Available online at www.sciencedirect.com
ScienceDirect

Statistical Methodology 3 (2006) 351–374

Statistical Methodology
Official Journal of the International Indian Statistical Association

www.elsevier.com/locate/stamet

# Two separate effects of variance heterogeneity on the validity and power of significance tests of location

Donald W. Zimmerman*

*Carleton University, Ottawa, Ontario, Canada*

Received 24 February 2005; received in revised form 26 September 2005; accepted 12 October 2005

**Abstract**

Heterogeneity of variances of treatment groups influences the validity and power of significance tests of location in two distinct ways. First, if sample sizes are unequal, the Type I error rate and power are depressed if a larger variance is associated with a larger sample size, and elevated if a larger variance is associated with a smaller sample size. This well-established effect, which occurs in $t$ and $F$ tests, and to a lesser degree in nonparametric rank tests, results from unequal contributions of pooled estimates of error variance in the computation of test statistics. It is observed in samples from normal distributions, as well as non-normal distributions of various shapes. Second, transformation of scores from skewed distributions with unequal variances to ranks produces differences in the means of the ranks assigned to the respective groups, even if the means of the initial groups are equal, and a subsequent inflation of Type I error rates and power. This effect occurs for all sample sizes, equal and unequal. For the $t$ test, the discrepancy diminishes, and for the Wilcoxon–Mann–Whitney test, it becomes larger, as sample size increases. The Welch separate-variance $t$ test overcomes the first effect but not the second. Because of interaction of these separate effects, the validity and power of both parametric and nonparametric tests performed on samples of any size from unknown distributions with possibly unequal variances can be distorted in unpredictable ways.

*Keywords:* Homogeneity of variance; Heterogeneity of variance; Student $t$ test; Wilcoxon–Mann–Whitney test; Van der Waerden test; Normal scores; Skewed distribution; Symmetric distribution; Nonparametric test; Type I error; Type II error; Power

It is well known that parametric tests of location, such as the Student $t$ test and the ANOVA $F$ test, depend on an assumption of homogeneity of variances of treatment groups. Violation of

* Tel.: +1 604 531 9313; fax: +1 604 535 5354.
*E-mail address:* dwzimm@telus.net.

1572-3127/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2005.10.002

the assumption alters statistical significance levels, especially when sample sizes are unequal. When a larger variance is associated with a smaller sample size, the probability of a Type I error exceeds the significance level, and when a larger variance is associated with a larger sample size, the probability of a Type I error falls below the significance level (see, for example, [5,8,15,23]). Investigators have reported changes in the Type I error rates of the Student *t* test when sample sizes are equal (e.g., [2,8,19,21]). These changes, although sometimes substantial, are relatively small compared to the much larger changes that occur when sample sizes differ.

It is also well known that nonparametric rank tests such as the Wilcoxon–Mann–Whitney test and Kruskal–Wallis test, protect the significance level for non-normal distributions and often have more power to detect differences than *t* and *F* tests. It is found that these significance tests are influenced by unequal variances combined with unequal sample sizes just like their parametric counterparts. Nevertheless, it turns out that the situation is more complicated in the case of nonparametric tests (see, for example, [3], 1972, [6,11,28]), and the properties of nonparametric methods under variance heterogeneity are still an unsettled issue.

This paper presents data indicating that heterogeneity of variance produces two distinct changes in Type I error rates and power of two significance tests based on ranks. The first kind occurs when unequal variances are combined with unequal sample sizes, and the second is observed even when sample sizes are equal. In the parametric case, the second effect is relatively small and diminishes as sample size increases, while in the nonparametric case it is substantial and becomes larger as sample size increases.

## 1. Method[1]

The random number generator used in this study was introduced by Marsaglia et al. [13] and has been described by Pashley [16, pp. 395–415]. Normal variates, $N(0, 1)$, were generated by the rejection method of Marsaglia and Bray [12] and were transformed to have various distribution shapes, using inverse distribution functions. Altogether, nine distribution shapes—six skewed and three symmetric—were included in the study. The six skewed distributions were: exponential, lognormal, chi-square, Gumbel (extreme value), power function, and mixed-normal. The three symmetric distributions were: normal, Laplace (double exponential), and logistic. The inverse distribution functions for generating the samples are given in Table A.1 in Appendix A.

Each replication consisted in obtaining two independent samples of $n_1$ and $n_2$ scores, respectively, from one of the distributions. All scores in one sample were multiplied by a constant, so that the ratio $\sigma_1/\sigma_2$ had a predetermined value. Throughout this paper, $\sigma_1$ denotes the standard deviation of the population from which the sample of $n_1$ scores was drawn and $\sigma_2$ that of the population from which the sample of $n_2$ scores was drawn. The sizes of each of the pairs of samples, $(n_1, n_2)$, were: (20,40), (30,30), (40,20), (40,80), (60,60), (80,40), (15,45), (45,15), (15,15), and (20,20). The ratios $\sigma_1/\sigma_2$ ranged from 1.0 to 2.5 in increments of 0.5 in some cases and from 1.0 to 3.0 in increments of 0.2 in others.

On each iteration, a two-sample Student *t* test was performed and evaluated at the 0.01 and 0.05 significance levels. Next, the large-sample normal-approximation form of the Wilcoxon–Mann–Whitney test was performed on the same scores at the same significance levels. Finally, a van der Waerden [24], or normal scores test, was performed on the same data and also evaluated at the same significance levels. For some distributions, entire power functions,

[1] The computer programs in this study were written in PowerBasic, version 3.5, PowerBasic, Inc., Carmel, CA. Listings of the programs can be obtained by writing to the author.