Robust tests on the equality of variances

by

Man-Po Lai

B.Sc. University of British Columbia, 1994

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF STATISTICS

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

1997

©Man-Po Lai, 1997

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Oct 15, 97

DE-6 (2/88)

Abstract

The classic F test for the hypothesis concerning the equality of two population variances is known to be non-robust. When we apply the classical F test to the non-normal samples, the actual size of the test can be different from its nominal level. Therefore, several robust alternatives have been introduced in the literature. In this thesis, I will present some of these alternatives, and illustrate their application with some examples. A new approach will also be introduced. The best feature of this method is that it seems to be able to overcome the adverse effect of outliers. A Monte Carlo study is used to compare the new test with the F test and the other methods. The results of this study are encouraging for the new test.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor, Dr. Ruben Zamer for introducing me such a interesting topic. Ruben was always willing to give his advice, ideas, and encouragement, which were much needed. Also, I would like to thank Dr. Paul Gustafson for his very useful comments and careful reading. Thanks also go to Daniel Ng for his invaluable help with latex, and various other problems along the way.

1 Introduction

It is a well-known fact that the two sample t-test is a reliable method to test the differences between population means because it is insensitive to the departures from normality in the populations. On the other hand, when testing the differences between population variances, the F test is known to be rather sensitive to the assumption of normality. As a result, it might be possible that the null hypothesis is rejected because of the fact that the random variables are not normally distributed rather than the fact that the variances are not equal. This chapter focuses on inferences about variances of two populations. Section 1.1 investigates the influence of non-normality on comparing the variation in two samples. Section 1.2 describes alternative robust methods which have been proposed to deal with the non-normality problem.

1.1 Non-normality

The classic F test was first proposed by Bartlett [1]. Unfortunately, the F test is very sensitive to the assumption that the underlying populations have normal distributions. Box [2] showed that when the underlying distributions are non-normal, this test can have an actual size several times larger than its nominal level of significance. To see the influence of non-normality on comparing the variation in two samples by a classical F test, we will look at the normally and non-normally distributed cases. Firstly, we will derive the asymptotic distribution of the classic F test statistic under the assumption of an underlying normal distribution. Secondly, we will investigate how this asymptotic distribution changes under departures from normality.

1.1.1 Normal Case

Let us consider a two sample problem. Let $y_{11}, ..., y_{1n_1}$ and $y_{21}, ..., y_{2n_2}$ be two independently distributed samples from the distributions $N(\mu, \sigma_1^2)$ and $N(\mu, \sigma_2^2)$ respectively. The asymptotic distribution of the test statistic in the classical F test will be derived, although the statistic has exact F distribution under the null hypothesis and normal assumption. We use the asymptotic distribution, because the distribution of the test statistic is hard to obtain when samples are non-normal. For simplicity, we assume first that $n_1 = n_2 = n$. The sample variances $S_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \overline{y_i})^2$, for $i = 1, 2$, are unbiased estimators of the corresponding population variances σ_i^2 for $i = 1, 2$ respectively, where $\overline{y_i}$ are the corresponding sample means. By the Central Limit Theorem,

$$\sqrt{n}\left[\begin{pmatrix} S_1^2 \\ S_2^2 \end{pmatrix} - \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}\right] \to N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right), \tag{1}$$

where

$$\Sigma = \begin{pmatrix} 2\sigma_1^4 & 0 \\ 0 & 2\sigma_2^4 \end{pmatrix}. \tag{2}$$

By the Delta Method,

$$\sqrt{n}\left(\frac{s_1}{s_2} - \frac{\sigma_1}{\sigma_2}\right) \to N\left(0, \nabla g' \Sigma \nabla g\right), \tag{3}$$

where $g(x, y) = \sqrt{x/y}$ and ∇g is the gradient of $g(x, y)$:

$$\nabla g = \begin{pmatrix} \frac{\partial}{\partial x} g(\sigma_1^2, \sigma_2^2) \\ \frac{\partial}{\partial y} g(\sigma_1^2, \sigma_2^2) \end{pmatrix} \tag{4}$$

$$\tag{5}$$

$$= \frac{1}{2}\begin{pmatrix} \sigma_1^{-1}\sigma_2^{-1} \\ -\sigma_1\sigma_2^{-3} \end{pmatrix}. \tag{6}$$

Therefore,

$$\nabla g' \Sigma \nabla g = \frac{\sigma_1^2}{\sigma_2^2}. \tag{7}$$

If the null hypothesis, $\mathrm{H}_0 : \sigma_1 = \sigma_2$, is true, and according to the equation (3)

$$\nabla g' \Sigma \nabla g = 1$$

and

$$T = \sqrt{n} \left(\frac{S_1}{S_2} - 1 \right) \rightarrow N(0,1). \tag{8}$$

So, we can use T to test the two-sided H_0, and would reject H_0 when T exceeds the upper $100(\alpha/2)$ percentile or falls below the lower $100(\alpha/2)$ percentile of the $N(0,1)$ distribution. Thus, H_0 is rejected when $|T| > z(1-\alpha/2)$. For instance, if $\alpha = 0.05$, then H_0 is rejected, when $|T|$ is greater than 1.96.

For the unequal sample size case, if $\frac{n_1}{n_2} \rightarrow d$, then

$$\sqrt{n_1 + n_2} \left(\frac{S_1}{S_2} - 1 \right) \rightarrow N \left(0, \frac{(1+d)^2}{2d} \right). \tag{9}$$

1.1.2 Non-normal Case

The method described in the last section is based on the assumption of normality. To see how this method is sensitive to departures from normality, we will look at the cases that the population of the variables follow other distributions: double exponential, t_5 ,t_{10}, χ^2_5,χ^2_{10}, and uniform. In addition, we will calculate their actual asymptotic significance levels.

Let us first look at the general case. If the observations $y_{11}, ..., y_{1n_1}$ and $y_{21}, ..., y_{2n_2}$ are independently distributed according to a general distribution $F(y)$, then

$$\mathrm{E}(S_i^2) \;=\; \sigma_i^2 \;, \tag{10}$$

$$\mathrm{Var}(S_i^2) \quad = \quad \sigma_i^4 \left(\frac{2}{n-1} + \frac{\gamma}{n} \right) \tag{11}$$

where

$$\gamma = \frac{\mathrm{E}(y-\mu)^4}{(\mathrm{E}(y-\mu)^2)^2} - 3. \tag{12}$$

γ is called the coefficient of kurtosis and measures the peakedness or flatness of the probability distribution function (pdf). For the normal case, $\gamma = 0$ and $\mathrm{Var}(S^2) = 2\sigma^4/(n-1)$. By the CLT,

$$\sqrt{n} \left[\begin{pmatrix} S_1^2 \\ S_2^2 \end{pmatrix} - \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} \right] \to N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right), \tag{13}$$

where

$$\Sigma = \begin{pmatrix} (2+\gamma)\sigma_1^4 & 0 \\ 0 & (2+\gamma)\sigma_2^4 \end{pmatrix}. \tag{14}$$

According to(4) and (14)

$$\nabla g' \Sigma \nabla g = \frac{2+\gamma}{2} \left(\frac{\sigma_1^2}{\sigma_2^2} \right). \tag{15}$$

By the Delta Method, if H_0 is true, we obtain

$$\sqrt{n} \left(\frac{S_1}{S_2} - 1 \right) \to N \left(0, \frac{2+\gamma}{2} \right), \tag{16}$$

and for the unequal sample size case, $\frac{n_1}{n_2} \to d$,

$$\sqrt{n_1+n_2} \left(\frac{S_1}{S_2} - 1 \right) \to N \left(0, \frac{(2+\gamma)(1+d)^2}{4d} \right). \tag{17}$$

If the normality assumption is met, $\gamma = 0$, so that equation 16 is equivalent to equation 8. However, for the non-normal cases, like t_5, γ won't be zero. So when we apply the classical F test to the non-normal samples, the actual size of the test would be different from its nominal level of significance, α. Table 1 displays the value of

distribution	γ	actual significance level
double exponential	3	0.215
t_5	6	0.327
t_{10}	1	0.110
χ_5^2	2.4	0.186
χ_{10}^2	1.2	0.121
uniform (a, b)	-1.2	0.002

Table 1: Actual asymptotic significance level of F test ($\alpha = 0.05$), with non-normal samples

γ and the actual significance level of the F test ($\alpha = 0.05$), for several non-normal distributions: double exponential,$t_5, t_{10}, \chi_5^2, \chi_{10}^2$, and uniform (a, b) . Note that the arguments in the uniform distribution do not affect the result since in this case γ is always equal to -1.2. Also, for a heavy-tailed distribution ($\gamma > 0$), the probability of rejecting H_0 exceeds 0.05; whereas, for a short-tailed distribution, the probability is less than 0.05.

1.2 Some Robust Methods

This section contains some discussion of other alternatives to the test based on T defined in (8). The six robust methods considered here are the Levene test [6], the Jacknife test [7], the Box test [2], the Box-Andersen test [3], the Moses test [9], and the Layard χ^2 test [5].

1.2.1 The Levene test

The idea of the Levene test [6] is to transform the original data y_{ij} into $z_{ij} = |y_{ij} - \overline{y_i}|$, $j = 1, ..., n_i$ for the two samples, $i = 1, 2$. Then, we just pretend that they are independently, identically, normal distributed under H_0, and use the usual t test on the two transformed samples: $z_{11}, ..., z_{1,n_1}$ and $z_{21}, ..., z_{2,n_2}$. Obviously, the z_{ij}'s do not satisfy the above assumptions. Normality is not met because the z_{ij}'s are absolute values. Independence is violated because of the common term $\overline{y_i}$ in each z_{ij}; also, they are not identically distributed unless $n_1 = n_2$. However, as mentioned at the beginning of this chapter, the t test is a reliable method to check the differences between means due to the fact that it is insensitive to non-normality. To apply the two sample t test we have a new statistic

$$T_l = (\bar{z}_1 - \bar{z}_2)/s$$

with

$$s = \left(\frac{var(z_1)}{n_1} + \frac{var(z_2)}{n_2} \right)^{1/2},$$

where $\bar{z}_1, \bar{z}_2, var(z_1)$, and $var(z_2)$ are the means and variances of the samples z_1 and z_2. Levene [6] showed that under the null hypothesis, the distribution of T_l can be approximated by a t distribution with degree of freedom

$$\frac{1}{\frac{c^2}{n_1-1} + \frac{(1-c)^2}{n_2-1}}$$

where

$$c = \frac{var(z_1)}{ns^2}.$$

For two side test, if $|T_l|$ is greater then $t_v(1-\alpha/2)$, where v is degree of freedom, H_0 would be rejected.

1.2.2 Modifications of Levene test

For skewed distributions, such as the χ^2 with 4 degrees of freedom (df), and heavy-tailed distributions, such as the Cauchy, the Levene test usually has too many rejections. That is, the actual rejection rate exceeds the nominal significance level.

For these settings, improved Levene-type procedures have been proposed by Brown and Forsythe [4] which modify the test statistic by replacing the central location $\overline{y_i}$ with more robust versions, such as the medians and the 10% trimmed means of the the two samples. Monte Carlo studies [4] show that all of these test statistics are robust for the very heavy tailed Cauchy distribution. For the $\chi^2(4)$ distribution, the statistics based on the median is robust but the 10% trimmed mean rejects too often. Usually the version based on the sample mean has the greatest power in situations when the three statistics are robust.

1.2.3 The Jacknife test

In [7], Miller proposed a procedure based on the Jacknife technique to test H_0 in the two-sample case. Let us first review the idea of the jacknife technique. Let θ be an unknown parameter, and let $(y_1, ..., y_N)$ be a sample of N independent observations with cumulative distribution function (cdf) G_θ. Suppose that we use $\hat{\theta}$ to estimate θ, and that the data is divided into n groups of size k. Let $\hat{\theta}_{-i}$, $i = 1, ..., n$, denote the estimation of θ obtained by deleting the i-th group and estimating θ from the $(n-1)k$ observations. Define $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$, and $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i$, $i = 1, ..., n$, then the statistics

$$(\tilde{\theta}-\theta)\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\theta}_i-\tilde{\theta})^2\right]^{-\frac{1}{2}} \tag{18}$$

should be approximately distributed as t with (n-1) df. The statistics 18 can be used to perform an approximate significance test on θ. To apply the jacknife technique to test $\mathrm{H}_0 : \ln\sigma_x^2 = \ln\sigma_y^2$ in the two-sample case, we first define

$$\begin{aligned}
\theta_x &= \ln\sigma_x^2\ , \quad \theta_y = \ln\sigma_y^2\ , \\
\hat{\theta}_x &= \ln S_x^2\ , \quad \hat{\theta}_y = \ln S_y^2\ , \\
{}_x\tilde{\theta}_i &= n_1 \ln S_x^2 - (n-1)\ln\ {}_xS_{-i}^2\ , \\
{}_y\tilde{\theta}_i &= n_2 \ln S_y^2 - (n-1)\ln\ {}_yS_{-i}^2\ ,
\end{aligned}$$

where n_i is the number of subsamples in the i^{th} sample. Since ${}_x\tilde{\theta}$ and ${}_y\tilde{\theta}$ are approximately independently distributed, Miller proposed to test H_0 by using a two sample t-test on the two samples: ${}_x\tilde{\theta}_1, ..., x\tilde{\theta}_{n_1}$, and ${}_y\tilde{\theta}_1, ...{}_y\tilde{\theta}_{n_2}$. To apply the two sample t test we have a new statistic

$$T_\theta = \frac{{}_x\bar{\tilde{\theta}} - {}_y\bar{\tilde{\theta}}}{s}$$

with

$$s = \left(\frac{var(\ {}_x\tilde{\theta})}{n_1} + \frac{var(\ {}_y\tilde{\theta})}{n_2}\right)^{1/2}$$

and ${}_x\bar{\tilde{\theta}}$, ${}_y\bar{\tilde{\theta}}, var(\ {}_x\tilde{\theta})$, and $var(\ {}_y\tilde{\theta})$ are the sample means and variances of the samples ${}_x\tilde{\theta}$ and ${}_y\tilde{\theta}$. He showed that under the null hypothesis, the distribution of T_θ can be approximated by a t distribution with degree of freedom

$$\frac{1}{\frac{c^2}{n_1-1}+\frac{(1-c)^2}{n_2-1}}$$

where

$$c = \frac{var(\,_x\tilde{\theta})}{ns^2}.$$

For the two side test, we first compute $|T_\theta|$. If $|T_\theta|$ is greater then $t_v(1-\alpha/2)$, we could reject H_0, and conclude that the two variances are different.

1.2.4 The Box test

The Box test [2] is the earliest robust test for equality of variances. For the two sample case, similar to Jacknife test, each sample is divided into subsamples of size $k(k>1)$. So there are n_1 subsamples for the fist sample $x_1, ..., x_{n_1}$,and n_2 subsamples for the second sample $y_1, ..., y_{n_2}$. Then $\ln S^2$ is obtained from each subsample. Let's define $G_{ij} = \ln S_{ij}^2$, $i = 1, 2$, and $j = 1, ..., n_i$. The G_{ij} are approximately distributed as $N\left[\ln\sigma_i^2, \frac{2}{m-1} + \frac{\gamma}{m}\right]$, and the Box procedure performs two sample t test on G_{ij} and to test H_0: $\ln\sigma_1^2 = \ln\sigma_2^2$. First, let's define $\bar{G_1}, \bar{G}2, var(G_1)$, and $var(G_1)$ as the sample means and variances of the two samples G_1 and G_2, and

$$T_G = \frac{\bar{G_1} - \bar{G_2}}{s}$$

with

$$s = \left(\frac{var(G_1)}{n_1} + \frac{var(G_2)}{n_2}\right)^{1/2}.$$

The null hypothesis can be approximated by a t distribution with degree of freedom

$$\frac{1}{\frac{c^2}{n_1-1} + \frac{(1-c)^2}{n_2-1}}$$

where

$$c = \frac{var(G_1)}{ns^2}.$$

For two sided test, if $|T_G|$ is greater then $t_v(1-\alpha/2)$, where v is degrees of freedom, H_0 would be rejected.

Also, Box suggested that the test statistics T_G will not have exactly a t distribution since $\ln S^2$ is not exactly normally distributed, but the level of significance should be closely approximate because of the robustness of the t statistics. The main disadvantage of the Box test is the loss of information in subdiving the samples, and different groups of the data within each sample have the potential to produce substantially different results.

1.2.5 The Moses test

The main idea of Moses test [9] is to apply the Wilcoxon two sample rank test to the value S^2 obtained from the subsamples as in the Box test. This method was studied in detail by Shorack [10]. Besides S^2, other measures of dispersion (e.g., the range, or the mean deviation about the sample mean) were also considered to be used in the subsamples. Moses pointed out that the following properties:(a) this test yields an exact significance level, and (b) the two population means can be left completely unspecified. However, like the Box test, this test still suffers from the loss of information due to the sample subdivision.

1.2.6 The Layard χ^2 test

Layard [5] suggested a χ^2 test statistic which is a function of the kurtosis γ. For large sample size n, the statistic approximately follows a $N[\ln \sigma^2, \tau^2]$ distribution, where $\tau^2 = 2 + [1 - (1/n)]\gamma$, and γ is the coefficient of kurtosis. Under H_0 the statistic

$$S = \sum(n_i - 1)\left[\ln S_i^2 - \frac{\sum(n_i-1)\ln S_i^2}{\sum(n_i-1)}\right]^2 / \tau^2$$

is asymptotically distributed like χ_1^2, and S_i^2 is the sample variance of the i_{th} sample.

However γ is unknown, so Layard suggested the use of

$$\hat{\gamma} = \frac{\sum(n_i)\sum\sum(X_{ij} - \overline{X}_i)^4}{[\sum\sum(X_{ij} - \overline{X}_i)^2]^2} - 3 \tag{19}$$

to estimate the kurtosis. Hence, we can use the estimate $\hat{\gamma}$ and base a test on $\hat{S} = \tau^2 S/\hat{\tau}^2$, where $\hat{\tau}^2 = 2 + [1 - \frac{1}{\bar{n}}]\hat{\gamma}$. If $\hat{S}$ exceeds the upper $100(\alpha/2)$ percentile or falls below the lower $100(\alpha/2)$ percentile of the χ_1^2 distribution, the null hypothesis would be rejected. Note that Layard [5] and Brown [4] have simulated sampling experiments which suggest that the χ^2 test compares favourably with Box test. A difficulty with this procedure is that quite large samples are needed to get a reasonable estimate of γ.

1.2.7 The Box-Andersen Test

Box and Andersen [3] applied permutation theory to construct an approximate robust test. The idea of this test is to adjust the degree of freedom for the statistic S_x/S_y , so that the mean and the variance of this distribution are equal to that under the permutation distribution.

Permutation theory assumes that the two samples have been randomly selected without replacement from $u_1 = y_{11}, ..., u_{n_1} = y_{1n_1}, u_{n_1+1} = y_{21}, ..., u_{n_1+n_2} = y_{2n_2}$, where $y_{ij} = x_{ij} - \mu_i$, and μ_i is the population mean of the i_{th} sample. For simplicity, μ_i's are assumed to be known. Each of the possible $\begin{pmatrix} n_1 + n_2 \\ n_1 \end{pmatrix}$ combinations is equally likely. Let

$$B = \frac{\sum_{j=1}^{n_1} y_{1j}^2}{\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}^2}.$$

The mean of B is the same under the normal and permutation distributions,

$$\mathrm{E}_N(B) = \mathrm{E}_P(B) = \frac{n_1}{N},$$

where $N = n_1 + n_2$. However, the variances differ. Under the normal distribution,

$$\mathrm{Var}_N(B) = \frac{2n_1n_2}{N^2(N+2)}.$$

Under the permutation distribution,

$$\mathrm{Var}_P(B) = \frac{2n_1n_2}{N^2(N+2)}\left[1 + \frac{1}{2}\left(\frac{N}{N-1}\right)(b_2 - 3)\right],$$

where

$$b_2 = \frac{(N+2)\sum_{i=1}^{2}\sum_{j=1}^{n_i} {y_{ij}}^4}{(\sum_{i=1}^{2}\sum_{j=1}^{n_i} {y_{ij}}^2)^2}.$$

By using new sample sizes, $\tilde{n_1}$, and $\tilde{n_2}$, we can make the two variances equal, where $\tilde{n_1} = dn_1$, $\tilde{n_2} = dn_2$, and

$$d = \left[1 + \frac{1}{2}\left(\frac{N+2}{N+2-b_2}\right)(b_2 - 3)\right]^2.$$

The mean of B is unchanged under this substitution. So, by redefinding the sample sizes, the normal theory distribution for B can be made to approximate the permutation distribution for B.

According to the discussion above, Shorack [10] suggested the following approximate Box-Andersen test. The test approximates the distribution of the usual F by an F distribution on degrees of freedom d_1, d_2, where

$$d_1 = \hat{d}\,(n_1 - 1) \text{ and } d_2 = \hat{d}\,(n_2 - 1)$$

with

$$\hat{d} = \left[1 + \frac{1}{2}(\hat{b}_2 - 3)\right]^{-1}$$

and

$$\hat{b}_2 = \frac{\left[\sum_{i=1}^{2} n_i\right]\left[\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^4\right]}{\left[\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2\right]^2}.$$

So, if the classic F statistic exceeds the upper $100(\alpha/2)$ percentile or falls below the lower $100(\alpha/2)$ percentile of the F_{d_1,d_2} distribution, the null hypothesis would be rejected.

1.3 Example

This section contains two examples, which are available on the internet at the address *http : //lib.stat.cmu.edu/DASL/allmethods.html.* The data file names are Clouds and Michelson.

1.3.1 The first example: Cloud

In the first example, clouds were randomly seeded or not with silver nitrate. Rainfall amounts were recorded from the clouds. The purpose of the experiment was to determine if cloud seeding increases rainfall. The side by side boxplots of the two logged variables Fig 1 indicate that the variances of the two groups are very similar after a log transformation.

To compare the significance levels of these six tests, two outliers, with the same value are added to the seeded sample, and the value of the outliers is increased until the results of these tests become steady. The side by side boxplots for each pair of samples are shown in Fig 2.

The results of these tests and the classic F test are displayed in Table 2. For the F, Levene, Layard, Jacknife, Box, Moses, and Box-Andersen tests, if the test result is 1 in the table, the test rejects H_0. For the Moses and Box tests, the test results may change due to different subsamples of the data within each sample. To see if these two tests are likely to reject the null hypothesis, for each pair of samples, each of these two tests is executed 100 times. The entries are the proportion of rejections.

As expected, the F test is very non-robust. It rejects H_0 as the two outliers 12 are added. In this example, of all the tests, the Moses and Box tests are less affected by the outliers. They do not reject the null hypothesis, even when the largest outliers 100 are added. In addition, the performance of the Box-Andersen test is quite good. The Levene test is not as good as the Box-Andersen test, but is better than the Layard test, and the Jacknife test is the worst one.

1.3.2 The second example: Michelson

In the Michelson's example, 100 determinations of the velocity of light in air using a modification of a method proposed by the French physicist Foucault. These measurements were grouped into five trials of 20 measurements each. The numbers are in km/sec, and have had 299,000 subtracted from them. The currently accepted 'true' velocity of light in vacuum is 299,792.5 km/sec. The side by side boxplots of the measurements in the first and fifth trials, Fig 3, reveal that their variances are very different.

To compare the power of the seven tests, one outlier is added to the sample with smaller sample variance, and the value of the outlier is increased until neither of these tests rejects H_0 . The results of these tests and the side by side boxplots of each

value of two outliers	F test	Levene test	Layard test	Jacknife test	Box test	Moses test	Box-Andersen test
no outlier	0	0	0	0	0.03	0.04	0
12	1	0	0	0	0	0.01	0
14	1	0	0	1	0.01	0.04	0
25	1	0	1	1	0	0.01	0
28	1	1	1	1	0	0.01	0
30	1	1	1	1	0	0.04	1
100	1	1	1	1	0	0	1

Table 2: Results of tests on variances for the Cloud data.

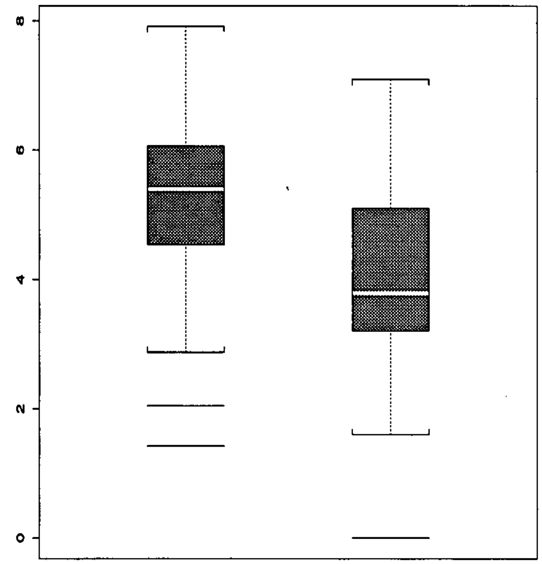

Figure 1: Side by Side Boxplots of the two logged variables in Cloud example

pair of samples are shown in Table 3 and Fig 4. Without the outlier, all tests except Box and Moses reject H_0, and these two tests have about 25% of results rejecting H_0. Hence, these two tests do not perform powerfully in this example. Surprisingly, the F test is not fooled by large outlier in this example. The Levene test is also very powerful. The Layard test is the worst. The Jacknife and Box-Andersen tests are about equally powerful.

According to the two examples, the power of the F test and the Jacknife test are not so affected by the outliers, but their significance levels are very sensitive to the outliers. The Layard test is not so powerful, but, in term of the significance

value of outlier	F test	Levene test	Layard test	Jacknife test	Box test	Moses test	Box-Andersen test
no outlier	1	1	1	1	0.28	0.25	1
950	1	1	0	1	0.21	0.25	1
980	1	1	0	0	0.13	0.12	0
1000	1	1	0	0	0.13	0.03	0
1100	0	0	0	0	0	0	0

Table 3: Results of tests on variance for the Michelson data.

level, it is better than the Jacknife test. The Levene test is the most powerful test in the Michelson's example, and its performance is better than the Layard test in the Cloud example. In addition, although, the Moses and Box tests are not affected by the largest outlier in the first example, they are not robust. They seem to be superior in the first example just because they are so conservative. Of all the tests, the Box-Andersen test is the best in these two examples.

2 A New Robust test

This chapter contains three sections. In the first section, a new robust method testing the equality of variances between two populations is presented. In the second section, the asymptotic distribution of the new test statistic described in the first section is derived. In the last section, the new method is applied to the two examples mentioned in the first chapter.

2.1 Robust Dispersion Estimates

First, an alternative measure of dispersion that is more resistant to outliers is introduced. The best feature of this new method is that it has superior ability to overcome the effect of outliers. This measure is insensitive to changes in the most extreme observations and therefore is resistant to outliers.

To start with, we just consider one sample, $x_1, ..., x_n$, with $x_i \sim N(\mu, \sigma^2)$, and x_i are independent. The alternative measure of dispersion, based on a sample $x_1, ..., x_n$, is called Sr. Notice that Sr satisfies the following equation

$$\sum_{i=1}^{n} \chi\left(\frac{x_i - T_n}{Sr}\right) = nb, \tag{20}$$

where T_n is the median of the sample. χ is defined as a function:

$$\chi(z) = \begin{cases} \frac{z^2}{c^2}\,, & \text{if } |z| \leq c; \\ 1\,, & \text{otherwise,} \end{cases} \tag{21}$$

where c is arbitrary. The value of b depends on the choice of c. To ensure consistency of Sr, we choose

$$b = E(\chi(z)), \tag{22}$$

with $z \sim N(0,1)$(i.e. $Sr \to \sigma$ as $n \to \infty$.) Observe that for $-c \leq z \leq c$, $\chi(z)$ equals the sample standard deviation score function.

For the two sample case, Sr_i is referred to as the new measure of dispersion in the i^{th} sample, $i = 1, 2$. The new test statistic for the H_0 will be based on the ratio

$$R = \frac{Sr_1}{Sr_2}. \tag{23}$$

The asymptotic distribution of R is derived in the next section.

In addition, Miller [8] also gave some references and mentioned the possibility of doing a test based on the ratio of MAD's, which is a particular case of robust scale estimate.

2.2 Asymptotic Distribution of R

In this section, the asymptotic distributions of the test statistic R for the normal and non-normal case are derived. To see the influence of non-normality when comparing the variation in two samples, we will look at the normally and non-normally distributed cases. Firstly, we will describe the statistical method based on the assumption of an underlying normal distribution. Secondly, we will investigate how this method is sensitive to the departure from normality.

2.2.1 Normal case

First, we need to compute the asymptotic distribution of $n(Sr - \sigma)$. Because R is location invariant, we can assume, without loss of generality, that $\mu = 0$. By the Taylor series expansion,

$$
\begin{aligned}
\frac{1}{n}\sum\left[\chi(\frac{x_i - T_n}{Sr})\right] - b &\approx \frac{1}{n}\sum\left[\chi(\frac{x_i}{\sigma})\right] - b - \frac{1}{n}\sum\left(\chi'(\frac{x_i}{\sigma})\frac{x_i}{\sigma^2}\right)(Sr-\sigma) - o(\frac{1}{\sqrt{n}}) \\
&\approx \frac{1}{n}\sum\left[\chi(\frac{x_i}{\sigma})\right] - b - \frac{1}{n}\sum\left(\chi'(\frac{x_i}{\sigma})\frac{x_i}{\sigma^2}\right)(Sr-\sigma). \qquad (24)
\end{aligned}
$$

So,

$$
\sqrt{n}(Sr-\sigma) \approx \frac{\frac{1}{\sqrt{n}}\sum\left(\chi(\frac{x_i}{\sigma})\right) - \sqrt{n}b}{\frac{1}{n}\sum\left(\chi'(\frac{x_i}{\sigma})\frac{x_i}{\sigma^2}\right)}. \qquad (25)
$$

By the Law of Large Numbers,

$$
\frac{1}{n}\sum\left[\chi'(\frac{x_i}{\sigma})(\frac{x_i}{\sigma^2})\right] \to \delta \qquad (26)
$$

with $\delta = E\left[\chi'(\frac{x}{\sigma})(\frac{x}{\sigma})\right] = E\left[\chi'(z)(z)\right]$. Also,

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum\left(\chi(\frac{x_i}{\sigma})\right) - \sqrt{n}b &= \sqrt{n}\left\{\frac{1}{n}\sum\left[\chi(\frac{x_i}{\sigma}) - b\right]\right\} \\
&= \sqrt{n}\left(\frac{1}{n}\sum y_i\right) \qquad (27)
\end{aligned}
$$

where

$$
y_i = \chi(\frac{x_i}{\sigma}) - b,
$$

and

$$
\mathrm{E}(y) = 0,\ \mathrm{Var}(y) = \mathrm{E}\{[\chi(\frac{x}{\sigma}) - b]^2\} = \tau^2.
$$

By the CLT,

$$
\frac{1}{\sqrt{n}}\sum\left(\chi(\frac{x_i}{\sigma})\right) - \sqrt{n}b \to N(0,\tau^2). \qquad (28)
$$

Therefore, by Slutsky's Theorem,

$$
\begin{aligned}
\sqrt{n}(Sr-\sigma) &\to \frac{\sigma N(0,\tau^2)}{\delta} \\
&= N(0, a\sigma^2), \qquad (29)
\end{aligned}
$$

c	a	b	EFF
1.041	0.989	0.500	0.51
1.7	0.625	1.294	0.80
2.07	0.555	0.218	0.90
2.3765	0.526	0.172	0.95

Table 4: relation between c, a, b and EFF

with

$$a = \frac{\tau^2}{\delta^2}.$$

The value of a depends on the choice of c. Table 4 shows how a, b and EFF, the relative efficiency of Sr to the classic sample standard deviation SD, varies with the value of c. The table shows that the efficiency of the dispersion estimate increases with c. We do not use larger c to obtain greater efficiency because as c increases, b will decrease, and the less the value of b is, the less robust the test is. In the next chapter, we will find a value of c, such that the test will be robust and efficient.

In the two sample case, suppose we have two independent samples, $x_{11}, ..., x_{1n_1}$ and $x_{21}, ..., x_{2n_2}$ from the populations, $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$. Suppose the x_{ij}, $j = 1, ..., n_i$ are independent. For simplicity, we assume $n_1 = n_2 = n$.

By the Central Limit Theorem,

$$\sqrt{n}\left[\begin{pmatrix} Sr_1 \\ rS_2 \end{pmatrix} - \begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix}\right] \to N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma\right), \tag{30}$$

where

$$\Sigma = \begin{pmatrix} a\sigma_1^2 & 0 \\ & \\ 0 & a\sigma_2^2 \end{pmatrix}. \tag{31}$$

Let us define a function $g(x, y) = x/y$. Thus, we have by the Delta Method,

$$\sqrt{n}\left(\frac{Sr_1}{Sr_2} - \frac{\sigma_1}{\sigma_2}\right) \to N\left(0, \nabla g' \Sigma \nabla g\right) \tag{32}$$

where

$$\nabla g = \begin{pmatrix} \frac{\partial}{\partial x} g(\sigma_1, \sigma_2) \\ \\ \frac{\partial}{\partial y} g(\sigma_1, \sigma_2) \end{pmatrix} \tag{33}$$

$$= \begin{pmatrix} \sigma_2^{-1} \\ \\ -\sigma_1 \sigma_2^{-2} \end{pmatrix} \tag{34}$$

and

$$\nabla g' \Sigma \nabla g = 2a\left(\frac{\sigma_1^2}{\sigma_2^2}\right). \tag{35}$$

If the null hypothesis, $\mathrm{H}_0 : \sigma_1 = \sigma_2$, is true,

$$\nabla g' \Sigma \nabla g = 2a$$

and

$$S = \frac{\sqrt{n}}{\sqrt{2a}}\left(R - 1\right) \to N(0, 1). \tag{36}$$

So, we can use S to test the two side H_0, and would reject H_0 when S exceeds the upper $100(\alpha/2)$ percentile or falls below the lower $100(\alpha/2)$ percentile of the $N(0, 1)$ distribution. Thus, H_0 is rejected when $|S| > z(1 - \alpha/2)$. For instance, if $\alpha = 0.05$, then H_0 is rejected, when $|S|$ is greater than 1.96.

Table 5 displays the upper and lower critical values (i.e. the acceptance regions) for the test statistic $R = Sr_1/Sr_2$, with $\alpha = 0.05$ based on both the asymptotic

		$n = 25$	$n = 50$
$c = 1.7$	Asymptotic distribution	(0.562, 1.438)	(0.690, 1.310)
	simulation	(0.631, 1.575)	(0.727, 1.383)
$c = 2.07$	Asymptotic distribution	(0.587, 1.413)	(0.707, 1.292)
	simulation	(0.648, 1.538)	(0.736, 1.356)
$c = 2.3765$	Asymptotic distribution	(0.598,1.402)	(0.708,1.292)
	simulation	(0.654, 1.535)	(0.745, 1.341)

Table 5: Acceptance regions of $R = Sr_1/Sr_2$ with $\alpha = 0.05$ obtained from asymptotic distribution and simulation with 10,000 repetitions.

distribution and generation of R from 10,000 random numbers in Splus. The larger the sample size is, the less difference between the acceptance regions obtained from the two methods. Fig 5 shows the simulated distribution of R, with sample sizes 25 and 50, $c = 1.7, 2.07, 2.3765$. The histogram for the smaller sample size is more skewed to right, but as the sample size increases it becomes more symmetric.

For unequal sample size case, if $\frac{n_1}{n_2} \to d$, then we obtain

$$\sqrt{n_1 + n_2}\,(R - 1) \to N(0, \frac{(1+d)^2}{d}a). \tag{37}$$

2.2.2 Non-normal case

To see how this new test is sensitive to departures from normality, we will look at the cases that the population of the variables follow other distributions: t_5 ,t_{10}, χ^2_5,χ^2_{10}, uniform(0,1), and uniform(0,10). In addition, we will estimate their actual significance levels by generating 10,000 numbers. Since we want to known if the two arguments in the uniform distribution affect the results, the uniform distributions with arguments (0,1), and (0,10) are investigated. The simulated significance levels ($\alpha = 0.05$) for the non-normal distributions are displayed in Table 6. The normal case is included in the table because we want to see how large the error is due to the generation of data. Note that the arguments in the uniform distribution do not affect the result. Also, for a heavy-tailed distribution , the probability of rejecting H_0 exceeds 0.05; whereas for a short-tailed distribution, the probability is less than 0.05. But, in general, the results are closer to 0.05 than the ones from classic F test. Also, the significance levels yielded by smaller c are closer to 0.05.

2.3 Examples

In this section, the new tests with $c = 1.7, 2.07, 2.3765$ are applied to the examples described in the first chapter. The test results and the the test statistics R's for each pair of samples are shown in Table 7 and 8. Table 9 displays the acceptance regions of R with $\alpha = 0.05$ for sample sizes n_1, n_2. The acceptance regions shown in the table are obtained from simulation with 1,000 repetitions. In the Cloud example, when no outlier is added, $n_1 = n_2 = 26$, and with $c = 1.7$, $R = 0.958$. Since R is within the acceptance region, [0.689, 1.457], shown in Table 9, the new test with

Distribution	$c = 1.7$	$c = 2.07$	$c = 2.3765$
$N(0,1)$	0.053	0.052	0.051
t_5	0.087	0.105	0.123
t_{10}	0.073	0.079	0.080
${\chi^2}_5$	0.074	0.127	0.150
${\chi^2}_{10}$	0.052	0.080	0.098
Uniform$(0,1)$	0.0005	0.001	0.002
Uniform$(0,10)$	0.0005	0.001	0.002

Table 6: Simulated actual significant level of the new test ($\alpha = 0.05$) from 10,000 generated data, with normal assumption for several non-normal distributions

value of two outlier	$c = 1.7$		$c = 2.07$		$c = 2.3765$	
	R	reject	R	reject	R	reject
no outlier	0.958	0	0.953	0	0.969	0
12	0.958	0	0.953	0	0.969	0
14	0.958	0	0.953	0	0.969	0
25	0.958	0	0.953	0	0.969	0
28	0.958	0	0.953	0	0.969	0
30	0.958	0	0.953	0	0.969	0
100	0.958	0	0.953	0	0.969	0

Table 7: Results of the new tests ($c = 1.7, 2.07, 2.3765$) on the Cloud example. If reject = 1, the test rejects H_0

$c = 1.7$ does not reject the null hypothesis. For all of the three tests, no matter how large the two outliers are, they still do not reject the null hypothesis. It means that the tests are not affected by the extremely large observations. Also, the value of R does not vary with the value of outliers for each test. Similarly, for the Michelson's example, the size of outlier does not make any influence on the results of the tests, and the value of R keeps constant with different values of outliers.

Based on these two examples, we can conclude that the new tests have superior ability to overcome the effect of outliers.

	$c = 1.7$		$c = 2.07$		$c = 2.3765$	
value of two outlier	R	reject	R	reject	R	reject
no outlier	1.841	1	1.882	1	1.781	1
950	1.841	1	1.882	1	1.589	1
980	1.841	1	1.882	1	1.589	1
1000	1.841	1	1.882	1	1.589	1
1100	1.841	1	1.882	1	1.589	1

Table 8: Results of the new tests ($c = 1.7, 2.07, 2.3765$) on the Michelson example. If reject = 1, the test rejects H_0

n_1	n_2	$c = 1.7$	$c = 2.07$	$c = 2.3765$
26	26	[0.689, 1.457]	[0.716, 1.464]	[0.698, 1.413]
26	28	[0.688,1.431]	[0.721, 1.407]	[0.710, 1.415]
20	20	[0.564, 1.774]	[0.599, 1.723]	[0.606, 1.679]
20	21	[0.654,1.592]	[0.661, 1.490]	[0.672, 1.470]

Table 9: Acceptance regions of $R = Sr_1/Sr_2$ with $\alpha = 0.05$ with sample sizes n_1, n_2 obtained from simulation with 1,000 repetitions.

3 Monte Carlo study

In this Chapter, we compare the new tests with the F test and the six robust tests described in the first chapter. Two types of Monte Carlo studies are presented. First we investigate the sensitivity of the tests to non-normality. Second we investigate the influence of outliers on the power and the significance level of the tests. The procedures for our first Monte Carlo study are the following:

(i) Generate one hundred and fifty pairs of samples; the sample size is 25, and the pseudo-random numbers represent samples from a uniform distribution.

(ii)Transform the pseudo-random numbers to obtain samples from a $N(0,1), {\chi^2}_{10}$, t_5, t_{10}, and t_{20} distributions.

(iii) After the transformation, the second sample was scaled by the factor Δ so that the ratio of the two variances is Δ^2 for each distribution. Different values of Δ are selected and applied to the samples.

(iv) Ten tests were applied to each of the 150 pairs of samples. The ten tests are the F test, the Box-Andersen test, the Levene test, the Jacknife test with subsample size $k = 1$, the Box and Moses tests both with subsample size $k = 5$, and the three new tests with $c = 1.7, 2.07, \text{and } 2.3765$.

(v) Repeat steps (i) to (iv) with sample size 50.

The entries in Tables 10 to 14 are the proportions of samples in 150 trials that the tests reject the null hypothesis $\sigma_x^2 = \sigma_y^2$ for the various distributions and Δ. For $\Delta = 1$ the proportions should be close to $\alpha = 0.05$. For $\Delta > 1$ the proportions are Monte Carlo estimates of the power of the tests at the particular selections of Δ for various distributions. The results of these tables reveal the following conclusion:

(i) The F test is extremely non-robust. It gives too many significant results for long tailed distributions.

(ii) The three new tests have about the same power, and in general they are the most powerful tests in the group. The three tests, when $\Delta = 1$, give more significant results than the other tests.

(iii) The new test with $c = 1.7$ is not as powerful as the new tests with $c = 2.07, 2.3765$, but its actual significance level is closer to 0.05.

(iv) The other tests are robust, but they are not as powerful as the new tests. In general, the Jacknife and Box-Andersen tests have about the same power. The Levene test is more powerful than these two tests.

(v)The Moses test is sightly less powerful than the Box test, and seems to be the least powerful of all the tests.

The second type of Monte Carlo studies includes two parts. The first part estimates the influence of outliers on the significance of the tests, and the second part estimates the influence of outliers on the power of the tests. The procedures for the first part are the following:

(i) Transform the first sample of each of the one hundred and fifty pairs of pseudo-random samples to obtain samples from $N(0,1)$ with different number of outliers from $N(5,0.1)$.

(ii) Transform the second sample of each of the one hundred and fifty pairs of pseudo-random samples to obtain samples from $N(0,1)$ without outlier.

(iii) Repeat steps (i) and (ii) with sample size 50.

The entries in Table 15 are the proportions of samples in 150 trials that the tests reject the null hypothesis $\sigma_x^2 = \sigma_y^2$ for the various numbers of outliers. Test with smaller values is less affected by the outliers, and seldom falsely rejects the null hypothesis. According to the results in the table, we have the following conclusions:

(i) The Moses test is less affected than the Box test. Both of these tests seem to be least affected by the outliers. However, it is probably due to the fact they are very conservative, and the result is consistent with the one obtained by Miller [7].

(ii)The new test with $c = 1.7$ is the second least affected one. When the sample size is 25 and less than 16% of observations in the first sample are outliers, the Levene test is slightly better than the new test with $c = 2.07$; the new test with $c = 2.3765$ is almost the worst one. Also, as the number of outliers increases, the new test with $c = 2.07$ becomes more affected by the outliers.

(iii)When the sample size is 50, the new test with $c = 2.07$ is better than the Levene test. In addition, the performance of the new test with $c = 1.7$ is almost the best in the group.

The second part is to test the effect of outliers on the power of the tests. The procedures are the following:

(i) Transform the first sample of each of the one hundred and fifty pairs of pseudo-random samples to obtain samples from $N(0,1)$ with different number of outliers from $N(5.5, 0.1)$.

(ii) Transform the second sample of each of the one hundred and fifty pairs of pseudo-random samples to obtain samples from $N(0,3)$ without outlier.

(iii) Repeat steps (i) and (ii) with sample size 50.

The entries in Table 16 are the proportions of samples in 150 trials that the tests reject the null hypothesis $\sigma_x^2 = \sigma_y^2$ for the various numbers of outliers. Tests with larger values are less affected by the outliers, and seldom falsely accepts the null hypothesis.

To estimate the influence of larger outlier, we repeat the procedures with larger outliers from $N(10, 0.1)$ distribution. The results are exhibited in Table 17. Based on these two tables, we have the following conclusions:

(i) The new test with $c = 1.7$ has the best performance.

(ii) When the sample contains less than 16% outliers, the new tests with $c = 2.07, 2.3765$ are the second best tests. Whereas, as the number of outliers increases, the new tests with higher values of c become the worst of the all.

	$n_1 = n_2 = 25$					$n_1 = n_2 = 50$				
ratio of standard deviation	1:1	1:1.5	1:2	1:2.5	1:5	1:1	1:1.5	1:2	1:2.5	1:5
F-test	0.053	0.513	0.927	0.987	1.000	0.053	0.767	1.000	1.000	1.000
Levene	0.047	0.460	0.893	0.973	1.000	0.080	0.733	0.987	1.000	1.000
Layard	0.040	0.407	0.827	0.953	1.000	0.067	0.707	0.980	1.000	1.000
Jacknife $k = 1$	0.027	0.493	0.900	0.973	1.000	0.040	0.740	1.000	1.000	1.000
Box $k = 5$	0.047	0.293	0.707	0.820	1.000	0.060	0.553	0.927	0.993	1.000
Moses $k = 5$	0.027	0.287	0.600	0.800	0.987	0.053	0.560	0.920	0.980	1.000
Box Andersen	0.033	0.487	0.913	0.973	1.000	0.053	0.747	0.993	1.000	1.000
New test $c = 1.7$	0.060	0.420	0.860	0.967	1.000	0.067	0.687	0.987	1.000	1.000
New test $c = 2.07$	0.047	0.453	0.893	0.980	1.000	0.080	0.740	0.993	1.000	1.000
New test $c = 2.3765$	0.060	0.453	0.900	0.973	1.000	0.067	0.760	1.000	1.000	1.000

Table 10: Monte Carlo Power Function for Tests on Variances for Normal distribution

ratio of standard deviation	$n_1 = n_2 = 25$					$n_1 = n_2 = 50$				
	1:1	1:1.5	1:2	1:2.5	1:5	1:1	1:1.5	1:2	1:2.5	1:5
F-test	0.153	0.540	0.860	0.960	1.000	0.093	0.707	0.987	1.000	1.000
Levene	0.060	0.487	0.833	0.953	1.000	0.087	0.653	0.973	1.000	1.000
Layard	0.040	0.353	0.740	0.900	1.000	0.027	0.507	0.913	1.000	1.000
Jacknife $k = 1$	0.080	0.440	0.760	0.900	0.993	0.053	0.600	0.940	1.000	1.000
Box $k = 5$	0.033	0.293	0.600	0.800	0.987	0.067	0.460	0.880	0.987	1.000
Moses $k = 5$	0.033	0.227	0.493	0.760	0.973	0.060	0.447	0.860	0.987	1.000
Box Andersen	0.053	0.407	0.780	0.913	1.000	0.047	0.600	0.933	1.000	1.000
New test $c = 1.7$	0.0737	0.427	0.847	0.967	1.000	0.073	0.653	0.980	1.000	1.000
New test $c = 2.07$	0.093	0.500	0.880	0.967	1.000	0.100	0.693	0.987	1.000	1.000
New test $c = 2.3765$	0.100	0.487	0.873	0.967	1.000	0.113	0.727	0.993	1.000	1.000

Table 11: Monte Carlo Power Functions for Tests on Variances for ${\chi_5}^2$ distribution

ratio of standard deviation	$n_1 = n_2 = 25$					$n_1 = n_2 = 50$				
	1:1	1:1.5	1:2	1:2.5	1:5	1:1	1:1.5	1:2	1:2.5	1:5
F-test	0.193	0.500	0.847	0.960	1.000	0.227	0.667	0.973	1.000	1.000
Levene	0.040	0.353	0.727	0.940	0.993	0.073	0.593	0.940	1.000	1.000
Layard	0.040	0.227	0.607	0.840	0.993	0.027	0.407	0.860	0.947	1.000
Jacknife $k = 1$	0.047	0.353	0.673	0.847	0.980	0.073	0.473	0.860	0.940	1.000
Box $k = 5$	0.040	0.240	0.547	0.780	0.980	0.053	0.433	0.860	0.967	1.000
Moses $k = 5$	0.020	0.200	0.467	0.660	0.980	0.060	0.427	0.847	0.960	1.000
Box Andersen	0.027	0.293	0.660	0.833	0.993	0.053	0.473	0.880	0.960	1.000
New test $c = 1.7$	0.087	0.427	0.840	0.953	1.000	0.120	0.667	0.973	1.000	1.000
New test $c = 2.07$	0.093	0.493	0.860	0.967	1.000	0.120	0.700	0.960	1.000	1.000
New test $c = 2.3765$	0.113	0.480	0.847	0.973	1.000	0.140	0.720	0.973	1.000	1.000

Table 12: Monte Carlo Power Functions for Tests on Variances for t_5 distribution

	$n_1 = n_2 = 25$					$n_1 = n_2 = 50$				
ratio of standard deviation	1:1	1:1.5	1:2	1:2.5	1:5	1:1	1:1.5	1:2	1:2.5	1:5
F-test	0.107	0.507	0.900	0.980	1.000	0.087	0.747	0.987	1.000	1.000
Levene	0.047	0.427	0.840	0.967	1.000	0.080	0.667	0.967	1.000	1.000
Layard	0.040	0.333	0.720	0.913	1.000	0.053	0.520	0.940	0.993	1.000
Jacknife $k = 1$	0.033	0.420	0.793	0.913	1.000	0.040	0.600	0.960	1.000	1.000
Box $k = 5$	0.033	0.260	0.613	0.807	1.000	0.073	0.507	0.920	0.993	1.000
Moses $k = 5$	0.020	0.200	0.560	0.700	0.973	0.053	0.480	0.860	0.980	1.000
Box Andersen	0.027	0.413	0.787	0.920	1.000	0.053	0.647	0.967	0.993	1.000
New test $c = 1.7$	0.067	0.427	0.847	0.967	1.000	0.093	0.673	0.987	1.000	1.000
New test $c = 2.07$	0.073	0.480	0.873	0.980	1.000	0.093	0.720	0.987	1.000	1.000
New test $c = 2.3765$	0.067	0.447	0.873	0.973	1.000	0.093	0.733	0.980	1.000	1.000

Table 13: Monte Carlo Power Functions for Tests on Variances for t_{10} distribution

ratio of standard deviation	$n_1 = n_2 = 25$					$n_1 = n_2 = 50$				
	1:1	1:1.5	1:2	1:2.5	1:5	1:1	1:1.5	1:2	1:2.5	1:5
F-test	0.073	0.493	0.900	0.980	1.000	0.060	0.767	0.993	1.000	1.000
Levene	0.047	0.433	0.873	0.967	1.000	0.080	0.720	0.973	1.000	1.000
Layard	0.053	0.367	0.773	0.927	1.000	0.060	0.647	0.967	1.000	1.000
Jacknife $k = 1$	0.033	0.453	0.833	0.953	1.000	0.040	0.707	0.987	1.000	1.000
Box $k = 5$	0.033	0.287	0.653	0.853	1.000	0.047	0.607	0.880	0.993	1.000
Moses $k = 5$	0.020	0.207	0.553	0.760	0.993	0.067	0.567	0.900	0.980	1.000
Box Andersen	0.027	0.440	0.860	0.953	1.000	0.053	0.713	0.980	1.000	1.000
New test $c = 1.7$	0.067	0.420	0.860	0.967	1.000	0.087	0.673	0.987	1.000	1.000
New test $c = 2.07$	0.053	0.480	0.880	0.980	1.000	0.087	0.733	0.993	1.000	1.000
New test $c = 2.3765$	0.060	0.447	0.887	0.973	1.000	0.067	0.747	0.993	1.000	1.000

Table 14: Monte Carlo Power Functions for Tests on Variances for t_{20} distribution

	$n_1 = n_2 = 25$					$n_1 = n_2 = 50$				
number of outliers	1	2	3	4	5	2	4	6	8	10
F-test	0.380	0.807	0.960	0.987	0.993	0.653	0.993	1.000	1.000	1.000
Levene	0.040	0.153	0.460	0.820	0.987	0.140	0.473	0.873	1.000	1.000
Layard	0.000	0.093	0.467	0.913	0.987	0.027	0.513	0.987	1.000	1.000
Jacknife $k = 1$	0.047	0.413	0.840	0.967	0.987	0.260	0.893	1.000	1.000	1.000
Box $k = 5$	0.013	0.073	0.233	0.400	0.547	0.100	0.207	0.560	0.740	0.907
Moses $k = 5$	0.047	0.033	0.173	0.267	0.400	0.067	0.227	0.420	0.653	0.840
Box Andersen	0.013	0.140	0.600	0.940	0.987	0.100	0.680	1.000	1.000	1.000
New test $c = 1.7$	0.073	0.100	0.240	0.400	0.700	0.080	0.187	0.407	0.727	0.933
New test $c = 2.07$	0.073	0.193	0.373	0.833	0.993	0.080	0.280	0.680	0.987	1.000
New test $c = 2.3765$	0.107	0.293	0.773	0.993	0.993	0.120	0.487	0.980	1.000	1.000

Table 15: Monte Carlo Power Functions for Tests on Variances for based on two samples from the $N(0,1)$ population with different number of outliers from the $N(5,0.1)$ in the first sample

	$n_1 = n_2 = 25$					$n_1 = n_2 = 50$				
number of outliers	1	2	3	4	5	2	4	6	8	10
F-test	0.953	0.713	0.407	0.153	0.067	1.000	0.967	0.820	0.540	0.273
Levene	0.947	0.760	0.473	0.240	0.087	1.000	0.987	0.867	0.613	0.273
Layard	0.747	0.407	0.180	0.087	0.040	1.000	0.833	0.613	0.333	0.187
Jacknife $k = 1$	0.073	0.093	0.093	0.087	0.067	0.947	0.807	0.613	0.440	0.307
Box $k = 5$	0.767	0.240	0.060	0.013	0.013	0.980	0.867	0.573	0.313	0.120
Moses $k = 5$	0.527	0.233	0.073	0.053	0.033	0.987	0.813	0.467	0.207	0.133
Box Andersen	0.820	0.507	0.280	0.140	0.080	1.000	0.907	0.740	0.480	0.293
New test $c = 1.7$	0.993	0.980	0.940	0.853	0.573	1.000	1.000	0.993	0.987	0.840
New test $c = 2.07$	0.993	0.980	0.880	0.567	0.013	1.000	1.000	0.993	0.820	0.067
New test $c = 2.3765$	1.000	0.960	0.700	0.080	0.027	1.000	1.000	0.940	0.333	0.127

Table 16: Monte Carlo Power Functions for Tests on Variances for based on two samples from the $N(0, 1)$ and $N(0, 3)$ populations with different number of outliers from the $N(5.5, 0.1)$ in the first sample

	$n_1 = n_2 = 25$					$n_1 = n_2 = 50$				
number of outliers	1	2	3	4	5	2	4	6	8	10
F-test	0.173	0.007	0.073	0.207	0.320	0.580	0.000	0.107	0.380	0.673
Levene	0.607	0.047	0.000	0.013	0.173	0.953	0.313	0.000	0.040	0.413
Layard	0.027	0.073	0.027	0.020	0.127	0.047	0.033	0.027	0.073	0.420
Jacknife $k = 1$	0.000	0.000	0.000	0.100	0.307	0.000	0.000	0.013	0.220	0.600
Box $k = 5$	0.127	0.000	0.000	0.000	0.020	0.847	0.167	0.013	0.000	0.020
Moses $k = 5$	0.007	0.000	0.000	0.000	0.080	0.787	0.047	0.000	0.027	0.153
Box Andersen	0.013	0.000	0.000	0.020	0.227	0.140	0.000	0.000	0.140	0.567
New test $c = 1.7$	0.993	0.980	0.940	0.853	0.573	1.000	1.000	0.993	0.987	0.840
New test $c = 2.07$	0.993	0.980	0.880	0.567	0.100	1.000	1.000	0.993	0.820	0.133
New test $c = 2.3765$	1.000	0.960	0.700	0.120	0.500	1.000	1.000	0.933	0.227	0.853

Table 17: Monte Carlo Power Functions for Tests on Variances for based on two samples from the $N(0,1)$ and $N(0,3)$ populations with different number of outliers from the $N(10,0.1)$ in the first sample

4 Conclusion

The classic F test for the hypothesis concerning the equality of two population variances is known to be non-robust. Let us consider a two sample problem. Suppose we have two samples, $y_{11}, ..., y_{1n_1}$ and $y_{21}, ..., y_{2n_2}$. Suppose the y_{ij}'s are independent and identically distributed with cdf $G((y_i - \mu_i)/\sigma_i)$. As $\frac{n_1}{n_2} \longrightarrow d$,

$$\sqrt{n_2}\left(\frac{S_1}{S_2} - 1\right) \rightarrow N\left(0, \frac{(2+\gamma)(1+d)}{4d}\right),$$

where γ is the coefficient of kurtosis. If normal assumption is met, $\gamma = 0$. However, for non-normal cases, like t_5, γ won't be zero. So, when we apply the classical F test to the non-normal samples, the actual size of the test would be different from its nominal level of significance, α. Therefore, several robust alternative procedures have been introduced in this century.

This paper presents a new robust method. The best feature of this new method is that it has superior ability to overcome the effect of outliers. First, an alternative measure of dispersion, Sr, that is more resistant to outliers was introduced.

The new test statistic was then defined using these robust dispersion estimates.

In Section 2.2.2, we estimated the actual significance levels of the new tests ($\alpha = 0.05$) for the non-normal case. We've found that for a heavy-tailed distribution the probability of rejecting H_0 exceeds 0.05; whereas for a short-tailed distribution, the probability is less than 0.05. But, in general, the results are closer to 0.05 than the ones from classic F test. Also, the significance levels yielded by smaller c are closer to 0.05.

According to the two examples described in the first two chapters, the performance of the new tests is obviously better than the other tests discussed in the first chapter. In these two examples, we can see that no matter how large the outliers are, the new tests are not affected by them. It can be explained by the fact that the test statistic R is not affected by the size of outliers but the number of outliers.

In addition, according to the first type of Monte Carlo study, the three tests have about the same power. In general, the new tests are most powerful in the group, although the true significance levels of the three tests are sightly more sensitive to the other tests. Also, the new test with $c = 1.7$ is just not as powerful as the new tests with $c = 2.07, 2.3765$, but its actual significance level is closer to the proposed significance level 0.05. Based on the second type of Monte Carlo study, the new test with $c = 1.7$ seems to have the superior power to overcome the effect of outliers.

On the whole, this paper has demonstrated that although the new test with $c = 1.7$ is just a little bit less powerful than those with $c = 2.07$, 2.3765, of all the tests, the new test with $c = 1.7$ has superior ability to overcome the effect of outliers.

References

[1] M.S. Bartlett. Properties of sufficiency and statistical tests. *Proceedings of the royal society A*, 160:262–282, 1937.

[2] G.E.P. Box. Non-normality and tests on variances. *Boimetrika*, 40:318–335, 1953.

[3] G.E.P Box and S.L. Andersen. Permuation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society*, B17:1–26, 1955.

[4] M.B. Brown and A.B. Forsythe. Robust test for the equality of variances. *Journal of American Statistical Association*, 69:364–367, 1974.

[5] M.W.J. Layard. Robust large-sample tests for homogeneity of variances. *Journal of American Statistical Association*, 68:105–198, 1974.

[6] H. Levene. Robust tests for equality of variance contributions to probability and statistics. pages 278–292. Stanford University Press, 1960.

[7] R.G. Jr Miller. Jacknifing variances. *Annals of Mathematical Statistics*, 39:567–582, 1968.

[8] Rupert G. Miller. Beyond anova, basis of applied statistics.

[9] L.E. Moses. Rank tests of dispersion. *Annals of Mathematical Statistics*, 34:973–983, 1963.

[10] G.R. Shorack. Nonparametric tests and estimation of scale in two sample problem. *Technical Report*, 10.

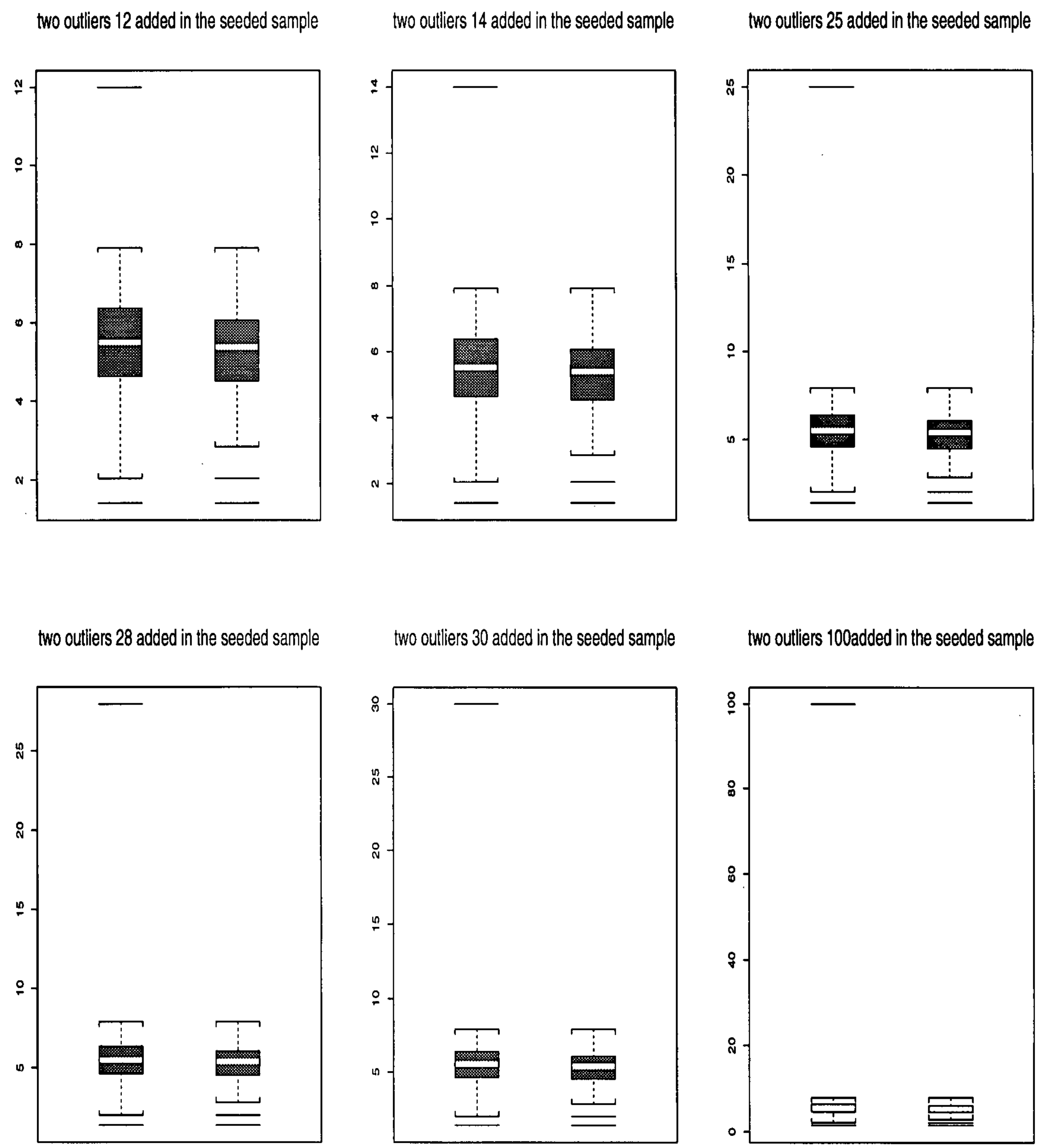

Figure 2: Side by Side Boxplots of the two logged variables with outliers in the seeded sample

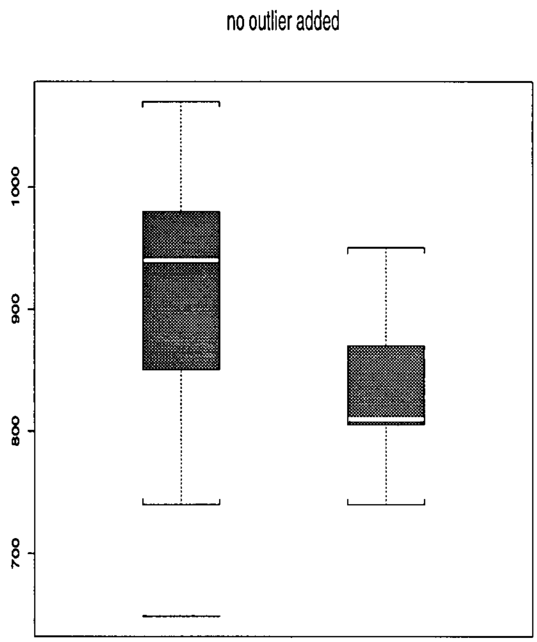

Figure 3: Side by Side Boxplots of the measurements in the first and fifth trials in Michelson's example

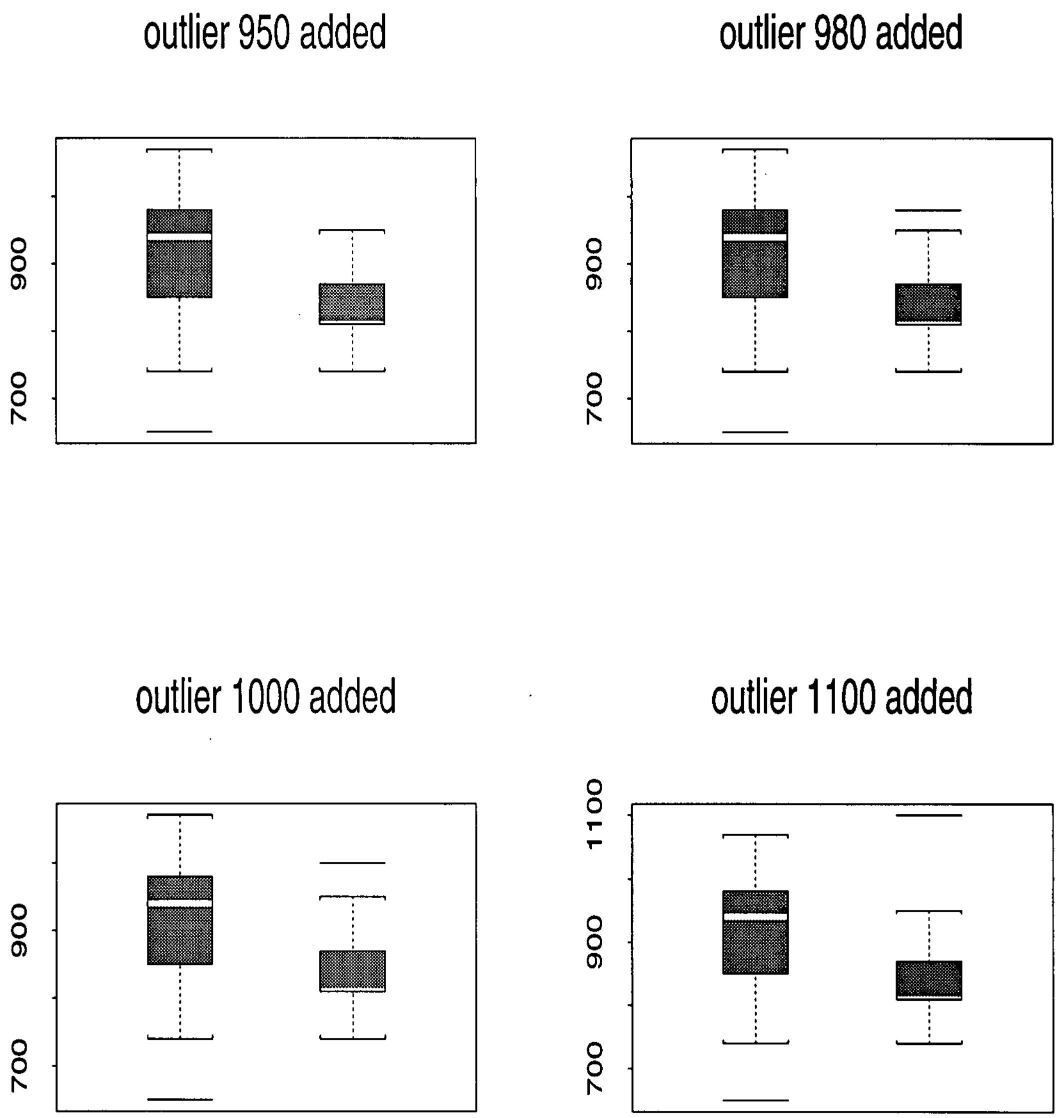

Figure 4: Side by Side Boxplots of the two variables with outliers in the fifth sample

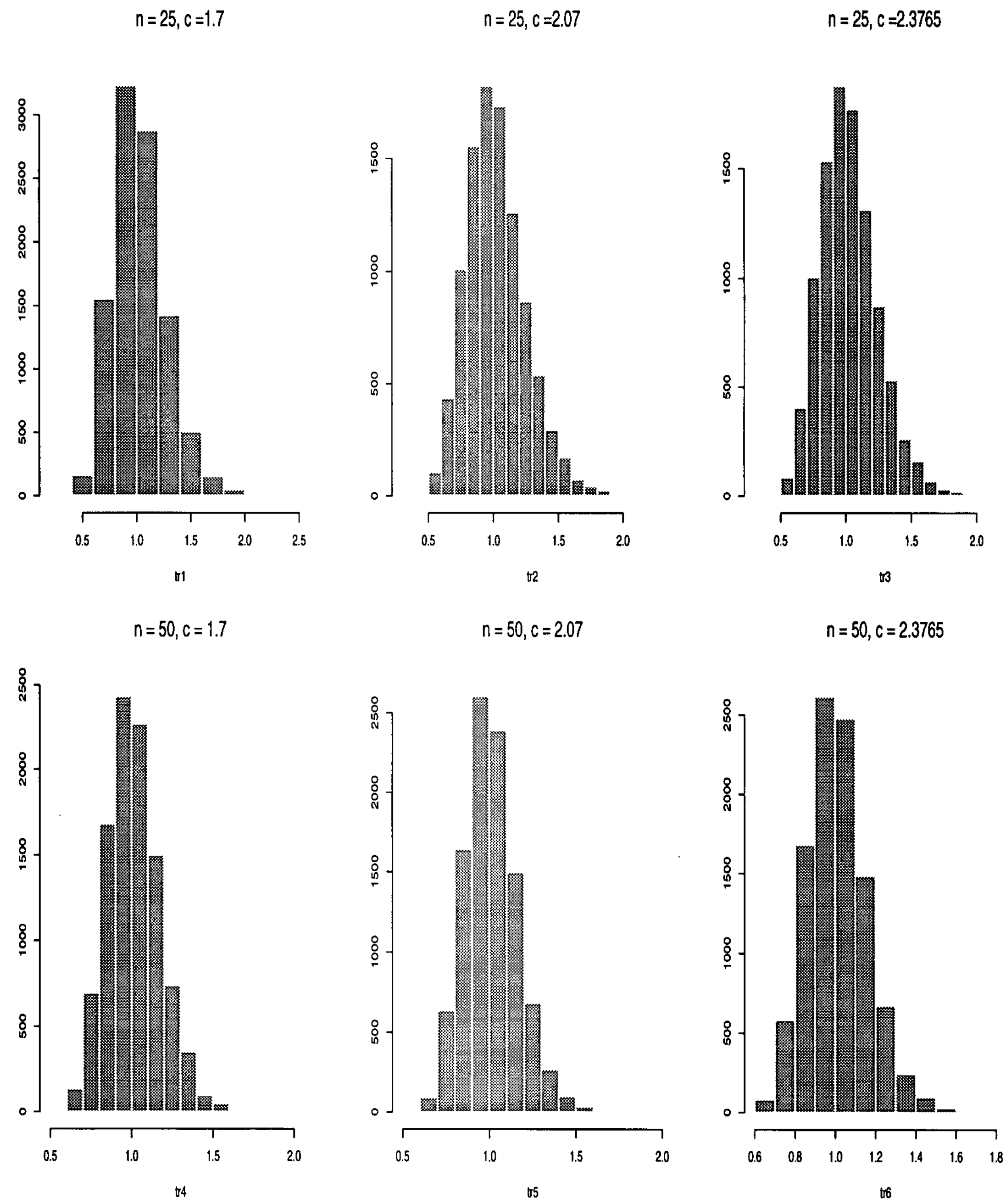

Figure 5: Histograms of R for different combinations of sample size n, and c