# EMPIRICAL LIKELIHOOD TESTS FOR CONSTANT VARIANCE IN THE TWO-SAMPLE PROBLEM

Paul Shen

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

May 2019

Committee:

Hanfeng Chen, Advisor

Wei Ning

Christopher M. Rump

# ABSTRACT

Hanfeng Chen, Advisor

In this thesis, we investigate the problem of testing constant variance. It is an important problem in the field of statistical influence where many methods require the assumption of constant variance. The question of constant variance has to be settled in order to perform a significance test through a Student *t*-Test or an *F*-test. Two of most popular tests of constant variance in applications are the classic *F*-test and the Modified Levene's Test. The former is a ratio of two sample variances. Its performance is found to be very sensitive with the normality assumption. The latter Modified Levene's Test can be viewed as a result of the estimation method through the absolute deviation from the median. Its performance is also dependent upon the distribution shapes to some extent, though not as much as the *F*-test. We propose an innovative test constructed by the empirical likelihood method through the moment estimation equations appearing in the Modified Levene's Test. The new empirical likelihood ratio test is a nonparametric test and retains the principle of maximum likelihood. As a result, it can be an appropriate alternative to the two traditional tests in applications when underlying populations are skewed. To be specific, the empirical likelihood ratio test of constant variance uses the optimal weights in summing the absolute deviations of observations from the median values, while the Modified Levene's test uses the simple averages. It is thus desired that the empirical likelihood ratio test is more powerful than the Modified Levene's test. Meanwhile, the empirical likelihood ratio test is expected to be as robust as the Modified Levene's test, as the empirical likelihood ratio test is also constructed via the same distance as the Modified Levene's test. A

real-life data set is used to illustrate implementation of the empirical likelihood ratio test with comparisons to the classic *F*-test and the Modified Levene's Test. It is confirmed that the empirical likelihood ratio test performs the best.

This thesis is dedicated to my mother, father, and sister.

# ACKNOWLEDGMENTS

I would like to express my sincere gratitude to Professor Hanfeng Chen, for his invaluable guidance, great patience, and encouragement throughout the project for writing the thesis and his great advice throughout my studies here. I learned a lot about the empirical likelihood methods and some other interesting statistical methodologies related to the topic of the thesis.

I would also like to thank Professors Wei Ning and Christopher M. Rump for their serving on my thesis committee, providing insights to the study, and making invaluable suggestions and comments on an earlier draft of the thesis. Professor Rump is also my academic advisor in the Department of Applied Statistics and Operations Research at Bowling Green State University. His encouragement and support has made it possible for me to achieve my academic ambitions in the program of M.S. in Applied Statistics.

I also want to acknowledge Dr. Ibrahim Çapar, my instructor for Linear Regression. His teaching of the course re-introduced me to the concept of constant variance and also aided my interest in the topic.

I also would like to thank the Department of Applied Statistics and Operations Research for funding me to complete my graduate studies.

Lastly, I want to acknowledge my family, my father Yinong Shen, my mother Yafang Miao and my lovely sister Amy for their love, understanding, and encouragement given to me to pursue graduate-level studies in Statistics.

TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1: INTRODUCTION

Constant variance is a key assumption that is tested in any multi-sample test, with two samples being a baseline case. More often than not, it is violated and there exists the need for transformation of variables due to the violation (Dielman, 2005). In the event of a violation in constant variance, the standard significance tests such as the *t*-Test and *F*-test are invalid as explained in Section 1.1. As a result, we typically check whether the assumption of constant variance is satisfied before doing a significance test, that is, to perform a test for constant variance. There are two popular tests for constant variance which are the classic *F*-test and the Modified Levene's test. The earlier version of the latter test was proposed in Levene (1960) and modified later by Brown and Forsythe (1974). Hence the Modified Levene's test is also often called the Brown-Forsythe test in literature.

In this thesis, we investigate the testing problem for constant variance and propose an innovative statistical test through the empirical likelihood method. We proceed with a brief review of the problem caused by non-constant variance and the two most commonly used tests for constant variance.

## 1.1 Pitfall of the Two-Sample *t*-Test

The two-sample *t*-Test is known as a significance test for equality of two means. It requires that the variances of the two normal populations from which the two samples are taken are identical. This is known as the constant variance assumption. Consider two samples of $X$ and $Y$ with respective sizes $n$ and $m$. The two-sample *t*-Test statistic is defined as

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{1.1.1}$$

(Utts and Heckard, 2011), where $\bar{X}$ is the sample mean of the *X*-sample and $\bar{Y}$ is the sample mean of the *Y*-sample, and $S_p$ represents the pooled standard deviation, i.e.,

$$S_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n + m - 2}} \tag{1.1.2}$$

with $s_X^2$ being the sample variance of the *X*-sample and $s_Y^2$ being the sample variance of the *Y*-sample. Under the assumption of constant variance, the test statistic *T* defined in equation (1.1.1) is the central-*t* distribution with $n + m - 2$ degrees of freedom. However, if the two populations have different variances, *T*'s distribution is no longer parameter-free, even if the two means are equal. This is because, if the two variances are different, the distribution of the pooled standard deviation used to standardize the test no longer belongs to a scale-parameter distribution family.

## 1.2 The Classic *F*-Test for Constant Variance

The classic *F*-test for constant variance is defined to be the ratio of the two sample variances $s_X^2$ and $s_Y^2$ such that

$$F^* = \frac{s_X^2}{s_Y^2}. \tag{1.2.1}$$

When the null hypothesis of constant variance holds true, *F** follows the F-distribution with $n - 1$ and $m - 1$ degrees of freedom. Note that *F** estimates the ratio of two variances. The null hypothesis of constant variance is rejected when the observed value of *F** is either very small or very large. This inference heavily relies on the normality assumption. When the normality assumption is problematic, this test is not useful and produces misleading inferential results. Thus, another test that is more robust against departure from the normality assumption should be used.

## 1.3 The Modified Levene's Test

In the event of a skewed distribution, outliers may affect the performance of the classic *F*-test dramatically. The Modified Levene's Test is proposed to be more robust against outliers as well as departure of normality. The Modified Levene's Test uses a pooled standard deviation based across the median of the residuals, which opens this test up to potential constant variance violations since the use of a pooled standard deviation suggests the population variances are assumed to be equal. Specifically, define

$$d_{1j} = |X_j - m_x| \text{ and } d_{2j} = |Y_j - m_y|, \tag{1.3.1}$$

where $m_x$ and $m_y$ represent the sample medians of the *X* and *Y* observations, respectively. Let $\bar{d}_1$ and $\bar{d}_2$ represent the means of $d_{ij}$ for each group $i$. Let $S_L$ represent the pooled standard deviation of the absolute value of the difference between the median of the residuals and the actual residual values, where

$$S_L = \sqrt{\frac{\sum_{j=1}^{n}(d_{1j} - \bar{d}_1)^2 + \sum_{j=1}^{m}(d_{2j} - \bar{d}_2)^2}{n + m - 2}} \tag{1.3.2}$$

Ultimately, the test statistic *L* is designed in the same manner of a *t*-Test statistic as

$$L = \frac{\bar{d}_1 - \bar{d}_2}{S_L\sqrt{\frac{1}{n} + \frac{1}{m}}}. \tag{1.3.3}$$

Under the null hypothesis of constant variance, the test statistic *L* is expected to follow a T-distribution with approximately $n + m - 2$ degrees of freedom. Yet, the performance of the Modified Levene's Test is somewhat dependent upon the shapes of the population distributions. In this thesis, we propose a non-parametric approach through the empirical likelihood method invented by Owen (1988).

The remainder of this thesis is organized as follows. In Chapter 2, the new empirical likelihood ratio test for constant variance is proposed, developed, and extended to cover more general application. Chapter 3 includes a real-life data analysis to illustrate the approach and compare it to the classic *F*-test and the Modified Levene's Test. Chapter 4 presents conclusions and recommendations.

# CHAPTER 2: EMPIRICAL LIKELIHOOD RATIO TEST FOR CONSTANT VARIANCE

## 2.1 The Empirical Likelihood Method

We proceed with a description of the empirical likelihood method that was invented by Professor Art Owen of Stanford University (Owen, 1988, 1990 and 2001). The empirical likelihood approach is a non-parametric approach that makes no model assumptions and is data-driven, and yet entertains the likelihood principle of statistics.

Consider a random sample of size $n$, say $X_1, \dots, X_n$, from an unknown population distribution function $F(x)$. The empirical likelihood is constructed by placing an unknown probability mass at each observation $X_i$. Define $p_i$ as

$$p_i = P(X = X_i) = F(X_i) - F(X_i^-)$$

and define the empirical log-likelihood function of $F$ as

$$l_n(F) = \sum_{i=1}^{n} \log(p_i).$$

It is clear that with the constraints $p_i \geq 0$ for all $i$ and $\sum_{i=1}^{n} p_i = 1$, $l_n(F)$ is maximized at the empirical function

$$F(x) = \frac{\#\{i : x_i \leq x\}}{n}.$$

The function $l_n(F)$ attains maximum value $-n \log(n)$ at $p_i = \frac{1}{n}$ which can be viewed as the likelihood maximum under the full nonparametric model.

Let a population functional characteristic, such as the mean $\theta = E(X)$, be of interest. With the true value $\theta_0$ of $\theta = E(X)$, the empirical log-likelihood maximum is to be obtained, subject to the constraint

$$\sum_{i=1}^{n} x_i p_i = \theta_0.$$

Let $\hat{l}_n(\theta_0)$ represent the likelihood maximum. Then the empirical log-likelihood

ratio test statistic is given by

$$R(\theta_0) = \hat{l}_n(\theta_0) - l_n(\hat{F}). \tag{2.1.1}$$

Equation (2.1.1) is similar to the likelihood ratio test statistic in a parametric model setup. $-2R(\theta_0)$ is distributed asymptotically as the central $\chi^2$ distribution with $k$ degrees of freedom, where $k$ is the dimension of $\theta$, i.e., $\chi^2_k$. Consequently, an asymptotic 100(1-$\alpha$)% confidence region for $\theta$ is constructed as

$$I_{1-\alpha} = \{\theta : -2R(\theta) \leq \chi^2_k(\alpha)\}$$

where $\chi^2_k(\alpha)$ is the upper $\alpha$ quantile of the $\chi^2_k$ distribution. To test $H_0$: $\theta = \theta_0$ versus $H_1$: $\theta \neq \theta_0$ at an asymptotic level of $\alpha$ for large sample size $n$, reject $H_0$ whenever $\theta_0$ does not belong to $I_{1-\alpha}$.

## 2.2 Empirical Likelihood Ratio Test for Constant Variance: The Two-Sample Case

In this section, an empirical likelihood ratio test for constant variance in a two-sample problem is proposed and developed. The test is nonparametric and enjoys the advantages of likelihood method.

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent samples, each from a population with a possibly different variance. One wishes to test $H_0$: the two populations have an identical variance versus $H_1$: the two populations have different variances.

Let $p_i$ and $q_j$ represent the probability masses for the observations $X_i$ and $Y_j$, respectively. The empirical likelihood function for $p_i$ and $q_j$ is given by

$$l^*(p, q) = \prod_i p_i \prod_j q_j$$

where $p_i$ and $q_j$ are subject to the constraints

$$p_i \geq 0, q_j \geq 0, \sum p_i + \sum q_j = 1. \tag{2.2.1}$$

The empirical log-likelihood function is

$$l(p,q) = \sum \log(p_i) + \sum \log(q_j).$$

Note that $l(p,q)$ defined above attains the maximum value, subject to the constraints of (2.2.1), at $p_i = q_j = 1/N$, where $N = n + m$. Therefore, the empirical log-likelihood ratio, denoted by $r(p,q)$, becomes

$$r(p,q) = \sum_{i=1}^{n} \log(Np_i) + \sum_{j=1}^{m} \log(Nq_j). \tag{2.2.2}$$

Following the idea of the Modified Levene's test, define

$$a_i = |X_i - m_x| \text{ and } b_j = |Y_j - m_y|,$$

where $m_x$ and $m_y$ represent the sample medians of the $X$ and $Y$ observations, respectively. Under the null hypothesis of constant variance, the probabilities $p$ and $q$ satisfy

$$\sum a_i p_i = \sum b_j q_j . \tag{2.2.3}$$

The maximum likelihood estimators for $\hat{p}_i$ and $\hat{q}_j$ subject to the constraints (2.2.1) and (2.2.3) can be found by the method of Lagrange multipliers. To do so, define $H(p,q,\mu,\tau)$ as

$$H(p,q,\mu,\tau) = l(p,q) + \mu\left(1 - \sum p_i - \sum q_j\right) + \tau\left(\sum a_i p_i - \sum b_j q_j\right).$$

Setting the partial derivatives of $H(p,q,\mu,\tau)$ to be zero results in the following equations:

$$\frac{\partial}{\partial p_i} H = \frac{1}{p_i} - \mu + \tau\alpha_i = 0, \tag{2.2.4}$$

$$\frac{\partial}{\partial q_j} H = \frac{1}{q_j} - \mu - \tau b_j = 0, \tag{2.2.5}$$

$$\frac{\partial}{\partial \mu} H = 1 - \sum_{i=1}^{n} p_i - \sum_{j=1}^{m} q_j = 0, \tag{2.2.6}$$

$$\frac{\partial}{\partial \tau} H = \sum a_i p_i - \sum b_j q_j = 0. \tag{2.2.7}$$

From the equations (2.2.4) and (2.2.5), we have

$$n = \mu \sum p_i - \tau \sum a_i p_i \tag{2.2.8}$$

$$m = \mu \sum q_j + \tau \sum b_j q_j. \tag{2.2.9}$$

Adding the two equations (2.2.8) and (2.2.9) and then inserting the equations (2.2.6) and (2.2.7) yields

$$\mu = N.$$

From the equations (2.2.4) and (2.2.5) again, we obtain the maximum likelihood estimators

$$\hat{p}_i = \frac{1}{N - \hat{\tau} a_i}, \qquad \hat{q}_j = \frac{1}{N + \hat{\tau} b_j},$$

where $\hat{\tau}$ is the root of

$$\sum_{i=1}^{n} \frac{a_i}{N - \tau a_i} = \sum_{j=1}^{m} \frac{b_j}{N + \tau b_j}. \tag{2.2.10}$$

It is proved (see, e.g., Owen, 2001) that $\hat{\tau}$ approaches 0 when the sample sizes go to ∞. As a result, $\hat{p}_i$ and $\hat{q}_j$ are positive when *N* is large so that they strictly satisfy the constraints of nonnegativity. Complementary slackness therefore ensures no need for including all-zero dual variables for the inequalities from constraints (2.2.1) in our Lagrangian formulation (Karush,

1939 and Kuhn and Tucker, 1951). Using the equation (2.2.2), we arrive at the empirical likelihood ratio test statistic for constant variance as follows:

$$T = -2\left\{\sum \log\left(\frac{N}{N - \hat{\tau}a_i}\right) + \sum \log\left(\frac{N}{N + \hat{\tau}b_j}\right)\right\}. \quad (2.2.11)$$

When the sample sizes go to ∞, the asymptotic null distribution of $T$ is the $\chi_1^2$-distribution (Owen, 2001). Consequently, the test is to reject the null hypothesis whenever $T$ is greater than the critical value of the $\chi_1^2$-distribution with a preset significance level such as $\alpha$. At the level of $\alpha$ = 5%, for example, reject the null hypothesis whenever $T > 3.84$.

Recall $a_i = |X_i - m_x|$ and $b_j = |Y_j - m_y|$. Compare the quantities

$$\sum |X_i - m_x|\hat{p}_i \text{ and } \sum |Y_j - m_y|\hat{q}_j$$

with the means of Levene's distance, which are defined, respectively, as

$$\bar{d}_1 = \sum |X_i - m_x|\frac{1}{n}, \qquad \bar{d}_2 = \sum |Y_j - m_y|\frac{1}{m}. \quad (2.2.12)$$

It is seen that the empirical likelihood ratio test uses optimal weights in summing the absolute deviations of observations from the median values, while the modified Levene's test uses the simple averages.

## 2.3 Extension

The empirical likelihood ratio test proposed in Section 2.2 in the setup of two-sample problem can be extended to one-way layout models or more generally to the linear regression models, without substantial difficulties.

In the case of a one-way layout model with $k$ levels, let $Y_{ij}$ be the response of the $i^{th}$ level and $j^{th}$ repetition, $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$. Let $N = n_1 + \cdots + n_k$ be the total sample size (Dielman, 2005). As before, define

$$d_{ij} = |Y_{ij} - M_i|$$

where $M_i$ is the sample median of the responses $Y_{i1}, \ldots, Y_{in_i}$. Let $\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}$ be the roots of the following equations:

$$\sum_{j=1}^{n_1} \frac{d_{1j}}{N + \tau_1 d_{1j} + \cdots + \tau_{k-1} d_{(k-1)j}} = \sum_{j=1}^{n_i} \frac{d_{ij}}{N - \tau_{i-1} d_{ij}}, i = 2, \ldots, k. \quad (2.3.1)$$

Define

$$\hat{p}_{1j} = \frac{1}{N + \tau_1 d_{1j} + \cdots + \tau_{k-1} d_{(k-1)j}}, \hat{p}_{ij} = \frac{1}{N - \tau_{i-1} d_{ij}}, j = 1, \ldots, n_i \; i = 2, \ldots, k, \quad (2.3.2)$$

and

$$T = -2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log\left(N \hat{p}_{ij}\right). \quad (2.3.3)$$

As the sample sizes go to ∞, the asymptotic null distribution of $T$ is the $\chi^2_{k-1}$ distribution. We reject the null hypothesis of constant variance whenever $T$ is greater than the critical value at significance level $\alpha$.

Note that when *k* is equal to 2, the setup reduces to the two-sample problem considered in earlier sections. Here we would like to remark that when *k* is moderate or large, the roots $\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}$ of the nonlinear equations in (2.3.1) may be difficult to obtain or approximate by any means, including the Newton-Raphson algorithm.

In the case of a linear (simple or multiple) regression model, we divide the data set into *k* groups, according to the values of the covariates (2005). Let $\varepsilon_{ij}$ be the $j^{\text{th}}$ residual from the $i^{th}$ group, and let $\tilde{\varepsilon}_i$ represent the median of the residuals in the $i^{\text{th}}$ group. Again, define

$$d_{ij} = |\varepsilon_{ij} - \tilde{\varepsilon}_i|,$$

for $j = 1, \ldots, n_i$ and $i = 2, \ldots, k$. With the group data, the likelihood ratio test for homogeneity versus heterogeneity under the linear regression model can be performed in the same way as the one-way layout model described earlier in this section.

# CHAPTER 3: APPLICATIONS TO REAL-LIFE DATA

To illustrate the proposed empirical likelihood ratio test for constant variance, we used Mortgage Defaulter data. A part of this data is shown below in Figure 3.1 for preview.

| | Bo_Age | Ln_Orig | Orig_LTV_Ratio_Pct | Credit_score | First_home | Tot_mthly_debt_exp | Tot_mthly_incm | orig_apprd_val_amt |
|---|---|---|---|---|---|---|---|---|
| 1 | 32 | 148000 | 100 | 759 | N | 0 | 4246 | 204872 |
| 2 | 40 | 168701 | 100 | 765 | N | 595 | 3200 | 118933 |
| 3 | 34 | 111000 | 100 | 691 | N | 1299 | 4150 | 130000 |
| 4 | 26 | 115500 | 98 | 665 | N | 969 | 2957 | 210000 |
| 5 | 42 | 96000 | 100 | 788 | N | 1072 | 6270 | 136000 |
| 6 | 26 | 86000 | 100 | 757 | N | 0 | 6940 | 127855 |

| | pur_prc_amt | DTI.Ratio | OUTCOME | State | Median_state_inc | UPB.Appraisal |
|---|---|---|---|---|---|---|
| 1 | 148000 | 0.000000 | non-default | FL | 40,171 | 0 |
| 2 | 118933 | 0.185938 | non-default | FL | 40,171 | 1 |
| 3 | 111000 | 0.313012 | non-default | TN | 38,550 | 0 |
| 4 | 115500 | 0.327697 | non-default | OH | 44,160 | 0 |
| 5 | 96000 | 0.170973 | non-default | NV | 46,984 | 0 |
| 6 | 86000 | 0.000000 | non-default | NC | 39,000 | 0 |

Figure 3.1 Preview of the dataset.

For illustration purposes, those who have taken out home loans were placed into two categories, where the loan amount is less than the appraised value of the home ($m$ = 8317, UPB.Appraisal = 0), or exceeds the appraised value of the home ($n$ = 6836, UPB.Appraisal = 1) (Bruce and Shmueli, 2018). The dataset contains 14 variables, which include

- *Borrower age at the time of the loan*
- *Value of the loan in US Dollars*
- *Total Monthly Debt*
- *Total Monthly Income*
- *Loan-To-Value*
- *Credit Score*
- *The appraised value of the home in US Dollars*
- *Purchase price of the home*
- *Debt-To-Income ratio*

- *Outcome*
- *The state of residence*
- *Median state income*
- *If loan amount exceeds home appraisal value (1 if Exceeds, 0 Otherwise).*

To conduct constant variance testing, we elected to use the *Loan-To-Value* variable. It has two groups, one where loan amount exceeds the home appraisal value and one where loan amount is less than the home appraisal value. In its entirety, the variable has a mean value of 93.08, where the amount loaned out for a home buyer, on average, is about 93% of the actual value of the home.

One sample of size 95 was taken on the variable *Loan-To-Value* from the group where loan amount exceeds the home appraisal value and another sample of size 95 was taken on the same variable but from the group where loan amount is less than the home appraisal value. The two samples were independently selected. We want to test constant variance with the two samples. For the samples drawn, the respective variances were 143.27 for the group where loan amount exceeds home appraisal value, and 93.22 for the group where loan amount was less than home appraisal value.

In applications, a Q-Q plot is often used to visually verify that the normality assumption is valid. A Q-Q plot is a probability plot of normal quantiles versus the sample quantiles. When the normality assumption holds, i.e., the samples are from a normal population, the Q-Q plot is expected to be linear. A curved Q-Q plot indicates violation of the normality assumption.

The data are displayed in Table 3.1 and Table 3.2, respectively. A Q-Q plot shown in Figure 3.2 clearly indicates serious violations of the normality assumption, confirming the appearance of skewed distributions in the sample histograms.

Figure 3.3 contains the histograms for the two samples.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 95 | 100 | 102 | 90 | 97 | 90 | 80 | 100 | 100 | 102 | 90 | 100 | 102 | 71 |
| 99 | 86 | 95 | 100 | 102 | 100 | 93 | 95 | 90 | 95 | 58 | 80 | 80 | 100 | 95 |
| 95 | 80 | 100 | 95 | 90 | 100 | 100 | 102 | 100 | 80 | 80 | 95 | 95 | 102 | 97 |
| 90 | 80 | 100 | 80 | 100 | 90 | 95 | 100 | 95 | 100 | 100 | 102 | 100 | 95 | 71 |
| 95 | 73 | 95 | 73 | 95 | 97 | 80 | 46 | 100 | 95 | 90 | 95 | 100 | 90 | 90 |
| 90 | 80 | 100 | 99 | 95 | 100 | 100 | 80 | 42 | 95 | 85 | 70 | 80 | 90 | 90 |
| 74 | 80 | 100 | 95 | 100 | | | | | | | | | | |

Table 3.1 A random sample of size 95 from the group where loan amount exceeds the home appraisal value on the variable *Loan-To-Value*.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 100 | 80 | 100 | 78 | 95 | 95 | 95 | 90 | 100 | 90 | 95 | 80 | 100 | 90 | 95 |
| 100 | 60 | 100 | 90 | 90 | 100 | 80 | 80 | 80 | 97 | 90 | 80 | 90 | 100 | 47 |
| 100 | 97 | 100 | 97 | 97 | 93 | 100 | 100 | 102 | 80 | 100 | 80 | 100 | 100 | 95 |
| 100 | 87 | 87 | 79 | 100 | 90 | 80 | 90 | 100 | 94 | 95 | 100 | 80 | 100 | 80 |
| 95 | 85 | 74 | 95 | 88 | 80 | 80 | 90 | 87 | 97 | 97 | 100 | 100 | 100 | 95 |
| 80 | 90 | 100 | 85 | 90 | 99 | 100 | 99 | 100 | 95 | 90 | 90 | 100 | 80 | 99 |
| 100 | 80 | 100 | 100 | 100 | | | | | | | | | | |

Table 3.2 A random sample of size 95 from the group where loan amount is less than the home appraisal value on the variable *Loan-To-Value*.

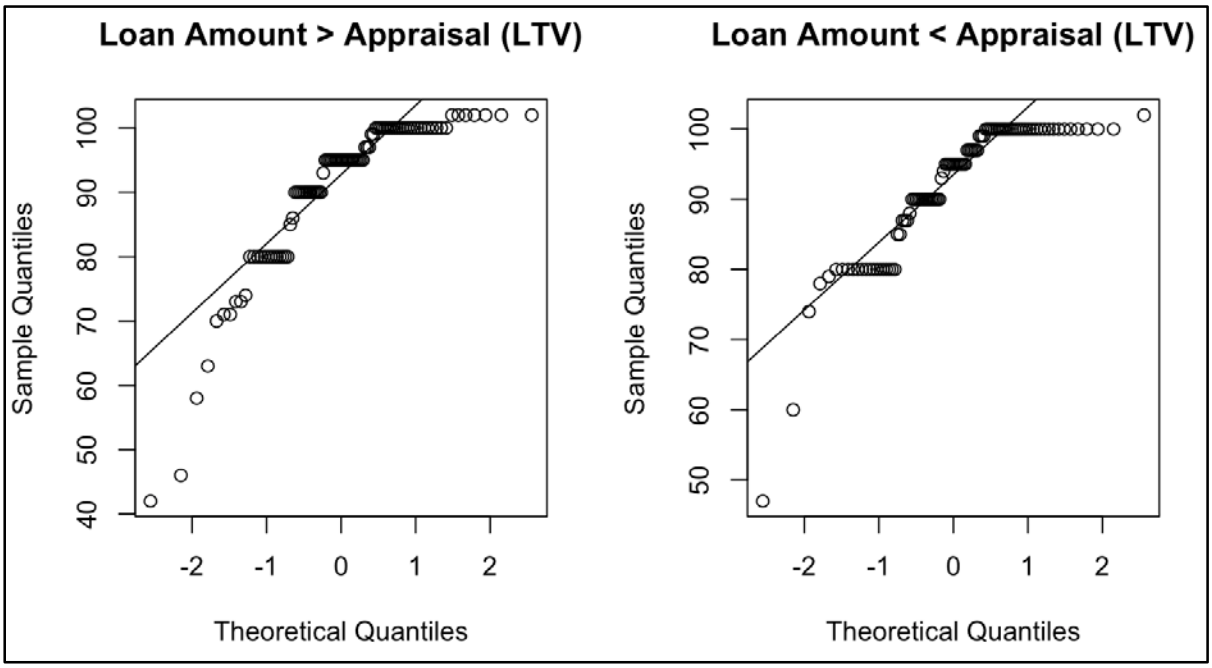


Figure 3.2 Q-Q Plots of the two samples drawn from the variable *Loan-To-Value*.

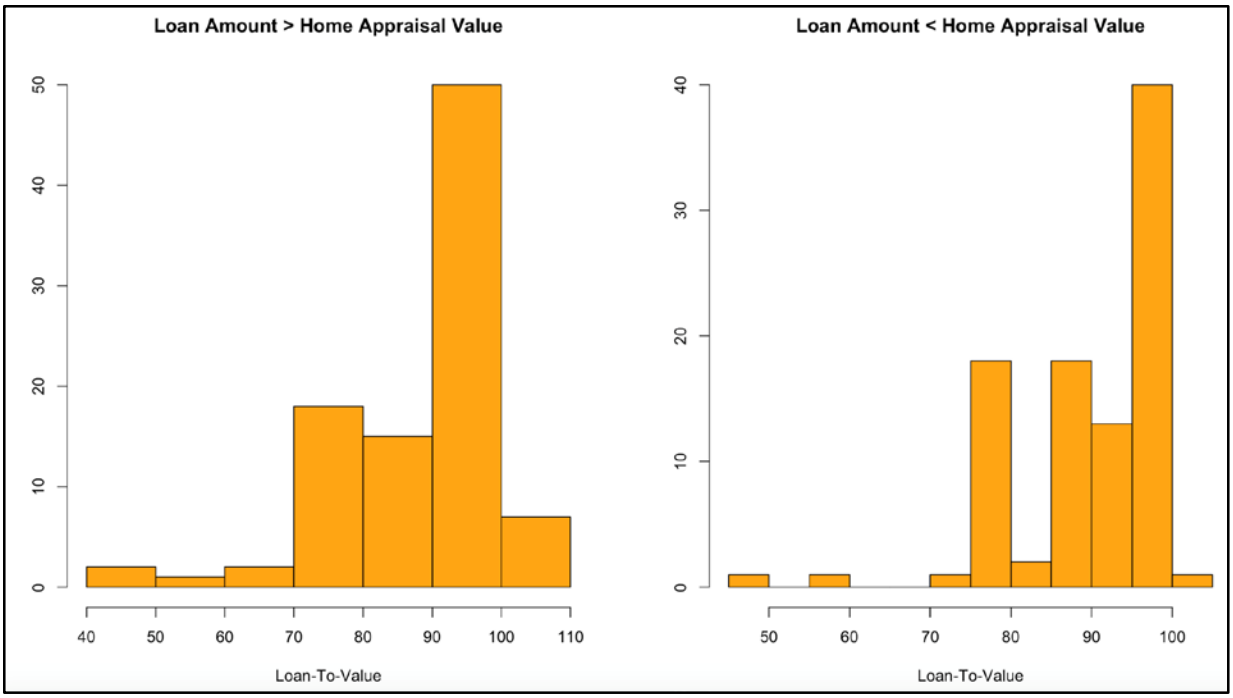


Figure 3.3 Histograms of the two samples drawn from the variable *Loan-To-Value*.

## 3.1 The *F*-Test

For the two-sided *F*-test, it has 94 and 94 degrees of freedom. At a significance level of 5%, the lower critical value is 0.67, and the upper critical value is 1.50. The calculated test statistic of 1.537 falls outside the interval of critical values, so the decision rule from the test is to reject the null hypothesis of constant variance. However, since the assumption of normality is violated, the result from the *F*-test cannot be trusted. The test results are shown in Table 3.3 below.

| Group | Degrees of Freedom | $s^2$ | *F** | *p*-value |
|---|---|---|---|---|
| Loan amount > Appraisal value | 94 | 143.269 | 1.537 | 0.0384 |
| Loan amount < Appraisal value | 94 | 93.218 | | |
| Total | 188 | | | |

Table 3.3 Two-sample results for the classic *F*-test.

## 3.2 The Modified Levene's Test

For the Modified Levene's Test, the *t*-distribution has 95 + 95 – 2 = 188 degrees of freedom. As a result, at a significance level of 5%, the critical value for the two-sided Modified Levene's Test is ±1.97. The observed test statistic $L$ of 0.7292 is well within the critical region for this test so there is not enough evidence to reject the null hypothesis of constant variance. With the two samples, calculations are given in Table 3.4 below.

| $\bar{d}_1$ | $\bar{d}_2$ | $S_L$ | $L$ | $p$-value |
|---|---|---|---|---|
| 8.1368 | 7.2316 | 8.5552 | 0.7292 | 0.4668 |

Table 3.4 Two-sample results for the Modified Levene's Test.

## 3.3 The Empirical Likelihood Ratio Test

For the empirical likelihood approach revealed that for this particular sample, we need to find the root $\hat{\tau}$ to the equation (2.2.10), which is shown below:

$$\sum_{i=1}^{n} \frac{a_i}{N - \tau a_i} = \sum_{j=1}^{m} \frac{b_j}{N + \tau b_j}.$$

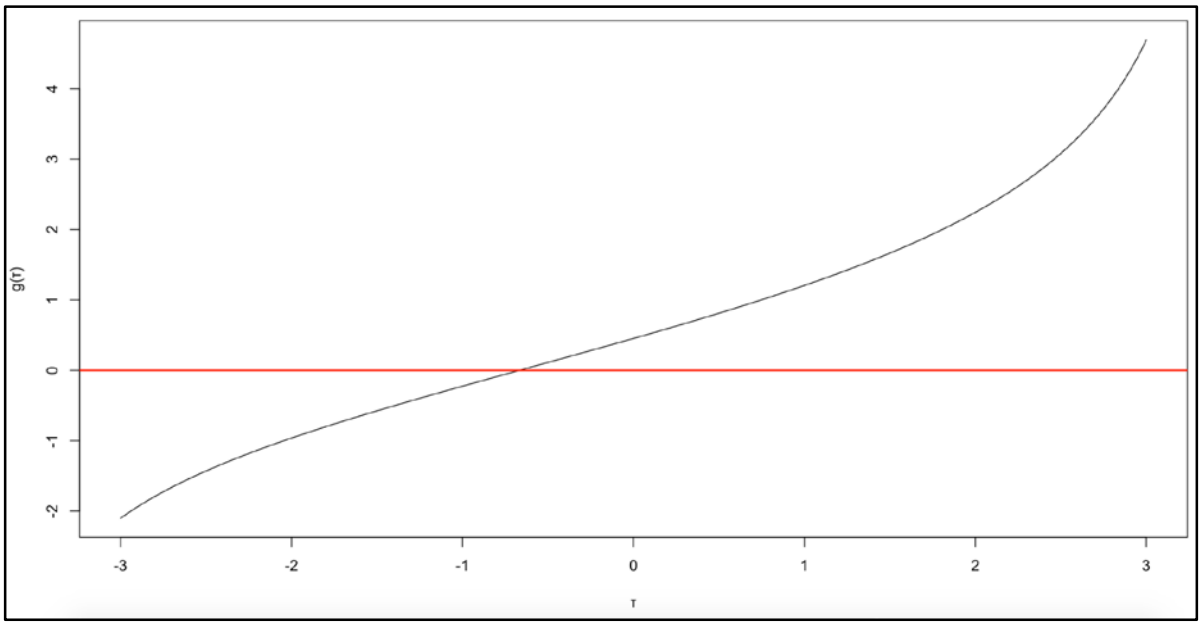


Figure 3.4 A graphical representation of equation (2.2.10).

At a value of -2/3, $\hat{\tau}$ satisfies the equation given in (2.2.10). Figure 3.4 represents the equation being evaluated at multiple values within the interval [-3, 3]. Likewise, with the Modified Levene's test, there was insufficient evidence from the empirical likelihood ratio test to reject the null hypothesis of constant variance. Results of the empirical likelihood approach are given in Table 3.5 below.

| $\hat{\tau}$ | *r(p,q)* | *T* | *p*-value |
|---|---|---|---|
| -2/3 | -0.1502 | 0.3004 | 0.5836 |

Table 3.5 Two-sample results for the empirical likelihood ratio test.

## 3.4 Summary of the Testing Results

Among the three tests performed, the *F*-test has a *p*-value of 0.0384 so that it rejects the null hypothesis at a significance as small as 3.9%. Both the Modified Levene's Test and the new empirical likelihood ratio test perform consistently with the *p*-values 0.4668 and 0.5836, respectively. As the Q-Q plot indicates, the normality assumption is seriously violated. We tend to believe that the results of the Modified Levene's Test and the new empirical likelihood ratio test are more trustful. In this sense, the proposed empirical likelihood ratio test does the best job.

# CHAPTER 4: CONCLUSION

The classic *F*-test and the Modified Levene's Test are two most popular tests for constant variance in applications. The former is formed under normality assumption, while the latter can be viewed as a result of moment estimation method. The performance of the Modified Levene's Test is also dependent upon the distribution shape in certain ways, though not as much as the *F*-test is. The empirical likelihood ratio test for constant variance is a nonparametric test and retains the principle of likelihood. To be specific, the empirical likelihood ratio test of constant variance uses the optimal weights in summing the absolute deviations of observations from the median values, while the Modified Levene's test uses the simple averages. It is thus desired that the empirical likelihood ratio test is more powerful than the Modified Levene's test. Meanwhile, the empirical likelihood ratio test is expected to be as robust as the Modified Levene's test, as the empirical likelihood ratio test is also constructed via the Levene's distance, i.e., the absolute deviation from the median.

As a result, it can be an appropriate alternative to the two traditional tests in applications with skewed underlying populations. With an analysis of a real-life data set, the empirical likelihood ratio test appears to perform the best.

# BIBLIOGRAPHY

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association,* 364–367.

Bruce, P. and Shmueli, G. (2018). "Mortgage Defaulters." *MortgageDefaulters* - www.coursehero.com/file/17306296/MortgageDefaulters/.

Dielman, T. E. (2005). *Applied Regression Analysis: A Second Course in Business and Economic Statistics*, *4th Ed*. Cole Thomson Learning.

Karush, W. (1939). "Minima of Functions of Several Variables with Inequalities as Side Constraints". M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois.

Kuhn, H. W.; Tucker, A. W. (1951). "Nonlinear programming". *Proceedings of 2nd Berkeley Symposium*. Berkeley: University of California Press. pp. 481–492. MR 0047303.

Levene, H. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling In Ingram Olkin; Harold Hotelling by et al*. Stanford University Press. pp. 278–292.

Owen, A.B. (1988). Empirical likelihood ratio confidence interval for a single functional. *Biometrika*, 237-249.

Owen, A.B. (1990). Empirical likelihood confidence regions. *The Annals of Statistics*, 90-120.

Owen, A.B. (2001). *Empirical Likelihood*. Chapman & Hall.

Utts, J. M. and Heckard, R. F. (2011). *Mind on Statistics*. Cengage Learning.