0.684	0.601
	0.8	0.000	0.877	0.827	0.000	0.883	0.843
g = 0.8, h = 0.1	-0.8	0.821	0.000	0.755	0.827	0.000	0.773
	-0.5	0.576	0.000	0.470	0.588	0.000	0.502
	-0.2	0.202	0.004	0.124	0.218	0.005	0.147
	0	0.046	0.044	0.038	0.052	0.051	0.051
	0.2	0.005	0.200	0.120	0.006	0.212	0.141
	0.5	0.000	0.583	0.469	0.000	0.594	0.502
	0.8	0.000	0.817	0.753	0.000	0.822	0.773
g = 0.8, h = 0.2	-0.8	0.885	0.000	0.844	0.892	0.000	0.865
	-0.5	0.715	0.000	0.628	0.729	0.000	0.666
	-0.2	0.287	0.001	0.189	0.314	0.002	0.229
	0	0.043	0.041	0.035	0.052	0.050	0.050
	0.2	0.002	0.282	0.186	0.002	0.305	0.224
	0.5	0.000	0.719	0.634	0.000	0.734	0.672
	0.8	0.000	0.884	0.842	0.000	0.891	0.863
g = 0.8, h = 0.3	-0.8	0.952	0.000	0.935	0.959	0.000	0.949
	-0.5	0.884	0.000	0.847	0.896	0.000	0.873
	-0.2	0.543	0.000	0.439	0.575	0.000	0.499
	0	0.042	0.038	0.031	0.052	0.050	0.050
	0.2	0.000	0.538	0.436	0.000	0.568	0.493
	0.5	0.000	0.883	0.847	0.000	0.895	0.874
	0.8	0.000	0.951	0.936	0.000	0.957	0.948

Table 3. Descriptive statistics for the difference in power between Gaussian and skewed, kurtotic and skewed/kurtotic distributions

		lower-tailed	upper-tailed	two-tailed
Distribution 2 and 3 (skewed)	Minimum	0.008	0.007	0.004
	Maximum	0.103	0.101	0.118
	Mean	0.040	0.039	0.050
Distribution 4 to 6 (kurtotic)	Minimum	0.037	0.033	0.026
	Maximum	0.103	0.101	0.118
	Mean	0.067	0.065	0.082
Distribution 7 to 12 (skewed and kurtotic)	Minimum	0.023	0.016	0.012
	Maximum	0.505	0.501	0.598
	Mean	0.180	0.179	0.210

For some of the conditions the gains in power are substantial. In general, those gains are greater for medium effect size, as we can see in Table 2.

4. Conclusions

In our simulation study, the power of the randomization test was superior in the case of the skewed and/or kurtotic distributions than in the case of the Gaussian distribution. The results of the Student-t test were similar to those of the randomization test. The latter test showed a slight advantage in the case of the two more strongly skewed and kurtotic distributions.

Thus our results suggest that if a researcher, in planning an experiment, chooses a sample size needed to detect a given population effect size, assuming a Gaussian distribution, he will be on the safe side, in terms of power, if his data come from a skewed and/or kurtotic distribution (within the range of values we have studied).

It is important to stress that, in this study, data for the two groups were simulated from the same distribution. It will be interesting, in future research, to evaluate power with data simulated from different distributions (e.g. data for a 'control' group simulated from a Gaussian distribution and data for an 'experimental' group simulated from skewed or/and kurtotic distributions).

It will be also interesting to use other values for the number of elements in each sample, to extend the range of values for the effect size and to simulate data from other distributions with different degrees of skewness and kurtosis.

REFERENCES

Badrianth S.G., Chatterjee S. (1988): On measuring skewness and elongation in common stock return distributions: The case of the Market Index. Journal of Business. 61(4): 451–72.

Badrianth S.G., Chatterjee S. (1991): A data-analytic look at skewness and elongation in common-stock return distributions. Journal of Business and Economic Statistics 9(9): 223–33.

Chung J.H., Fraser D.A.S. (1958): Randomization tests for a multivariate two-sample problem. Journal of the American Statistical Association 53: 729–735.

Cohen J. (1962): The statistical power of abnormal-social psychological research: a review. Journal of Abnormal and Social Psychology 65(3): 145–153.

Cohen J. (1969): Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen J. (1988): Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, Lawrence Erlbaum Associates.

Dutta K.K., Babbel D.F. (2002): On measuring skewness and kurtosis in short rate distributions: The case of the US Dollar London Inter Bank offer rates. Wharton Financial Institutions Center Working Paper.

Dwass M. (1957): Modified randomization tests for non-parametric hypotheses. Annals of Mathematical Statistics 28: 181–187.

Edgington E.S. (1964): Randomization tests. Journal of Psychology 57: 445–449.

Edgington E.S. (1966): Statistical inference and non-random samples. Psychological Bulletin 66: 485–487.

Edgington E.S. (1969a): Approximate randomization tests. Journal of Psychology, 72: 143–149.

Edgington E.S. (1969b): Statistical inference: The distribution-free approach. New York: McGraw–Hill.

Edgington E.S. (1995): Randomization tests (3rd ed.). New York: Marcel Dekker.

Edgington E.S., Onghena P. (2007): Randomization tests (4th ed.). Boca Raton: Chapman & Hall/CRC.

Field C.A., Genton M.G. (2006): The multivariate g-and-h distribution. Technometrics 48: 104–111.

Fisher R.A. (1935): The design of experiments. Edinburgh: Oliver & Boyd.

Hays W.L. (1994): Statistics (4th ed.). Fort Worth: Harcourt Brace.

Headrick T.C., Kowalchuk R.K., Sheng Y. (2008): Parametric probability densities and distribution functions for Tukey g-and-h transformations and their use for fitting data. Applied Mathematical Sciences 2(9): 449–462.

Hinkelmann K., Kempthorne O. (1994): Design and analysis of experiments, Volume I: Introduction to experimental design. New York: Wiley.

Hoaglin D.C. (1983): g-and-h distributions. In S. Kotz & N. L. Johnson (Eds.), Encyclopaedia of Statistical Sciences, Vol. 3 (pp. 298–301). New York: Wiley.
Hoaglin D.C. (1985): Summarizing shape numerically: the g-and-h distributions. In D.C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.) Exploring data tables, trend, and shapes (pp .417–513). New York: Wiley.
Kempthorne O. (1952): Design and analysis of experiments. New York: Wiley.
Kempthorne O. (1955): The randomization theory of experimental inference. Journal of the American Statistical Association 50: 946–967.
Lehmann E.L. (1986): Testing statistical hypotheses (2nd ed.). New York: Springer.
Manly B.F.J. (1997): Randomization, bootstrapping and Monte Carlo methods in Biology (2nd ed). London: Chapman & Hall.
Marascuilo L.A., Serlin R.C. (1988): Statistical methods for the social and behavioral sciences. New York: Freeman.
Martinez J., Iglewicz B. (1984): Some properties of the Tukey g-and-h family of distributions. Communications in Statistics – Theory and Methods 13(3): 353–69.
Mills T. C. (1995). Modelling skewness and kurtosis in the London Stock Exchange FT-SE Index Return Distributions. The Statistician 44(3): 323–32.
Mood A.M., Graybill F.A., Boes D.C. (1974): Introduction to the theory of statistics (3rd ed.). New York: McGraw-Hill.
Neyman J., Pearson E. (1928a): On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. Biometrika 20A: 175–240.
Neyman J., Pearson E. (1928b): On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. Biometrika 20A: 263–294.
Neyman J., Pearson E. (1933): On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society, Series A 231: 289–337.
Pitman E.J.G. (1937a): Significance tests that may be applied to samples from any population. Journal of the Royal Statistical Society, Suppt. 4: 119–130.
Pitman E.J.G. (1937b): Significance tests that may be applied to samples from any population, II. The correlation coefficient test. Journal of the Royal Statistical Society, Suppt. 4: 225–232.
Pitman E.J.G. (1938): Significance tests that may be applied to samples from any population, III. The analysis of variance test. Biometrika 29: 322–335.
R Development Core Team (2008): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org
Tukey J.W. (1977): Modern techniques in data analysis. NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA.
Westfall P.H., Young S.S. (1993): Resampling-based multiple testing: Examples and methods for p-value adjustment. New York: Wiley.
Winer B.J., Brown R., Michels K.M. (1991): Statistical principles in experimental design (3rd ed.). New York: McGraw-Hill.