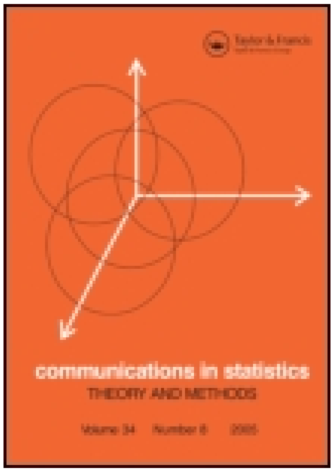

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

A Modified Hall-Padmanabhan Test for the Homogeneity of Scales

Marco Marozzi [a]

[a] Dipartimento di Economia e Statistica , Universitá della Calabria , Rende , Italy

Published online: 02 Aug 2012.

Communications in Statistics—Theory and Methods, 41: 3068–3078, 2012

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2011.606483

A Modified Hall-Padmanabhan Test for the Homogeneity of Scales

MARCO MAROZZI

Dipartimento di Economia e Statistica, Universitá della Calabria, Rende, Italy

We consider a simple adaptive test for the homogeneity of scales due to Hall and Padmanabhan (1997) which is based on the ratio of adaptively trimmed sample variances. We develop a modification of this test that has the form of a combined bootstrap test and that is obtained within the nonparametric combination of dependent tests framework (Pesarin, 2001). We considered other approaches for the combination of tests concluding that the Liptak method is the most suitable one for the problem at hand. We compare the modified test with the original one in terms of robustness of significance level and power in a simulation study which considers distributions that range from symmetric to skewed and from light to heavy tailed ones. We show that the modified test is more powerful than the original one under heavier than normal tailed and very skewed distributions. A practical application to detect difference in scale of energy intake of lactating and non pregnant and not lactating women is presented.

Keywords Nonparametric combination; Nonparametric testing; Resampling; Scale testing.

Mathematics Subject Classification 62G10; 62G09; 62P10.

1. Introduction

The practitioner might be interested in whether two samples differ in scale on some response variable for numerous reasons. For example, in analytical chemistry an appropriate scale test should be used to evaluate significant differences in measurement precision or as a preliminary to equivalence testing (Limentani et al., 2005) taking into account the multiplicity issue while performing preliminary tests to the other ones. Scale equality is also an important assumption underlying the pooled variance t test for the two sample location problem. It is well known that if this assumption is not fulfilled then the effects on the type-one error rate of the t test may be negative. Whether two samples differ in scale may also be of interest by itself, for example in quality control and in analysis of outer continental

Received November 24, 2010; Accepted July 15, 2011
Address correspondence to Marco Marozzi, Dipartimento di Economia e Statistica, Universitá della Calabria, Via P. Bucci Cubo 0c, Rende 87036, Italy; E-mail: marozzi@unical.it

shelf bidding on oil and gas (Conover et al., 1981). It is important to underline that the practitioner, before testing scale equality with a particular test, should address two questions: whether the test is robust in size and how powerful it is. We define a test for scale comparison to be robust in size if its maximum estimated significance level (MESL) does not exceed .075 when the nominal significance level is .05. Marozzi (2011) suggested this criterion which is more stringent than Conover et al. (1981) criterion of not exceeding .1. The literature shows that there is no test which is the uniformly most powerful one for the homogeneity of scales. According to Conover et al. (1981), a test which shows good overall performance in terms of power and robustness against non normality is the modified Levene $W50$ test. However, Pan (1999) showed that the $W50$ test has a structural problem that arises in certain situations where its power never reach 1, no matter how different the scales are. Recently, Marozzi (2011) compared ten Levene type tests, including some resampling tests never considered before, as well as the F test, the modified Fligner and Killeen (1976) test, an adaptive test due to Hall and Padmanabhan (1997), and two tests due to Shoemaker (1995, 1999). In light of his findings, the computationally simplest robust test is $W50$, and higher power, while preserving robustness, is achieved by considering some resampling Levene type tests.

Along with the test considered in Marozzi (2011), Hall and Padmanabhan (1997) proposed another adaptive test, much simpler than the other one, that is based on the ratio of adaptively trimmed sample variances. Let $\underline{X}_j = (X_{1j}, \ldots, X_{n_j j})$ denote independent samples from distributions F_j, where $F_j(u) = F((u - \mu_j)/\sigma_j)$, $j = 1, 2$, F is a distribution with location 0 and scale 1, with μ_j e σ_j unknown. We would like to test

$$H_0 : \sigma_1 = \sigma_2 \quad \text{against} \quad H_1 : \sigma_1 > \sigma_2.$$

Following Hall and Padmanabhan (1997), we define the β trimmed V_β ratio as

$$V_\beta = S^2_{1,\beta}/S^2_{2,\beta},$$

where for $0 < \beta < .5$, $S^2_{j,\beta}$ denotes the sample variances of the β trimmed $\underline{X}_j$ sample, $j = 1, 2$. The p-value of the test based on V_β is estimated via the bootstrap by drawing with replacement (Boos and Brownie, 1989) from $\underline{Y} = (X_{ij} - \overline{X}_j, i = 1, \ldots, n_j, j = 1, 2)$, where $\overline{X}_j$ is the mean of the jth sample, D couples of samples of size n_1 and n_2, respectively, and computing V_β for each couple obtaining ${}_dBV_\beta$, $d = 1, \ldots, D$. Let ${}_0BV_\beta$ be the observed value of the test statistic, i.e., the V_β statistic computed on $\underline{Y}$. The BV_β test suggests to reject H_0 if

$$M_{BV_\beta}\left({}_0BV_\beta\right) = \frac{1}{D}\sum_{d=1}^{D} I\left({}_dBV_\beta > {}_0BV_\beta\right) < \alpha,$$

where $M_{BV_\beta}\left({}_0BV_\beta\right)$ denotes the bootstrap p-value. The adaptive Hall and Padmanabhan HP test statistic is defined as

$$HP = \begin{cases} BV_{.10} & \text{when } W < 2.5 \\ BV_{.25} & \text{otherwise} \end{cases},$$

where $W = \frac{n_1W_1+n_2W_2}{n_1+n_2}$, and W_j is the Hogg et al. (1975) tailweight measure computed on the $\underline{X}_j$ sample, $j = 1, 2$

$$W_j = \frac{\overline{U}_{j,.05} - \overline{L}_{j,.05}}{\overline{U}_{j,.5} - \overline{L}_{j,.5}},$$

where $\overline{U}_{j,\gamma}$ and $\overline{L}_{j,\gamma}$ are, respectively, the averages of the largest and smallest $\gamma 100\%$ elements of the $\underline{X}_j$ sample, $\gamma = .05$ and $.5$. Note that fractional items are used in these averages, for example if $n_j = 30$ then $.05n_j = 1.5$, $\overline{L}_{j,.05} = (X_{(1)j} + 0.5X_{(2)j})/1.5$ and $\overline{U}_{j,.05} = (X_{(n_j)j} + 0.5X_{(n_j-1)j})/1.5$, where $X_{(i)j}$ is the ith order statistic of the $\underline{X}_j$ sample.

In this article, within the nonparametric combination of dependent tests framework (Pesarin, 2001), we develop in Sec. 2 a modification of the original test by Hall and Padmanabhan for the homogeneity of scales. This modification has the form of a combined resampling test. We consider other approaches for the combination of tests concluding that the Liptak method is the most suitable one for the problem at hand. To evaluate how this test performs in comparing scale parameters, we carry out in Sec. 3 a simulation study which considers a range of distributions from symmetric to skewed, and from light- to heavy-tailed ones. We show that the modified Hall-Padmanabhan test is more powerful than the original one under heavier than normal-tailed distributions. In Sec. 4, we apply the tests to detect difference in scale of energy intake of lactating women and non pregnant and not lactacting women in the Third World.

2. A Modification of the Hall and Padmanabhan Test

Pesarin (2001) proposed the nonparametric combination of several dependent tests as a general framework for complex hypothesis testing when a single test is not directly available, or difficult to find and/or justify. When many parameters are involved, or different aspects are of interest, a two-step approach is suggested which is often referred to as the Multi-Aspect procedure (Pesarin, 2001).

Step 1. Analyze the data set by means of a finite number of different tests, each being used to assess a particular aspect of interest and called for this reason partial test.

Step 2. Take into account all available information through the combination of the partial tests into a combined test which assesses the global problem.

Combinations of p-values can be used to address situations when either the observed values of test statistics that need to be combined or their distributions are unknown, or this information is available but the distributions are such that a method to obtain a single global test is not available (Birnbaum, 1954). It should be also noted that combinations of p-values allow the design of complex experiments by giving the possibility to weigh the individual tests according to how informative they are or might be. Combinations of p-values are generally used for multi-stage analyses in which inferences are combined using the same statistic from different samples and for multi-phase analyses in which inferences are combined using different statistics from the same sample, as in genetic epidemiology (Won et al., 2009). If the partial test statistics were independent, the combined test can be easily obtained

(Loughin, 2004). In practice, however, this hypothesis is generally not fulfilled because often the test statistics are functions of the same data set. Moreover, the underlying dependence structure among the partial tests is generally unknown or, when known, it is often difficult to be handled. Therefore, Pesarin (2001) proposed a procedure which is nonparametric with respect to the underlying dependence relations among partial tests. It is generally difficult to act within the parametric framework when testing for restricted alternatives or when dealing with categorical or mixed variables. Moreover, there exist problems that are impossible to be solved within a parametric framework (Pesarin and Salmaso, 2010). An example is when testing for multivariate paired observations when some data are missing (unless missing data are missing completely at random). For a non theoretical introduction to this approach, see Marozzi (2002). Even if we are not facing a complex hypothesis testing problem, we can use the framework of the nonparametric combination because it is very flexible and led us to develop a modification of the *HP* test. In different contexts, the usefulness of a combined test is shown, for example, in Marozzi (2004a,b, 2007) and Salmaso and Solari (2005).

In our case, as partial tests we consider $BV_{.10}$ and $BV_{.25}$. The computation of $M_{BV_{.10}}$ and $M_{BV_{.25}}$ accomplishes the first step of the combination. Note that the bootstrap resamplings are carried out in a synchronized way, that is the $BV_{.10}$ and $BV_{.25}$ statistic are computed at each d iteration of the bootstrap procedure on the same resampling of the data set, in order to maintain the dependence between the partial tests, which is implicitly used while performing the second step of nonparametric combination of the partial tests. In the second step, the modified Hall-Padmanabhan test statistic is obtained by nonparametric combination of $BV_{.10}$ and $BV_{.25}$. Partial p-values rather than partial test statistics are combined because they are one to one decreasingly related. Let φ be a proper combining function (see Pesarin, 2001), the modified MHP_φ statistic is

$$MHP_\varphi = \varphi(M_{BV_{.10}}, M_{BV_{.25}}).$$

The observed value of MHP_φ is estimated as ${}_0MHP_\varphi = \varphi(M_{BV_{.10}}({}_0BV_{.10}), M_{BV_{.25}}({}_0BV_{.25}))$ and its distribution is estimated using the bootstrap samples of the first step. For example, the dth bootstrap value of MHP_φ is computed as ${}_dMHP_\varphi = \varphi(M_{BV_{.10}}({}_dBV_{.10}), M_{BV_{.25}}({}_dBV_{.25}))$. Since large values of the observed test statistic are evidence against the null hypothesis, then H_0 is rejected if $M_{MHP_\varphi}({}_0MHP) = \frac{1}{D}\sum_{d=1}^{D} I({}_dMHP_\varphi > {}_0MHP_\varphi) < \alpha$. The most familiar combining functions are:

(i) the Fisher combining function $\varphi_F = \ln\left(\frac{1}{M_{BV_{.10}}}\right) + \ln\left(\frac{1}{M_{BV_{.25}}}\right)$;
(ii) the Liptak combining function $\varphi_L = \Phi^{-1}\left(1 - M_{BV_{.10}}\right) + \Phi^{-1}\left(1 - M_{BV_{.25}}\right)$, where $\Phi(\cdot)$ denotes the standard normal distribution function;
(iii) the Tippett combining function $\varphi_T = \max\left(1 - M_{BV_{.10}}, 1 - M_{BV_{.25}}\right)$.

The optimal combination of p-values is the one that gives rise to the uniformly most powerful global test. However, Birnbaum (1954) showed that since the most powerful test is different according to the situation it does not exist. Won et al. (2009) derived the most powerful global test for a simple alternative hypothesis when the practitioner can specify the expected effect size for each alternative as well as an approximation of the most powerful test if only the effect size ratios are available. Firstly, we emphasize that their method is for independent tests with underlying normal distributions. Secondly, we emphasize that information about the expected effect sizes is necessary. Moreover, it has been shown that their method gives rise to the most powerful test, but its power may not be substantially larger than that of other methods (Won et al., 2009, p. 1551). Therefore, we do not follow this approach because we do not assume normality, we have strongly dependent tests and we do not assume to have a priori information on the scale change. Zaykin et al. (2002) modified the Fisher combination by considering only those p-values less than some specified cut-off point, and by extending the procedure to situations when the tests are correlated. Their truncated product method for combining p-values is especially useful when dealing with microarray or genome-wide scan data (when the p-values to be combined may be more than 1,000). For our problem we do not use this method because we have only two p-values and therefore it is somewhat the same as applying the original Fisher method (which we will consider here).

There exist situations in which one combining function is preferable to the others (Pesarin, 2001). The combining function might be chosen to have good power against alternatives that might be of interest for, or be expected to arise by, the practitioner, provided that he/she has this knowledge. For example, the Tippett combining function has a good power behavior when only one among the partial null hypotheses is false. The Liptak combining function is good when the partial null hypotheses are jointly false. The Fisher combining function has an intermediate behavior and then it is useful when nothing is expected about the partial null hypotheses. Loughin (2004) compared heuristically and through simulation several combining functions for independent p-values and made similar suggestions. Note that in our case, the decomposition of the global null hypothesis is purely "fictitious" being $H_0 = H_{0a} \cap H_{0b}$, where $H_{0a} = H_{0b} = H_0$. Therefore, we expect that the Liptak function is the most suitable one for the problem at hand. However, for the purpose of comparison we also consider as well the other two combining functions. Let MHP_F, MHP_L, and MHP_T denote the modified Hall-Padmanabhan test based on the Fisher, Liptak, and Tippett combining function, respectively. It is important to note that different combining functions give rise to different global p-values but the corresponding tests are asymptotically equivalent in the alternative (Pesarin and Salmaso, 2010). To by-pass the problem of which combining function choose, an iterated strategy may be followed by applying a certain number of combining functions to the partial tests. The second-stage global p-values are combined again into a third-stage global p-value. According to Pesarin and Salmaso (2010), the third-stage global p-value is almost invariant with respect to the third-stage combining function. We do not iterate the combination because in our case it is clear that the Liptak combining function is likely the most suitable one. In general, however, instead of trying to guess the study results in advance (and then select the corresponding best combining function) combination of combinations might be considered. Note that in our case it makes no sense to use a weighted combining function since the two partial null hypotheses are identical. However, it is important to emphasize that in some contexts (like quality control and clinical trials) the partial hypotheses may have different degrees of importance and the nonparametric combination of dependent test framework is more useful than the parametric one because weighted testing within the latter is practically very difficult, especially when following the likelihood ratio principle (Pesarin, 2001).

3. The Monte Carlo Study

To evaluate how the modified Hall-Padmanabhan tests perform in detecting scale changes, we carried out a Monte Carlo study. Random samples varying in size were taken from two populations distributed as either standard normal, uniform between $-\sqrt{3}$ and $\sqrt{3}$, Laplace double exponential with scale parameter of $1/\sqrt{2}$, Student's t with 2 df, chi-squared with 8 df and scale parameter of .25, exponential with scale parameter of one, and generalized Pareto distribution with 0 location, 1 scale, and .25 shape parameter. We used `R` and the functions `rnorm, runif, rt, rchisq, rexp`, and `rgpd` (the last one is included in the `EVD` contributed package, whereas the other ones are included in the base distribution of `R`) respectively for generating samples from the normal, uniform, t, chi squared, exponential, and Pareto distribution. The double exponential was generated by sampling from an exponential distribution and then randomly assigning a positive or negative sign to each observation in the sample. The distributions range from symmetric to skewed and from light-to heavy-tailed ones. The null hypothesis tested was that the two populations had equal scales against the directional alternative hypothesis that the first population has a greater scale. After generating the two samples under the null hypothesis (i.e., with equal scales), two alternative hypotheses have been simulated by multiplying the first sample elements respectively by 2 and 3. The sample size configurations were $(n_1, n_2) = (10, 10)$, $(10, 20)$, $(20, 10)$, and $(20, 20)$. For each condition, we drew 10,000 couples of samples and 1,000 resamplings were considered for the bootstrap tests.

The proportion of rejection (at $\alpha = .05$) of the null hypothesis is reported in Tables 1–4. As it can be seen, both the original and the modified Hall-Padmanabhan

Table 1
Type I error rate and power estimates for $(n_1, n_2) = (10, 10)$

	Norm	Unif	D. Exp	t_2	Chisq	Exp	Pareto
σ_1/σ_2	1	1	1	1	1	1	1
	2	2	2	2	2	2	2
	3	3	3	3	3	3	3
F	0,049	0,015	0,122	0,213	0,083	0,167	0,236
	0,628	0,684					
	0,930	0,978					
HP	0,033	0,032	0,039	0,032	0,038	0,052	0,058
	0,335	0,429	0,260	0,221	0,318	0,279	0,242
	0,639	0,755	0,479	0,421	0,586	0,508	0,410
MHP_L	0,032	0,031	0,040	0,034	0,037	0,052	0,060
	0,315	0,378	0,273	0,245	0,313	0,279	0,257
	0,620	0,698	0,511	0,481	0,575	0,509	0,443
MHP_F	0,030	0,030	0,038	0,032	0,035	0,050	0,059
	0,308	0,367	0,264	0,234	0,302	0,272	0,251
	0,609	0,688	0,499	0,465	0,564	0,499	0,433
MHP_T	0,023	0,023	0,028	0,023	0,027	0,038	0,046
	0,249	0,311	0,206	0,178	0,250	0,229	0,213
	0,532	0,638	0,412	0,365	0,486	0,433	0,376

Table 2
Type I error rate and power estimates for $(n_1, n_2) = (10, 20)$

	Norm	Unif	D. Exp	t_2	Chisq	Exp	Pareto
σ_1/σ_2	1	1	1	1	1	1	1
	2	2	2	2	2	2	2
	3	3	3	3	3	3	3
F	0,049	0,007	0,113	0,207	0,086	0,152	0,214
	0,774	0,878					
	0,977	0,993					
HP	0,038	0,036	0,042	0,036	0,045	0,048	0,050
	0,503	0,693	0,328	0,258	0,454	0,328	0,235
	0,797	0,916	0,600	0,497	0,749	0,575	0,421
MHP_L	0,039	0,034	0,047	0,043	0,044	0,050	0,053
	0,496	0,594	0,402	0,333	0,452	0,346	0,276
	0,807	0,874	0,705	0,625	0,767	0,616	0,505
MHP_F	0,036	0,033	0,045	0,040	0,042	0,048	0,052
	0,488	0,593	0,391	0,322	0,445	0,339	0,269
	0,802	0,873	0,695	0,607	0,760	0,609	0,496
MHP_T	0,025	0,026	0,031	0,029	0,032	0,038	0,040
	0,441	0,592	0,328	0,243	0,394	0,285	0,223
	0,755	0,862	0,610	0,488	0,706	0,541	0,421

Table 3
Type I error rate and power estimates for $(n_1, n_2) = (20, 10)$

	Norm	Unif	D. Exp	t_2	Chisq	Exp	Pareto
σ_1/σ_2	1	1	1	1	1	1	1
	2	2	2	2	2	2	2
	3	3	3	3	3	3	3
F	0,047	0,014	0,134	0,269	0,082	0,175	0,292
	0,730	0,806					
	0,979	0,999					
HP	0,033	0,035	0,038	0,039	0,042	0,055	0,069
	0,426	0,564	0,285	0,262	0,390	0,353	0,307
	0,758	0,912	0,552	0,536	0,724	0,619	0,539
MHP_L	0,034	0,037	0,041	0,038	0,040	0,056	0,074
	0,417	0,492	0,345	0,324	0,394	0,376	0,345
	0,765	0,855	0,640	0,638	0,741	0,662	0,601
MHP_F	0,033	0,037	0,040	0,036	0,039	0,055	0,072
	0,408	0,484	0,334	0,312	0,383	0,369	0,337
	0,758	0,852	0,627	0,621	0,733	0,656	0,593
MHP_T	0,027	0,032	0,031	0,027	0,032	0,047	0,059
	0,347	0,436	0,269	0,242	0,326	0,324	0,286
	0,703	0,831	0,555	0,512	0,675	0,595	0,531

Table 4
Type I error rate and power estimates for $(n_1, n_2) = (20, 20)$

	Norm	Unif	D. Exp	t_2	Chisq	Exp	Pareto
σ_1/σ_2	1	1	1	1	1	1	1
	2	2	2	2	2	2	2
	3	3	3	3	3	3	3
	0,055	0,008	0,134	0,259	0,095	0,172	0,267
F	0,905	0,974					
	0,999	1,000					
HP	0,039	0,040	0,041	0,034	0,048	0,051	0,053
	0,618	0,849	0,393	0,380	0,567	0,422	0,323
	0,878	0,988	0,701	0,694	0,850	0,710	0,564
MHP_L	0,041	0,040	0,045	0,041	0,048	0,052	0,058
	0,633	0,729	0,519	0,497	0,587	0,467	0,404
	0,927	0,963	0,846	0,830	0,896	0,795	0,700
MHP_F	0,039	0,039	0,043	0,039	0,046	0,051	0,055
	0,631	0,734	0,508	0,483	0,580	0,457	0,397
	0,927	0,966	0,840	0,822	0,895	0,790	0,692
MHP_T	0,032	0,030	0,034	0,028	0,038	0,040	0,047
	0,604	0,761	0,445	0,397	0,547	0,415	0,350
	0,922	0,976	0,792	0,739	0,879	0,751	0,639

tests are robust in size because their MESL do not exceed .075 being .069, .074, .072, and .059 for the HP, MHP_L, MHP_F, and MHP_T tests, respectively. It should be noted that the MHP_T test is the most conservative one, with a degree of conservativeness that decreases as skewness of the distribution increases. The other modified Hall-Padmanabhan tests and the original one have similar sizes and are quite conservative with the exception of the skewed distributions (in particular for the Pareto). The tables show that the F test is not robust in size for heavier than normal tailed and skewed distributions and therefore in these cases its power is meaningless and omitted in the tables. Without surprises, the F test shows good power under normality. The less powerful test among the modified Hall-Padmanabhan tests is MHP_T. The reason is that the combined test based on the Tippett function has a good power behavior when only one among the two partial null hypotheses is false. This is not our case because the partial null hypotheses are identical and therefore the situation in which the Tippett function gives rise to a competing power behavior is impossible. For the same reason, we note that among the modified Hall-Padmanabhan tests the MHP_L test is the most powerful one with the MHP_F test performing similarly. Under the normal and chi squared distribution, the MHP_L and the HP test perform similarly. Under the uniform distribution the HP test performs better than the MHP_L test. Under the double exponential, the t_2, the exponential and the Pareto distribution the MHP_L test performs better than the HP test; the unique exception is for the exponential distribution with $n_1 = n_2 = 10$ where they perform similarly. It should be underlined that the power of all tests increases as the sample sizes and the scale ratio increase.

We conclude that the modified MHP_L test should be preferred to the original one under heavier than normal tailed and very skewed distributions, but so does

Table 5
24-h energy expenditure (kj/d) of a group of lactating women and of a group of non pregnant and non lactating women

Lactating women							
9920	8581	9305	10765	8079	9046	7134	8736
10230	7121	8665	5167	8527	7791	8782	6883
Non pregnant and non lactating women							
5289	6209	6054	6665	6343	7699	5678	6954
6916	4770	5979	6305	6502	6113	6347	5657

not for lighter than normal tailed distributions. We caution that our conclusions are based on simulation results. Moreover, caution should be made in interpreting a test for scale difference when the location and the scale of the distribution are dependent (as for the exponential). In fact, in such situations a difference in scale is accompanied by a difference in location and therefore the alternative hypothesis concerns a location-scale change rather then a pure scale change.

4. Application to Energy Intake of Lactating Women

Many studies done in communities of Third World women reported energy intakes far below the FAO dietary recommendations, especially when lactating. It is not clear if this is due to an adaptation of energy metabolism during lactation or to an underestimation of the true energy intake in lactating women. Since food intake measurement may be unreliable to assess energy requirements, Frigerio et al. (1991) measured the combined 24-h energy expenditure (kj/d) for lactating women and their babies in a respiratory chamber (a indirect calorimeter). The study took place in Keneba, a rural Gambian village. Table 5 shows the 24-h energy expenditure of 16 lactating women and of 16 non pregnant and non lactating women (control group).

The comparison of the locations of the populations underlying the two groups is the main problem of the study. However, before assessing the main problem it should be evaluated if the scale equality assumption is fulfilled and then the multiplicity issue should be taken into account. We applied the original Hall-Padmanabhan test as well as the modified Hall-Padmanabhan tests. By estimating the p-values through 100,000 resamplings we obtain the following results: .09829, .05624, .05922, and .07386 for the HP, MHP_L, MHP_F and MHP_T, respectively. The tests are not significant at the $\alpha = .05$ nominal level, but since the p-values are less than .1, we suggest to test for the location problem by using a test which is robust against non equal scales (see, e.g., Wilcox, 2005) and to take into account the multiplicity issue. Note that MHP_T needs a higher level than MHP_L and MHP_F in order to reject H_0. This result might suggest a gain in power of the MHP_L and MHP_F tests with respect to MHP_T in accordance with the results found in the previous section.

5. Conclusion

We developed a modification of a simple adaptive test due to Hall and Padmanabhan (1997) for the homogeneity of scales. This modification has the form of a combined bootstrap test and has been obtained within the framework of the nonparametric combination of dependent tests (Pesarin, 2001). It should be noted that the uniformly most powerful combined test does not exist. We considered other approaches for p-value combination, but they assume stringent assumptions (e.g., independent tests with underlying normal distributions and/or a priori information on the scale change) and/or give rise to much more complex global tests and without substantial power improvement than simpler methods. We compared the modified test with the original one in terms of robustness of significance level (the ability to maintain the type-one error rate close to the nominal level of the test) and power (the ability to reject a false null hypothesis) in a simulation study. The simulation results led us to conclude that the modified Hall-Padmanabhan test based on the Liptak combining function should be preferred to the original one under heavier than normal-tailed and very skewed distributions. Generalizations require caution because our conclusions are based on simulation experiments.

References

Birnbaum, A. (1954). Combining independent tests of significance. *J. Amer. Statist. Assoc.* 49:559–575.

Boos, D. D., Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics* 31:69–82.

Conover, W. J., Johnson, M. E., Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* 23:351–361.

Fligner, M. A., Killeen, T. J. (1976). Distribution-free two-sample tests for scale. *J. Amer. Statist. Assoc.* 71:210–213.

Frigerio, C., Schutz, Y., Whitehead, R., Jequier, E. (1991). A new procedure to assess the energy requirements of lactation in Gambian women. *Amer. J. Clin. Nutrit.* 54:526–533.

Hall, P., Padmanabhan, A. R. (1997). Adaptive inference for the two-sample scale problem. *Technometrics* 23:351–361.

Hogg, R. V., Randles, R. H., Fisher, D. M. (1975). A two-sample adaptive distribution-free test. *J. Amer. Statist. Assoc.* 70:656–661.

Limentani, G. B., Ringo, M. C., Ye, F., Berquist, M. L., McSorley, E. O. (2005). Beyond the t-test: statistical equivalence testing. *Analyt. Chem.* 77:221A–226A.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computat. Statist. Data Anal.* 47:467–485.

Marozzi, M. (2002). Some notes on nonparametric inferences and permutation tests. *Metron* LX:141–153.

Marozzi, M. (2004a). A bi-aspect nonparametric test for the two-sample location problem. *Computat. Statist. Data Anal.* 44:639–648.

Marozzi, M. (2004b). A bi-aspect nonparametric test for the multi-sample location problem. *Computat. Statist. Data Anal.* 46:81–92.

Marozzi, M. (2007). Multivariate tri-aspect non-parametric testing. *J. Nonparametric Statist.* 19(6):269–282.

Marozzi, M. (2011). Levene type tests for the ratio of two scales. *J. Statist. Computat. Simul.* 87(7):815–826.

Pan, G. (1999). On a Levene type test for equality of two variances. *J. Statist. Computat. Simul.* 63:59–71.

Pesarin, F. (2001). *Multivariate Permutation Tests with Applications in Biostatistics*. Chichester: Wiley.

Pesarin, F., Salmaso, L. (2010). *Permutation Tests for Complex Data*. Chichester: Wiley.

Salmaso, L., Solari, A. (2005). Multiple aspect testing for case-control designs. *Metrika* 62:331–340.

Shoemaker, L. H. (1995). Tests for differences in dispersion based on quantiles. *Amer. Statist.* 49(2):179–182.

Shoemaker, L. H. (1999). Interquantile tests for dispersion in skewed distributions. *Commun. Statist. Simul. Computat.* 28:189–205.

Wilcox, R. (2005). *Robust Estimation and Hypothesis Testing*. New York: Academic Press.

Won, S., Morris, N., Lu, Q., Elston, R. C. (2009). Choosing an optimal method to combine *p*-values. *Statist. Med.* 28:1537–1553.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., Weir, B. S. (2002). Truncated product method for combining *p*-values. *Genet. Epidemiol.* 22:170–185.