DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

TYPE-M ROBUST MEASURES OF CORRELATION

by

Kübranur ERGÜN

February, 2018

İZMİR

TYPE-M ROBUST MEASURES OF CORRELATION

A Thesis Submitted to the

Graduate School of Natural and Applied Sciences of Dokuz Eylül University

In Partial Fulfillment of the Requirements for the Master of Science in

Statistics, Statistics Program

by

Kübranur ERGÜN

February, 2018

İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"TYPE-M ROBUST MEASURES OF CORRELATION"** completed by **KÜBRANUR ERGÜN** under supervision of **ASSOC. PROF. DR. A. FIRAT ÖZDEMİR** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. A. Fırat ÖZDEMİR

Supervisor

Yrd. Doç. Dr. Burcu Aytaçoğlu

(Jury Member)

Asst. Prof. Dr. Engin Yıldıztepe

(Jury Member)

Prof. Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENT

I would like to express my gratitude to my advisor Assoc. Prof. Dr. A. Fırat ÖZDEMİR for his patience, motivation and support throughout my thesis.

Besides my advisor, I would like to thank all of academic staff in Dokuz Eylül University Department of Statistics for their support.

Finally, I want to thank to my mother Ayşe ERGÜN, my father Ercan ERGÜN, my brother Eray ERGÜN and my love Dr. İbrahim ABDELLA for their endless love, support and patience during my thesis. I am also grateful to Tuğçe TARIM, Şebnem KABAL, Merve MAVİ, Evrim SINAK and Tuğçe IRMAK for their moral and support.

Kübranur ERGÜN

TYPE-M ROBUST MEASURES OF CORRELATION

ABSTRACT

Correlation analysis determine the degree and direction of the relationship between two variables. The well known correlation is Pearson’s correlation. But it is not robust when normality and homoscedasticity assumptions are violated. Therefore, robust methods are improved to minimize effect of violation of assumptions. Robust measures of correlations are classified into two types. First type protects against outliers among the marginal distributions without taking into account the overall structure of the data (Type-M) and second type that take into account the overall structure of the data when dealing with outliers (Type-O).

In this study, five Type-M robust correlations (percentage bend correlation, Winsorized correlation, biweight midcorrelation, Kendall’s Tau and Spearman’s rho) are analyzed and compared when they are used with a heteroscedastic test procedure in terms of actual significance level. The percentage bend, Kendall’s tau and Spearman’s rho correlations saved nominal significance level much better Winsorized and Biweight correlations with using percentile bootstrap method according to simulation study

Keywords: Type-M robust correlation, percentage bend correlation, biweight correlation, winsorized correlation, Kendall’s tau, Spearmans’s rho

M-TİPİ DAYANIKLI İLİŞKİ ÖLÇÜLERİ

ÖZ

Korelasyon analizi, iki değişken arasındaki ilişkinin derecesini ve yönünü belirler. En iyi bilinen korelasyon Pearson korelasyonudur. Fakat normallik ve homojen varyans olma varsayımları sağlanmadığında dayanıklı değildir. Bu nedenle, sağlanmayan varsayımların etkisini en aza indirgemek için dayanıklı yöntemler geliştirilmiştir. Dayanıklı ilişki ölçüleri iki tipte sınıflandırılmıştır. Birinci tip (M tipi), verilerin genel yapısını dikkate almadan marjinal dağılımlar arasında uç değerlere karşı korur ve ikinci tip (O tipi), verilerin genel yapısını da dikkate alarak uç değerlere karşı korur.

Bu çalışmada 5 M-tipi dayanıklı ilişki yöntemi (yüzdelik kırılma korelasyonu, Winsorized korelasyon, ikili korelasyon, Kendall’s Tau and Spearman’s rho) analiz edilmiştir ve heterojen varyans test prosedürü ile gerçek anlamlılık düzeyleri yönünden karşılaştırılmıştır. Simülasyon çalışmasına göre; yüzdelik kırılma, Kendall’s tau ve Spearman’s rho yüzdelik bootstrap metodu kullanılarak gerçek anlamlılık düzeyini korumada Winsorized ve ikili korelasyona göre daha iyi sonuç vermiştir.

Anahtar kelimeler: M tipi dayanıklı korelasyon, yüzdelik kırılma korelasyonu, Winsorized korelasyon, ikili korelasyon, Kendall’s tau, Spearman’s rho

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE

INTRODUCTION

Correlation analysis is the statistical method used to determine the degree and direction of the relationship between two variables. For instance, an increment of weight with height level or, as the house's square meters increase, its price also increases. Correlation analysis investigates the relationship between the two variables whether there is a relationship between them or whether they changed in the opposite direction. In order to manage and see the relationship between variables, scatter plot can be drawn. The scatter plot also indicates whether relation exists among variables, as well as its direction and degree. The order of the points in the scatter from the bottom to top indicates the relationship between variables are in the same direction, while from top to bottom it indicates the relationship is in opposite directions. The closer the points are to the straight line, the stronger the relationship between the variables.

It is important that the strength of the relationship is strong or weak. Because strong relationship allows you to make better predictions than a weak relationship (Higgins, 2005). The most well known correlation coefficient is Pearson's product moment correlation coefficient, or Pearson's r that was developed by Karl Pearson from a connected idea introduced by Sir Francis Galton at the end of 1800's. (Chee, 2015). Pearson's correlation is a measure of the linear relationship between two interval or ratio variables and can take a value between -1 and 1, and is indicated by "ρ (rho)" (Cramer, 1998). The formula for Pearson Correlation Coefficient;

$$\rho=\frac{cov(X,Y)}{\sigma_X\sigma_Y} \tag{1.1}$$

$$cov(X,Y)=E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right] \tag{1.2}$$

Then the formula is also written:

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1.3}$$

Pearson's correlation coefficient when applied to a sample is commonly indicated by the letter r (Higgins, 2005). Let X and Y are random variables and the observed sample $(x_1, y_1), \ldots, (x_n, y_n)$ of a bivariate random variable (X, Y) the classical estimator of the correlation coefficient ρ is given by the sample correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{1/2}} \tag{1.4}$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ and $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ are the sample means (Shevlyakov and Smirnov, 2011).

If we observe positive relationship (Figure 1.1), it means that when one variable increases, the other variable also increases. Similarly, when both decreasing are also true.

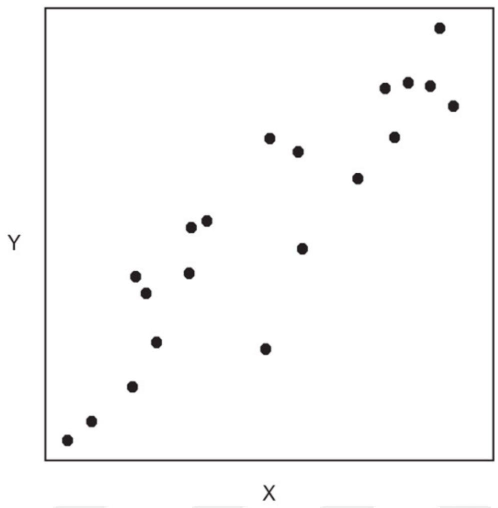

Figure 1.1 Positive correlation

If we observe negative relationship (Figure 1.2), it means that when one variable increases, the other variable decreases. The variables tend to move in the opposite direction.

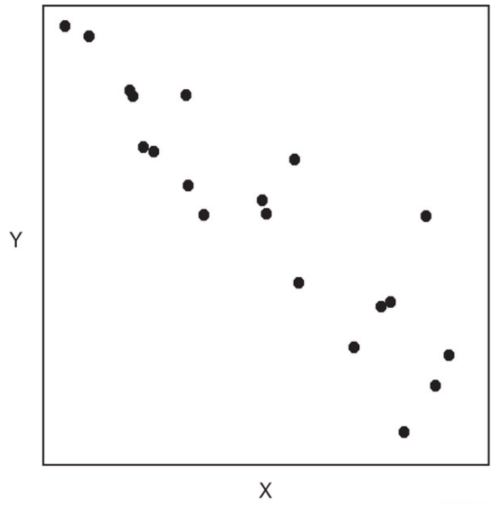

Figure 1.2 Negative correlation

If there is no relationship (Figure 1.3), it means that no predictable change between variables.

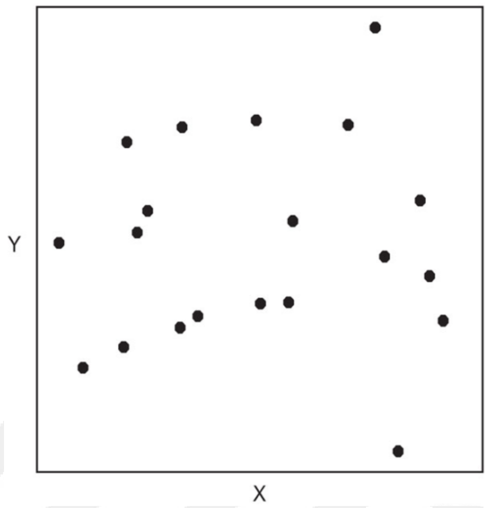

Figure 1.3 No relationship

If we observe data set like Figure 1.4, it means that there is a nonlinear relationship between two variables.

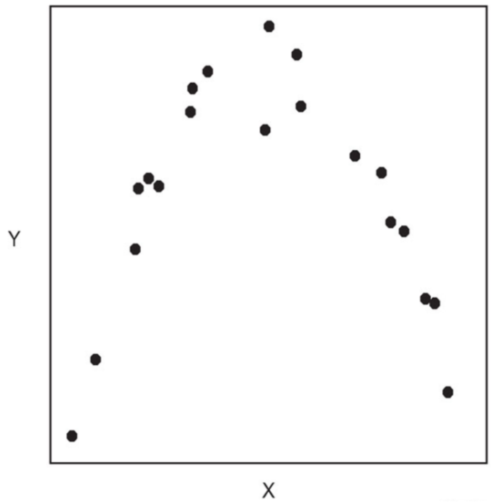

Figure 1.4 Nonlinear relationship

Assumptions

Pearson's r have some assumptions: linearity and normality. When we calculate the correlation coefficient, we consider that the correlation is linear. If the relationship is curvilinear, we use correlation index instead of correlation coefficient. Data set has to be normally distributed and we use interval data to determine a linear relationship (Field, 2009). Homoscedasticity assumes that the variance is the same for all values of the explanatory variable (X) along the regression line. If we don't take into account the assumption of homoscedasticity, the chances of obtaining a statistically significant result increases even though H_0 is true (Chee, 2015).

Significance Test of Pearson's r

The aim of testing the hypothesis is to decide whether or not the association between two random variables is zero. Surely, the best-known method is to test the hypothesis that Pearson's correlation is zero, using Student's T.

$$H_0: \rho=0 \tag{1.5}$$

$$T=r\sqrt{\frac{n-2}{1-r^2}} \tag{1.6}$$

If H_0 is true, T has a Student's t – distribution with $v = n - 2$ degrees of freedom (Muirhead, 1982). Particularly, reject H_0 if $|T| > t_{1-\alpha/2}$, the 1-α/2 quantile of Student's t-distribution with $n - 2$ degrees of freedom When X and Y are independent, there are general conditions under which $E(r) = 0$ and $E(r^2) = 1/(n - 1)$ (Huber, 1981).

If one of the marginal distributions is changed, the magnitude of ρ can be changed largely. That's why Pearson's correlation is not robust. The test of independence, based on T, will be relatively robust in terms of type 1 errors. But the case can depend on a variety of situations. Especially problems originate in three situations:

1. When $\rho = 0$, but X and Y are dependent (e.g, Edgell & Noon, 1984).
2. When acting one-sided tests (Blair & Lawson, 1982).
3. When considering the more general main aim of testing for independence among all pairs of ρ random variables (Wilcox, 2012). There is also the problem of computing a confidence interval for ρ.

Although many methods have been proposed, simulations do not support their use, any particular method give good probability coverage because it is not known how large of a sample size is needed (Wilcox, 1991a). A modified percentile method seems well in terms of probability coverage provided ρ is relatively close to zero. But if ρ gets close to one, it begins to fail (Wilcox & Muska, 2001).

Pearson correlation and test of independence (T) have some disadvantages. They are affected by (Wilcox, 2012);

- Outliers

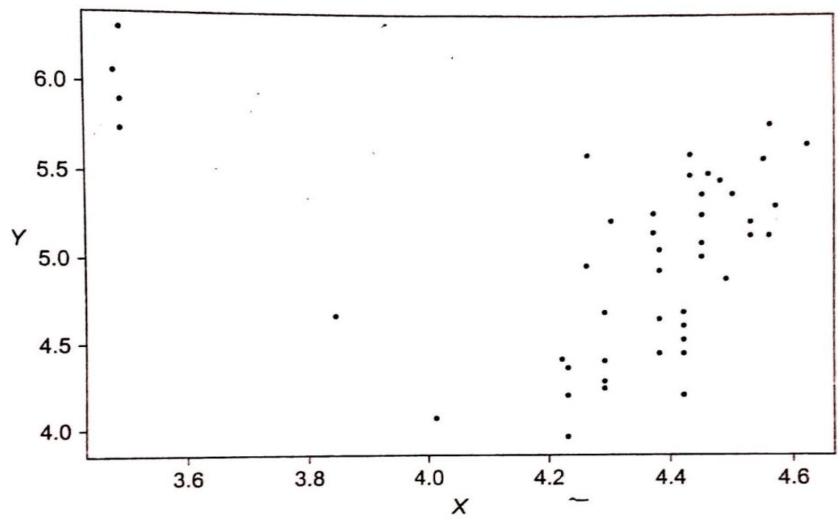

Figure 1.5 Outliers

- The magnitude of the slope around which are clustured

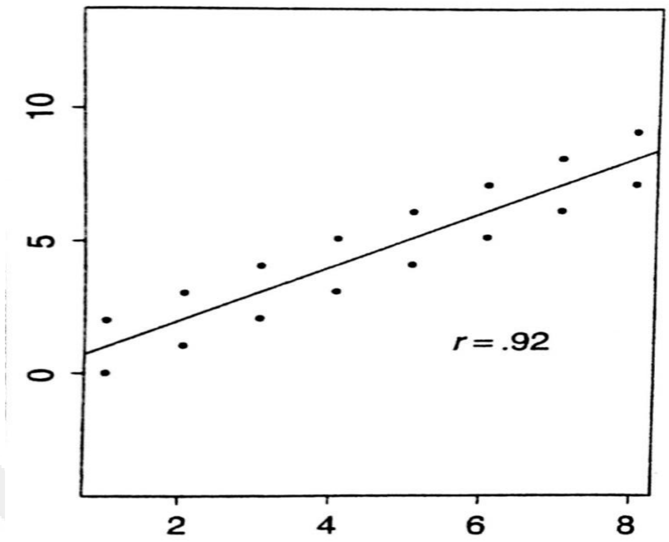

Figure 1.6 The magnitude of slope

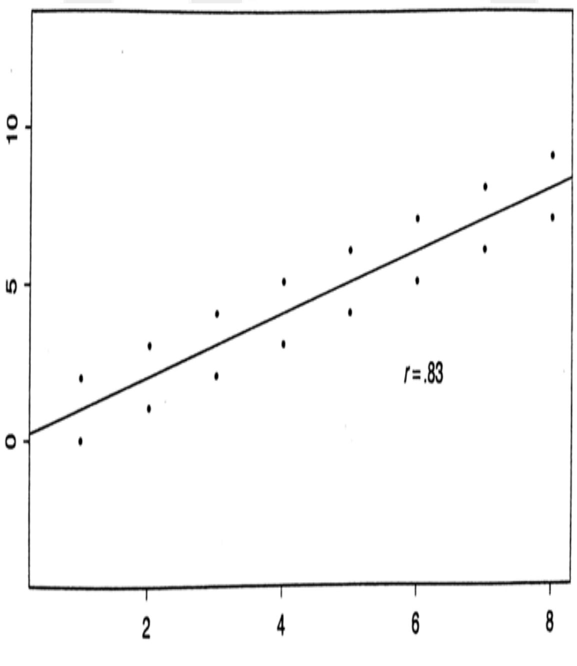

Figure 1.7 The magnitude of slope

- Curvature

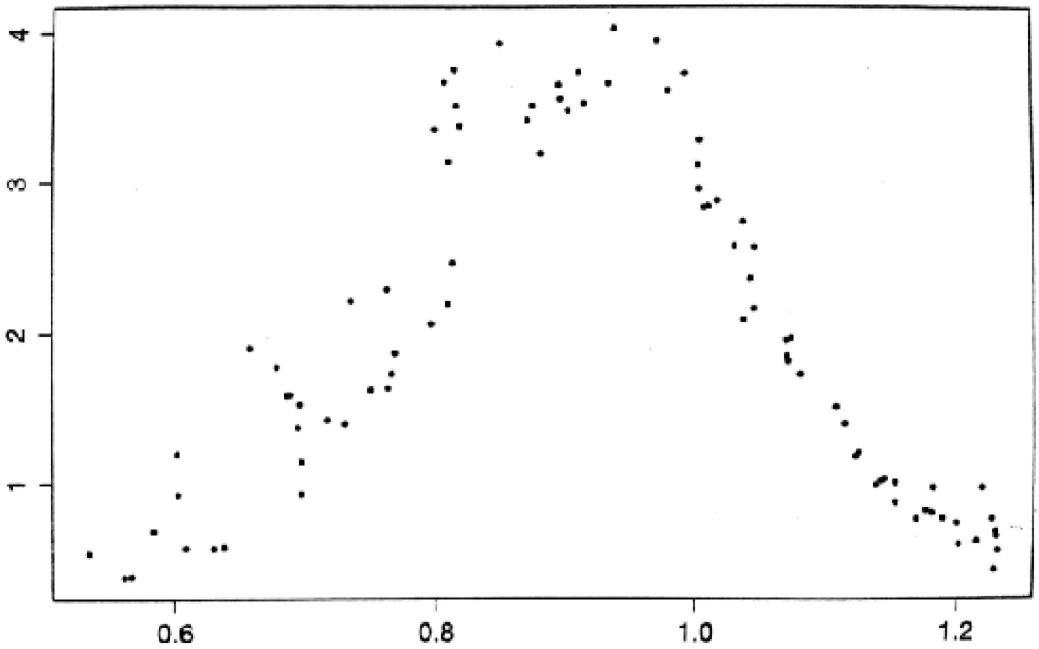

Figure 1.8 Curvature

- The magnitude of residuals

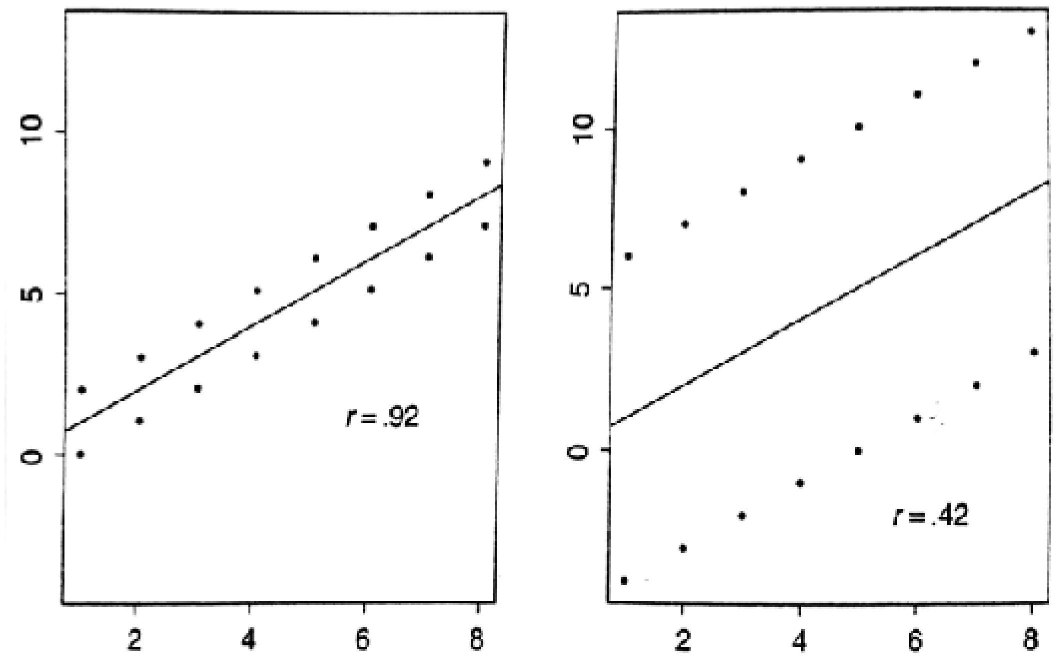

Figure 1.9 The magnitude of residuals

- Restriction of range

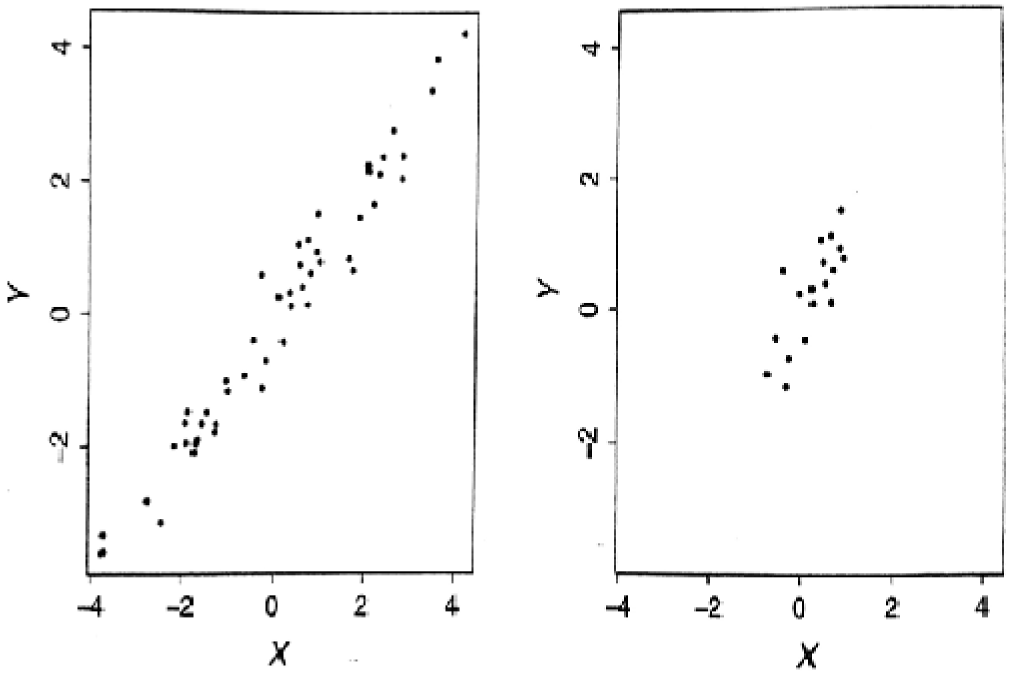

Figure 1.10 Restriction of range

Another disadvantage is also heteroscedasticity which refers to a situation where the conditional variance of Y varies with X. Heteroscedasticity is related to the test of H_0. Because the derivation of the test statistic T is based on the assumption that X and Y are independent. Even if $\rho = 0$, but there is heteroscedasticity, the wrong standard error is being used by T. (Wilcox, 2012).

When normality and homoscedasticity assumptions are violated, it can lead to increase rate of Type 1 Error. Robust methods are designed to perform well when assumptions are violated (Erceg-Hurn&Mirosevich, 2008).

Although there are many robust correlation methods, none of them has been chosen as the best one among the others so far. The aim of this study is to examine and compare five Type-M robust correlation methods. These methods are; percentage bend correlation, Winsorized correlation, biweight midcorrelation, Kendall's Tau and Spearman's rho. Traditional tests of zero correlation of all these correlations are sensitive to heteroscedasticity. We applied a percentile bootstrap method that is insensitive to heteroscedasticity to test the hypothesis of zero correlation for all these Type-M robust correlations Actual Type 1 Error probabilities (actual significance levels) are compared for different heteroscedasticity patterns and distributions. Results and recommendations are given.

This thesis is organized in four chapters. General information about the types of robust correlation and detailed information about Type-M robust measures of correlations (the percentage bend correlation, Winsorized correlation, biweight midcorrelation, Kendall's Tau and Spearman's rho) are given in the second chapter. In the third chapter; the design and the results of the simulation study are given. Finally, conclusions and some recommendations are given in the fourth chapter.

CHAPTER TWO
TYPE-M ROBUST MEASURES OF CORRELATION

2.1 Types of Robust Correlation

Pearson's correlation is the most well known relationship measure. When there is a small change in one of the marginal distributions, Pearson's correlation is significantly affected by this situation. At the same time, if there is a single outlier in the distribution, Pearson's correlation is again affected. Therefore, Pearson's correlation is not robust and it does not give accurate results in case where the assumptions are violated.

Robust correlations are classified into two types. Type-M robust correlations protect against outliers without taking into account overall structure of the data. Type-O robust correlations take into account the overall structure of data when dealing with outliers. Working with the Type-M is easier to develop tests of the independence between two random variables. But working with Type-O robust correlations sometimes are preferable according to progress.

Five Type-M robust correlations methods are described in this section.

2.1.1 The Percentage Bend Correlation

When testing the hypothesis of independence, percentage bend correlation is relatively successful in terms of Type 1 error probabilities. It is indicated by "ρ_{pb}".The steps of ρ_{pb} is performed as belows:

(1) Let (X_1, Y_1), …, (X_n, Y_n) be a random sample.

(2) M_x is median for the observations X_1,…,X_n. $(1-\beta)n$ is rounded down to the nearest value and $0 \leq \beta \leq 0.5$.

$$W_i = |X_i - M_x| \tag{2.1}$$

$$m = [(1-\beta)n] \tag{2.2}$$

(3) We sort W_i values in ascending order.

$$W_{(1)} \leq \ldots \leq W_{(n)} \tag{2.3}$$

$$\widehat{\omega}_{(x)} = W_{(m)} \tag{2.4}$$

(4) Let i_1 be the number of $X_{(i)}$ values such that $(X_i - M_x)/\widehat{\omega}_{(x)} < -1$ and i_2 be the number of $X_{(i)}$ values such that $(X_i - M_x)/\widehat{\omega}_{(x)} > 1$.

$$S_x = \sum_{i=i_1+1}^{n-i_2} X_i \tag{2.5}$$

$$\widehat{\phi}_x = \frac{\widehat{\omega}_{(x)}(i_2 - i_1) + S_x}{n - i_1 - i_2} \tag{2.6}$$

(5) Set $U_i = (X_i - \widehat{\phi}_x)/\widehat{\omega}_{(x)}$. Repeat these computations for the Y_i values yielding $V_i = (Y_i - \widehat{\phi}_y)/\widehat{\omega}_{(y)}$.

$$\psi(x) = max[-1, min(1,x)] \tag{2.7}$$

(6) Set $A_i = \psi(U_i)$ and $B_i = \psi(V_i)$. The percentage bend correlation between X and Y is estimated to be (Wilcox, 2012)

$$r_{pb}=\frac{\sum A_i B_i}{\sqrt{\sum A_i^2 \sum B_i^2}} \tag{2.8}$$

$\rho_{pb} = 0$ in case X and Y are independent. ρ_{pb} and ρ have similar values under normality, but ρ_{pb} is more robust than ρ. When their population values can differ substantially, there is little apparent difference between the bivariate distributions (Wilcox, 1994d). ρ_{pb} is not overly sensitive slight changes in distribution. ρ_{pb} is a robust measure of the linear relationship between two random variables. Therefore, ρ_{pb} is designed so that its value is not overly sensitive small part of the population.

When X and Y are independent, $r_{pb} = 0$. To test the H_0: $\rho_{pb} = 0$,

$$T_{pb}=r_{pb}\sqrt{\frac{n-2}{1-r_{pb}^2}} \tag{2.9}$$

reject H_0 if $|T_{pb}| > t_{1-\alpha}$, the 1-α quantile Student's t-distribution with $v = n - 2$ degrees of freedom. This test provides reasonably good control over the probability of Type 1 Error (Wilcox, 1994d).

H_0: $\rho_{pb} = 0$ is sensitive to heteroscedasticity like conventional T test of H_0: $\rho = 0$. If there is heteroscedasticity, the wrong standard error is being used and so, the probability of rejecting can increase H_0 with the sample size.

2.1.2 Winsorized Correlation

The population Winsorized correlation between X_1 and X_2, is given by

$$\rho_{\omega}=\frac{E_{\omega}\left[\left(X_1-\mu_{\omega 1}\right)\left(X_2-\mu_{\omega 2}\right)\right]}{\sigma_{\omega 1}\sigma_{\omega 2}} \tag{2.10}$$

where $\sigma_{\omega j}$ is the population Winsorized standard deviation of X_j, and $E_\omega(X)$ is the Winsorized expected value of X.

When X_1 and X_2 are independent, $\rho_\omega = 0$ and $-1 \leq \rho_\omega \leq 1$. In order to estimate ρ_ω, based on the random sample $(X_{11}, X_{12}), \ldots, (X_{n1}, X_{n2})$, first Winsorize observations by computing the Y_{ij} values. We estimate ρ_ω that compute the Pearson's correlation with the Y_{ij} values.

$$r_\omega = \frac{\sum(Y_{i1}-\bar{Y}_1)(Y_{i2}-\bar{Y}_2)}{\sqrt{\sum(Y_{i1}-\bar{Y}_1)^2(Y_{i2}-\bar{Y}_2)^2}} \tag{2.11}$$

To test $H_0: \rho_\omega = 0$,

$$T_\omega = r_\omega\sqrt{\frac{n-2}{1-r_\omega^2}} \tag{2.12}$$

reject H_0 if $|T_\omega| > t_{1-\alpha}$, the 1-α quantile Student's t-distribution with $v = h - 2$ degrees of freedom. $h = n - 2g, g = [\gamma n]$, is the number of observations left after trimming. When testing the hypothesis of zero correlation, the Winsorized correlation works well under independence in terms of Type 1 error probabilities. But percentage bend correlation is still better, when at the least $\beta = 0.1$ (Wilcox, 2012). Heteroscedasticity is a problem for T_ω due to being sensitive like all of the hypothesis testing methods. As for power, best method depends on values of ρ, ρ_{pb} and ρ_ω.

2.1.3 The Biweight Midcorrelation

Biweight midcorrelation is considered to be a good alternative to Pearson correlation because it is more robust to outliers. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from bivariate distribution and $K = 9$;

$$U_i = \frac{X_i - M_x}{9 * MAD_x} \tag{2.13}$$

where M_x and MAD_x are the median and the value of MAD for the X values. We calculate similarly;

$$V_i = \frac{Y - M_y}{9 * MAD_y} \tag{2.14}$$

If $-1 \leq U_i \leq 1$, $a_i = 1$ otherwise $a_i = 0$. Similarly, if $-1 \leq V_i \leq 1$ $b_i = 1$, otherwise $b_i = 0$.

The sample biweight midcovariance between X and Y is:

$$s_{bxy} = \frac{n \sum a_i (X_i - M_x)\left(1 - U_i^2\right)^2 b_i \left(Y_i - M_y\right)\left(1 - V_i^2\right)^2}{\left[\sum a_i \left(1 - U_i^2\right)\left(1 - 5U_i^2\right)\right]\left[\sum b_i \left(1 - V_i^2\right)\left(1 - 5V_i^2\right)\right]} \tag{2.15}$$

The biweight midcorrelation is also between X and Y is:

$$r_b = \frac{s_{bxy}}{\sqrt{s_{bxx} s_{byy}}} \tag{2.16}$$

2.1.4 Kendall's Tau

The other well-known type M correlation is Kendall's Tau which is between two ranked variables. Consider two pairs observations,(X_1, Y_1) and (X_2, Y_2). Assumes tied values never occur and $X_1 < X_2$. If $Y_1 < Y_2$, two pairs of observations are said to be concordant. On the other hand, if $Y_1 > Y_2$, two pairs of observations are said to be discordant. When i_{th} and j_{th} points are concordant, $K_{ij} = 1$, otherwise $K_{ij} = -1$. The formula of Kendall's Tau:

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i<j} K_{ij} \tag{2.17}$$

$\hat{\tau}$ is zero under independence.$H_0: \tau = 0$ is rejected when $|Z| \geq z_{1-\alpha/2}$, where $Z = \frac{\hat{\tau}}{\sigma^2}$ and $\sigma_\tau^2 = \frac{2(2n+5)}{9n(n-1)}$.

Although Kendall's tau provides protection against outliers, outliers can alter its value.

2.1.5 Spearman's rho

Spearman's rho is based on the resulting ranks. It provides protection against outliers among the X values, ignoring Y, as well as outliers among Y values ignoring X. But outliers can alter its value substantially (Wilcox, 2012).

$$r_s=1-\frac{6\sum d_i^2}{n(n^2-1)} \tag{2.18}$$

(d_i: is the difference in rank between paired values of X and Y).

When X and Y are independent, $\rho_s = 0$. To test H_0: $\rho_s = 0$

$$T=\frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \tag{2.19}$$

H_0 is rejected if $|T| \geq t$, where t is the $1-\alpha/2$ quantile of Student's t-distribution with $n-2$ degrees of freedom.

Spearman's rho is affected by heteroscedasticity even H_0 is true, like all of the hypothesis testing methods.

2.2 Heteroscedastic Tests of Zero Correlation

Tests of zero correlations of Type M correlations, including Pearson's correlation, are sensitive to heteroscedasticity. If there is heteroscedasticity, the probability of rejecting H_0 can increase as the sample size gets large even correlations are equal to zero.

To test the hypothesis of any type M correlation is equal to zero, percentile bootstrap method can be used since the Percentile bootstrap method performs well in terms of controlling the probability of a Type 1 error under heteroscedasticity.

Percentile bootstrap method (Wilcox, 2012);

(1) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample.

(2) Generate a bootstrap sample by resampling with replacement n pairs of points yielding $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$.

(3) Compute any of the robust estimators and label the result r^*.

(4) Repeat this B times yielding $r_1^*, \dots, r_B^*$.

(5) Let $l = \alpha B/2$ (rounded to the nearest integer) and $u = B - l$.

(6) The hypothesis of a zero correlation is rejected, if $r_{(l+1)}^* > 0$ or $r_{(u)}^* < 0$, where $r_{(1)}^* \leq \cdots \leq r_{(B)}^*$ are the values $r_1^*, \dots, r_B^*$ is written in ascending order.

CHAPTER THREE

DESIGN OF SIMULATION STUDY AND RESULTS

3.1 Design of Simulation

In this chapter of the thesis, five Type-M robust measures of correlations methods the percentage bend correlation, Winsorized correlation, biweight midcorrelation, Kendall’s Tau and Spearman’s rho are compared with a simulation study. The distributions of X and the error term ε were taken as standard normal distribution, g-and-h distribution, gamma distribution and lognormal distribution in simulation study.

g-and-h distribution was introduced by Tukey (1977). Hoaglin (1985) described the Tukey (1977) parameter estimation approach which observe how a distribution differs from normal distribution with the parameters g which is about skewness, and with a parameter h which is about kurtosis. When g and h are both equal to zero, the g-and-h distribution is equivalent to standard normal distribution (Hoaglin, 1985).

Let Z be a random variable which is generated from standard normal distribution. When $g \neq 0$, with the transformation

$$X=\frac{(exp(gZ)\text{-}1)exp\left(\frac{hZ^2}{2}\right)}{g} \tag{3.1}$$

and when $g = 0$, with the transformation

$$X=Z\ exp\left(\frac{hZ^2}{2}\right) \tag{3.2}$$

the data is generated from g-and-h distribution (Hoaglin, 1985).

Three different g-and-h distributions are used in this thesis in addition to standard normal distribution:

- Symmetric heavy tailed (SH) g-and-h distribution with g=0 and h=0.5
- Asymmetric light tailed (AL) g-and-h distribution with g=0.5 and h=0
- Asymmetric heavy tailed (AH) g-and-h distribution with g=0.5 and h=0.5

The skewness and kurtosis values for g and h values are given in the Table 3.1.

Table 3.1 The skewness and kurtosis values of g-and-h distributions

g	h	$\widehat{k}_1$ (Skewness)	$\widehat{k}_2$ (Kurtosis)
0	0	0	3
0	0.5	0	11986.2
0.5	0	1.81	9.7
0.5	0.5	120.1	18393.6

The shapes of the distributions for g and h values are given below:

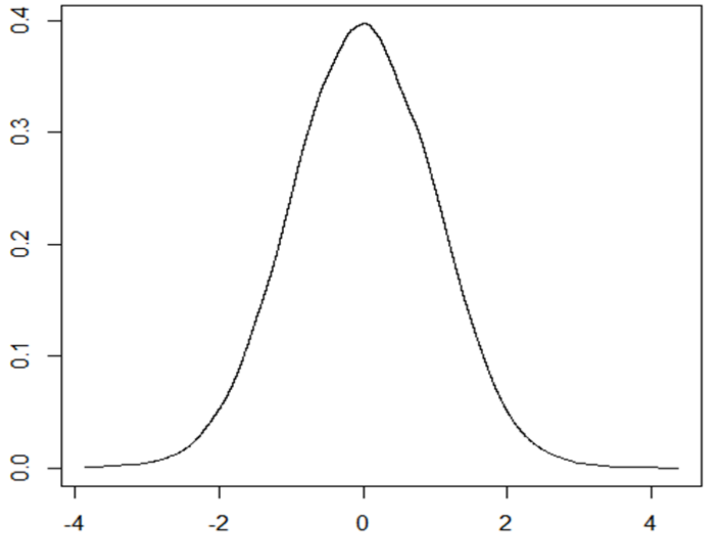

Figure 3.1 The shape of the g-and-h distribution when g=0 and h=0

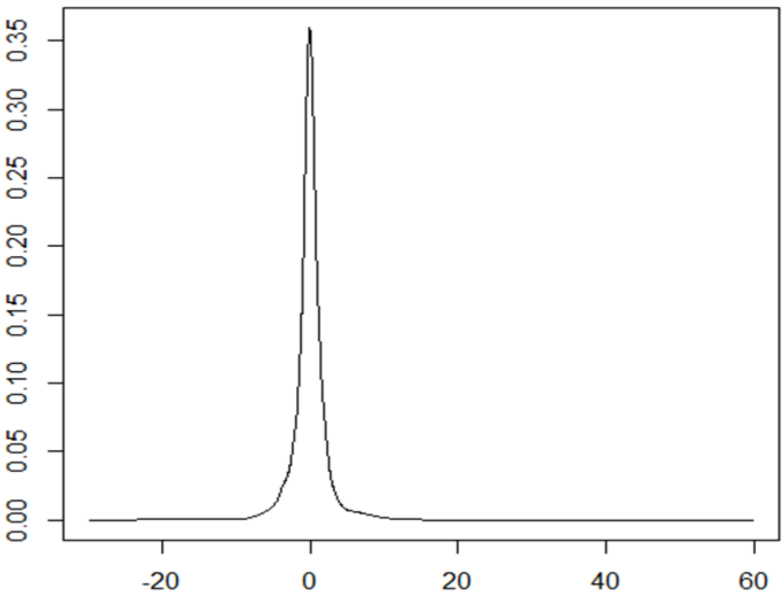

Figure 3.2 The shape of the g-and-h distribution when g=0 and h=0.5

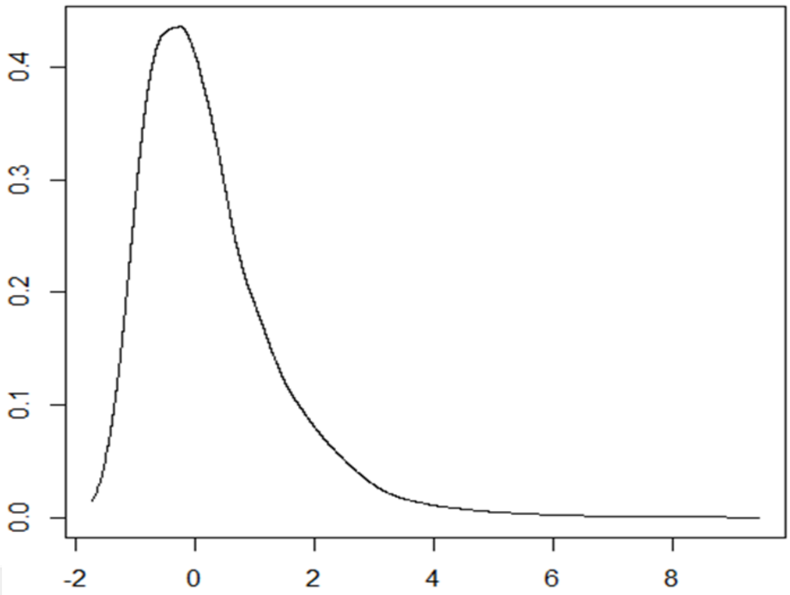

Figure 3.3 The shape of the g-and-h distribution when g=0.5 and h=0

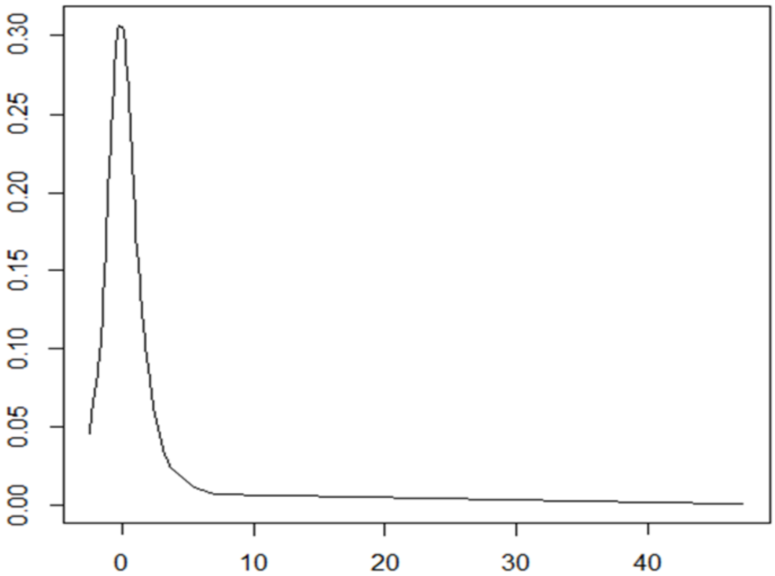

Figure 3.4 The shape of the g-and-h distribution when g=0.5 and h=0.5

We also use gamma distribution and lognormal distribution. Let X be a continuous random variable and have the gamma distribution with parameters α and $\beta > 0$. Gamma distribution's pdf is;

$$f(x;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} . x^{\alpha-1} . e^{-x/\beta} \tag{3.3}$$

where $x \geq 0, \beta > 0$ and $\alpha > 0$ (Freund, 1992). α is called the shape parameter that determines the basic shape of the graph. β is a scale parameter and influenced by the spread of the distribution. In this simulation study, we take the value of α as 1. Furthermore we take the value of β as 1 and 2.

Lognormal distribution is used when the distribution is skewed. The probability density function of the lognormal distribution with parameters μ and σ is given by

$$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} . exp\left(-\frac{(ln(x) - \mu)^2}{2\sigma^2}\right), x>0 \tag{3.4}$$

Simirlarly, μ is taken as 0 and 1; σ is taken as 1 for lognormal distributions in this simulation study.

The shapes of gamma and lognormal distributions are given below:

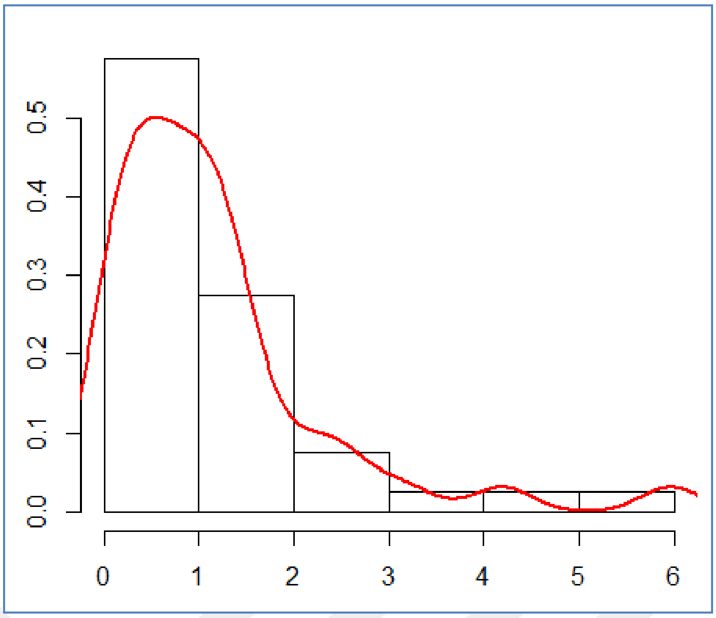

Figure 3.5 Gamma Distribution for α=1, β=1

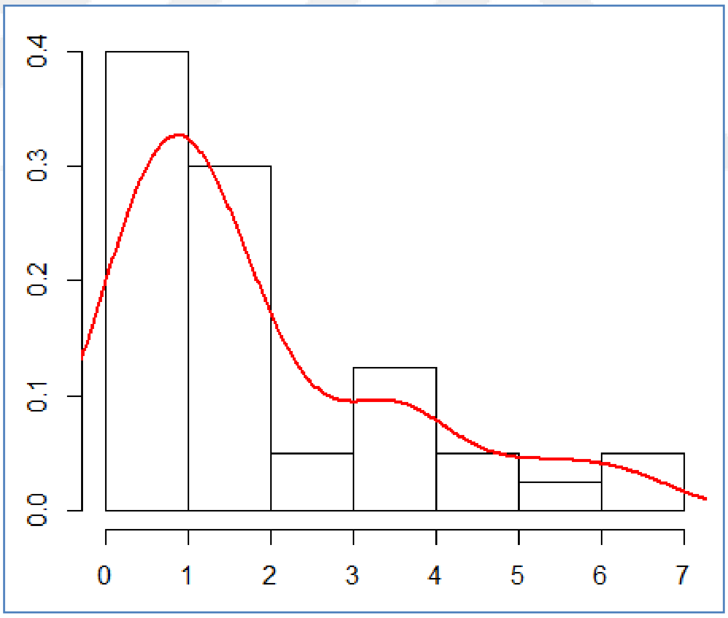

Figure 3.6 Gamma Distribution for α=1, β=2

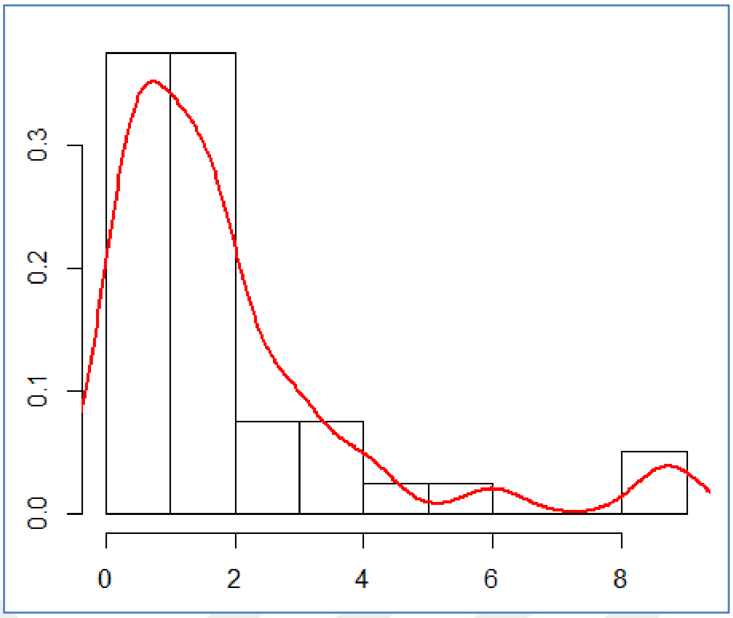

Figure 3.7 Lognormal Distribution for μ=0, σ=1

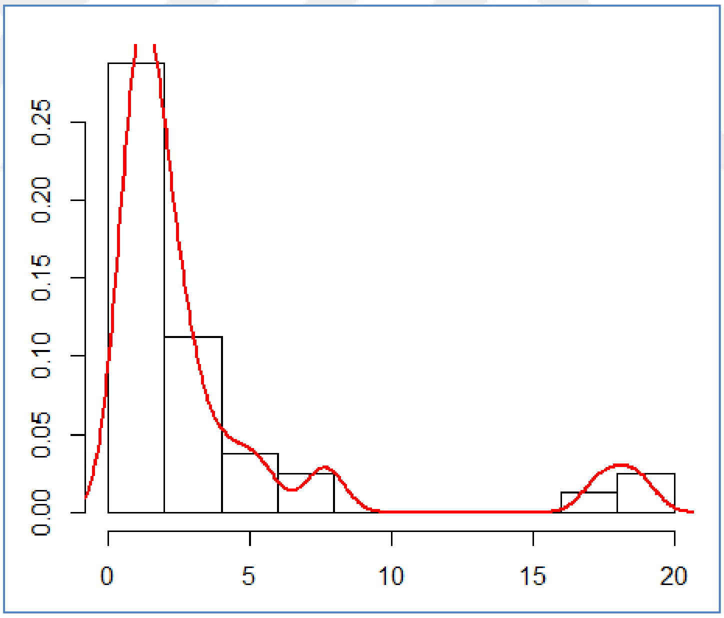

Figure 3.8 Lognormal Distribution for μ=1, σ=1

The Type-M measures of robust correlations methods are compared according to actual Type 1 error probabilities (actual significance level). At first, to compare Type-M measure of robust correlations methods in terms of actual significance level($\hat{\alpha}$), the observations are generated with the model (Wilcox, 2012);

$$y = \lambda(x)\varepsilon \quad (3.5)$$

Three variance patterns (VP) are used. VP1 corresponds to $\lambda(x) = 1$ (homoscedastic case), VP2 corresponds to $\lambda(x) = |x| + 1$ (Var(Y/X) smallest when x≅E(X)) and VP3 corresponds to $\lambda(x) = \frac{1}{|x|+1}$ (Var(Y/X) largest when x≅E(X)).

The sample size of n=40 are taken from each distribution. Nominal type I error (significance level) is taken as $\alpha = 0.05$. R programming language (3.4.0) is used. Simulations are based on 10000 replications. Included Type-M robust measures of correlations, the percentage bend correlation, Winsorized correlation, biweight midcorrelation, Kendall's Tau and Spearman's rho are denoted as PB, WI, BW, TAU and SP.

3.2 Actual Significance Levels Results

The results are given in Table 3.2, Table 3.3, Table 3.4, Table 3.5, Table 3.6, Table 3.7, Table 3.8 and Table 3.9 in terms of actual significance levels. Although the importance of a Type I error depends on the situation, Bradley (1980) has suggested that ideally, the actual level should be between 0.045 and 0.055 when testing at the 0.05 level. And at a minimum the actual level should be between 0.025 and 0.075. Here we focused on a comprimise among these two suggestions and used the interval (0.040, 0.060) as a criterion of robustness. Tables and figures show that how the actual significance levels of Type-M robust correlations methods spread around the nominal 0.05 level.

Correlation methods were put in order in terms of closeness of their actual significance level to the nominal significance level under the preference column in each table.

In Table 3.2, where X has normal distributions, actual significance levels fall in the interval in almost all cases. PB, BW, TAU and SP are preferred with VP1 and VP2. PB and BW perform better with VP3.

Table 3.2 Actual significance levels, X has normal distribution

x	**ε**	**VP**	**PB**	**BW**	**WI**	**TAU**	**SP**	**Preference**
N	**N**	**1**	0.0459	0.0496	0.0319	0.0429	0.0426	**BW, PB, TAU, SP**
		2	0.0470	0.0498	0.0333	0.0498	0.0484	**BW, TAU, SP, PB**
		3	0.0413	0.0434	0.0265	0.0348	0.0379	**BW, PB**
N	**SH**	**1**	0.0423	0.0433	0.0296	0.0429	0.0426	**BW, TAU, SP, PB**
		2	0.0433	0.0422	0.0308	0.0495	0.0472	**TAU, SP, PB, BW**
		3	0.0427	0.0383	0.0235	0.0371	0.0394	**PB**
N	**AL**	**1**	0.0452	0.0440	0.0307	0.0429	0.0426	**PB, BW, TAU, SP**
		2	0.0467	0.0459	0.0328	0.0498	0.0502	**TAU, SP, PB, BW**
		3	0.0403	0.0390	0.0263	0.0346	0.0360	**PB**
N	**AH**	**1**	0.0434	0.0444	0.0276	0.0429	0.0426	**BW, PB, TAU, SP**
		2	0.0445	0.0431	0.0288	0.0482	0.0480	**TAU, SP, PB, BW**
		3	0.0410	0.0399	0.0227	0.0376	0.0388	**PB**

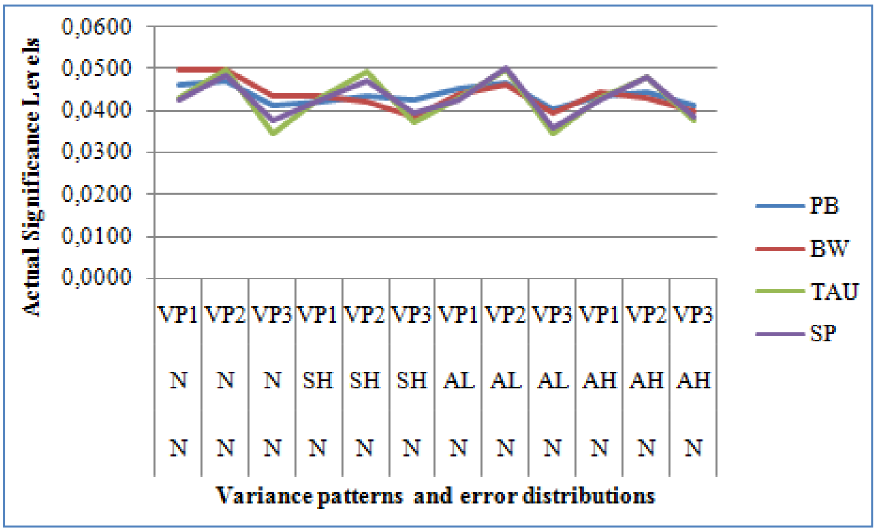

Figure 3.9 Actual significance levels of the four best Type-M robust correlation methods when X has normal distribution

In Table 3.3, where X has symmetric heavy tailed distributions, actual significance levels fall in the (0.04 - 0.06) interval for three correlations. PB, TAU and SP perform better with VP1and VP2.

Table 3.3 Actual significance levels, X has symmetric heavy tailed distribution

x	ε	VP	PB	BW	WI	TAU	SP	Preference
SH	**N**	**1**	0.0442	0.0421	0.0286	0.0429	0.0426	**PB, TAU, SP, BW**
		2	0.0454	0.0386	0.0302	0.0515	0.0519	**TAU, SP, PB**
		3	0.0341	0.0355	0.0156	0.0275	0.0328	
SH	**SH**	**1**	0.0413	0.0358	0.0253	0.0429	0.0426	**TAU, SP, PB**
		2	0.0434	0.0326	0.0280	0.0506	0.0499	**SP, TAU, PB**
		3	0.0376	0.0333	0.0154	0.0341	0.0371	
SH	**AL**	**1**	0.0439	0.0365	0.0277	0.0429	0.0426	**PB, TAU, SP**
		2	0.0457	0.0365	0.0300	0.0522	0.0534	**TAU, SP, PB**
		3	0.0335	0.0327	0.0162	0.0270	0.0326	
SH	**AH**	**1**	0.0412	0.0358	0.0250	0.0429	0.0426	**TAU, SP, PB**
		2	0.0445	0.0325	0.0265	0.0509	0.0511	**TAU, SP, PB**
		3	0.0374	0.0312	0.0154	0.0343	0.0367	

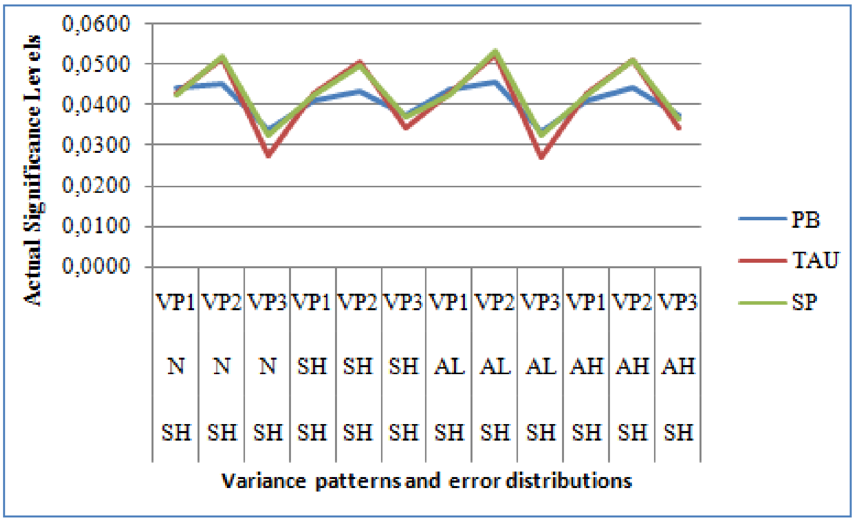

Figure 3.10 Actual significance levels of the three best Type-M robust correlation methods when X has symmetric heavy tailed distribution

In Table 3.4, where X has asymmetric light tailed distribution, actual significance levels fall in the interval in four correlations. PB,TAU and SP are more preferable than BW. Because when we look at the overall situation, BW only is marked 3 times and not marked in other cases. PB, TAU and SP perform better with VP1 and VP2.

Table 3.4 Actual significance levels, X has asymmetric light tailed distribution

x	ε	VP	PB	BW	WI	TAU	SP	Preference
AL	**N**	**1**	0.0456	0.0465	0.0305	0.0429	0.0429	**BW, PB, TAU, SP**
		2	0.0472	0.0441	0.0332	0.0484	0.0484	**TAU, SP, PB, BW**
		3	0.0391	0.0387	0.0262	0.0332	0.0369	
AL	**SH**	**1**	0.0454	0.0396	0.0294	0.0484	0.0489	**SP, TAU, PB**
		2	0.0446	0.0376	0.0298	0.0488	0.0476	**TAU, SP, PB**
		3	0.0395	0.0333	0.0233	0.0366	0.0393	
AL	**AL**	**1**	0.0446	0.0384	0.0298	0.0429	0.0426	**PB, TAU, SP**
		2	0.0465	0.0422	0.0340	0.0480	0.0508	**SP, TAU, PB, BW**
		3	0.0389	0.0336	0.0257	0.0333	0.0380	
AL	**AH**	**1**	0.0424	0.0393	0.0393	0.0267	0.0426	**SP, PB**
		2	0.0437	0.0379	0.0379	0.0299	0.0470	**SP, PB**
		3	0.0392	0.0392	0.0331	0.0235	0.0394	

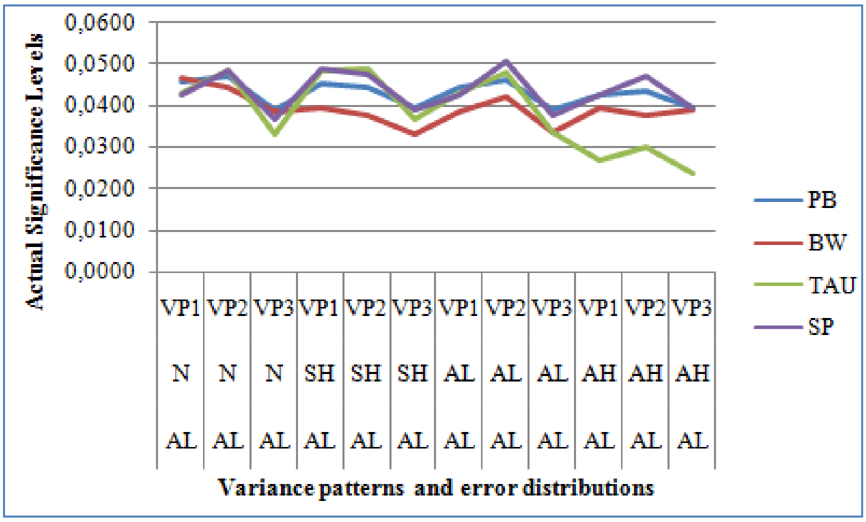

Figure 3.11 Actual significance levels of the three best Type-M robust correlation methods when X has asymmetric light tailed distribution

In Table 3.5 where X has asymmetric heavy tailed distributions, actual significance levels fall in the interval in three correlations. PB, TAU and SP are more preferable than other methods. Because when we look at the overall situation, none of the values from BW and WI are not marked. PB, TAU and SP perform better with VP1 and VP2.

Table 3.5 Actual significance levels, X has asymmetric heavy tailed distribution

x	ε	VP	PB	BW	WI	TAU	SP	Preference
AH	**N**	**1**	0.0445	0.0394	0.0277	0.0429	0.0426	**PB, TAU, SP**
		2	0.0448	0.0362	0.0310	0.0520	0.0510	**SP, TAU, PB**
		3	0.0339	0.0354	0.0179	0.0294	0.0336	
AH	**SH**	**1**	0.0413	0.0328	0.0246	0.0429	0.0426	**TAU, SP, PB**
		2	0.0443	0.0315	0.0281	0.0506	0.0498	**SP, TAU, PB**
		3	0.0357	0.0310	0.0167	0.0336	0.0369	
AH	**AL**	**1**	0.0332	0.0138	0.0120	0.0429	0.0426	**TAU, SP**
		2	0.0446	0.0345	0.0293	0.0516	0.0519	**TAU, SP, PB**
		3	0.0318	0.0305	0.0177	0.0279	0.0346	
AH	**AH**	**1**	0.0412	0.0335	0.0231	0.0429	0.0426	**TAU, SP, PB**
		2	0.0440	0.0296	0.0270	0.0495	0.0512	**TAU, SP, PB**
		3	0.0348	0.0291	0.0163	0.0329	0.0369	

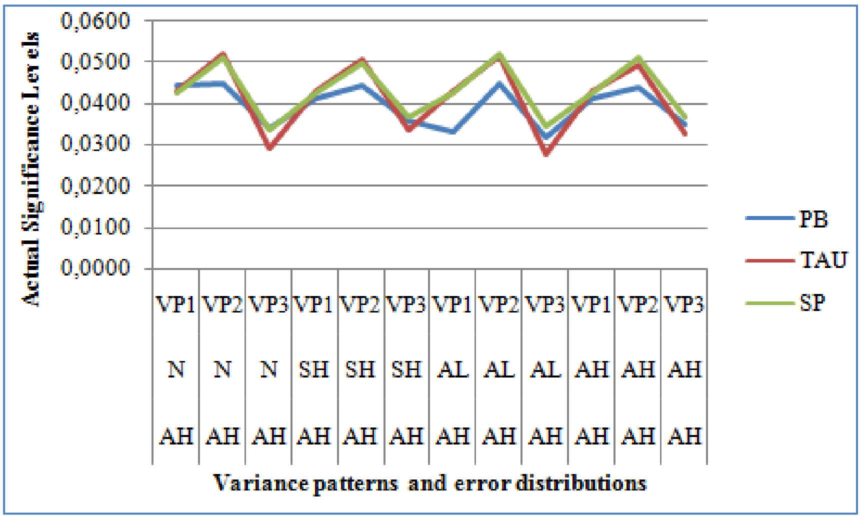

Figure 3.12 Actual significance levels of the three best Type-M robust correlation methods when X has asymmetric heavy tailed distribution

In Table 3.6 and Table 3.7, where X has gamma distribution and **ε** has g&h distribution, actual significance levels fall in he interval in three correlations. PB, TAU and SP are more preferable than BW. Because when we look at the overall situation, BW only is marked 2 times in Table 3.6 and 1 time in Table 3.7 and not marked in other cases for both tables. PB, TAU and SP perform better with all variance patterns.

Table 3.6 Actual significance levels, X has gamma distribution and **ε** has g&h distribution

x	ε	VP	PB	BW	WI	TAU	SP	Preference
G (40,1,1)	**N**	**1**	0.0416	0.0343	0.0304	0.0434	0.0430	**TAU, SP, PB**
		2	0.0452	0.0330	0.0356	0.0469	0.0483	**SP, TAU, PB**
		3	0.0441	0.0401	0.0329	0.0448	0.0456	**SP, TAU, PB, BW**
G (40,1,1)	**SH**	**1**	0.0400	0.0302	0.0264	0.0434	0.0430	**TAU, SP, PB**
		2	0.0430	0.0284	0.0310	0.0453	0.0473	**SP, TAU, PB**
		3	0.0419	0.0343	0.0291	0.0432	0.0448	**SP, TAU, PB**
G (40,1,1)	**AL**	**1**	0.0405	0.0304	0.0298	0.0434	0.0430	**TAU, SP, PB**
		2	0.0459	0.0371	0.0351	0.0503	0.0512	**TAU, SP, PB**
		3	0.0468	0.0484	0.0324	0.0498	0.0497	**TAU, SP, BW, PB**

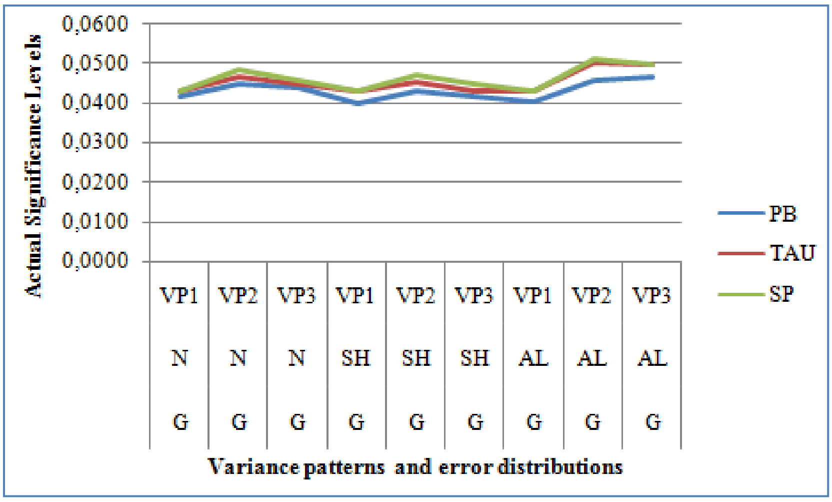

Figure 3.13 Actual significance levels of the three best Type-M robust correlation methods when X has gamma distribution and εhas g&h distribution

Table 3.7 Actual significance levels, X has gamma distribution and εhas g&h distribution

x	ε	VP	PB	BW	WI	TAU	SP	Preference
G (40,1,2)	**N**	**1**	0.0416	0.0334	0.0304	0.0434	0.0430	**TAU, SP, PB**
		2	0.0456	0.0331	0.0340	0.0463	0.0478	**SP, TAU, PB**
		3	0.0430	0.0389	0.0310	0.0424	0.0443	**SP, PB, TAU**
G (40,1,2)	**SH**	**1**	0.0400	0.0302	0.0264	0.0434	0.0430	**TAU, SP, PB**
		2	0.0413	0.0291	0.0321	0.0469	0.0468	**TAU, SP, PB**
		3	0.0457	0.0337	0.0304	0.0472	0.0490	**SP, TAU, PB**
G (40,1,2)	**AL**	**1**	0.0453	0.0355	0.0325	0.0466	0.0480	**SP, TAU, PB**
		2	0.0454	0.0344	0.0346	0.0510	0.0520	**TAU, SP, PB**
		3	0.0456	0.0431	0.0342	0.0516	0.0528	**TAU, SP, PB, BW**

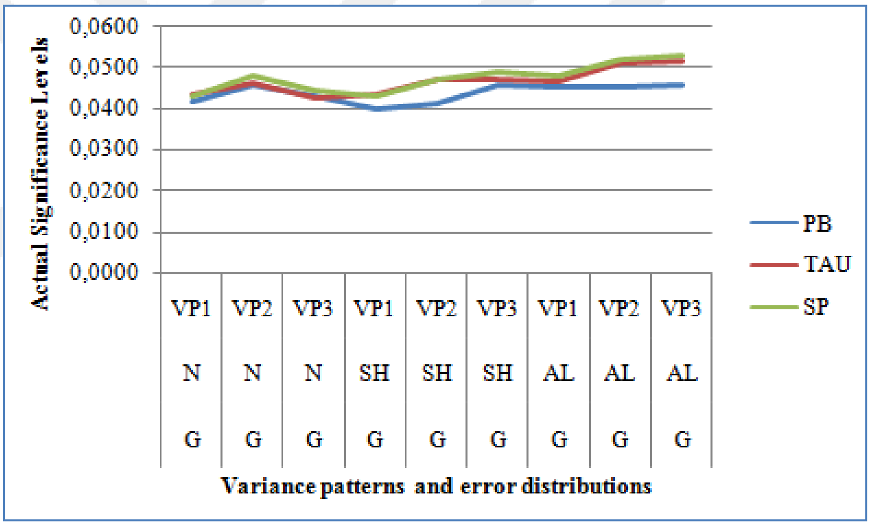

Figure 3.14 Actual significance levels of the three best Type-M robust correlation methods when x has gamma distribution and εhas g&h distribution

Similarly, in Table 3.8 and Table 3.9, where X has lognormal distribution and **ε** has g&h distribution, actual significance levels fall in the interval in three correlations. PB, TAU and SP are more preferable than BW. Because when we look at the overall situation, BW only is marked 2 times in Table 3.8 and 1 time in Table 3.9 and not marked in other cases for both tables. PB, TAU and SP perform better with all variance patterns.

Table 3.8 Actual significance levels, X has lognormal distribution and **ε** has g&h distribution

x	ε	VP	PB	BW	WI	TAU	SP	Preference
LogN (40,0,1)	**N**	**1**	0.0431	0.0325	0.0301	0.0429	0.0426	**PB, TAU, SP**
		2	0.0444	0.0335	0.0365	0.0505	0.0513	**TAU, SP, PB**
		3	0.0443	0.0365	0.0323	0.0406	0.0459	**SP, PB, TAU**
LogN (40,0,1)	**SH**	**1**	0.0399	0.0281	0.0266	0.0429	0.0426	**TAU, SP**
		2	0.0420	0.0257	0.0337	0.0477	0.0478	**SP, TAU, PB**
		3	0.0417	0.0335	0.0302	0.0462	0.0479	**SP, TAU, PB**
LogN (40,0,1)	**AL**	**1**	0.0430	0.0320	0.0341	0.0484	0.0486	**SP, TAU, PB**
		2	0.0482	0.0403	0.0380	0.0579	0.0580	**PB, TAU, SP, BW**
		3	0.0428	0.0487	0.0309	0.0501	0.0505	**TAU, SP, BW, PB**

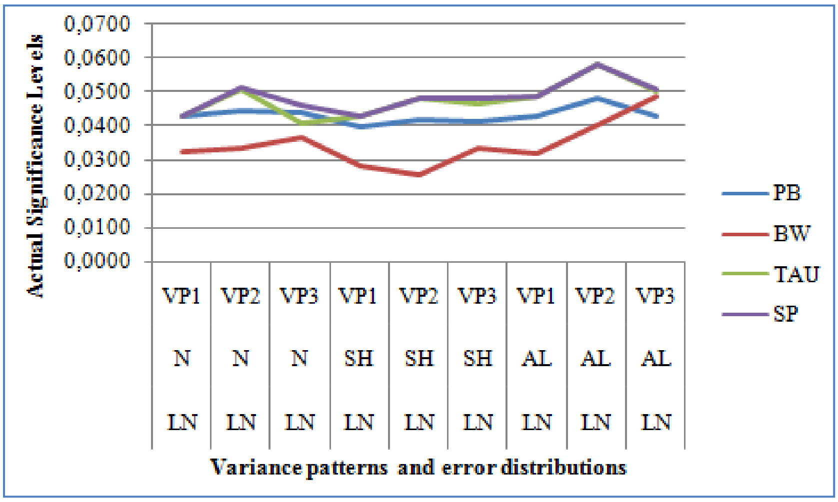

Figure 3.15 Actual significance levels of the three best Type-M robust correlation methods when X has lognormal distribution and **ε** has g&h distribution

Table 3.9 Actual significance levels, X has lognormal distribution and **ε** has g&h distribution

x	ε	VP	PB	BW	WI	TAU	SP	Preference
LogN (40,1,1)	**g&h (40,0,0)**	**1**	0.0431	0.0325	0.0301	0.0429	0.0426	**PB, TAU, SP**
		2	0.0446	0.0273	0.0378	0.0506	0.0522	**TAU, SP, PB**
		3	0.0468	0.0348	0.0362	0.0465	0.0501	**SP, PB, TAU**
LogN (40,1,1)	**g&h (40,0,0.5)**	**1**	0.0399	0.0281	0.0266	0.0429	0.0426	**TAU, SP**
		2	0.0436	0.0251	0.0345	0.0487	0.0505	**SP, TAU, PB**
		3	0.0412	0.0322	0.0315	0.0436	0.0464	**SP, TAU, PB**
LogN (40,1,1)	**g&h (40,0.5,0)**	**1**	0.0424	0.0272	0.0291	0.0429	0.0426	**TAU, SP, PB**
		2	0.0458	0.0395	0.0386	0.0548	0.0537	**SP, PB, TAU**
		3	0.0476	0.0585	0.0362	0.0519	0.0541	**TAU, PB, SP, BW**

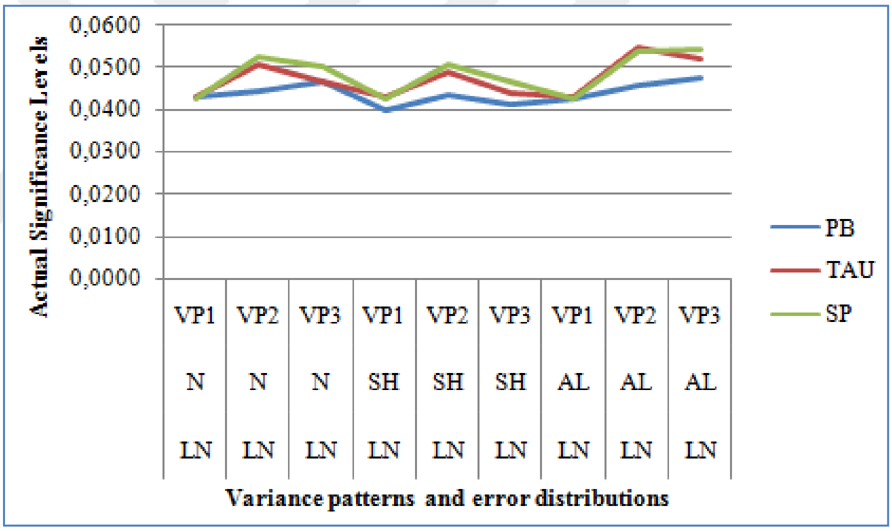

Figure 3.16 Actual significance levels of the three best Type-M robust correlation methods when X has lognormal distribution and **ε** has g&h distribution

CHAPTER FOUR
CONCLUSION

Pearson's correlation is the most suitable option when there is a Normally distributed population distribution and homoscedasticity. When assumptions are violated, Pearson's correlation loses its accuracy. Because Pearson's correlation is sensitive to heteroscedasticity. And outliers, the magnitude of slope, curvature, restriction of range and the magnitude of residuals are also other problems.

To minimize the effects of these problems, some robust methods are suggested. One of these methods is called as Type-M robust measures of correlation.

In this thesis, 5 Type-M robust correlation methods are examined and compared using different distribution and variance patterns. The conventional hypothesis testing method used withType-M correlations is also sensitive to heteroscedasticity. Therefore, we used a heteroscedastic test of zero population correlation that is based on a Percentile Bootstrap approach.

84 experimental cases are considered in this thesis. Number of actual significance levels that fall in the selected 0.04 – 0.06 interval was 69 for percentage bend correlation. Corresponding numbers were 68 for Spearman correlation, 66 for Kendal's Tau, 19 for biweight midcorrelation and 0 for winsorized correlation.

Under 86 cases out of 140, there was no actual significance level for any correlation methods within the 0.04 – 0.06 interval when VP3 is used while generating data. This number is about 50 for the other two variance patterns. This means that the VP3 generates observations in a way that the included correlation methods have some difficulties in controlling the nominal significance level.

In the first simulation study, X has a normal distribution. When X has a normal distribution, PB, BW, TAU and SP are preferred. The values of PB, BW, TAU and SP methods are within 0.04-0.06 with VP1 and VP2.

When X has a symmetric heavy tailed distribution, PB, TAU and SP methods are preferred with VP1 and VP2.

When X has an asymetric light tailed distribution, preferred methods can change according to the distribution of ε. When X has an AL distribution and ε has a normal distribution, PB, BW, TAU and SP methods are preferred with VP1 and VP2. When X has an AL distribution and ε has a SH or AL distribution, PB, TAU and SP methods are preferred with VP1 and VP2. When X has an AL distribution and ε has an AH distribution, PB and SP methods are preferred with VP1 and VP2.

When X has asymmetric heavy tailed distribution, PB, TAU and SP methods are generally preferred with VP1 and VP2.

When X has a gamma distrubution, PB, TAU and SP methods are preferred with VP1, VP2 and VP3.

When X has a lognormal distribution, PB, TAU and SP methods are preferred with VP1, VP2 and VP3.

According to the simulation results obtained using different distributions, generally PB, TAU and SP methods are preferred with VP1 and VP2.

In general, the use of percentile bootstrap method with robust type-M correlations for testing zero population correlation hypothesis has given conservative results. Almost all of the actual significance levels were less than nominal significance level 0.05.

The percantage bend, Kendall’s tau and Spearman’s rho correlations saved nominal significance level much better than Winsorized and Biweight correlations. The use of Winsorized correlation with this percentile bootstrap method might create drastic problems in terms of saving nominal significance level. When testing the hypothesis of zero correlation among two random variables, the use of heteroscedastic test procedure which is based on a percentile bootstrap approach is recommended with robust correlation coefficients such as percentage bend, Spearman’s rho and Kendall’s tau.

REFERENCES

Blair, R. C., & Lawson, S. B. (1982). Another look at the robustness of the product-moment correlation coefficient to population non-normality. *Florida Journal of Educational Research, 24,* 11-15.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Chee J. D. (2015). Pearson's Product-Moment Correlation: Sample Analysis. Retrieved April 20, 2017, from https://www.researchgate.net/publication/277324930_Pearson's_Product-Moment_Correlation_Sample_Analysis

Cramer, D. (1998). *Fundamental Statistics for Social Research*. London: Routledge.

Edgell, S. E., & Noon, S. M. (1984). Effect of violation of normality on the t test of the correlation coefficient. *Psychological Bulletin, 95,* 576-583.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), Oct 2008, 591-601.

Field, A. P. (2009). *Discovering statistics using SPSS*. London, England: SAGE.

Freund, J. E. (1992). *Mathematical Statistics* (5th ed.) Prentice-Hall International, Inc.

Georgy Shevlyakov and Pavel Smirnov, Robust Estimation of the Correlation Coefficient: *An Attempt of Survey Austiian Journal of Statistics, 40*(1&2), 2011, 147–156.

Higgins J. (2005). Introduction to Multiple Regression. Retrieved May 16, 2017, from http://www.biddle.com/documents/bcg_comp_chapter4.pdf

Huber, P. (1981). *Robust Statistics*. Wiley.

Muirhead R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley-Interscience.

Tukey J. W. (1977). Modern techniques in data analysis. NSF-sponsored regional research conference; North Dartmouth, MA: Southeastern Massachusetts University

Wilcox, R. R. (1991a). Bootstrap inferences about the correlation and variance of paired data. *British Journal of Mathematical and Statistical Psychology*, *44*, 379-382.

Wilcox, R.R. (1994d). The Percentage Bend Correlation Coefficient. *Psychometrika*, *59*(4), 601-616.

Wilcox, R. R., & Muska, J. (2001). Inferences about correlations when there is heterocedasticity, *British Journal of Mathematical and Statistical Psychology, 54*, 39-47.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Amsterdam: Elsevier/ Academic Press