Methods to Test For Equality of Two Normal Distributions

Julian Frank
Department of Mathematics, Karlsruhe Institute of Technology (KIT)
and
Bernhard Klar
Department of Mathematics, Karlsruhe Institute of Technology (KIT)

September 24, 2015

Abstract

Statistical tests for two independent samples under the assumption of normality are applied routinely by most practitioners of statistics. Likewise, presumably each introductory course in statistics treats some statistical procedures for two independent normal samples. Often, the classical two-sample model with equal variances is introduced, emphasizing that a test for equality of the expected values is a test for equality of both distributions as well, which is the actual goal.

In a second step, usually the assumption of equal variances is discarded. The two-sample t-test with Welch correction and the F-test for equality of variances are introduced. The first test is solely treated as a test for the equality of central location, as well as the second as a test for the equality of scatter. Typically, there is no discussion if and to which extent testing for equality of the underlying normal distributions is possible, which is quite unsatisfactorily regarding the motivation and treatment of the situation with equal variances.

It is the aim of this article to investigate the problem of testing for equality of two normal distributions, and to do so using knowledge and methods adequate to statistical practitioners as well as to students in an introductory statistics course. The power of the different tests discussed in the article is examined empirically. Finally, we apply the tests to several real data sets to illustrate their performance. In particular, we consider several data sets arising from intelligence tests since there is a large body of research supporting the existence of sex differences in mean scores or in variability in specific cognitive abilities.

Keywords: Fisher combination method, minimum combination method, likelihood ratio test, two-sample model

1 Introduction

Statistical tests for two independent samples under the assumption of normality are applied routinely by most practitioners of statistics. Likewise, statistical inference for two independent normal samples is of great relevance in every introductory statistics course. There, the approach is often quite similar: First, the importance of shift models is stated, motivating the classical two-sample model with equal variances (see, e.g., Bickel and Doksum (2006, page 4)). The ultimate aim is to compare both distributions. If normality is assumed, this corresponds to a test for equality of the expected values, i.e. Student's t-test. In a second step, usually the assumption of equal variances is discarded. The two-sample t-test with Welch correction is introduced, however, at most times without going into details of Welch's distribution approximation. The introduction and adjacent discussion on the F-test for equality of variances often varies in the level of detail. Welch's t-test is solely treated as a test for the equality of central location, as well as the F-test as a test for the equality of scatter. Typically, there is no discussion if and to which extent testing for equality of the underlying normal distributions is possible. Not only is this astonishing looking at the motivation of the classical t-test, but also due to (at least) two other reasons: For one thing lectures continue with general procedures for testing nested parametric models, including in particular likelihood-ratio tests. For another, when it comes to dealing with the one-way anova, you rarely fail to see the problem of multiple testing being mentioned, along with suitable corrections including, most of the times, the Bonferroni correction.

In some textbooks testing for equality of variances is merely left as an exercise if not outright skipped. A possible reason for this could be seen in the non-robustness of this particular test against deviances from the normal distribution. Still, as no alternative tests are at least alluded, students get the impression that differences in scatter are more or less irrelevant - variance is a statistical Cinderella. Yet, everybody actually applying statistical procedures knows very well that differences in variance and location are of comparable importance.

Summing up, it can be said that from a practical point of view, given a two-sample model under normality, the aim has to be to judge whether the two samples originate from basically similar distributions or not. However, in many cases the classical and, of

course, very comfortable assumption of equal variances has no grounding. In the midst of these considerations, discussion in lectures and textbooks stops without further ado and the students (and maybe some lecturers as well) are left without a clue how to deal with this situation.

It is the aim of the following article to investigate the problem of testing for the equality of two normal distributions, and to do so using knowledge and methods adequate to statistical practitioners as well as to students in an introductory course in mathematical statistics. Mathematically speaking, the following testing problem will be considered: Let $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ be independent normally distributed random variables, where $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for all $i = 1, \ldots, m$ and $Y_j \sim \mathcal{N}(\nu, \tau^2)$ for all $j = 1, \ldots, n$. In contrast to Student's t-test, we do not make further assumptions about the parameters, so that $(\mu, \nu, \sigma^2, \tau^2) \in \Theta = \mathbb{R}^2 \times (0, \infty)^2$ is arbitrary. It is the objective to test if the two samples stem from identical distributions. The corresponding testing problem is given by the following hypothesis and alternative:

$$\begin{aligned} & H_0: \quad \vartheta = (\mu, \nu, \sigma^2, \tau^2) \in \Theta_0 = \{\vartheta \in \Theta : \mu = \nu,\ \sigma^2 = \tau^2\} \\ \text{vs.} \quad & H_1: \quad \vartheta \in \Theta \setminus \Theta_0. \end{aligned} \tag{1}$$

The first classical approach is to develop a likelihood-ratio test. Doing so is a simple way to obtain an asymptotically valid test. In Section 2 the likelihood-ratio test statistic is derived, and different approximations of the distribution of the test statistic under H_0 found in the literature are summed up. Among them one can find an asymptotic expansion proposed by Muirhead (1982), as well as a recently developed method to derive the exact distribution by numerical integration (Zhang et al., 2012).

A further approach is to combine different p-values as illustrated in Section 3. For this procedure the hypothesis H_0 is obtained by combining the hypotheses of both t- and F-test. Performing both tests using the same data $(x_1, \ldots, x_m)$ and $(y_1, \ldots, y_n)$, the resulting p-values can be combined yielding a new test statistic and, thus, a test result for (1). Most combination methods require the tests to be combined being independent under H_0 which holds in the case under consideration. In the specific case of Fisher's method, the same approach, but applied in a slightly different way, can be found in Perng and Littell (1976).

In Section 4, power of the different tests is compared empirically. The ability of each

method to correctly detect the alternative differs with respect to whether there is a difference in expectation, variance, or both. Loughin (2004) compares the method of combining the p-values without regard to a specific testing problem. However, it is instructive to apply these methods directly to the problem at hand and compare them with the likelihood-ratio tests in Section 2.

Situations where one is interested in differences in variability as well as in means can be found almost everywhere. A long list of such applications is compiled in Gastwirth (2009). We discuss in Section 5 several examples from two subject areas, namely engineering and psychology. In particular, we consider several data sets arising from mental or intelligence tests since there is a large body of research supporting the existence of sex differences in specific cognitive abilities, some favouring men, some favouring women, sometimes differences are found in mean scores, or in variability, or in both.

2 The Likelihood Ratio Test

A classic approach in order to construct a test for H_0 is the application of the maximum likelihood method. The unrestricted maximum likelihood estimator $\hat{\vartheta}$ is given by

$$\hat{\vartheta} = \left(\hat{\mu}, \hat{\nu}, \hat{\sigma}^2, \hat{\tau}^2\right) = \left(\bar{X}, \bar{Y}, \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})^2, \frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})^2\right),$$

with $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \bar{Y} = \sum_{j=1}^{n} Y_j$, while the maximum likelihood estimator $\hat{\vartheta}_0$ under H_0 is given by

$$\hat{\vartheta}_0 = \left(\hat{\mu}_0, \hat{\mu}_0, \hat{\sigma}_0^2, \hat{\sigma}_0^2\right)$$

with $\hat{\mu}_0 = \frac{m\bar{X}+n\bar{Y}}{m+n}$ and $\hat{\sigma}_0^2 = \frac{1}{m+n}\left(\sum_{i=1}^{m}\left(X_i - \hat{\mu}_0\right)^2 + \sum_{j=1}^{n}\left(Y_j - \hat{\mu}_0\right)^2\right)$. Denoting the likelihood function by $L(\vartheta)$, the likelihood ratio statistic $\Lambda_{m,n}$ is equal to

$$\begin{aligned}
\Lambda_{m,n} &= \frac{L(\hat{\vartheta}_0)}{L(\hat{\vartheta})} \\
&= \frac{\left(2\pi\hat{\sigma}_0^2\right)^{-\frac{m+n}{2}} \exp\left(-\frac{1}{2\hat{\sigma_0}^2}\left(\sum_{i=1}^{m}\left(x_i - \hat{\mu}_0\right)^2 + \sum_{j=1}^{m}\left(y_j - \hat{\mu}_0\right)^2\right)\right)}{\left(2\pi\hat{\sigma}^2\right)^{-\frac{m}{2}} \exp\left(-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{m}\left(x_i - \hat{\mu}\right)^2\right) \cdot \left(2\pi\hat{\tau}^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\hat{\tau}^2}\sum_{j=1}^{n}\left(y_j - \hat{\nu}\right)^2\right)} \\
&= \frac{\left(\hat{\sigma}^2\right)^{\frac{m}{2}} \cdot \left(\hat{\tau}^2\right)^{\frac{n}{2}}}{\left(\hat{\sigma}_0^2\right)^{\frac{m+n}{2}}}.
\end{aligned}$$

Assuming $\frac{m}{m+n} \to p$ for $m+n \to \infty$ and some $p \in (0,1)$, it follows from the general theory of likelihood ratio tests, given that Θ_0 and Θ have dimensions 2 and 4, that

$$-2\log\Lambda_{m,n} \xrightarrow{\mathcal{D}} \chi^2_2 \qquad \text{for } m+n \to \infty \text{ under } H_0 \tag{2}$$

(Hogg et al., 2005, pp. 351-353). Hence, an asymptotic level α test rejects H_0 if

$$-2\log\Lambda_{m,n} \geq \chi^2_{2;1-\alpha}, \tag{3}$$

where $\chi^2_{2;p}$ denotes the p-quantile of the χ^2-distribution with 2 degrees of freedom. Typically, fairly large sample sizes are needed to use these asymptotic results for finite samples. However, there are several approaches available to transform the test statistic or to determine a more exact distribution in order to improve the finite sample behaviour.

Pearson and Neyman (1930) directly considered $\Lambda_{m,n}$, showing that under H_0, the limiting distribution is the uniform distribution $U(0,1)$ (note that, if Z is uniformly distributed, $-2\log Z$ is exponentially distributed with mean 2, or χ^2_2-distributed; hence, this result is in agreement with (2)). They proposed to approximate the exact distribution of $\Lambda_{m,n}$ for finite n and m by a beta distribution matching the first two moments.

Muirhead (1982) considered an asymptotic expansion of the distribution of the likelihood ratio test statistic under multivariate normality; in the univariate case, we obtain the following corollary.

Corollary 2.1

Let $F_{\chi^2_q}$ denote the distribution function of the χ^2_q-distribution. It holds under H_0:

$$P_{H_0}\left(-2\rho\log\Lambda_{m,n} \leq u\right) = F_{\chi^2_2}(u) + \frac{\gamma}{\rho^2\,(m+n)^2}\left(F_{\chi^2_6}(u) - F_{\chi^2_2}(u)\right) + O\left((m+n)^{-3}\right),$$

with

$$\rho = 1 - \frac{22}{24(m+n)}\left(\frac{m+n}{m} + \frac{m+n}{n} - 1\right),$$
$$\gamma = \frac{1}{2}\left(\left(\frac{m+n}{m}\right)^2 + \left(\frac{m+n}{n}\right)^2 - 1\right) - \frac{121}{96}\left(\frac{m+n}{m} + \frac{m+n}{n} - 1\right)^2.$$

Hence, the function $F_{m,n}$, defined by

$$F_{m,n}(u) = F_{\chi^2_2}(u) + \frac{\gamma}{\rho^2\,(m+n)^2}\left(F_{\chi^2_6}(u) - F_{\chi^2_2}(u)\right),$$

is an approximation of the distribution function of $-2\rho \log \Lambda_{m,n}$ under the hypothesis. Then, an approximate test of H_0 against H_1 rejects H_0 if

$$-2\rho \log \Lambda_{m,n} \geq F_{m,n}^{-1}(1-\alpha). \tag{4}$$

The improvement achieved by this expansion is illustrated in Table 1. There, the quantiles of $F_{\chi_2^2}$ and $F_{10,20}$ are compared with the simulated quantiles of $-2\log \Lambda_{10,20}$ and $-2\rho \log \Lambda_{10,20}$ (based on 10^5 replications) for sample sizes $m = 10$ and $n = 20$. Table 1 indicates that the empirical and theoretical levels are much closer for the Muirhead-approximation than the test based on asymptotic χ^2 results. For practical purposes, the empirical level of the test based on the expansion is sufficiently close to the theoretical level even for small sample sizes.

	Asymptotic χ^2		Muirhead approximation	
p	$F_{\chi_2^2}^{-1}(p)$	p-quantile of $-2\log \Lambda_{10,20}$	$F_{10,20}^{-1}(p)$	p-quantile of $-2\rho \log \Lambda_{10,20}$
0.75	2.77	3.13	2.70	2.79
0.90	4.61	5.19	4.46	4.64
0.95	5.99	6.74	5.77	6.02
0.99	9.21	10.33	8.74	9.22
0.999	13.82	15.48	12.78	13.83

Table 1: Comparison of the χ^2- and Muirhead-approximation for $m = 10$ and $n = 20$

There have also been several approaches to determine the exact distribution of $\Lambda_{m,n}$ in more or less computable form. Jain et al. (1975) developed computable but complicated series representations for the density and distribution function. Nagar and Gupta (2004) tabulated the distribution of $\Lambda_{m,n}$ for the balanced case $m = n$. Zhang et al. (2012) determined the exact distribution of $\Lambda_{m,n}$ as

$$P\left(\Lambda_{m,n} \leq u\right) = 1 - C \int_{r_1}^{r_2} w_1^{(m-3)/2-1} \left(\int_{\frac{u^{2/n} m^{m/n}}{(m+n)^{(m+n)/n} w_1^{m/n}}}^{1-w_1} \frac{w_2^{(n-3)/2}}{\sqrt{1-w_1-w_2}} \mathrm{d}w_2 \right) \mathrm{d}w_1 \tag{5}$$

for $u \in (0,1)$, with $C = \frac{\Gamma\left(\frac{m+n-1}{2}\right)}{\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)}$ and $r_1 < r_2$. Hereby r_1 and r_2 denote the two

roots of the function $g(w_1) = 1 - w_1 - \frac{u^{2/n} m^{m/n} n}{(m+n)^{(m+n)/n} w_1^{m/n}}$. The double integral in (5) can be evaluated by any numerical quadrature method.

Remark

In many cases, likelihood ratio tests exhibit some kind of optimality. Hsieh (1979) has shown that for the testing problem under consideration, the likelihood ratio test is asymptotically optimal in the sense of Bahadur efficiency.

3 Combination of *p*-values

Another method to obtain a test for H_0 is to combine the p-values of Student's t-test and the F-test for equality of variances. For this purpose, the hypothesis H_0 has to be rephrased as a multiple one. Using the hypothesis and alternative of the t-test

$$H_0' : \vartheta \in \Theta_0' = \left\{\vartheta \in \Theta : \mu = \nu, \sigma^2 = \tau^2\right\} \text{ vs.}$$
$$H_1' : \vartheta \in \Theta_1' = \left\{\vartheta \in \Theta : \mu \neq \nu, \sigma^2 = \tau^2\right\}$$

and the F-test

$$H_0'' : \vartheta \in \Theta_0'' = \left\{\vartheta \in \Theta : \sigma^2 = \tau^2\right\} \text{ vs.}$$
$$H_1'' : \vartheta \in \Theta_1'' = \left\{\vartheta \in \Theta : \sigma^2 \neq \tau^2\right\},$$

H_0 can be reformulated as

$$H_0 : H_0' \text{ and } H_0'' \text{ are true vs.}$$
$$H_1 : \text{ At least one of the alternatives } H_1' \text{ or } H_1'' \text{ is true.}$$

The statistic of the t-test

$$T = \frac{\sqrt{\frac{mn}{m+n}}(\bar{X} - \bar{Y})}{S}$$

with $S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$, $S_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$ and the pooled variance estimator

$$S^2 = \frac{1}{m+n-2}((m-1)S_X^2 + (n-1)S_Y^2),$$

has a t-distribution with $m+n-2$ degrees of freedom under H_0', while the F-test statistic

$$Q = \frac{S_X^2}{S_Y^2}$$

is $F_{m-1,n-1}$-distributed under H_0''. Methods of combining p-values like Fisher's method (Fisher, 1932) presume the independence of the corresponding p-value under the hypothesis. To prove the independence of T and Q under H_0, one can invoke Basu's theorem to prove the independence of the sum and the quotient of two independent χ^2-distributed random variables (Lehmann and Romano, 2005, pp. 152-153). This result is applied to $(m-1)S_X^2/\sigma^2$ and $(n-1)S_Y^2/\sigma^2$ to derive the independence of Q and S^2 under H_0. Since S_X^2, S_Y^2, $\bar{X}$ and $\bar{Y}$ are independent, T and Q are independent as well. After having shown the independence of t- and F-test statistics, we can combine the p-values

$$G_1(t) = P_{H_0'}(|T| \geq |t|) \text{ and } G_2(q) = 2\min\left\{P_{H_0''}(Q \geq q), P_{H_0''}(Q \leq q)\right\}$$

in order to obtain a test for H_0. Note that $G_2(q)$ corresponds to the p-value of the usual two-sided F-test with equal tail probabilities.

The crucial fact in the following examples of combination methods proposed in the literature is the following: if the distribution of the test statistic under H_0 is unique and continuous, then the p-value, considered as a random variable, follows the uniform distribution on the unit interval under H_0 (this fact is often stated only for simple null hypotheses which is overly restrictive).

3.1 Combination method due to Fisher

Fisher (1932) proposed the combined statistic

$$M_1 = -2\log(G_1(T)G_2(Q)) = -2\log G_1(T) - 2\log G_2(Q).$$

Since $-2\log G_1(T)$ and $-2\log G_2(T)$ are χ_2^2-distributed under H_0, the decision to

$$\text{reject } H_0 \text{ if } M_1 \geq \chi_{4;1-\alpha}^2$$

leads to an exact level α-test.

For the testing problem given in (1), Perng and Littell (1976) proved that the test based on M_1, just as the likelihood ratio test in Section 2, is asymptotically optimal in the sense of Bahadur efficiency (see also Singh (1986)).

This is not too astonishing regarding the close relation between both tests. Fisher's method combines both tests as a product with equal weights under H_0. On the other

hand, as shown by Pearson and Neyman (1930), the squared t-test statistic and the F-statistic are one-to-one correspondences of the likelihood ratio statistics for testing H_0' against H_1' and H_0'' against H_1'', respectively. They showed further that the likelihood ratio for testing H_0 against H_1 can be expressed as the product of the likelihood ratio for testing H_0' against H_1' and the likelihood ratio for testing H_0'' against H_1''. Hence, the likelihood ratio test combines both tests as a product with approximately equal weights under H_0 (since they have the same limiting distribution under H_0).

3.2 Minimum combination method and Bonferroni correction

Another classical approach is the minimum combination method proposed by Tippett (1931) with test statistic

$$M_2 = \min(G_1(T), G_2(Q)).$$

Since the distribution function of the minimum of two independent and uniformly distributed random variables is $u(2-u)$ for $u \in (0,1)$, an exact test at level α can be stated as

$$\text{reject } H_0 \text{ if } M_2 \leq 1 - \sqrt{1-\alpha}. \tag{6}$$

This method is closely related to the simplest method to control the familywise error rate, the Bonferroni correction. Using this method, the overall hypothesis is rejected at a significance level of α if either of the individual hypotheses is rejected at a level of $\alpha/2$, which corresponds to

$$\text{reject } H_0 \text{ if } M_2 \leq \alpha/2. \tag{7}$$

Generally, the Bonferroni correction is considered as rather conservative, which can indeed be the case for stronly dependent statistics. However, for independent test statistics, both methods are equal for practical purposes since $\alpha/2$, the critical value in (7), is just the first order Taylor expansion of the critical value in (6).

3.3 Maximum and sum combination methods

Two further very simple approaches use maximum and sum of the two p-values. Using the maximum statistic $M_3 = \max(G_1(T), G_2(Q))$, the corresponding test of level α is given by

$$\text{reject } H_0 \text{ if } M_3 \leq \sqrt{\alpha}. \tag{8}$$

The use of the sum $M_4 = G_1 + G_2$ was proposed by Edington (1972). Since the distribution function of the convolution of two uniform random variables is $u^2/2$ for $u \in (0, 1)$, an exact test at level α is

$$\text{reject } H_0 \text{ if } M_4 \leq \sqrt{2\alpha}. \tag{9}$$

3.4 Combination methods due to Stouffer-Lipták and due to Mudholkar and George

Two further more often used methods are based on the statistic

$$M_5 = \left(\Phi^{-1}(1 - G_1) + \Phi^{-1}(1 - G_2)\right)/\sqrt{2},$$

where Φ^{-1} denotes the inverse of the standard normal distribution function, and the logit statistic

$$M_6 = -\sqrt{7/(4\pi^2)}\left(\text{logit}(G_1) + \text{logit}(G_2)\right),$$

with $\text{logit}(p) = \log\left(p/(1-p)\right)$. Early references for the combination method using M_5 are Stouffer et al. (1949) and Lipták (1958), and the method is found in the literature under both names. For brevity, we use the name Stouffer's method in the following. Since $\Phi^{-1}(1 - G_i), i = 1, 2$, is normally distributed under H_0, the decision

$$\text{reject } H_0 \text{ if } M_5 \geq \Phi^{-1}(1 - \alpha) \tag{10}$$

defines an exact test of level α.

Since $\text{logit}(G_i), i = 1, 2$, follows a standard logistic distribution under H_0, direct calculations yield

$$P_{H_0}(\text{logit}(G_1) + \text{logit}(G_2) > u) = u\left(\frac{e^{-u}}{1 - e^{-u}}\right)^2 + (u - 1)\frac{e^{-u}}{1 - e^{-u}}, \quad -\infty < u < \infty,$$

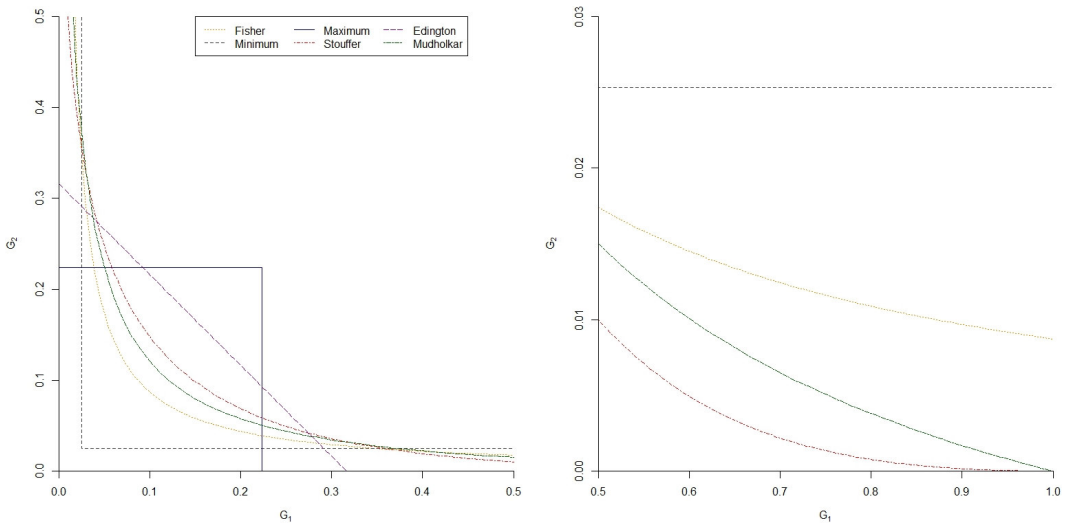

Figure 1: Rejection regions of the different combination methods for $\alpha = 5\%$.

which can be used to perform an exact test based on M_6. However, Mudholkar and George (1979) (see also George and Mudholkar (1983)) proposed to approximate the distribution of M_6 by a t_{14}-distribution, leading to the approximative level α test

$$\text{reject } H_0 \text{ if } M_6 \geq t_{14;1-\alpha}. \tag{11}$$

The proposed approximation is indeed very accurate, the exact and approximate 0.95-quantiles given by 1.7649 and 1.7613, respectively.

The logit statistic has the same exact Bahadur slope under H_0 as Fisher's statistic. Hence, it is also optimal in the sense of Bahadur efficiency (Mudholkar and George, 1979; Berk and Cohen, 1979).

Figure 1 shows the different rejection regions of the discussed combination methods. Each region covers 5% of the unit square. In the right figure, the abscissa is continued from 0.5 to 1, but the scaling of the ordinate is different. Similar displays are shown, for example, in Loughin (2004), who concludes that the main feature of the maximum method and the combination method due to Edington is the inability to reject the hypothesis if one p-value is large, regardless how small the other is. He also states that these combination methods can be useful, given the circumstance that both p-values are equally significant. Thus, it is crucial to compare the tests empirically under different alternatives.

It should be noted that the presented combination rules can also be used in other situations, for example for combining dependent tests. A description and comparison of combination rules in a nonparametric context can be found in Pesarin and Salmaso (2010, pp. 128-134).

4 Empirical level and power of the tests

In the following, the tests are compared at level $\alpha = 0.05$, whereby the exact 0.95-quantile of M_6 was used as well as the exact 0.95-quantile of $\Lambda_{m,n}$, calculated by (5), which is 6.93032 for $m = n = 10$, 6.430465 for $m = n = 20$ and 6.657326 for $m = 10,\ n = 30$. The values for $m = n$ coincide with the values given in Nagar and Gupta (2004). To achieve such a high accuracy, we used very high numbers of quadrature points in evaluating the double integral in (5) by Gauss-Legendre quadrature together with an extrapolation method.

First, we consider the case of equal variances but different expected values. To this end, the p-values G_1 and G_2 were simulated 10^5 times with sample sizes of $m = n = 20$ and under fixed parameters $\sigma^2 = \tau^2 = 1,\ \mu = 0$. Table 2 shows the empirical power of the tests for varying expectation ν. Clearly, the power should only depend on the absolute value of ν, which is the case within the simulation accuracy.

The maximum and the sum combination methods are often incapable of rejecting H_0, since with increasing ν the expected value of G_1 is decreasing, but G_2 remains uniformly distributed over $(0, 1)$ under H_0''. This leads to a minimal type II error of $1 - \sqrt{\alpha} = 0.776$ for the maximum method and a minimal error of $1 - \sqrt{2\alpha} = 0.683$ for the combination method due to Edington. This is illustrated in Figure 2, where the left plot shows the absolute power, whereas the plot on the right hand side shows the power relative to the best test. From the remaining tests, the minimum method performs best, followed closely by the LQ-test and Fisher's method, and, in some distance, Mudholkar's and Stouffer's methods.

Next, the case of equal expectations but different variances is considered. Here, the parameters $\mu = \nu = 0,\ \sigma^2 = 1$ are fixed, and τ varies between 0.2 and 2.6. As Table 3 and Figure 3 show, the results are similar to the preceding setup. This is perhaps surprising at first sight, since the t-test is not a level α-test if the variances differ. However, it is

ν	Fisher	Minimum	Maximum	Edington	Stouffer	Mudholkar	LQ-test
-1.5	0.986	0.991	0.223	0.314	0.903	0.962	0.988
-1	0.769	0.798	0.216	0.291	0.613	0.698	0.781
-0.5	0.252	0.258	0.143	0.164	0.215	0.233	0.256
0	0.050	0.050	0.050	0.050	0.050	0.050	0.050
0.5	0.251	0.257	0.142	0.163	0.215	0.232	0.255
1	0.770	0.799	0.218	0.292	0.618	0.701	0.782
1.5	0.986	0.990	0.222	0.313	0.903	0.962	0.988

Table 2: Empirical power for varying ν, $\sigma^2 = \tau^2 = 1, \mu = 0$ and $m = n = 20$.

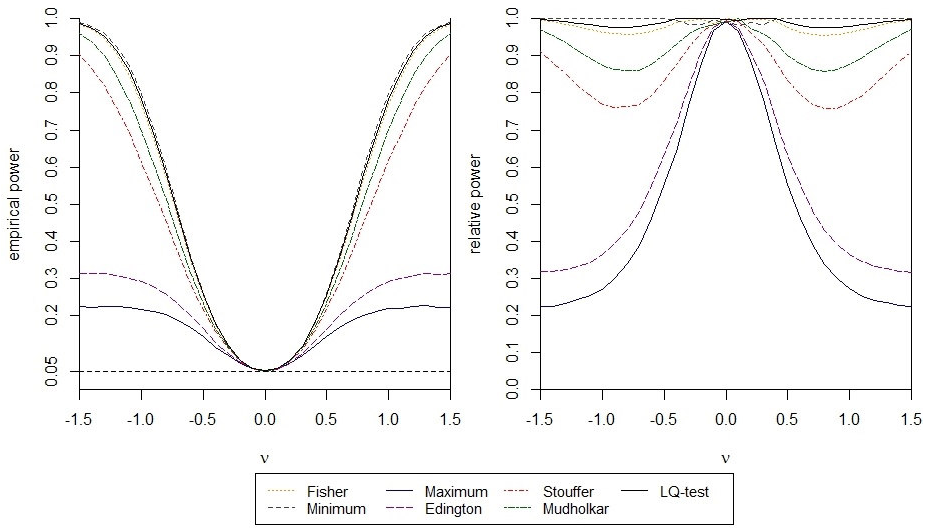

Figure 2: Empirical absolute (left figure) and relative (right figure) power for varying ν.

well-known that the t-test is quite robust against the violation of homogeneity of variances; a quantitative statement in this direction can be found in Perng and Littell (1976, p. 970).

τ	Fisher	Minimum	Maximum	Edington	Stouffer	Mudholkar	LQ-test
0.2	1.000	1.000	0.230	0.321	0.998	1.000	1.000
0.6	0.459	0.480	0.180	0.224	0.365	0.411	0.473
1	0.050	0.050	0.049	0.050	0.049	0.049	0.050
1.4	0.216	0.222	0.126	0.142	0.183	0.198	0.221
1.8	0.585	0.609	0.198	0.253	0.459	0.523	0.600
2.2	0.853	0.870	0.223	0.301	0.701	0.789	0.864
2.6	0.959	0.966	0.226	0.313	0.849	0.921	0.964

Table 3: Empirical power for varying τ, $\sigma^2 = 1, \mu = \nu = 0$ and $m = n = 20$.

Our main interest lays in situations where the expected values as well as the variances differ. Here, the parameters ν and τ have been chosen so that the p-values of the t- and F-test are nearly equal on average (see Table 4 and Figure 4). To this end, the statistics T and Q were simulated 10^5 times in order to estimate $E_\vartheta(G_1)$ and $E_\vartheta(G_2)$ by their arithmetic means.

In the present case, the methods of Mudholkar, Fisher, Stouffer and the LQ-test behave very similar, the first two having the edge over the latter (see Table 5). The minimum combination method exhibit a slightly lower power, as already stated by Loughin (2004). However, the results contradict Loughins statement that the maximum method and the one due to Edington perform in this case better than the other tests. As illustrated in Figure 5, both methods can recognise the alternative now, but their empirical power is still lower.

We rerun all simulations with smaller sample sizes $m = n = 10$. Apart from generally lower values of the power, all of the aforementioned conclusions remain unchanged. Further, we rerun all simulations with unequal sample sizes, namely $m = 10$ and $n = 30$, hence maintaining the total sample size. Under this scenario the empirical power decreases under every alternative compared to the balanced case $m = n = 20$. Thereby, the power under a changing τ declines more than under a changing ν. Given the alternative that both

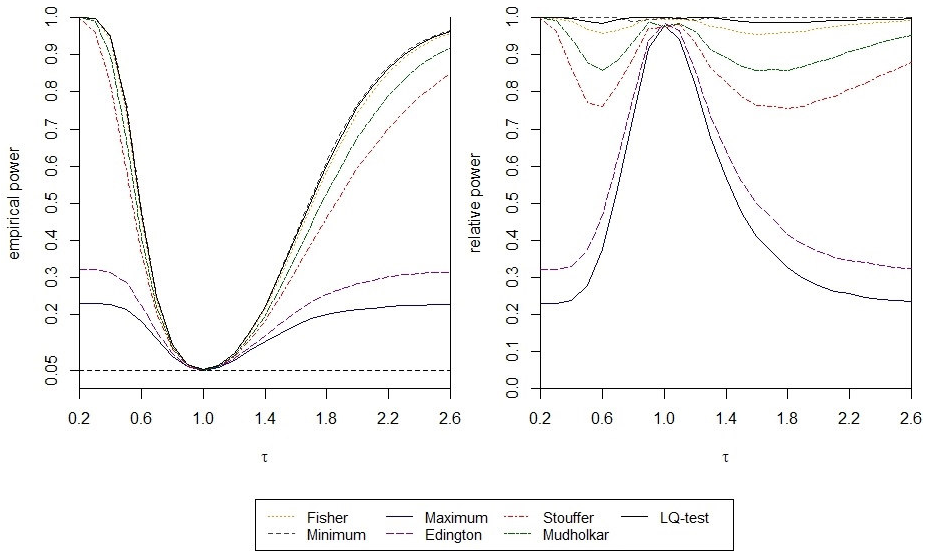

Figure 3: Empirical absolute (left figure) and relative (right figure) power for varying τ.

no.	Fisher	Minimum	Maximum	Edington	Stouffer	Mudholkar	LQ-test
1	0.050	0.050	0.049	0.049	0.049	0.049	0.049
3	0.076	0.074	0.069	0.071	0.074	0.075	0.076
5	0.159	0.148	0.129	0.139	0.153	0.156	0.158
7	0.294	0.260	0.234	0.257	0.287	0.293	0.291
9	0.455	0.393	0.364	0.401	0.449	0.454	0.448
11	0.621	0.539	0.509	0.555	0.617	0.624	0.612
13	0.758	0.671	0.637	0.689	0.757	0.763	0.749
15	0.859	0.783	0.742	0.791	0.855	0.862	0.852
17	0.922	0.864	0.822	0.864	0.921	0.925	0.917

Table 5: Empirical power for varying ν and τ, $\mu = 0, \sigma^2 = 1$ and $m = n = 20$.

no.	ν	τ	$\bar{G}_1$	$\bar{G}_2$
1	0.000	1.00	0.50	0.50
2	0.075	1.05	0.49	0.49
3	0.150	1.10	0.47	0.48
4	0.225	1.15	0.44	0.45
5	0.300	1.20	0.40	0.41
6	0.375	1.25	0.36	0.37
7	0.450	1.30	0.31	0.33
8	0.525	1.35	0.27	0.30
9	0.600	1.40	0.24	0.26
10	0.675	1.45	0.20	0.23
11	0.750	1.50	0.17	0.20
12	0.825	1.55	0.15	0.17
13	0.900	1.60	0.13	0.14
14	0.975	1.65	0.11	0.12
15	1.050	1.70	0.09	0.10
16	1.125	1.75	0.08	0.09
17	1.200	1.80	0.07	0.07

Table 4: Average p-values for different choices of ν and τ.

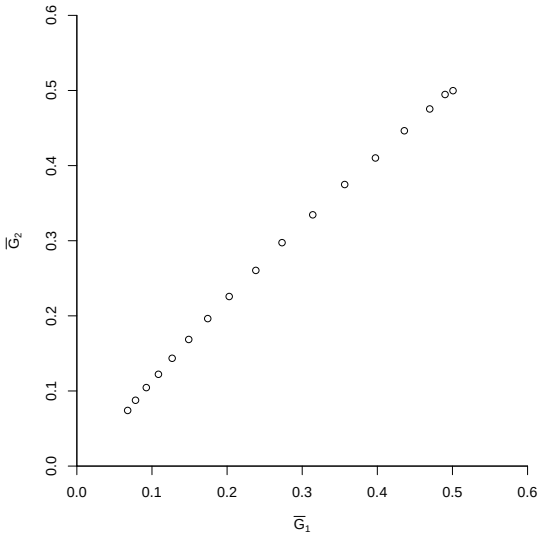

Figure 4: Visualisation of Table 4

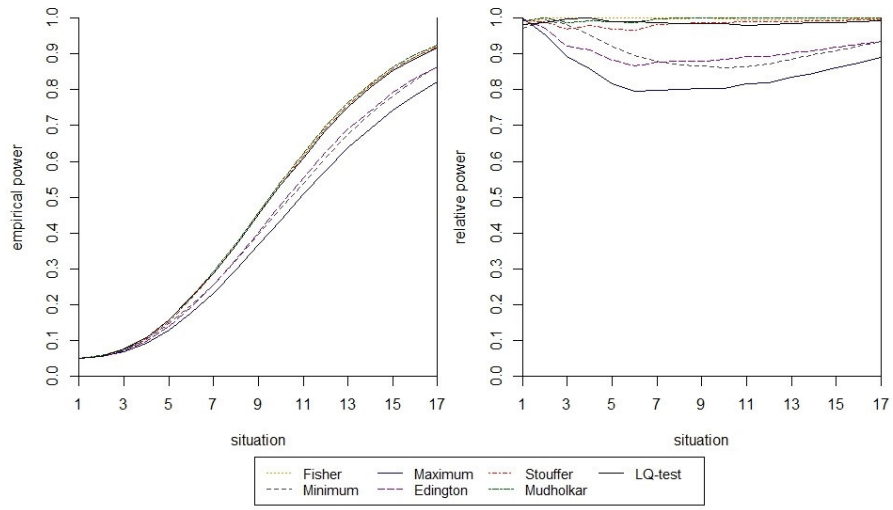

Figure 5: Empirical absolute (left figure) and relative (right figure) power for varying ν and τ, $\mu = 0, \sigma^2 = 1$.

parameters change, the LQ-test performs considerably best and the minimum method worst.

Remark

The statements about the asymptotic optimality of the likelihood ratio test or the combination method of Fisher are at first sight surprising. Consider, for example, the first simulation scenario with a difference between means but equal variances. Then, the optimality property of, say, Fisher's method means that we lose no power in a specific asymptotic sense if we use this method instead of the t-test which is well-known to be optimal in this situation for any finite sample size.

Relative efficiency of a sequence of tests $\{S_n\}$ *with respect to another sequence* $\{T_n\}$ *is the ratio of the sample sizes necessary for* $\{S_n\}$ *and* $\{T_n\}$ *in order to attain the power* β *under the level* α *for a specific alternative. Bahadur asymptotic relative efficiency considers the limit of this ratio for a sequence of levels decreasing to zero keeping* β *and the alternative fixed.*

To compute relative efficiencies exemplarily we fix power $\beta = 0.6$ *and mean* $\nu = 0.5$*, and consider four decreasing values of* α*, namely* $0.187, 0.027, 0.0012, 3 \cdot 10^{-6}$*. For the balanced*

case $m = n$, the t-test needs sample sizes 20,50,100 and 200 to reach power $\beta = 0.6$, whereas the combination method of Fisher has the same power for sample sizes 27,64,122 and 230. Hence, the relative efficiencies for the four different levels are 0.74, 0.78, 0.82 and 0.87. Indeed, relative efficiency increases, but even for large sample sizes (and, correspondingly, small levels) it is far away from 1, the limit for $\alpha \to 0$ given by theory.

5 Data Examples

We applied the tests to several data sets exemplifying the behaviour of the presented methods.

Example 1. First, we consider a small data set discussed by Nair (1984) giving the times in minutes to breakdown of an insulating fluid under elevated voltage stresses of 32 and 36 kV, respectively, for the first and second sample. 15 observations are taken at each voltage. This data set is also considered in Shoemaker (1999), Marozzi (2011) and Marozzi (2012) where the question of interest is if the variability of times is significantly higher for the 32KV power voltage. However, in practice, it would certainly be interesting as well if the mean times differ significantly between the two samples. Since the observations are failure times, the distributions of the raw data are highly skewed. However, after a log-transformation, the data look rather symmetric and short-tailed, as Figure 6 shows. Hence, the proposed tests should be appropriate for the transformed data. Means and standard deviations of the transformed samples are

$$32 \text{ kV}: \ \bar{x} = 2.229, \ s_x = 0.902, \quad 36 \text{ kV}: \ \bar{y} = 2.198, \ s_y = 1.110.$$

For testing the equality of the underlying distributions, we can use any of the test discussed above. The corresponding p-values are given in the second line of Table 6; there, we added the usual significance codes to the p-values for a fast overview: p^{***} if $p \leq 0.001$, p^{**} if $0.001 < p \leq 0.01$, p^{*} if $0.01 < p \leq 0.05$, and p° if $0.05 < p \leq 0.10$.

For the data at hand, the p-value of the minimum combination method is considerably larger than for all other methods, whereas all remaining p-values are comparable and smaller than 0.01. Since the tests find a significant difference between the two underlying distributions, one could informally proceed as proposed in Section 3 of Zhang et al. (2012),

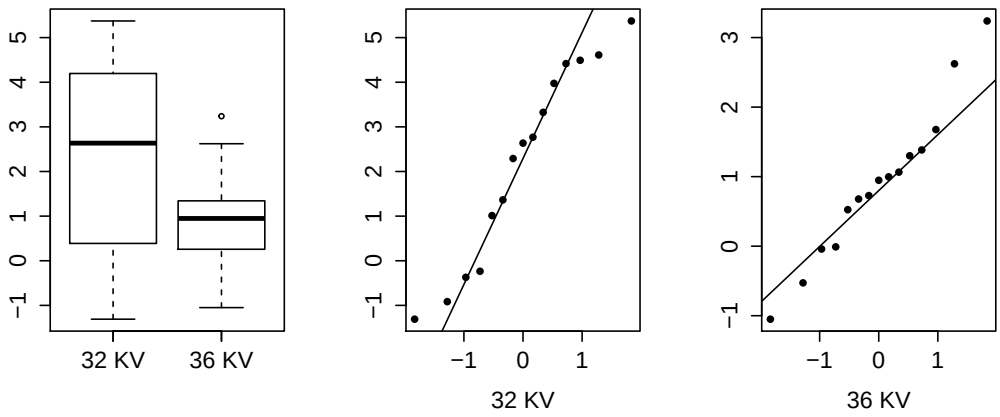

Figure 6: Box-plots (left figure) and QQ-plots for normality (middle and right plot) for the log-transformed samples of example 1.

Ex.	Fisher	Minimum	Maximum	Edington	Stouffer	Mudholkar	LQ-test
1	0.0058**	0.030*	0.0021**	0.0019**	0.0033**	0.0045**	0.0072**
2a	0.0007***	0.0017**	0.0059**	0.0030**	0.0006***	0.0006***	0.0008***
2b	0.11	0.045*	1	0.52	1	1	0.074°
3	0.0098**	0.040*	0.0039**	0.0034**	0.0057**	0.0075**	0.012*

Table 6: p-values for the different testing methods applied to examples 1-3.

first applying the F-test ($p = 0.015$) followed by Welch's t-test ($p = 0.050$, cf. Bickel and Doksum (2006, pp. 264), or Aspin and Welch (1949)). Clearly, one has to be careful when formally reporting the results of such follow-up tests. The results of this example are in agreement with the simulation results: the minimum method has comparably low power when there exist differences in the means as well as in the variances.

Examples 2a,b. There is a long scientific dispute about whether there are sex differences in cognitive abilities. Deary et al. (2007) used a novel design, comparing 1292 pairs of opposite-sex siblings who participated in the US National Longitudinal Survey of Youth 1979. The mental test applied was divided in several subtasks (Deary et al. (2007), Table 1). Here, we consider the results of two subtests of the test battery, namely word knowledge and mathematics knowledge. Means $\bar{x}, \bar{y}$ and standard deviations s_x, s_y for males and females for the word knowledge subtest are

$$\bar{x} = 22.3, \quad s_x = 9.0, \quad \bar{y} = 22.9, \quad s_y = 8.2.$$

The third line in Table 6 shows the p-values for the tests for equality of distributions. Here, the values for maximum, sum and minimum method are larger than the remaining values, but all tests yield a significant result on the 0.01-level. Since there is a significant difference between the two distributions, we applied the two-sample t-test ($p = 0.077$) and the F-test ($p = 0.0008$), indicating a significantly larger variability in the results of males compared to females.

Means and standard deviations for the mathematics knowledge subtest for males and females are given by

$$\bar{x} = 11.9, \quad s_x = 6.5, \quad \bar{y} = 11.9, \quad s_y = 6.1.$$

It is a somewhat extreme example showing no difference between the sample means. The p-values of the tests can be found in line 4 of Table 6. Here, only the minimum combination method shows a significant result on the 0.05-level, followed by the LQ-test and Fisher's combination method resulting in p-values of 0.074 and 0.11. The remaining p-values are larger than 0.5. Here, the F-test yields a p-value of 0.023. Clearly, the minimum method is not affected by the large p-value of the t-test, and hence, performs best in this example.

It should be noted that in this and the following example we are dealing with large samples where it is possible that very small differences are statistically significant but may be of not much scientific or practical importance.

Example 3. Table 1 in Steinmayr (2010) shows the results for various subtests of the German Intelligence-Structure-Test 2000-R performed by 426 male and 551 female students attending 11th or 12th grade with age ranging from 16 to 18 years. Means and standard deviations for the matrices knowledge subtest for males and females are reported as

$$\bar{x} = 11.06, \quad s_x = 2.90, \quad \bar{y} = 11.39, \quad s_y = 2.61.$$

The results in line 5 of Table 6 show that, as in example 1, the p-value of the minimum method is considerably larger than for all other methods, followed by the LQ-test and Fisher's method with p-values around 0.01. Since there are significant differences between the underlying distributions, we applied the two sample t-test ($p = 0.062$) and F-test ($p = 0.020$), again indicating a larger variability in the test scores of male students. In this example, it is clearly noticeable that the p-values of all combination tests except the minimum method can be much smaller than the p-values of the individual t- and F-tests.

6 Discussion

- To sum up the results of the simulation study, it is clear that the maximum and the sum combination methods should not be used due to its inability to detect many alternatives. From Figure 1, one could expect that both methods might be useful if differences in location come along with differences in variability which is the rule rather than the exception in many biometrical applications. However, the simulations show that these methods are inferior to other combination methods even in the case of location-scale differences. Furthermore, the minimum combination method is not really recommendable due to its comparably low power when there exist differences in the means as well as in the variances. There is not much to choose from the remaining tests. In terms of power, the likelihood ratio test has the edge over the other methods at least in unbalanced situations, whereas the Fisher combination method stands out due to its simplicity.

- Even if the data examples corroborate the previous findings, the performance of the different tests for specific data sets may be astonishing. As always in such situations, there is a danger that one performs several tests, and chooses a specific one afterwards for reporting.

- Like the F-test for the homogeneity of variances, all tests previously described are sensitive to the assumption that the data are drawn from underlying Gaussian distributions. This assumption should be checked by diagnostic plots. There are various more robust (and less efficient) competitors to the F-test available (see, e.g., Marozzi (2011)), but combination of these tests with tests for equality of location are not straightforward since the test statistics are not independent. The same holds for combinations of nonparametric tests like the Lepage test (Lepage (1971), Marozzi (2013)). There also exists nonparametric location-scale tests like the Cucconi rank test (Cucconi, 1968) which are not combination tests. Marozzi (2009) shows that the Cucconi test is a powerful alternative to the Lepage test and suggests to carry out the test as permutation test. Clearly, such tests can be preferable in specific applications.

- It is certainly possible to cover one or more of the presented methods in classroom. At least, it should be made clear that the combination of the t- and the F-test using a Bonferroni correction leads to a valid test for H_0 against H_1 at level α.

  If one accepts (or proves) the independence of the tests, it is possible to discuss the more refined combination methods. Such a treatment accentuates the randomness of p-values, an important fact which is often obscured in classroom (Murdoch et al., 2008).

  Determining the likelihood ratio statistic in (2) is a worthwhile exercise, while a more or less sophisticated implementation of the likelihood ratio test of Muirhead in Corollary 2.1 is an interesting task for an accompanying statistical computing lab.

- One caveat: strictly speaking, none of the tests is a diagnostic test, insofar as it is not possible to deduce differences in means or variances from a rejection of the overall hypothesis (this would be possible using Welch's t-test and the F-test with

Bonferroni correction). However, nothing speaks against an informal approach as in Section 3 of Zhang et al. (2012).

Since the minimum methods corresponds to the Bonferroni correction, and since the t-test is robust against violations of variance homogeneity, the minimum methods is closest to a diagnostic test.

Acknowledgement. The authors thank the Editor and two anonymous referees for their valuable comments on the original version of the manuscript.

References

Aspin, A.A., and Welch, B.L., 1949. *Tables for use in comparisons whose accuracy involves two variances, separately estimated.* Biometrika 36, 290-296.

Berk, R.H., and Cohen, A., 1979. *Asymptotically Optimal Methods of Combining Tests*, Journal of the American Statistical Association 74, 812-814.

Bickel, P.J., and Doksum, K.A., 2006. *Mathematical Statistics, Basic Ideas and Selected Topics, Vol. 1*, Pearson, 2nd edition.

Cucconi, O., 1968. *Un nuovo test non parametrico per il confronto tra due gruppi campionari*, Giornale degli Economisti XXVII, 225-248.

Deary, I.J., Irwing, P., Der, G., and Bates, T.C., 2007. *Brother-sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979*, Intelligence 35, 451-456.

Edington, E.S., 1972. *An additive method for combining probability values from independent experiments*, The Journal of Psychology 80, 351-363.

Fisher, R.A., 1932. *Statistical Methods for Research Workers*, Oliver & Boyd, 4th edition.

The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice, Gastwirth, J.L., Gel, Y.R., and Miao, W., 2009, Statistical Science, 24, 343-360.

George, E.O., and Mudholkar, G.S., 1983. *On the convolution of logistic random variables* Metrika 30, 1-13.

Hogg, R.V., McKean, J.W., and Craig, A.T., 2005. *Introduction to Mathematical Statistics*, Pearson Education, 6th edition.

Hsieh, H.K., 1979. *On asymptotic optimality of likelihood ratio tests for multivariate normal distributions*, Ann. Statist. 7, 592-598.

Jain, S.K, Rathie, P.N., and Shah, M.C., 1975. *The exact distributions of certain likelihood ratio criteria.* Sankhya Ser.A. 37, 150-163.

Lehmann, E.L., and Romano, J.P., 2005. *Testing Statistical Hypotheses*, Springer, 3rd edition.

Lepage, Y., 1971. *A combination of Wilcoxon's and Ansari-Bradley's statistics*, Biometrika 58, 213-217.

Lipták, T., 1958. *On the combinationn of independent tests*, Magyar Tudományos Akadémia Matematikai Kuatató Intezetenek Kozlemenyei 3, 1971-1977.

Loughin, T.M., 2004. *A systematic comparison of methods for combining p-values from independent tests*, Computational Statistics & Data Analysis 47, 467-485.

Marozzi, M., 2009. *Some notes on the location-scale Cucconi test*, Journal of Nonparametric Statistics 21, 629-647.

Marozzi, M., 2011. *Levene type tests for the ratio of two scales*, Journal of Statistical Computation and Simulation 81, 815-826.

Marozzi, M., 2012. *A combined test for differences in scale based on the interquantile range*, Stat. Papers 53, 61-72.

Marozzi, M., 2013. *Nonparametric simultaneous tests for location and scale testing: A comparison of several methods*, Communications in Statistics - Simulation and Computation, 42, 1298-1317.

Mudholkar, G.S., and George, E.O., 1979. *The logit statistic for combining probabilities*, in: Rustagi, J. (Ed.), Symposium on Optimizing Methods in Statistics, 345-366, Academic Press.

Muirhead, R.J., 1982. *On the Distribution of the Likelihood Ratio Test of Equality of Normal Populations*, The Canadian Journal of Statistics 10, 59-62.

Murdoch, D.J., Tsai, Y., and Adcock, J., 2008. *P-Values are Random Variables*, The American Statistician 62, 242-245.

Nagar, D. K., and Gupta, A. K., 2004. *Percentage Points for Testing Homogeneity of Several Univariate Gaussian Populations*, Applied Mathematics and Computation 156, 551-561.

Nair, V.N., 1984. *On the behaviour of some estimators from probability plots*, Journal of the American Statistical Association 79, 823-830.

Pearson, E. S., and Neyman, J., 1930. *On the Problem of Two Samples*, in: Joint Statistical Papers, eds. J. Neyman and E. S. Pearson, Cambridge University Press, pp. 99-115, 1967.

Perng, S.K., and Littell, R.C., 1976. *A Test of Equality of Two Normal Population Means and Variances*, Journal of the American Statistical Association 71, 968-971.

Pesarin F., and Salmaso L., 2010. *Permutation Tests for Complex Data: Theory, Applications and Software*, Wiley.

Shoemaker, L.H., 1999. *Interquantile tests for dispersion in skewed distributions*, Commun. Stat. Simul. Comput. 28, 189-205.

Singh, N., 1986. *A Simple and Asymptotically Optimal Test for the Equality of Normal Populations: A Pragmatic Approach to One-Way Classification*, Journal of the American Statistical Association 81, 703-704.

Steinmayr, R., Beauducel, A., and Spinath, B., 2010. *Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied?*, Intelligence 38, 101-110.

Stouffer, S., Suchman, E., DeVinnery, L., Star, S., and Williams, R., 1949. *The American Soldier, volume I: Adjustment during Army Life*, Princeton University Press.

Tippett, L.H.C., 1931. *The Method of Statistics*, Williams and Norgate.

Zhang, L., Xu, X., and Chen, G., 2012. *The Exact Likelihood Ratio Test for Equality of Two Normal Populations*, The American Statistician 66, 180-184.