Distribution-Free Confidence Intervals for Functions of Quantiles

Submitted by

Chandima N.P.G. Arachchige

Bachelor of Science (Honours)

A thesis submitted in total fulfilment
of the requirements for the degree of
Doctor of Philosophy

School of Engineering and Mathematical Sciences

College of Science, Health and Engineering

La Trobe University

Victoria

Australia

December 16, 2019

Contents

List of Figures

List of Tables

List of Publications

The part II of this thesis consists with the following four accepted or submitted papers. The author contributed to developing theories, simulations, examples and finally manuscript writing under the valuable guidance received from two supervisors throughout the whole process.

I Arachchige, Chandima NPG, Cairns, Maxwell, & Prendergast, Luke A. 2019a. *Interval estimators for ratios of independent quantiles and interquantile ranges.* Communications in statistics-simulation and computation, 1-17.

II Arachchige, Chandima NPG, & Prendergast, Luke A. 2019. *Confidence intervals for median absolute deviations.* arxiv preprint arxiv:1910.00229.

III Arachchige, Chandima NPG, Prendergast, Luke A, & Stuadte, Robert G. 2019b. *Robust analogues to the coefficient of variation.* arxiv preprint arxiv:1907.01110.

IV Arachchige, Chandima NPG & Prendergast, Luke A 2019b. *Mean skewness measures* arXiv:submit/2971320

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my principal supervisor, Prof. Luke A. Prendergast for giving me the opportunity to do a PhD and providing excellent supervision throughout my candidature. His invaluable guidance helped me in all aspects, research, preparing manuscripts and finally writing this thesis. He was an excellent advisor and mentor during my PhD and I am extremely grateful for his friendship, patience and a great sense of humour. I'm very grateful to my co-supervisor Prof. Robert G. Staudte for his valuable guidance, advice and enormous support throughout my PhD in relation to research, preparing manuscripts and writing this thesis. I could not have imagined better supervisors.

Many thanks to all the academic and non-academic staff of the Department of Mathematics and Statistics at La Trobe University for their invaluable support and friendliness. I would also like to thank La Trobe University for significant financial support in the form of a Postgraduate Research Scholarship. Special thanks goes to La Trobe security staff for providing me with a secure transport service during the last period of my PhD and their friendliness. Thank you so much for all the Sri Lankan friends and other colleagues who did PhD at La Trobe for their kindness and support which made me happy and feel like La Trobe is my home. Thanks so much for the "Broadway" and "Kingsbury" housemates for their friendliness and numerous help during my PhD candidature. My sincere thanks also goes to all the academic and non-academic staff of the Department of Statistics, the University of Colombo for their support and valuable encouragement before and during my PhD.

Many thanks to my family, my mother, father, brother, sister, brother-in-law and their two lovely children for believing that I would do my best throughout my PhD studies and which made me feel like that I'm with them, even though I'm thousands of miles away. Finally, for everyone who helped me, even with just from a single encouraging word and for making me more independent and stronger during my PhD candidature and throughout my whole life.

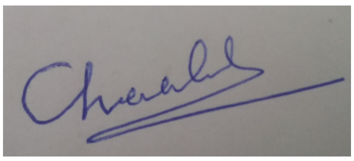
Statement of Authorship

Except where reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis submitted for the award of any other degree or diploma.

No other person's work has been used without due acknowledgement in the main text of the thesis.

The thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

Chandima N.P.G. Arachchige



16th December 2019

Abbreviations

EDF Empirical Distribution Function

ECDF Empirical Cumulative Distribution Function

EQF Empirical Quantile Function

QOR Quantile Optimality Ratio

GLD Generalized Lambda Distribution

IC Influence Curve

IF Influence Function

TIF Theoretical Influence Function

SIF Sample Influence Function

EIF Empirical Influence Function

PIF Partial Influence Function

ASV Asymptotic Variance

ASE Asymptotic Standard Error

ASD Asymptotic Standard Deviation

SD Standard Deviation

IQR Inter Quantile Range

MAD Median Absolute Deviation

PB Price and Bonett

LN Log Normal

EXP Exponential

PAR Pareto

cp coverage probability

CI Confidence Interval

w width of the confidence interval

Est. Estimator

CV Coefficient of Variation

iid independent and identically distributed

Summary

The objective of this thesis is to provide new distribution free point and interval estimators for measures of spread, relative spread and skewness involving quantiles. The main advantages of these new quantile-based measures are that they are comparatively efficient to compute with minimal distributional assumptions. In the two-samples case, the location and scale of two independent samples can be compared using ratios of linear combinations of quantiles. In the single-sample case, distribution-free inferences can be made using quantile versions of the coefficient of variation and measures of skewness. The new estimators can be directly applied to many areas, such as economics, medical statistics, bio-statistics and social sciences. This thesis consists of four papers, either published or submitted at the time of thesis submission, and also includes some introductory material.

In paper I, the distribution-free point and interval estimators were introduced for ratios of independent quantiles to compare the location of two independent samples. The interquantile range, which is the most natural quantile-based estimator of scale, was considered to compare the scale of two independent samples. The distribution-free point and interval estimators were introduced to the squared ratios of interquantile ranges. The best choice of the probability to achieve the minimum asymptotic variance for the squared ratio of interquantile ranges was proposed. Robustness properties of the estimators were investigated using partial influence functions. The simulation results reveal that all the intervals provide excellent coverage probabilities even for small sample sizes and for a wide range of distributions. An R shiny web application was developed and is publicly available to readers to run the simulations as they desire. Some real-world data examples were considered, and the results suggest that new estimators perform really well compared to the classical parametric tests such as t-test and F-test.

In paper II, the median absolute deviation which is the most robust estimator of scale with respect to the breakdown point was considered. The distribution-free point and interval estimators were introduced to the median absolute deviation and the difference and squared ratio of median absolute deviations to make inferences on spread of a single sample and

to compare the spread of two populations respectively. Robustness properties of the new estimators were investigated using an influence function and partial influence functions. Simulations were conducted to check the performance of the new estimators and the results suggest that the coverage probabilities are very close to the nominal coverage even with small sample sizes and for a variety of distributions. A real-world data example was considered, and the results suggest that the new estimators are more robust to the outliers when compared to the F-test.

In paper III, two robust versions of the coefficient of variation based on linear combinations of quantiles were considered to make distribution-free inferences of relative dispersion for a single sample. The first measure was the interquartile range divided by the median and the second measure was the median absolute deviation divided by the median. The distribution-free point and interval estimators were constructed to the two robust versions of the coefficient of variations and the robustness properties were investigated using influence functions. The performance of the new estimators was compared with several existing estimators via simulations and the results suggest that the new methods perform well for a wide variety of distributions. The R shiny web application was developed and is publicly available to compare the performance of the new estimators of spread with some existing estimators. The interval estimators were introduced to the ratios of the robust coefficient of variations as an extension to compare the relative dispersion between two independent samples. The examples reveal that different conclusions can be made based on robust and non-robust versions of the coefficient of variation.

In paper IV, some integrated versions were constructed to existing measures of skewness based on ratios of linear combinations of quantiles. These existing skewness measures are some generalizations of Bowley's well-known skewness coefficient. The validity of the properties that any measures of skewness should satisfy was tested for new measures of skewness. The distribution-free point and interval estimators were introduced for new measures of skewness to make inferences about the skewness of a single sample. A simulation study was conducted to compare the performance of the new estimators with the existing estimators and the results suggest that the new measures perform well for wide range of distributions. The R shiny web application was developed and is publicly available to compare the performance of new skewness estimators. The interval estimators were introduced to the difference of the measures of skewness to compare the skewness between two independent samples. Some real-world examples were used and different conclusions were observed based on different methods.

Part I

Introduction

1. Background

There is a vast body of literature on interval estimators for population quantiles and ratios of linear combinations of population quantiles. However, it is relatively rare to find a distribution-free inference regarding linear combinations of quantiles. The two-sample case includes a comparison of location and scale of the linear combination of quantiles between two independent samples. The one-sample case includes making distribution-free inferences for quantile versions of the coefficient of variation and the measures of skewness. One of the main concerns is to find the best choice of quantile combinations to obtain efficient robust tests and confidence intervals. These results can be directly applied in many areas such as economics, medical statistics, bio-statistics and social sciences. The objective of this thesis is to provide new point and interval estimators for functions of linear combinations of quantiles. The main advantage of these new quantile-based measures is that the intervals are shown to have good coverage properties with minimal distributional assumptions, while being comparatively efficient to compute. Additionally, the robustness of quantiles to extreme outliers that often arise from highly skewed population distributions is another advantage. The results are supported by theory, extensive simulation studies and applications to real-world data sets. New R shiny web applications were developed to run the simulations and are publicly available.

This thesis is divided into two sections: Part I and Part II. Part I comprises four chapters and Part II comprises four publications. This chapter consists of a brief discussions of previous research work related to the location and scale comparisons of two independent populations, the coefficient of variation and measures of skewness for a single population. The second chapter summarises the theory of quantile estimators and the influence function which was used to build the new methods that are introduced later in all four publications in Part II. The third chapter consists of some additional works which were not included in the four publications but are still worth discussing. Finally, the fourth chapter contains the main results of the thesis, conclusions and the further works. The most important R programs used can be found in the Appendix.

1.1 Comparing location and scale of two independent populations

1.1.1 Comparing locations of two independent populations

The t-test is the most commonly used test to compare the differences of means between two independent populations under the assumption of normality. However, when the distributions are skewed or highly non-normal, the median may be a more appropriate measure of location. If two independent samples are coming from identical shape distributions, the Mann-Whitney-Wilcoxon (Hollander & Wolfe, 1999) method can be used to construct confidence intervals for differences and ratios of medians, but is limited in use since it is not robust to violations of the identical shape assumption (Pratt, 1964). The Mood (1950) and the Hettmansperger (1984) methods also can be used to construct the confidence interval for a difference of medians under the identical shape assumptions. Howevere, the Mood (1950) and the Hettmansperger (1984) methods were more robust to a violation of the identical shape assumption compared to the Mann-Whitney-Wilcoxon method, although the coverage probabilities of the 95% confidence interval for differences of medians be lower than the nominal coverage (Pratt, 1964). Hettmansperger (1984) introduced a two-sample test and a confidence interval for differences in population medians based on the one sample confidence intervals which he calls "sign-intervals". This confidence interval is constructed by subtracting the endpoints of the two sign-intervals. McKean & Schrader (1984) showed through a simulation study that this method is highly unstable for small to moderate sample sizes, even for normal distributions. However, Hettmansperger (1984) stated that this method has well-defined asymptotic distribution theory and the same asymptotic efficiency as the Mood (1950) test.

Wang & Hettmansperger (1990) derived two sample tests and confidence intervals for differences of median survival times for the censored survival data following the method described in Hettmansperger (1984) which is for uncensored data. In the presence of censored data, Wang & Hettmansperger (1990) modified the sign-interval to a new interval called "quantile-interval" whose endpoints constitute the quantiles of the Kaplan & Meier (1958)'s product-limit estimator. Wei & Gail (1983) derived a non-parametric confidence interval to compare the ratio of scales of two populations based on the method described by Hodges Jr & Lehmann (1963) which estimates the location in terms of rank tests. Wang &

Hettmansperger (1990) stated that the method described in Wei & Gail (1983) can also be used to obtained interval estimators of differences of locations. Su & Wei (1993) showed the main drawback of the Wang & Hettmansperger (1990)'s method is the difficulty of estimating the density function in the presence of censored data. To overcome this limitation, Su & Wei (1993) introduced a simple and purely non-parametric interval estimation procedure for the difference or ratio of two median failure times of censored data which does not require any non-parametric density estimation. Since the identical shape assumption is unrealistic (Lehmann & D'Abrera, 1975), the main advantage of this method is that it is valid asymptotically even when the distributions are different in shape. Even though the method described by Su & Wei (1993) is restricted to survival data, Price & Bonett (2002) stated that this method is worthy since it does not require the unrealistic identical shape assumption.

Price & Bonett (2002) introduced asymptotic distribution-free confidence intervals for the difference and ratio of medians to compare the locations of two populations. The main advantage of this method is that it does not require the identical shape distribution assumption of the populations which was the main drawback of the existing classical non-parametric confidence intervals for the difference or ratio of medians. Price & Bonett (2002) suggested that the ratio of medians is more meaningful than the differences in medians when the response variable is measured on a ratio scale. Price & Bonett (2002)'s proposed method is easy to compute, performs well even for small samples, is valid for unequal sample sizes and can be applied to any continuous random response variable including survival times. We discussed Price & Bonett (2002)'s methods in detail in Arachchige *et al.* (2019a). Motivated by this method, as an improvement, we propose the ratio of quantiles to compare the locations of two populations and our completed work can be found in Arachchige *et al.* (2019a).

1.1.2 Comparing scales of two independent populations

The F-test is the most commonly used test to compare the differences of variances in two independent populations when the populations are normally distributed. However, if the normality assumption is violated, then the F-test can be highly unreliable and the level of significance (size of the test) can be very different from the assumed level. Pearson (1931) first highlighted the sensitivity of the F-test to departures from normality and concerns were also confirmed later by Geary (1947) and Gayen (1950). The F-test is highly sensitive to the fourth moment of the distribution, which is kurtosis (Miller Jr, 1997), and therefore the size of the test can be very high (>0.05) for heavy-tailed distributions and low (<0.05) for light-tailed distributions. Recently, Hosken *et al.* (2018) noted that "do not use F-test to compare variances" due to its sensitivity to deviations from the normality. Shoemaker (2003) showed that ln$F$ is approximately distributed as normal when the two samples are coming from two identically distributed populations with possibly different in locations and spreads.

Shoemaker (2003) introduced two adjustments to the F-test which improve its robustness properties and power. In the first adjustment of the F-test (the F_1-test), the distribution of F is approximated by Snedecor's F distribution using an appropriate adjustment to the degrees of freedom. To approximate the degrees of freedom, the matching of the moment technique was applied for the moment of Snedecor's F distribution and to the moment of the F-statistic. Since this is a complex procedure, as an alternative approach, Shoemaker (2003) suggested applying a matching of the moment technique to match the moment of the normalized F-statistic to the moment of the normal distribution. In the second adjustment to the F-test (F_2 - test), Shoemaker (2003) directly matched the mean and variance between the F-statistic and Snedecor's F distribution again using the matching of the moment technique to approximate the degrees of freedom. Shoemaker (2003) compared the level and power of the F_1-test and the F_2-test to the F-test, Levene (1961) and Brown & Forsythe (1974) tests (discussed later) via a simulation study and concluded that the F_1-test is the often best choice out of all the selected tests for comparisons based on both level and power.

There are several parametric alternatives to the F-test in the literature. The Levene (1961) test, Box-Anderson (Box & Andersen, 1955) test, Brown & Forsythe (1974) test, O'Brien test (O'brien, 1979; O'Brien, 1981) test and Smith's test (GBÜNEBERG *et al.*, 1966; Van Valen, 2005) are amongst the most popular. In the Levene test, the original values are replaced by their mean absolute deviations. Initially, the Levene test was designed only for equal sample sizes and Draper & Hunter (1969) generalized the test for unequal sample sizes. The Levene test is robust even for extreme changes of the skewness and kurtosis of the distribution. Also, the Levene test is very straightforward and easy to implement since the test function is readily available in R. The Box-Anderson test is an robust test which was developed based on permutation theory. The Brown-Forsythe test replaces the mean in the Levene test by 10% trimmed mean and hence it is more robust compared to the Levene test for non-normality and better for heavy- tailed distributions. The O'Brien test is another way to test the homogeneity of variances which uses a transformation of the original values of the distribution. The O'Brien test minimises both type I and type II errors and is often better for skewed distributions. The Smith test is a rarely used robust test which does not require the normality assumption.

There are several re-sampling tests such as the bootstrap, jackknife (Miller, 1968) and permutation tests which are an alternative to the F-test. The bootstrap method is a sometimes used method for testing differences in variances. However, not all the bootstrap methods are robust or perform well (Hall & Wilson, 1991). The jackknife method performs well for samples of reasonable size and is robust for both skewness and kurtosis, except for extreme kurtosis. For extreme kurtosis, the type I error of the jackknife method is less than 0.05. An advantage of the jackknife method is that there is an interval estimator associated with the test (Bissell & Ferguson, 1975). However, the jackknife method produces the unstable type I error for unequal sample sizes (Conover *et al.*, 1981). The test statistic of the permutation tests is distribution-free and hence power is low for small samples.

Hosken *et al.* (2018) compared the performance of a few parametric and re-sampling tests to testing the differences between two population variances. Under the parametric tests, Hosken *et al.* (2018) compared Levene (1961)'s test, the Box-Anderson (Box & Andersen, 1955) test and Smith's (Van Valen, 2005) test. Under the re-sampling tests, Hosken *et al.* (2018) compared the Jackknife (Miller, 1974), bootstrap and permutation tests. Hosken *et al.* (2018) formed several conclusions based on their simulation results such as (i) the Smith test is resistant to extreme skewness and kurtosis and type I error is less than 0.05, (ii) permutation tests are more powerful than the jackknife test for samples of a small sizes and is robust and powerful for large sample sizes and (iii) Levene's test performs best compared to the other tests for variance comparisons.

There is a vast body of literature on non-parametric alternative tests to the F-test. The Mood (1954) test, Siegel-Tukey (Siegel & Tukey, 1960) test and Klotz (Klotz *et al.*, 1962) test are a few of the more popular tests. The Mood test has a higher efficiency (Mood, 1954) than the other methods. The Siegel-Tukey test assumes the locations of the two populations are equal or approximately equal, and can be applied when the two samples are different in size. It is also preferred for distributions which have heavy tails (Siegel & Tukey, 1960). Klotz *et al.* (1962) compared the asymptotic relative efficiencies of the Siegel-Tukey test, Mood test and the Normal Scores Rank test (Klotz test) which uses the more convenient normal quantiles. Raghavachari (1965) introduced a modified version of the Klotz test which can be used when the locations of the distributions are completely unknown. Several non-parametric tests were introduced by Sukhatme (1957, 1958); Sukhatme *et al.* (1958). Moses (1963) also introduced a class of non-parametric tests called "Rank Like" tests. The Ansari-Bradley (Ansari *et al.*, 1960) test is another non-parametric test which has a modified version when the differences of locations of the populations are unknown. The efficiency of the Ansari-Bradley is lower than the efficiency of some of the other tests, but its advantage is that it is simple to apply.

The $W50$ test is a popular non-parametric test which replaces the mean of the Levene's (Levene, 1961) test by the median. The $W50$ test is an asymptotically distribution-free (Miller, 1968), simple, powerful (Brown & Forsythe, 1974), efficient and robust (Brown & Forsythe, 1974; Conover *et al.*, 1981) test for testing differences in variation, when the underlying distributions are long-tailed distributions. Hall (1972); Geng *et al.* (1979); Balakrishnan & Ma (1990) also recommended the $W50$ test based on their simulation studies. Lim & Loh (1996) showed that the bootstrap version of the $W50$ test has high power. However, the $W50$ test has a hidden structural problem. In particular, for some underlying distributions, the power of the test never reaches the highest value of 1, even when the population variances are very different (Pan, 1999). The $W50$ test consists of an interval estimator, however, for some underlying distributions, this interval becomes non-informative by providing intervals such as $(0, +\infty)$.

To overcome the limitations of the $W50$ test, Pan (1999) introduced two modifications, these being the $M50$ and $L50$ tests. In Pan (1999)'s simulation study, the simulated sizes of the $W50$ test were less than the assumed significance level of 0.05. Therefore, Pan (1999) changed the cut-off value of the $W50$ test from $t_{\alpha/2}$ to the normal upper quantile $Z_{\alpha/2}$ to become closer to the size of the $W50$ test up to a nominal level and hence introduced the $M50$ test. However, the main drawback of the $W50$ test remains in the $M50$ test, but is less severe. The $M50$ test is more powerful than the $W50$ test since $Z_{\alpha/2} < t_{\alpha/2}$. The $L50$ test was calculated using logarithms of the mean absolute deviations from the medians and it completely overcomes the limitation of the $W50$ test while retaining all the beneficial properties of the $W50$ test. For symmetric or approximately symmetric distributions with moderate to heavy tails, the sizes of the $L50$ test are becoming closer to the nominal 0.05 level (Pan, 1999). The $L50$ test is more powerful than the $M50$ test. Balakrishnan & Ma (1990) showed that the $L50$ test is more robust than the Jackknife test (Miller, 1968). All three tests, $W50$, $M50$ and $L50$ are associated with interval estimators (Pan, 1999).

The R-test (O'brien, 1979) is another popular non-parametric test which is most suitable for light tailed distributions. Conover *et al.* (1981) modified the FK test by using the median instead of the mean (Fligner & Killeen, 1976) and this is one of the best tests of the linear rank tests proposed by Fligner & Killeen (1976) due to its robustness properties and high power. Since the FK test performs poorly for skewed distributions, Hall & Padmanabhan

(1997) introduced the HP test, which is an adaptive version of the FK test. Balakrishnan & Ma (1990) compared the power performance of the FK and $W50$ tests and concluded that they are similar for most of the underlying distributions. Hall & Padmanabhan (1997) showed the FK test and HP test have similar power performance. Therefore, the $L50$ is competitive against the modified FK test based on power performance (Pan, 1999).

Another popular test is the Squared Rank test (Conover & Conover, 1980) which is a non-parametric version of Levene's test for the equality of variances. Shoemaker & Hettmansperger (1982) proposed a test based on the asymptotic variance of an M-estimator of location, which is called " midvariance" . Lax (1985) studied the robustness of the family of A-estimators for long-tailed, symmetric and uni-modal distributions. However, the associated measure of dispersion of most of the aforementioned tests cannot be easily calculated. Also, most of these tests can handle a violation of normality only in heavy tailed distributions and are unreliable for skewed distributions.

Shoemaker (1995) introduced a test for differences in dispersion based on the easy-to-compute and the more natural measure of variability known as "interquantile range (IQR_p, where $0 < p < 1/2$)" which is a generalization of the commonly used measure of the dispersion the "interquartile range ($\text{IQR}_{0.25}$)" (Chu *et al.*, 1957). Shoemaker (1995) investigated the properties of the interquantile range for symmetric distributions such as breakdown point, influence functions and asymptotic variance and concluded that the estimators have bounded influence functions and a nonzero breakdown point while variance has a zero breakdown point. The breakdown point is the proportion of bad points that is needed to be able to render estimator non-sensical and more details can be found in Hampel (1974). Shoemaker (1995) showed that the interquartile range performs poorly for both heavy-tailed and light-tailed distributions while the interquantile range performs well compared to using the sample standard deviation, for light-tailed distributions. Shoemaker (1995) further demonstrated the excellent performance of the interquantile ranges which were calculated using more extreme quantiles for a wide range of distributions. The Shoemaker (1995)'s test is based on the p^{th} and $(1-p)^{th}$ quantiles based on the order statistics of the combined sample related to two populations with the same location (median) but which are different in scale. Comparisons have been made of the proposed test which is based on the interquantile test to the classical F-test, normal scores test (Klotz *et al.*, 1962) and the squared rank test (Conover & Conover, 1980). The two interquantile tests (10th and 16th) have been studied for a wide variety of distributions such as Exponential, Laplace, t, standard normal, Beta, uniform and the triangle distribution for three different sample sizes,

$n = \{10, 25, 50\}$ with 1000 simulation trials. According to Shoemaker (1995), the proposed test based on the interquantile range performed well for light-tailed distributions and small sample sizes. Shoemaker (1995) highlighted some limitations of the selected tests, including the interquantile test based on simulations. One of the main limitations is that none of the tests were able to reach the assumed level of significance for the exponential distribution and some highly skewed distributions. Shoemaker (1995) also noted that further investigation is needed to find a test which can compare the dispersion of highly skewed distributions with unknown location parameters.

Shoemaker (1999) introduced a class of new tests as an improvement of his previous test (Shoemaker, 1995), based on the interquantile range which is valid for both skewed and symmetric distributions with known or unknown locations. First, Shoemaker (1999) compared the asymptotic relative efficiencies of $\mathrm{IQR}_{0.1}$ and $\mathrm{IQR}_{0.25}$ to the standard deviation based on the ratios of the squared coefficient of variation for a wide variety of distributions and concluded that $\mathrm{IQR}_{0.1}$ performed well for skewed distributions compared to the standard deviation and $\mathrm{IQR}_{0.25}$ performed comparatively poorly for all the selected distributions except the log-normal distribution. We discussed the Shoemaker (1999)'s test including all conditions in detail in Section 2.2 of Arachchige *et al.* (2019a). Shoemaker (1999) compared the new test with the F-test and the squared rank test (Conover & Conover, 1980) for finite samples through a simulation study for a wide range of distributions such as normal, log-normal, chi-square, Weibull, Beta, and exponential with different parameter choices with 1000 simulation trials. Shoemaker (1999) studied the effect of both equal $((n_1, n_2) = \{(10, 10), (25, 25), (50, 50)\})$ and unequal $((n_1, n_2) = \{24, 49\})$ sample sizes and used different interquantile ranges for different sample sizes for comparisons. The results show that the proposed test performs well over a variety of distributions compared to the F-test and squared rank test and more extreme interquantile tests are more powerful compared to the less extreme interquantile tests except for very highly skewed distributions. Shoemaker (1999) also noted that the power can be improved by choosing p between 0.1 and 0.25. Shoemaker (1999) further suggested that more than one interquantile range can be calculated as an alternative approach.

Marozzi (2011) compared the performance of some non-parametric, and parametric tests with quantile-based tests of Shoemaker (1995) and Shoemaker (1999) considering their robustness and power. Marozzi (2011) selected $W50$, $L50$, $M50$, R tests and their permutation and bootstrap versions, modified FK test and HP test under non-parametric tests and F-test as the parametric test. Marozzi (2011) used a wide range of distributions such as normal, bimodal, uniform, double exponential, t, chi-squared and exponential distribution with sample sizes of 10, 30 for 10000 simulation trials for his simulation study. Marozzi (2011) considered a test to be robust if the estimated type I error is at most 0.075 which is 1.5 times the significance level of 0.05. Marozzi (2011) found that the $W50$ test is computationally the simplest robust test which has a higher power and the re-sampling versions of the $W50$, $L50$, $M50$, R tests are more robust and powerful than the original tests.

Marozzi (2012) noted that the main drawbacks of Shoemaker (1999)'s test are that it requires finite population variances and an awareness of whether the underlying distributions are highly skewed or not. Marozzi (2012) firstly introduced a permutation version of Shoemaker (1999)'s two selected interquantile range tests such that one test is more suitable for highly skewed distributions while the other test is for less skewed distributions which do not require finite population variances. Marozzi (2012) secondly introduced a combined interquantile test based on a combination of two of Marozzi (2012)'s permutation versions of Shoemaker (1999)'s interquantile range tests which does not require an awareness of whether or not the underlying distributions are highly skewed. Marozzi (2012) compared the results of his new test with a few different tests such as the $W50$ test (Brown & Forsythe, 1974), $M50$ test and $L50$ test (Pan, 1999), the R test (O'brien, 1979) and the modified FK test (Conover *et al.*, 1981) by avoiding the F-test and squared rank test for comparisons since these two tests are not suitable for differences in scales (Conover *et al.*, 1981). Marozzi (2012) conducted a simulation study which has settings similar to Shoemaker (1999)'s simulation study and compared the performance of the permutation version and combined interquantile test. Marozzi (2012) concluded, based on the simulation results, that the overall most powerful test is $M50$ except for exponential and log-normal distributions. Marozzi (2012) also stated the combined interquantile test is the most powerful test for for exponential and log-normal distributions and the $W50$ and R tests are less powerful than the $M50$ test. Marozzi (2012) found that the permutation test is robust and more powerful than Shoemaker (1999)'s test and the combined interquantile test is robust and more powerful than both the traditional and the permutation version of Shoemaker (1999)'s test.

Motivated by Shoemaker (1995, 1999, 2003)'s tests which are based on a more natural measure of dispersion which is the interqunatile range, we introduce a ratio of interquantile ranges with the interval estimator to compare the variation between two populations. Our complete work can be found in Arachchige *et al.* (2019a) and some extra work can be found in Section 3.1 in Chapter 3.

Median absolute deviation (MAD) is another robust alternative to measure the dispersion of a distribution (Hampel, 1974). MAD is the median of the absolute residuals from the median and hence it is the natural scale counterpart of the median (Hampel *et al.*, 1986). Gauss (1816) had considered MAD much earlier and this was rediscovered by Hampel (1968). Later, Hampel (1974) called MAD "median deviation" . Hampel (1974) showed that the MAD is the " most robust" estimator of scale with regard to both the break-down point and gross error sensitivity. Gross error sensitivity is the supremum of the absolute value of the influence function and can be found more details in (Hampel *et al.*, 1986). The MAD consists of the highest possible value of the breakdown point for a scale estimator which is equal to 1/2. Therefore, it appears to be the most useful single estimate of scale and is the symmetrised version of the interquantile range (Huber, 1981). Rousseeuw & Croux (1993) stated that MAD has a bounded influence function with the sharpest bound compared to all the other estimators of dispersion, and hence it is very robust. MAD is the non-parametric natural estimator of the "probable error" of a single observation which was widely used in the history of statistics. Furthermore, MAD is widely used in regression analysis as a measure of dispersion due to its high breakdown property (Rousseeuw & Croux, 1993). Mean absolute deviation from the median is another measure of variability which is simple to understand and easy to compute (David, 1998). Pan (1999) introduced an interval estimator for the ratio of median absolute deviation from the median. However, Pan (1999)'s method did not perform well for small sample sizes and very unequal samples (David, 1998). Bonett & Seier (2003) introduced approximate interval estimators for mean absolute deviation from the median for a single sample and ratios of mean absolute deviation by median for two samples. Bonett & Seier (2003) also noted that mean absolute deviation from a target value instead of the median is another good alternative.

Since MAD is highly resistant to outliers, it has many applications including outlier detection in areas such as medical statistics, pharmaceutical and bio-pharmaceutical sciences and bio-statistics. Wellmann & Gather (2003) used MAD to identify the outliers in a one-way random effects model. Mishra *et al.* (2008) applied MAD to determine the effect of outliers in the estimation of coefficients of linear regression using the ordinary least squares method. Leys *et al.* (2013) used MAD with the median to create a new rejection criterion for detecting outliers by avoiding traditional criteria based on mean and standard deviations which are very sensitive to outliers. Vastrad *et al.* (2013) use MAD to identify the outliers in Oxazolines and Oxazoles molecular description data. Rajput *et al.* (2011) selected the most relevant dimensions from high dimensional data by minimizing MAD in high dimensional data clustering. Chung *et al.* (2008) used MAD as a robust measure to improve hit selection rather than the popular measure standard deviation in genome scale RNAi screens. Wu *et al.* (2002) used MAD as an estimator of standard deviation in Shewhart control charts. These are just some examples of where and how the MAD has been used.

Bonett & Seier (2003) suggested a bootstrap confidence interval for median absolute deviation from the median as a good choice. Bonett & Seier (2003) suggested distribution-free confidence intervals for median absolute deviation from the target (instead of median) using an approach given on page no. 137 in Snedecor & Cochran (1980) for a single sample. Bonett & Seier (2003) also suggested ratios of median absolute deviations from the target (instead of median) using an approach given in Price & Bonett (2002) as other possible alternatives. In both cases, the target is a known point and so if the median is used, then this differs from the standard MAD since the median must be known, and not estimated. Motivated by the aforementioned reasons, we propose a distribution-free interval estimator for median absolute deviation to make inference on the spread of a single population and ratios and differences of MAD with their interval estimators as robust alternatives to compare the spread of two populations. Our complete work regarding this can be found in Arachchige & Prendergast (2019) and some additional work can be found in Section 3.2 in Chapter 3.

1.2 Measuring relative dispersion

1.2.1 Measuring relative dispersion using a coefficient of variation

Pearson (1896) introduced the coefficient of variation (CV) which is the ratio of the standard deviation to the mean. The CV has broad applications for measuring relative spread since it is expressed in absolute units. The CV is very useful in many areas such as engineering, physics, chemistry, climatology, economics, business and finance for reliability or quality assurance studies. Since the CV is dimensionless and therefore does not vary with changes in measurement units, it is broadly used to express the precision and repeatability of an assay in analytical chemistry (e.g. Reed *et al.*, 2002). In finance, the CV is used as a measure of relative risk (e.g. Brief & Owen, 1969; Miller & Karson, 1977; Boyle & Rao, 1988; Weinraub & Kuhlman, 1994; Worthington & Higgs, 2003). For example, the equality of the CVs for two stocks can be tested to find out whether the two stocks produce the same risk or not. In economics, CV is considered as a summary measure of inequality (e.g. Atkinson, 1970; Chen & Fleisher, 1996). In medicine, the CV has been used to assess the homogeneity of bone test samples produced from a specific method to determine the effect of external treatments on the bones (Hamer *et al.*, 1995). In fault tree analysis, the CV has been used to analyse the uncertainty Ahn (1995). In climatology, CV has also been used to analyse rainfall data (Singh *et al.*, 1987; Ananthakrishnan & Soman, 1989; Ma & Zhang, 1991). In business, jobs are scheduled to minimize the CV (e.g. De *et al.*, 1996). Gong & Li (1999) used the CV to assess the measured strength of ceramics. Cole *et al.* (2000) used the CV as a summary measure, when developing age and sex specific cut-off points for the body mass index for overweight and obesity in children. Hillier & SO (1991) studied the effect of the CV of operation times on the allocation of storage space optimally in production line systems. Another advantage of CV is that the squared ratio of CVs can be used to compare the efficiency between scale estimators (Shoemaker, 1999). D'Alvise *et al.* (1999) introduced new techniques which estimate the statistical properties of SAR images by estimating CV. The reciprocal or inverse of the of CV, called the signal to noise ratio (McGibney & Smith, 1993), has important applications in quality control and reliability. A measure of relative dispersion is more meaningful if the response variable can be measured on a ratio scale and the variability is proportional to centrality (Zar, 1984, pg. 32). One of the main drawbacks of CV is it assumes the measurement error is proportional to the mean (Eisenberg, 2016). The CV statistic is an appropriate measure only for ratio variables, that is continuous variables with natural zero points (Eisenberg, 2016). The CV is not an efficient measure of spread when the distributions depart from normality (Fisher, 1922; Norris, 1938; Eisenberg, 2016). The standard deviation and mean are not the most efficient estimators of dispersion and location for skewed distributions. The statistics based on sample moments can be affected by the presence of a few outliers (Hampel *et al.*, 1986) and so some supplementary estimates of relative variation have been proposed in the literature.

1.2.2 Interval estimators for the coefficient of variation

McKay (1932) derived an interval estimator for CV and suggested its use when CV < 0.33. Pearson (1932) and Fieller (1932) showed that the McKay interval is very accurate for CV < 0.33. Umphrey (1983) also showed that the McKay interval is very accurate for CV < 0.33. Iglewicz & Myers (1970) compared the McKay interval with an exact interval under the asumption of normal data based on the non-central t distribution and concluded that the McKay interval is efficient for $n \geq 10$ and CV < 0.33. Later this interval was modified by Vangel (1996) which is known as the "modified McKay interval" and is shown to be nearly exact under normality and more accurate compared to McKay's original interval. Hendricks

& Robey (1936) studied the distribution of the sample CV when the data is sampled from a normally distributed population. Koopmans *et al.* (1964) constructed confidence intervals for the CV for normal and log normal distributions. Iglewicz (1967) derived some properties of the CV estimator such as, mean, variance and the exact distribution of the sample CV under normality and independent and identically distributed (iid) assumptions. The classical interval estimators for the CV are not robust to violation of the normality assumption. Sharma & Krishna (1994) introduced point and interval estimators for the inverse of the CV without assuming the normality of the parent population, based on the asymptotic sampling distribution of the inverse of the CV. Chaturvedi & Rani (1996) developed a sequential procedure to construct a confidence interval of fixed-width for the inverse of the CV of a normal population. Edward Miller (1991) introduced a one sample test statistic for a CV for a normal population and a two sample test statistic for CVs to compare the dispersion in two normal populations. Edward Miller (1991) showed that the McKay interval is reasonably accurate for $0.33 < \mathrm{CV} < 0.67$. Verrill (2003) also constructed an exact confidence interval for the CV for normal and log normal distributions and developed web-based computer programs to easily compute the intervals. Pang *et al.* (2005) constructed the point and interval estimators of the CV for log-normal, gamma and Weibull distributions. Panichkitkosolkul (2009) introduced an interval estimator for the CV for a normal distribution by replacing the sample CV of Vangel (1996)'s interval by a maximum likelihood estimator. Ng (2006) compared the performance of the McKay (1932), Sharma & Krishna (1994) and Edward Miller (1991) intervals and concluded that the McKay interval is more suitable for $n \geq 15$ and $\mathrm{CV} < 0.33$ and Miller interval is more suitable for $\mathrm{CV} > 0.33$. Panichkitkosolkul (2009) conducted Monte Carlo simulations to compare the performance of the new method with the McKay (1932) method and Curto & Pinto (2009) derived the asymptotic sampling distribution of the CV for non iid random variables as a special application to finance data and introduced two tests to compare the CVs of two populations. Mahmoudvand & Hassani (2009) introduced a new unbiased point estimator for the population CV from a normal distribution which has lower variance than the usual sample CV. Mahmoudvand & Hassani (2009) constructed two new interval estimators for the population CV based on a new point estimator of the CV Mahmoudvand & Hassani (2009) compared the performance with McKay (1932), Edward Miller (1991), Vangel (1996) and Sharma & Krishna (1994). Mahmoudvand & Hassani (2009) concluded that the proposed two interval estimators are simple to use and perform well even for small sample sizes. Banik & Kibria (2011) compared the performance of Hendricks & Robey (1936), McKay (1932), Edward Miller

(1991), Sharma & Krishna (1994) and Curto & Pinto (2009) intervals and some proposed bootstrap intervals such as non-parametric, parametric-t and bootstrap Miller (Bootstrap interval for Edward Miller, 1991) for symmetric and positively skewed distributions. Banik & Kibria (2011) concluded that McKay (1932), Edward Miller (1991), Sharma & Krishna (1994) and bootstrap Miller intervals performed well for symmetric distributions and McKay (1932) performed well for highly skewed distributions.

Groeneveld (2011a) re-investigated a confidence interval method for the CV which was originally proposed by Sharma & Krishna (1994) based on the inverse of the CV (1/CV which is called the signal to noise ratio). Groeneveld (2011a) also derived the influence function Hampel (1974) of the CV (see the next chapter for discussion and examples of the influence function), the inverse of the CV and the difference of two CVs as well. Later, we use this method for our comparisons and more details can be found in Arachchige *et al.* (2019b). Gulhar *et al.* (2012) compared 15 parametric and non-parametric interval estimators of the population CV. These 15 different interval estimators of CV consist of six existing interval estimators including three parametric and three non-parametric, three median modified interval estimators, four bootstrap intervals following Banik & Kibria (2011) and Gulhar *et al.* (2012)'s newly introduced interval estimator of the CV. The selected interval estimators include Edward Miller (1991), Sharma & Krishna (1994), Curto & Pinto (2009), McKay (1932), modified McKay Vangel (1996), Panichkitkosolkul (2009) intervals and their median modifications. Gulhar *et al.* (2012) introduced median modification to estimate the variance instead of the mean under the modified intervals since the median performed well compared to the mean for skewed distributions (Kibria, 2006; Shi & Golam K, 2007). In addition, Gulhar *et al.* (2012) proposed a new interval estimator based on the interval estimator for the population variance (σ^2). Gulhar *et al.* (2012) selected sample sizes as $n = \{15, 25, 50, 100, 500\}$, both symmetric (normal) and skewed (chi-square and gamma) distributions with different parameter choices, CV $= 0.1, 0.3, 0.5$ with 2000 simulation trials and 1000 bootstrap samples for their simulations. Finally, Gulhar *et al.* (2012) concluded that eight methods perform really well compared to the other methods, these being Miller, MacKay, modified MacKay and their median modifications, Panich and their proposed method. Later, we summarise in detail a few of the methods which preformed reasonably well, at least for some distributions considered. Later, we use a few of these intervals for our comparisons in Arachchige *et al.* (2019b). Albatineh *et al.* (2014) compared the performance of McKay (1932), Edward Miller (1991), Vangel (1996) and Mahmoudvand & Hassani (2009) intervals for rank-based samples by avoiding more common simple random sampling procedures.

There are several methods in literature to compare two populations based on the CV. Lohrding (1975) introduced a likelihood ratio test to compare the CV in two populations. Rao & Bhatt (1995) introduced tests based on jackknife and bootstrap procedures to compare the CV in two populations. Cabras *et al.* (2006) introduced a statistical test to compare the CV of two populations based on non-parametric bootstrapping. Cabras *et al.* (2006)'s test is a modified version of Sachs (n.d.)'s test statistics which is to compare the CV in two populations. Cabras *et al.* (2006) compared the performance of his test with the performance of Lohrding (1975)'s parametric bootstrap test and the Fisher F-test. Cabras *et al.* (2006) also introduced test statistics based on the ratios of two population CVs and log transformed ratios of two population CVs to compare the CV in two populations and stated that they had not satisfied the relevant properties compared to their other new method. Amiri & Zwanzig (2010) introduced another test to compare the equality of CVs between two populations based on Sachs (n.d.)'s test statistic, which works even when CV > 0.33. Amiri & Zwanzig (2010) stated that the new test is quick and easy to implement compared to the other existing bootstrap tests.

1.2.3 Alternatives to the coefficient of variation

There are several alternatives to the CV in the literature such as ratios of mean deviation from the mean or median and ratios between sums and differences of upper and lower quartiles (Q_3 and Q_1) which is called the "coefficient of variability" defined as $(Q_3 - Q_1)/(Q_3 + Q_1)$; for example see page no. 134 of Lovitt & Holtzclaw (1929), page no. 41 of Arkin & Colton (1935), page no. 153 of Sorenson (1936). The coefficient of variability is more popular with these researchers as it is a robust measure compared to the other measures. But there were no attempts to to find the usage of the considerations of the minimum sampling variances of these methods (Norris, 1938). Zwillinger & Kokoska (1999) called this measure as the "coefficient of quartile variation" . Bonett (2006) stated that the coefficient of quartile variation is more appropriate when data come from non-normal distributions. Bonett (2006) introduced an interval estimator for the coefficient of quartile variation which exhibited good coverage even for small samples and even in highly non-normal distributions. Bulent & Hamza (2018) constructed the percentile (non-parametric) and t-bootstrap (parametric) confidence intervals to coefficient of quartile variation and compared with the Bonett (2006)'s confidence interval. Bulent & Hamza (2018) concluded that the bootstrap intervals typically provide more conservative coverages for small samples ($n \leq 15$) compared to the Bonnet's (Bonett, 2006) confidence interval and both methods performed similarly and well for large samples ($n = 20, \ldots, 100$). Bulent & Hamza (2018) further noted that the average width of the (Bonett, 2006)'s were high for small sample sizes.

Another unit free alternative measure to the CV is called as " coefficient of dispersion" and defined as τ/m where, τ is the mean absolute deviation from median and m is the median. The coefficient of dispersion is widely used in tax assessments (Gastwirth, 1988, pg.28,29) and in biological applications since τ and m are more suitable as location and scale estimators compared to mean and standard deviation when the distribution is not normal. Bonett & Seier (2005) constructed distribution-free confidence intervals for τ/m based on the asymptotic distribution of the estimator of τ/m which was derived by Gastwirth (1982). Bonett & Seier (2005) compared the performance of the interval estimator of τ/m with the BC_a bootstrap interval (page no. 180 of Efron & Tibshirani, 1994) and concluded that the confidence interval for τ/m performed well compared to the BC_a bootstrap method for small samples and non-normal distributions. Bonett & Seier (2005) mentioned that the τ/m is more preferred compared to the CV for non-normal distributions. Bonett & Seier (2005) 's method does not perform well for extremely non-normal distributions unless the sample sizes are large (Bonett, 2006).

Shapiro (2005) introduced a robust alternative to the CV based on the interquartile range (IQR) and the median which we will label here as RCV_Q for convenience. Reimann *et al.* (2008) and Varmuza & Filzmoser (2009) considered another robust version of the CV, which we will denote RCV_M, which uses the median absolute deviation (MAD) in place of the standard deviation and the median. However, to the best of our knowledge, there has been little, if any, research into interval estimators for either the RCV_Q or RCV_M and it is this fact that motivates our research. Given recent findings highlighting excellent coverage for estimators based on ratios of quantiles, we introduce asymptotic interval estimators of RCV_Q and RCV_M and also the both non-parametric and parametric bootstrap interval estimators for RCV_M. In addition, we introduce interval estimators for ratios of RCV_Qs and RCV_Ms to compare the relative dispersion between two independent populations. We compared the performance of our new interval estimators with some selected existing interval estimators from Groeneveld (2011b) and Gulhar *et al.* (2012) as discussed above. Our complete work regarding to this can be found in Arachchige *et al.* (2019b) and some extra works can be found in Section 3.3 in Chapter 3.

1.3 Measuring skewness

Skewness measures the shape and the asymmetry of a continuous distribution. Skewness is zero for a symmetric distribution. If there is a long right tail, then the distribution is positively skewed and if the tail is to the left, then the distribution is negatively skewed. Therefore, the measures of skewness provide the degree and the direction of the skewness. There is a vast body of literature on measures of skewness. Two popular measures of skewness are Pearson's first skewness coefficient (Pearson's mode skewness) and Pearson's second skewness coefficient (Pearson's median skewness) which are defined as $\mathrm{SK}_1 = (\mu - M)/\sigma$ and $\mathrm{SK}_2 = 3(\mu - m)/\sigma$, where μ, M and m are the mean, mode and median respectively (Pearson, 1894, 1895). In addition, the standardised third central moment, introduced by Charlier (1905) and Edgeworth (1908), is often used to measure skewness of a random variable X and is defined as $\gamma_1 = \mu_3/\sigma^3$ where μ_3 is the third central moment and σ is the standard deviation. The main drawback of these three measures of skewness is that they can be strongly affected by outliers and one single outlier can highly influence the skewness coefficient. In addition, γ_1 can be arbitrarily large and hence hard to interpret. Benjamini & Krieger (1996) described the measures of skewness in terms of skewness and spread functions.

Ngatchou-Wandji (2006) introduced three tests for measures of skewness of an unknown distribution and compared the level and power of these tests with some existing tests. Boshnakov (2007) introduced some measures of asymmetry in terms of mode rather than mean or median. Boshnakov (2007) measured the asymmetry of an absolute continuous distribution using confidence transformation which is described in Boshnakov (2003).

1.3.1 Density-based measures of skewness

Critchley & Jones (2008) introduced the density based functional measure of skewness called "asymmetry functions " and discussed the properties and the method of estimation. Critchley & Jones (2008)'s prototype asymmetry function was similar to Boshnakov (2007)'s odds asymmetry function and their preferred function was obtained by replacing the quantile functions in γ_p (Hinkley, 1975) by asymmetric density functions. Since the density-based asymmetry functions measure asymmetry with respect to mode, they were naturally defined only for uni-modal densities. Critchley & Jones (2008) stated that their asymmetric function consists of natural location and scales measures which are median (m) and $f^{-1}(m)$ which can be easily interpreted in terms of the probability density function (f). In addition, Critchley & Jones (2008) introduced an integrated version of γ_p with asymmetric density functions over p and concluded that it is more robust.

1.3.2 Quantile-based measures of skewness

The well-known Bowley's coefficient (pg. 162 Yule, 1912; Bowley, 1920) defined as $B_1 = (Q_3 + Q_1 - 2m)/(Q_3 - Q_1)$, is an entirely quantile-based and robust measure of skewness where Q_3 and Q_1 represent the 3rd and 1st quartiles respectively. A more general case of the Bowley's coefficient, denoted γ_p for $p \in (0, 0.5)$, has been defined by David & Johnson (1956) and considered by Hinkley (1975) and Groeneveld & Meeden (1984). The $\gamma_p \in [-1,\ 1]$ and $\gamma_p = 0$ for symmetric distributions. γ_p becomes equal to Bowely's coefficient (B_1) when $p = 0.25$.

Van Zwet (1964) introduced a method ordering two distributions based on skewness. Oja (1981) introduced four properties any measure of skewness should satisfy. MacGillivray *et al.* (1986) also introduced some skewness ordering. Groeneveld & Meeden (1984) tested whether γ_p satisfies the four properties of the measure of skewness. Groeneveld (1991) described positive skewness form another angle, that is, as a location and scale-free movement of the probability mass of the symmetric distribution and used the influence function to compare the γ_1 and γ_p measures of skewness. Groeneveld (1991) stated that the influence function describes changes in the skeweness measures with deviations from symmetry. Arnold & Groeneveld (1995) introduced $\gamma_M = 1 - 2F(M_x)$ where M_x is the mode of the assumed uni-modal distribution of X and γ_M is a function of the distribution function (F) at M_x. Arnold & Groeneveld (1995) compared γ_M with γ_1 using the influence function. The γ_M is an analogous scalar skewness measure of the natural average of Critchley & Jones (2008)'s asymmetry function.

Brys *et al.* (2003) introduced four new measures of skewness called “medcouple”, “medtriple”, “repeated medcouple”and “repeated medtriple”, which are based only on ranks and are robust against outliers. These new measures are obtained by replacing some or all of the quantiles in γ_p with actual data points. The medcouple replaces the other two quantiles, except the median, while the medtriple replaces all three quantiles of γ_p with actual data points. The repeated versions of medcouple and medtriple are obtained by using repeated medians instead of considering the medians of all couples or triples of actual data points. Brys *et al.* (2003) stated that the repeated versions are computationally more complex compared to the original measures even though they have similar or higher breakdown points. Brys *et al.* (2003) carried out a simulation study to compare the performance of these four new skewness measures with γ_p when $p = 0.25, 0.125$ for both contaminated and non-contaminated as well as symmetric and asymmetric distributions and concluded that medcouple is the best measure of the six measures. Brys *et al.* (2003) stated that γ_p when $p = 0.25$ is less sensitive to outliers while γ_p when $p = 0.125$ is more suitable for capturing the asymmetry in the data since it uses more information from the tails of the distribution. Subrahmanya N & Aruna Rao (2003) also compared the robustness properties of the new estimators using breakdown points and found that all four measures of skewness positive points which are 25%, 20.6%, 25%, 50% and repeated medtriple has the highest point.

Brys *et al.* (2004) further studied medcouple which has a 25% breakdown point and showed that, it satisfies the properties which should be followed by any measure of skewness. Medcouple belongs to Hössjer *et al.* (1996)'s class of incomplete generalized L-statistics and hence is asymptotically normally distributed. Nevertheless, Brys *et al.* (2004) computed the influence function of medcouple and compared this with the influence functions of γ_p when $p = 0.25, 0.125$ and found that the influence function of medcouple is continuous, except at the median, while the other two are step functions and all three are bounded. The influence function of medcouple is a smoothed version of the influence functions for γ_p when $p = 0.25, 0.125$ and its gross error sensitivity is approximately equal to the gross error sensitivity of γ_p when $p = 0.25$. Brys *et al.* (2004) derived the asymptotic variance of medcouple and compared this with the finite sample variances of samples of different sizes. Brys *et al.* (2004) finally concluded that medcouple has a combined strength of γ_p when $p = 0.25, 0.125$ in terms of sensitivity to detect skewness and robustness towards outliers.

1.3.3 Further extensions of the γ_p

Groeneveld *et al.* (2009) introduced an improved version of the γ_p called λ_p for right skewed distributions for $p \in (0,\ 0.5)$ for which good point and interval estimators can be found easily. Later, we define γ_p and λ_p properly with their conditions. The λ_p measure is appropriate only when the direction of the skewness is known. Groeneveld *et al.* (2009) stated that λ_p is easy to interpret compared to γ_p and influence functions have been used to compare the sensitivity of λ_p and γ_p to right skewness. Groeneveld *et al.* (2009) showed that the influence function of λ_p is more sensitive to right skewness with decreasing p and its sensitivity to right skewness is at least four times higher than the influence function of γ_p. Groeneveld *et al.* (2009)

estimated the γ_p and λ_p based on the usual "plug in" method and the Bayesian bootstrap method (Rubin, 1981). They stated that the reason for selecting the Bayesian bootstrap method is that Meeden (1993) showed the Bayesian bootstrap method is better than the usual "plug in" method to estimate the quantiles. Groeneveld *et al.* (2009) constructed a 95% bootstrap confidence interval of the usual "plug in" method and 0.95 credible interval of the Bayesian bootstrap method for γ_p and λ_p when $p = \{0.05, 0.1, 0.15\}$. Based on their results, they recommended $p = 0.05$ is a good choice since it does not ignore the tail behaviour of the distribution while showing some robustness against outliers. In addition, they concluded that the Bayesian bootstrap interval is narrower than the usual "plug in" interval even though the coverage tends to be high. Groeneveld *et al.* (2009) have introduced another measure of skewness denoted η_p for left-skewed distributions and derived the influence function.

Staudte (2014) constructed distribution-free tests and confidence intervals for γ_p. These confidence intervals are distribution-free since their average width depends only on p, the confidence level and the sample size n for a variety of symmetric and asymmetric models. Staudte (2014) stated that the intervals provide good coverage for moderate to large sample sizes. In addition, Staudte (2014) derived a simple formula to find the required sample size to obtain a good coverage for the intervals with a pre-specified width. In addition, Staudte (2014) conducted a power comparison between tests including two distribution-free tests from Ngatchou-Wandji (2006).

One drawback of γ_p and λ_p is that p must be chosen. To overcome this limitation in γ_p, Groeneveld & Meeden (1984) introduced another skewness coefficient by integrating the numerator and denominator of the γ_p with respect to p on (0, 0.5) and defined it as $b_3 = (\mu - m)/E \mid X - m \mid$. In addition, Groeneveld & Meeden (1984) derived the influence function of b_3 to make comparisons with the influence functions of γ_1 and γ_p. One of our main objectives is to integrate γ_p and λ_p over p rather than integrating the numerator and denominator separately to overcome the limitation of the need to choose p in both γ_p and λ_p. Nevertheless, we introduce another two powerful skewness measures, integrated $p\gamma_p$ and $p\lambda_p$ over p as alternatives to the existing measures of skewness. Staudte (2014) provided the distribution-free confidence intervals for γ_p. Therefore, we introduce the distribution free interval estimators for λ_p, integrated versions of γ_p, λ_p, $p\gamma_p$ and $p\lambda_p$ and conduct a simulation study to make comparisons between these estimators, before applying them to some real-world examples. Our complete work related to this can be found in Paper 4 and Section 3.4 in Chapter 3.

1.4 Major contribution of this thesis

The main objective of this thesis is to provide new point and interval estimators for some function(s) of linear combinations of quantiles. This thesis is divided into two parts. Part I consists of four chapters and an Appendix. Part II consists of four publications.

Part I

Chapter 1 - Provides the background of the thesis. Previous research works related to the main research problems are briefly discussed.

Chapter 2 - Discusses the theory of quantile estimators and their influence function which were used in the four publications to develop new theories and methods.

Chapter 3 - Provides additional works related to publications which was not included in the papers but is worth discussing.

Chapter 4 - Presents the main results and findings of the thesis, conclusions and future work.

Appendix - Provides the most important R programs that we developed to run the simulations related to the works in the four publications.

Part II

Part II of this thesis comprises the four accepted or submitted papers.

1.4.1 Summaries of original papers

Paper I

The t-test and the F-test are the most commonly used methods to compare the location and scale of two populations under the normality assumption. However, these tests are highly unreliable for skewed distributions. The difference and the ratio of medians are two existing non-parametric alternatives to the t-test for location comparisons. Researchers suggested non-parametric tests to compare the scale based on the most natural estimator of scale which is the interquantile range (IQR). The ratios of independent quantiles and the squared ratios of IQRs with their point and interval estimators were proposed as alternatives to the existing methods to compare the location and scales of two independent populations respectively. The performance of the new measures was compared with the existing methods via simulations and the results suggest that the new methods provide very good coverage probabilities which are very close to nominal even for small sample sizes and a wide range of distributions. The R shiny web application was developed to run the simulations efficiently. The best choice of p was suggested for the squared ratio of IQRs by minimizing asymptotic variance. The robustness properties of the estimators were investigated using partial influence functions. Finally, some real-world data examples were used to compare the performance of the new measures and the results reveal that the new estimators provide different conclusions compared to existing methods.

Paper II

The median absolute deviation (MAD) is a robust measure of dispersion which is applied in many areas, including pharmaceutical and bio-pharmaceutical research. Motivated by this, an interval estimator for MAD was introduced to make inferences on the dispersion of a single population. The difference and squared ratios of MAD with their point and interval estimators were introduced to compare spread for two independent populations. Simulations were conducted to check the performance of the new estimators and the resulting coverage probabilities were very close to nominal coverage, even for small sample sizes and a wide variety of distributions. The robustness properties of the estimators were investigated using the influence function and partial influence functions. The example reveals that the difference and squared ratio of MAD are robust to outliers compared to the F-test.

Paper III

The coefficient of variation (CV) is the most commonly used method to measure relative dispersion. However, CV is highly unreliable for outliers and skewed distributions since it is based on mean and standard deviation. The IQR divided by median and the MAD divided by median are two existing quantile-based robust alternatives to the CV. The main objective of this paper is to further investigate these two robust versions of the CV. First, comparisons were made between all three measures of relative dispersion for some distributions. Then several properties which should satisfy any measure of relative dispersion were tested for CV and the two robust alternatives to CV. The distribution-free interval estimators were proposed for the two robust alternatives of CV and a simulation study was conducted to compare the performance of the new interval estimators with several existing interval estimators of CV. The results suggest that coverage probabilities are very close to nominal even for small sample sizes and for a wide range of distributions. The R shiny web application was developed to run the simulations to make comparisons between the point and interval estimators of CV and two robust CV's for a wide range of distribution. As an extension to the two populations, interval estimators were introduced for ratios of CVs and the two robust alternatives and examples reveal that different conclusions can be formed by using robust alternatives compared to CV.

Paper IV

Skewness measures the shape and the asymmetry of a continuous distribution. There are several skewness measures available in the literature and of these, Bowley's well-known skewness coefficient is a quantile-based measure of skewness. There are some generalised versions of Bowley's skewness coefficient. However, the main drawback of these existing measures is that p must be chosen. Therefore, the objective of this paper is to introduce more powerful alternatives to the existing measures of skewness which remove the "need to choose p" requirement with their interval estimators. The integrated versions are introduced to the generalized versions of Bowley's skewness coefficient. First, it was tested for the validity of the four properties for the new measures of skewness which should be satisfied by any of the measures of skewness. Then, distribution-free confidence intervals were derived, and the performance of the intervals was compared via a simulation study. The observed coverage probabilities were very close to nominal, even for small sample sizes and a wide range of distributions considered. The point and interval estimators for the difference of the new measures of skewness were introduced to compare the skewness of two populations. Some real-world data examples were used to compare the conclusions obtained by different measures of skewness.

1.4.2 My contribution in all four papers

Paper I

In Paper I, research related to the ratio of independent quantiles to compare the location of two independent quantiles was undertaken by one of my colleagues, Maxwell Cairns. My contribution was to develop theories, simulations and finally examples related to the ratio of interquantile ranges. First, I investigated the existing methods to compare the scale of two independent populations. Then, I contributed to the construction of the confidence intervals, partial influence function derivations and comparisons, asymptotic variance derivations and comparisons of the squared ratio of IQRs and ratio of variances. The best choice of p to obtain the minimum asymptotic variance for the squared ratio of IQRs was investigated. Simulations were conducted for a wide range of distributions, including skewed distributions, to compare the performance of the existing estimators and the new estimators. R programs and the R shiny web application were developed to run the simulations. I was able to find two very nice data sets which were named Melbourne house price data and prostate cancer data to compare the conclusions made by the existing and new estimators. Finally, I contributed to the manuscript preparation.

The main results of this paper were presented at the 1st Victorian Research Students Meeting in Probability and Statistics (VRSMiPS) which was held on 5th June 2017 at Latrobe University, Melbourne, Australia and the International Robust Statistics Conference (ICORS) 2017 which was held on 2nd to 7th of July 2017 at the University of Wollongong, Sydney, Australia. Positive feedback, suggestions and comments regarding the research works were received.

Paper II

For Paper II, I first conducted a literature search to find early work related to median absolute deviations. Then, distribution-free interval estimators for MAD, difference and squared ratios of MADs were constructed. Partial influence function derivations and comparisons were conducted. R programs were developed to conduct the simulations for a wide range of distributions. Then, different conclusions provided by the estimators were compared using the prostate cancer data. Finally, I contributed to the manuscript writing.

The main results of this paper were presented at the 3rd Victorian Research Students Meeting in Probability and Statistics (VRSMiPS) which was held on 2nd October 2019 at the University of Melbourne, Melbourne, Australia.

Paper III

For Paper III, I first conducted a literature search for existing interval estimators of CV and alternatives to CV. Two robust quantile-based alternatives, IQR divided by median (RCV_Q)

and MAD divided by median (RCV_M) were investigated and numerical comparisons were conducted between the measures for a wide range of distributions. Then, the robustness properties of RCV_Q and RCV_M were investigated using influence functions and I conducted an influence function comparison of CV, RCV_Q and RCV_M. Asymptotic variances of the estimators were derived and relative asymptotic standard deviation comparisons were conducted. The distribution-free point and interval estimators were constructed for RCV_Q and RCV_M. In addition, parametric and non-parametric bootstrap intervals were constructed for RCV_M. Interval estimators for the ratios of RCV_Qs and RCV_Ms were constructed to compare the relative spread between two independent populations. R programs and the R shiny web application were developed to run the simulations. Simulations were conducted for a wide range of distributions. Two real-world data sets were used to compare the different conclusions provided by CV and two robust alternatives to the CV. Finally, I contributed to the manuscript writing.

The main results of this paper were presented at the 2nd Victorian Research Students Meeting in Probability and Statistics (VRSMiPS) which was held on 25th September 2018 at Monash University, Melbourne, Australia and the International Society for Clinical Biostatistics and Australian Statistical Conference 2018 (ISCB ASC 2018) which was held on 26th to 30 August 2018, Melbourne, Australia.

Paper IV

For Paper IV, I first conducted a literature search on the existing measures of skewness. Then, γ_p and λ_p the two generalized versions of Bowley's well-known skewness coefficient were further investigated. Then, the new measures of skewness were checked to ensure they satisfied the four properties which should be satisfied by any measures of skewness. Since closed form expressions are not available, numerical approximations were used to estimate the measures of skewness and asymptotic variances. A relative asymptotic variance comparison was conducted graphically. Distribution-free confidence intervals were constructed for new measures of skewness. R programs and the R shiny web application were developed to run the simulations. Simulations were conducted for a wide range of distributions. Two very nice new data sets, called "computer data" and "income data" were compared with an existing data set to compare the conclusions achieved by different measures of skewness. Finally, I contributed to the manuscript writing.

2. Theory of quantile estimators

In this chapter, first we discuss the key definitions related to the quantile function, quantile density function and quantile density estimation. Most of these definitions can be found in, for example, Tukey (1965), and Parzen (1962, 1977, 1979). Then we discuss the theory behind the influence function and the partial influence functions in detail, including some examples and their application to quantiles. Finally, we discuss the theory behind asymptotic variances, the connection between asymptotic variance and the influence function, point and interval estimations including their application to functions of quantiles. These results can be found in more detail, for example, in Hampel (1968, 1974); Hampel *et al.* (1986), Huber (1981), Staudte & Sheather (1990) and Clarke (2018). In addition to these references, some early works of my supervisors which are related to our research topic can be found in Prendergast & Staudte (2017a,b, 2016b,a).

This chapter details the methodology that is used to develop our new theorems and the results related to the point and interval estimators of the functions of quantiles that we introduce in four publications. Therefore, readers who are already familiar with these concepts can omit this chapter.

2.1 Key definitions

This section contains key definitions related to quantiles such as the quantile function, quantile density function and several quantile density estimators. The methods defined in this section are used in subsequent chapters to develop the new quantile-based estimators and to discuss their properties.

2.1.1 The quantile function

Let $X_1,...,X_n$ be n independent sample realisations of random variable $X \sim F$, where we assume a continuous distribution function $F(x) = Pr(X \leq x)$ with positive domain. The probability density function is $f(x) = F'(x)$. Then the quantile function, $\mathcal{G}(F;p) = G(p)$, is (e.g. Parzen, 1979)

$$G(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\} \tag{2.1}$$

where $0 \leq p \leq 1$. The quantile function was earlier called the "representing function" by Tukey (1965). When $-\infty < x < \infty$ and $0 < p < 1$, $G(p)$ has a fundamental property which is $F(x) \geq p$ if and only if $G(p) \leq x$.

When F is continuous and uniformly distributed on $[0,\ 1]$, the quantile function $G(p)$ satisfies,

$$G(p) = \inf\{x : F(x) = p\}, \tag{2.2}$$

$$F(G(p)) = p \qquad \text{for } 0 \leq p \leq 1, \tag{2.3}$$

$$1 - F(G(p)) = 1 - p. \tag{2.4}$$

Note that x_p and $\mathcal{G}(\cdot, p)$ are other commonly used notations to represent the quantile function $\mathcal{G}(F;p) = G(p)$.

2.1.2 Estimation of the quantile function

In this section, we provide several definitions for the empirical quantile function and sample quantile function.

Empirical quantile function

Tukey (1965) called a sample quantile function an "empirical representing function". Wilk & Gnanadesikan (1968) called a sample quantile function an "empirical cumulative distribution function (ECDF)" and explained its advantages. Wilk & Gnanadesikan (1968) stated that a sample quantile function provides the nonparametric estimator of the quantile function. In addition, the ECDF can be applied even in situations where the sample is non-random.

Parzen (1979) called the sample quantile function an "empirical quantile function" and introduced several definitions for this. Let $X_1, \ldots, X_n$ be an independent sample from a continuous random variable X, then the empirical distribution function (EDF) can be defined as,

$$F_n(x) = \widetilde{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}\ , \tag{2.5}$$

where I_a is the indicator function equal to one if its argument, a, is true and zero otherwise, and this is the usual formula where $F_n(x) = \widetilde{F}(x)$ is the empirical distribution that assigns probability mass $1/n$ to each observation X_i. Then, Parzen (1979)'s first definition of the empirical quantile function (EQF) is,

$$\widetilde{G}(p) = \widetilde{F}^{-1}(p) = \inf\{x : \widetilde{F}(x) \geq p\} \ . \tag{2.6}$$

Let $X_{(1)}, X_{(2)}, ..., X_{(n)}$ be the order statistics of the random sample. Then, Parzen (1979)'s second definition of the empirical quantile function is based on order statistics. The EQF is a piece-wise constant function and its values are the order statistics $X_{(1)} < X_{(2)} < ... < X_{(n)}$ and defined as,

$$\widetilde{G}(p) = X_{(i)} \ \text{for} \ (i-1)/n < p \leq i/n, \ i = 1, ..., n \ . \tag{2.7}$$

Parzen (1979)'s third definition of the empirical quantile function is as a piece-wise linear function which is his preferred definition out of the three definitions. In the third definition, the empirical quantile function $\widetilde{G}(p)$ is defined as,

$$\widetilde{G}(p) = n\left(\frac{i}{n} - p\right) X_{(i-1)} + n\left(p - \frac{i-1}{n}\right) X_{(i)} \tag{2.8}$$

for $\frac{i-1}{n} \ \leq p \ \leq \ \frac{i}{n}$ and $i = 1, ..., n$.

Sample quantile function

Parzen (1979) stated that there are three different ways to estimate quantiles. These estimation methods are parametric, non-parametric and non-parametric pre-flattened. For the parametric method, to estimate the quantile function, $G(p)$, one can assume the form

$$G(p) = \mu + \sigma G_0(p) \tag{2.9}$$

and this is analogous to the classical location and scale parameter model of $f(x)$ which is

$$f(x) = \frac{1}{\sigma} f_0 \left[\frac{(x-\mu)}{\sigma}\right] \tag{2.10}$$

and where,

$$\mu = \int_0^1 G(p)dp \ \text{and} \ \sigma^2 = \int_0^1 \{G(p) - \mu\}^2 dp \ . \tag{2.11}$$

Then the sample quantile function, $\widehat{G}(p)$, is

$$\widehat{G}(p) = \widehat{\mu} + \widehat{\sigma} G_0(p) \tag{2.12}$$

where G_0 is the quantile function associated with f_0 and $\widehat{\mu}$ and $\widehat{\sigma}$ the estimates of μ and σ from (2.11).

For the non-parametric approach to estimate the quantile, the estimator is written in the form of a kernel estimator by averaging over the values of $\widetilde{G}(u)$ for u in a neighbourhood of p. Then, the kernel estimator of the $\widehat{G}(p)$ is defined as,

$$\widehat{G}(p) = \int_0^1 \widetilde{G}(u) \frac{1}{b} k\left(\frac{p-u}{b}\right) du \tag{2.13}$$

where k is a suitable kernel, b is the bandwidth or smoothing parameter and $\frac{i-1}{n} \leq u \leq \frac{i}{n}$ for $i = 1, ..., n$. For a specific sample size n, the amount of smoothness in the estimator is controlled by the bandwidth.

If k is differentiable, $\widehat{G}(p)$ can be approximated as (e.g. Welsh, 1988),

$$\widehat{G}(p) \approx \sum_{i=1}^{n} X_{(i)} \frac{1}{b} k\left(\frac{p-u}{b}\right) . \tag{2.14}$$

Sheather & Marron (1990) stated that considering a linear combination of order statistics, the efficiency of the sample quantiles can be improved. An appropriate weight function can be used to form a weighted average of all the order statistics and a class of such estimators are called kernel quantile estimators. Sheather & Marron (1990) also noted that for a given kernel, k, the asymptotically optimal bandwidth of the kernel quantile function depends on the first and second derivative of the quantile function, $G'(p)$ and $G''(p)$, and if the first and second derivative of the kernel function k exist, one can estimate $G'(p)$ and $G''(p)$ using the first and second derivative of the kernel quantile function. Sheather & Marron (1990) also suggested a method to find the optimal bandwidth which minimizes the asymptotic mean square error of the kernel quantile estimator.

Sample quantiles in statistical software R

The sample quantiles which are used in statistical software R (Development Core Team, 2018) are based on one or two order statistics. There are nine definitions for the sample quantile in the R function `quantile`. Hyndman & Fan (1996) compared these nine definitions and recommended definition 8 to be the best choice since it provides median-unbiased estimates of the quantile function $G(p)$ regardless of the distribution. Therefore, we used definition 8 to estimate the quantiles in our simulations and real-world data applications that follow. In addition, Hyndman & Fan (1996) found six desirable properties for a sample quantile and checked whether the sample quantile types satisfy these properties.

Let $\{X_1, ..., X_n\}$ be the set of independent observations from distribution F and let the order statistics be denoted $\{X_{(1)}, ..., X_{(n)}\}$. Then the sample quantile of type l ($l = 1, \ldots, 9$) denoted $\widehat{G}_l(p)$ is written as (Hyndman & Fan, 1996),

$$\widehat{G}_l(p) = (1 - \delta_l) X_{(j)} + \delta_l X_{(j+1)} \tag{2.15}$$

where and $\frac{j-t_l}{n} \leq p < \frac{j-t_l+1}{n}$ for some t_l ε R and $0 \leq \delta_l \leq 1$. The value of δ_l is a function of $j = [pn + t_l]$ and $v_l = np + t_l - j$ where t_l is a constant determined by the type of sample quantile.

Sample quantile definition 8

Let $X_{(j)}$ be the jth order statistic where j is the nearest integer to np. The median position is $mF(X_{(j)}) \approx (j - \frac{1}{3})/(n + \frac{1}{3})$, where m denotes the median. Therefore, sample quantile type 8, $\widehat{G}_8(p)$, is defined by setting $p_j = (j - \frac{1}{3})/(n + \frac{1}{3})$ and then $p_j \approx mF(X_{(j)})$. $\hat{G}_8(p)$ is a continuous function of p with $\delta = v$, $t = (\frac{p+1}{3}) p_j = (k - 1/3)/(n + 1/3)$ and p_j is the plotting position in a quantile plot in which $X_{(j)}$ is plotted against p_j (Hyndman & Fan, 1996). The resulting sample quantile is approximately median-unbiased of order $o(n^{-1/2})$ (pg. 248 Reiss, 1989).

2.1.3 Density quantile function

The density quantile function is defined as,

$$q(p) = f(G(p)) = f(x_p) \tag{2.16}$$

where x_p, the pth quantile, will be used for simplicity when necessary in what follows.

2.1.4 Quantile density function

Assuming F has a positive and continuous derivative $F'(x) = f(x)$ on its domain, the quantile density function is defined as (Parzen, 1979)

$$g(p) = \mathcal{G}'(F;p) = G'(p) = 1/f(x_p) \tag{2.17}$$

and it was earlier called the "sparsity function" by Tukey (1965). Note that the density quantile function $f(x_p)$ and quantile density function $g(p)$ are reciprocals of each other.

2.1.5 Quantile density estimation

There is much literature on quantile density estimation. The empirical quantile density function $\widetilde{g}(p)$ can be obtained using the empirical quantile function in (2.8) and is defined as

$$\widetilde{g}(p) = n(X_{(i)} - X_{(i-1)}) \tag{2.18}$$

for $(i-1)/n \leq p < i/n$ where the $n(X_{(i)} - X_{(i-1)})$ are called spacings of the sample (Pyke, 1965; Pyke *et al.*, 1972) and $\widetilde{g}(p)$ is asymptotically exponentially distributed with mean $g(p)$ (Parzen, 1979). The quantile density can be estimated using three approaches, namely parametric, non-parametric and non-parametric pre-flattened (Parzen, 1979). The kernel density estimation approach is a non-parametric approach to estimate the quantile density. If the kernel estimator is defined as in (2.13) and if one chooses the $\widetilde{G}(p)$ as a piece-wise linear function as in (2.8), then by differentiating (2.13), a smooth estimator of the quantile density function can be obtained as

$$\widehat{g}(p) = \int_0^1 \widetilde{g}(u)\frac{1}{b}k\left(\frac{p-u}{b}\right)du. \tag{2.19}$$

Parzen (1979) noted that the quantile density estimator which is defined in (2.19) provides good properties only for a fixed value of p. The non-parametric pre-flattened approach overcomes this issue by multiplying $g(p)$ by a factor $f_0(G_0(p))$ and, by then letting $b(p) = f_0(G_0(p))g(p)$, the kernel density estimator is defined as

$$\widehat{b}(p) = \int_0^1 \widetilde{b}(u)\frac{1}{b}k\left(\frac{p-u}{b}\right)bu \tag{2.20}$$

Bofinger (1975) also described the estimation of quantile density functions in the form of (2.19) based on the sample quantile function under nonparametric density estimation. If $F(x)$ is a strictly increasing, unknown and continuous distribution function in the neighbourhood of $x = a$ and $f = F'$ is the density function, then to estimate $f(a)$ at a particular quantile, Moore & Yackel (1977) investigated the properties of nearest neighbour density function estimators which are in the form of (2.19) and stated that the nearest neighbour estimators and the bandwidth estimators with same kernel k have the same consistency properties. In addition, Moore & Yackel (1977) noted that the nearest neighbour estimators are extensively used, and many researchers prefer them to bandwidth estimators. Jones (1992) compared the two methods, the first derivative of the kernel quantile estimator and the reciprocal of the kernel density estimator to estimate the quantile density and concluded that the first method is preferable than to the second method.

Falk (1997), Welsh (1988) and Jones (1992) studied the asymptotic properties of the quantile density estimator $\widehat{g}(p)$ which is given as, for bandwidth b,

$$\widehat{g}(p) = \sum_{i=1}^{n} X_{(i)} \left\{ k_b \left(p - \frac{i-1}{n} \right) - k_b \left(p - \frac{i}{n} \right) \right\} , \tag{2.21}$$

where $k_b(.)$ is an even function on [-1, 1] with variance $\sigma_k^2 = \int x^2 k(x)dx$ and roughness $K = \int k^2(y)dy$. It is this estimate that we favor in our work to follow.

Epanechnikov (1969) examined some properties of multivariate probability density with an arbitrary kernel as an improvement of univariate (Rosenblatt, 1956; Parzen, 1962) and bivariate (Maniya, 1961; Nadaraya, 1964) kernel density estimations. Researchers often use the Epanechnikov (1969) kernel as the kernel function in kernel density estimation. Therefore, we also used the kernel density estimator in (2.21) with Epanechnikov (1969)'s kernel in estimation.

Parzen (1979) noted that the main drawback of the kernel density estimation approach is the complication of optimally choosing bandwidth b. To avoid this limitation, as an alternative approach Parzen (1979) suggested an autoregressive quantile density estimator which uses a sequence of autoregressive densities of order l, where $l = 1, 2, \ldots$ to estimate the true density. The critical thing in this approach is to decide a suitable value for l to estimate the optimal density. A graphical approach is currently the most reasonable approach to determine the value of l. Parzen (1979) stated that it is easy to estimate the autoregressive quantile density for different values of order l compared to the estimation of kernel density for different values of bandwidth b.

Jones (1992) compared the two methods, the first derivative of the kernel quantile estimator and the reciprocal of the kernel density estimator which used to estimate the quantile density and concluded that the first method is preferred.

Falk (1986) and Welsh (1988) suggested estimating $g(p)$ and the second derivative of quantile density, $g''(p)$, separately and use the ratio of $g(p)/g''(p)$ to find the optimal bandwidth by minimizing the asymptotic mean square error of the kernel estimator. The b determines the connection between the amount of smoothness and how close the estimation is to the true distribution. Jones (1992) also derived the asymptotic mean square error of the kernel density estimator and found that the optimal choice of b to minimize the mean square error to estimate the quantile density depends on $g(p)/g''(p)$. Cheng *et al.* (2006) found that the kernel smoothed quantile estimators are more efficient than the empirical quantile estimators, especially for small samples and have been used for the ratio of $g(p)/g''(p)$ to calculate the mean squared error of the quantile estimators.

Quantile optimality ratio

Prendergast & Staudte (2016a) introduced the quantile optimality ratio (QOR) approach to choose an optimal bandwidth for a kernel density estimator that is needed to estimate the quantile densities. We used the QOR approach to find the optimal bandwidth to estimate the quantile density for the purpose of computing the asymptotic standard errors of the estimators in our interval estimators in what follows. The QOR is defined as,

$$QOR = \frac{g(p)}{g''(p)} \tag{2.22}$$

and it is location and scale invariant. The quantile density can be estimated using the kernel density estimator and the kernel density estimator can be written as a linear combination of order statistics as in (2.21). Jones (1992) derived the asymptotic mean square error of the quantile density in (2.21) as,

$$MSE[\widehat{g}(p)] = \frac{b^4\sigma_k^4[g''(p)^2]}{4} + \frac{Kg^2(p)}{bn} \tag{2.23}$$

where K and σ_k are defined following (2.21). Prendergast & Staudte (2016a) showed that the minimum $MSE[\widehat{g}(p)]$ occurs when

$$b = a(p)n^{(-1/5)} \tag{2.24}$$

where

$$a(p) = \left\{ \frac{K}{\sigma_k^4} \left[\frac{g(p)}{g''(p)} \right]^2 \right\}^{1/5}. \tag{2.25}$$

Therefore, considering (2.24) and (2.25), it can be clearly seen that the asymptotically optimal choice of bandwidth b to estimate the quantile density only depends on the underlying distribution through the QOR. The best choice of bandwidth b in (2.21) depends on the p and the underlying distribution. Since the performance of the interval estimators depends on the choice of b, we use the QOR to select our b although other choices of b are also possible.

Su (2009) proposed two methods called "Normal-GLD" and "Analytical maximum likelihood-GLD" to find the interval estimators of quantiles based on the generalised lambda distribution (GLD). Alternatively, Prendergast & Staudte (2016a) proposed "GLD-QOR" and "LN-QOR (Lognormal QOR)" to select good values for the bandwidth to find the interval estimators of quantiles. Furthermore, Prendergast & Staudte (2016a) conducted comparisons between their proposed methods and the methods in Su (2009) and concluded that both "GLD-QOR" and "LN-QOR" performed well. Considering the efficiency in simulation time, we select "LN-QOR" to construct the interval estimators for our newly introduced functions of quantiles.

2.2 Influence function

Hampel (1968, 1974) introduced the influence function (IF) which was originally called the "influence curve (IC)". It is an important tool which is used to describe and measure the robustness of an estimator. Later, it was called the "influence function " because of its generalisation to higher dimensional spaces. The IF measures the infinitesimal robustness of an estimator and it provides a measurable understanding of how the estimator responds to a small proportion (ε) of contamination x_0. The IF provides information about on the rate of change of an estimator with respect to the proportion of observations which are not from distribution F. Therefore, the IF can be expressed as an approximation to the relative change of an estimator due to the addition of a small proportion of contamination at any point x_0. Campbell (1978) viewed the influence function from another angle and considered it to be a random variable. Campbell (1978) justified this by saying the influence function is a mathematical transformation of a random variable X so that then it has a probability distribution.

2.2.1 Theoretical influence function

Throughout, let F denote a distribution function. Then, define the contamination distribution, a mixture distribution, as

$$F_{\varepsilon} = (1-\varepsilon)F + \varepsilon\Delta_{x_0} \tag{2.26}$$

where F is the uncontaminated distribution, $\varepsilon \in [0,1]$ is the proportion of contamination and Δ_{x_0} has all of its mass at the contaminant x_0. The mixture distribution F_{ε} is an appropriate model to represent contamination and it consists of observations from both F and point x_0 with high probability $1-\varepsilon$ and small probability ε, respectively. Therefore, sampling from F_{ε} runs the risk of getting a bad observation at x_0 with small probability ε and a good observation from distribution F with probability $1-\varepsilon$.

Suppose that for F, there is a parameter of interest, θ (where $\theta \in \Theta$ and Θ is an open convex subset in real line $\mathfrak{R}$), and associated estimator with the statistical functional $\mathcal{T}$ such that $\mathcal{T}(F) = \theta$ and $\mathcal{T}(F_n) = \widehat{\theta}$. Here F_n denotes the empirical distribution function, assuming a sample of size n from F, which gives $1/n$ probability at each of the n data points. Hence, $\mathcal{T}(F_n)$ is the estimate of θ and $\mathcal{T}(F_n)$ weakly converges to $\mathcal{T}(F)$ when n goes to ∞. The relative influence on estimator functional $\mathcal{T}(F)$ of ε proportion of contaminated observations at x_0 is given by,

$$\frac{\mathcal{T}(F_{\varepsilon}) - \mathcal{T}(F)}{\varepsilon} \ ,$$

where F is the uncontaminated distribution, $\varepsilon \in [0,1]$ is the proportion of contamination and Δ_{x_0} has all of its mass at the contaminant x_0.

Hampel (1974) defined the influence function as,

$$\mathrm{IF}(x_0; \mathcal{T}, F) \equiv \lim_{\varepsilon \to 0} \frac{\mathcal{T}(F_{\varepsilon}) - \mathcal{T}(F)}{\varepsilon} \tag{2.27}$$

which provides the rate of change in estimator functional $\mathcal{T}$, at F, caused by introducing an infinitesimal amount of contamination at x_0.

Another definition of the influence function is that it is the first derivative of the estimator functional at F_{ε} when $\varepsilon = 0$ (e.g. p48 Clarke, 2018). That is

$$\mathrm{IF}(x_0; \mathcal{T}, F) \equiv \frac{\partial}{\partial \varepsilon} \mathcal{T}(F_{\varepsilon}) \bigg|_{\varepsilon=0} \tag{2.28}$$

Therefore, the influence function is a directional derivative of $\mathcal{T}$ at F in the direction of Δ_{x_0} (Hampel *et al.*, 1986). As an example, here we illustrate how to obtain the influence function for the mean and variance estimators. Let $\mathcal{M}$ denote the functional for the mean parameter μ, we have $\mu = \mathcal{M}(F) = \int x dF$, where $dF = f(x)dx$. Then the mean estimator of the mixture distribution F_ε can be written as,

$$\begin{aligned}
\mathcal{M}(F_\varepsilon) &= \int x dF_\varepsilon \\
&= \int x d[(1-\varepsilon)F + \varepsilon\Delta_{x_0}] \\
&= (1-\varepsilon)\int x dF + \varepsilon \int x\Delta_{x_0} \\
&= (1-\varepsilon)\mu + \varepsilon x \\
&= \mu + \varepsilon(x_0 - \mu).
\end{aligned} \tag{2.29}$$

Then, using the definition of the influence function in (2.27), the influence function of the mean estimator can be obtained as

$$\begin{aligned}
\mathrm{IF}(x_0; \mathcal{M}, F) &= \lim_{\varepsilon \to 0} \frac{[\mu + \varepsilon(x_0 - \mu)] - \mu}{\varepsilon} \\
&= x_0 - \mu.
\end{aligned} \tag{2.30}$$

Note that the IF for the mean is unbounded in that influence grows the further x is moved away from μ. Hence, this supports the common notion that outliers are highly influential on the sample mean.

Similarly, let $\mathcal{V}$ denote the functional for the variance parameter σ^2, we have $\sigma^2 = \mathcal{V}(F) = \int (x-\mu)^2 dF$. Then the variance of the mixture distribution F_ε can be written as

$$\begin{aligned}
\mathcal{V}(F_\varepsilon) &= \int [x - \mathcal{M}(F_\varepsilon)]^2 dF_\varepsilon \\
&= \int \{x - [\mu + \varepsilon(x_0 - \mu)]\}^2 d[(1-\varepsilon)F + \varepsilon\Delta_{x_0}] \\
&= \int [(x-\mu)^2 - 2\varepsilon(x_0-\mu)(x-\mu) + \varepsilon^2(x_0-\mu)^2][(1-\varepsilon)dF + \varepsilon d\Delta_{x_0}] \\
&= (1-\varepsilon)\int (x-\mu)^2 dF - 2\varepsilon(1-\varepsilon)(x_0-\mu)\int (x-\mu)dF + \varepsilon^2(1-\varepsilon)(x_0-\mu)^2 \int dF \\
&\quad + \varepsilon \int (x-\mu)^2 d\Delta_{x_0} - 2\varepsilon^2(x_0-\mu)\int (x-\mu)d\Delta_{x_0} + \varepsilon^3(x-\mu)^2 \int d\Delta_{x_0} \\
&= (1-\varepsilon)\sigma^2 + \varepsilon^2(1-\varepsilon)(x_0-\mu)^2 + \varepsilon(x_0-\mu)^2 - 2\varepsilon^2(x_0-\mu)^2 + \varepsilon^3(x_0-\mu)^2 \\
&= \sigma^2 + \varepsilon\left[(x_0-\mu)^2 - \sigma^2\right] - \varepsilon^2(x_0-\mu)^2.
\end{aligned} \tag{2.31}$$

Then the influence function for the variance estimator is,

$$\begin{aligned}\mathrm{IF}(x_0;\mathcal{V},F) &= \lim_{\varepsilon\to 0}\frac{[(1-\varepsilon)\sigma^2+\varepsilon(x_0-\mu)^2-\varepsilon^2(x_0-\mu)^2]-\sigma^2}{\varepsilon}\\ &= \lim_{\varepsilon\to 0}\frac{\varepsilon[(x_0-\mu)^2-\sigma^2]-\varepsilon^2(x_0-\mu)^2]}{\varepsilon}\\ &= (x_0-\mu)^2-\sigma^2. \end{aligned} \tag{2.32}$$

Considering (2.31) and (2.32) the $\mathcal{V}(F_\varepsilon)$ can be approximated as,

$$\mathcal{V}(F_\varepsilon) \approx \sigma^2 + \varepsilon\mathrm{IF}(x_0;\mathcal{V},F) \tag{2.33}$$

since the result is that in (2.31) but ignoring the $o(\varepsilon^2)$ term. This is another way of thinking about influence functions and power series expansions of $\mathcal{T}(F_\varepsilon)$. That is, in general,

$$\mathcal{T}(F_\varepsilon) = \mathcal{T}(F) + \varepsilon\mathrm{IF}(x_0;\mathcal{T},F) + o(\varepsilon^2)$$

so that a large $\mathrm{IF}(x_0;\mathcal{T},F)$ means that $\mathcal{T}(F_\varepsilon)$ will be very different from $\mathcal{T}(F)$.

Hampel (1974) and Hampel *et al.* (1986) introduced the influence function for Fisher-consistent (Fisher, 1922) functionals. Let $X_1, X_2, \ldots.., X_n$ be a random sample with probability distribution function F, then we say that any estimator $\widehat{\theta}_n$ is a Fisher consistent estimator of θ, if $\widehat{\theta}_n = \mathcal{T}[F_n]$ and it satisfies $\theta = \mathcal{T}[F]$, where $\mathcal{T}$ is the estimator functional (e.g. pg.12 Jurečková *et al.*, 2019). Rousseeuw & Ronchetti (1981) extended the Hampel (1974); Hampel *et al.* (1986)'s influence function for non-Fisher-consistent functionals with the purpose of investigating the infinitesimal robustness of more general statistics.

2.2.2 Partial influence functions

Campbell (1978), Hampel *et al.* (1986) and Critchley & Vitiello (1991) explained how to extend the theoretical influence function for more than one population. Of these, Campbell (1978) showed how to expand the theoretical influence function for a number k of populations in multivariate between-group studies. Hampel *et al.* (1986) considered the partial influence function when defining the influence function for two sample tests. However, Pires & Branco (2002) gave a proper definition for influence functions when there are two populations, introducing the name "partial influence function" and noted that this can be extended to more than two populations. When there is more than one population, the influence function is determined by contaminating each of the populations, one at a time, while the other population remains uncontaminated. The partial influence functions of the estimator functional $\mathcal{T}$ at (F_1, F_2) with relation to F_1 and F_2, respectively, are given by

$$\begin{aligned}
\mathrm{PIF}_1(x_0; \mathcal{T}, F_1, F_2) &= \lim_{\varepsilon \to 0} \left[\frac{\mathcal{T}[(1-\varepsilon)F_1 + \varepsilon \Delta_{x_0}, F_2] - \mathcal{T}(F_1, F_2)}{\varepsilon} \right] \\
&= \lim_{\varepsilon \to 0} \left[\frac{\mathcal{T}(F_{1,\varepsilon}, F_2) - \mathcal{T}(F_1, F_2)}{\varepsilon} \right] \\
\mathrm{PIF}_2(x_0; \mathcal{T}, F_1, F_2) &= \lim_{\varepsilon \to 0} \left[\frac{\mathcal{T}[F_1, (1-\varepsilon)F_2 + \varepsilon \Delta_{x_0}] - \mathcal{T}(F_1, F_2)}{\varepsilon} \right] \\
&= \lim_{\varepsilon \to 0} \left[\frac{\mathcal{T}(F_1, F_{2,\varepsilon}) - \mathcal{T}(F_1, F_2)}{\varepsilon} \right]. \qquad (2.34)
\end{aligned}$$

Campbell (1978) explained how the sample versions of the influence function works when there is more than one population. In the sample version of the partial influence function, eliminate an observation from only one of the samples. According to Pires & Branco (2002) the interpretation of the partial influence function, e.g. $\mathrm{PIF}_1(x_0; \mathcal{T}, F_1, F_2)$, is that it measures the n_1 times the change on the relevant component of $\mathcal{T}$ that is caused by an additional observation in x, in the first sample, when $\mathcal{T}$ is applied to a combined sample of size $(n_1 + n_2)$ and similarly can be interpreted for $\mathrm{PIF}_2(x_0; \mathcal{T}, F_1, F_2)$. As an example, here we illustrate how to obtain the partial influence functions for the ratio of independent variance estimators in detail (a shortened version of this can be found in Equation 8, Section 3.2 of Publication I).

Recall $\mathcal{V}(F_\varepsilon) \approx \sigma^2 + \varepsilon \mathrm{IF}(x_0; \mathcal{V}, F)$ and $\mathrm{IF}(x_0; \mathcal{V}, F) = (x_0 - \mu)^2 - \sigma^2$ in (2.33) and (2.32) respectively. Let $\mathcal{M}(F_j) = \mu_j$, $\mathcal{V}(F_j) = \sigma_j^2$ $(j = 1, 2)$ and $\mathcal{R}$ is the functional for the ratio of variances where $\mathcal{R}(F_1, F_2) = \mathcal{V}(F_1, F_2) = \mathcal{V}(F_1)/\mathcal{V}(F_2) = \sigma_1^2/\sigma_2^2 = \rho$. Let, $z_j = (x_0 - \mu_j)/\sigma_j$ $(j = 1, 2)$ and then we can derive the PIF of the ratio of variances as,

$$
\begin{aligned}
\mathrm{PIF}_1(x_0;\mathcal{R},F_1,F_2) &= \lim_{\varepsilon\to 0}\left[\frac{\mathcal{V}(F_{1,\varepsilon},F_2)-\mathcal{V}(F_1,F_2)}{\varepsilon}\right] \\
&= \lim_{\varepsilon\to 0}\left[\frac{\mathcal{V}(F_{1,\varepsilon})/\mathcal{V}(F_2)-\mathcal{V}(F_1)/\mathcal{V}(F_2)}{\varepsilon}\right] \\
&= \lim_{\varepsilon\to 0}\left[\frac{[\sigma_1^2+\varepsilon \mathrm{IF}(x_0;\mathcal{V},F_1)]/\sigma_2^2-\sigma_1^2/\sigma_2^2}{\varepsilon}\right] \\
&= \frac{\mathrm{IF}(x_0;\mathcal{V},F_1)}{\sigma_2^2} \\
&= \frac{(x_0-\mu_1)^2-\sigma_1^2}{\sigma_2^2} \\
&= \frac{(x_0-\mu_1)^2}{\sigma_2^2}-\frac{\sigma_1^2}{\sigma_2^2} \\
&= \frac{\sigma_1^2}{\sigma_2^2}\left[\frac{(x_0-\mu_1)^2}{\sigma_1^2}-1\right] \\
&= \rho[z_1^2-1].
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\mathrm{PIF}_2(x_0;\mathcal{R},F_1,F_2) &= \lim_{\varepsilon\to 0}\left[\frac{\mathcal{V}(F_1,F_{2,\varepsilon})-\mathcal{V}(F_1,F_2)}{\varepsilon}\right] \\
&= \lim_{\varepsilon\to 0}\left[\frac{\mathcal{V}(F_1)/\mathcal{V}(F_{2,\varepsilon})-\mathcal{V}(F_1)/\mathcal{V}(F_2)}{\varepsilon}\right] \\
&= \lim_{\varepsilon\to 0}\left[\frac{\sigma_1^2/[\sigma_2^2+\varepsilon \mathrm{IF}(x_0;\mathcal{V},F_2)]-\sigma_1^2/\sigma_2^2}{\varepsilon}\right] \\
&= \lim_{\varepsilon\to 0}\left\{\frac{\sigma_1^2}{\sigma_2^2}\left[\frac{\sigma_2^2-[\sigma_2^2+\varepsilon \mathrm{IF}(x_0;\mathcal{V},F_2)]}{\varepsilon[\sigma_2^2+\varepsilon \mathrm{IF}(x_0;\mathcal{V},F_2)]}\right]\right\} \\
&= \lim_{\varepsilon\to 0}\left\{-\frac{\sigma_1^2}{\sigma_2^2}\left[\frac{\mathrm{IF}(x_0;\mathcal{V},F_2)}{\sigma_2^2+\varepsilon \mathrm{IF}(x_0;\mathcal{V},F_2)}\right]\right\} \\
&= -\rho\left[\frac{\mathrm{IF}(x_0;\mathcal{V},F_2)}{\sigma_2^2}\right] \\
&= -\rho\left[\frac{(x_0-\mu_2)^2-\sigma_2^2}{\sigma_2^2}\right] \\
&= -\rho\left[\frac{(x_0-\mu_2)^2}{\sigma_2^2}-1\right] \\
&= -\rho[z_2^2-1]
\end{aligned}
$$

Rousseeuw & Ronchetti (1981) extended the influence function to more than one population for non-Fisher consistent functionals.

2.2.3 Influence function for quantile estimators

Let $x_p = F^{-1}(p)$ for $p \in [0,1]$ be the pth quantile. Assuming F has a continuous and positive density f at x_p, the influence function of the x_p estimator is defined as, e.g. see page 59 of Staudte & Sheather (1990) and page 21 of Rieder (2012),

$$\mathrm{IF}(x_0, x_p, F) = \{p - I[x_p \geq x_0]\}g(p) \tag{2.35}$$

where, $g(p)$ is the quantile density as in (2.17).

Staudte & Sheather (1990) explained how to derive the influence function of the pth quantile. Let the mixture distribution be F_ε as in (2.26) and $g(\varepsilon) = F_\varepsilon^{-1}(p)$ defined in (2.36). Since

$$F_\varepsilon^{-1}(p) = \begin{cases} F^{-1}\left(\frac{p}{1-\varepsilon}\right), & p \leq (1-\varepsilon)F \\ x_0, & (1-\varepsilon)F \leq p < (1-\varepsilon)F + \varepsilon \\ F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right), & (1-\varepsilon)F + \varepsilon \leq p \end{cases}$$

(Staudte & Sheather, 1990, p.56). Let, $(1-\varepsilon)F + \varepsilon \leq p$, then

$$g(\varepsilon) = F_\varepsilon^{-1}(p) = F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right). \tag{2.36}$$

Hence

$$g'(\varepsilon) = \frac{d}{d\varepsilon}F_\varepsilon^{-1}(p) = \frac{d}{d\varepsilon}F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right) = \frac{\left(\frac{d}{d\varepsilon}\right)\left(\frac{p-\varepsilon}{1-\varepsilon}\right)}{f\left[F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right)\right]}.$$

For simplicity, let $\mathrm{IF}(x_0, x_p, F) = \mathrm{IF}_p(x)$ and using the definition of IF in (2.28),

$$\mathrm{IF}_p(x) = \frac{d}{d\varepsilon}g(\varepsilon)\bigg|_{\varepsilon=0} = \frac{d}{d\varepsilon}F^{-1}\left(\frac{p-\varepsilon}{1-\varepsilon}\right)\bigg|_{\varepsilon=0} = g'(\varepsilon)\big|_{\varepsilon=0}$$

$$\text{Therefore,}\quad \mathrm{IF}_p(x) = g'(\varepsilon)\big|_{\varepsilon=0} = \frac{p-1}{f(x_p)} = (p-1)g(p), \quad x < x_p$$

Similarly, we can prove the rest and finally,

$$\mathrm{IF}_p(x) = \begin{cases} \frac{p-1}{f(x_p)} = (p-1)g(p) & x < x_p \\ 0 & x = x_p \\ \frac{p}{f(x_p)} = pg(p) & x > x_p \end{cases} \tag{2.37}$$

which is the expression also found on page 53 of Clarke (2018).

Therefore, the contamination at x_p has zero influence on the pth quantile of F. The contamination at any $x > x_p$ has a fixed positive influence on the pth quantile of F, since for any proportion of contamination ε the contamination shifts the pth quantile to the right by $F^{-1}(p/(1-\varepsilon))$ without considering the value of x. Similarly, the contamination at any $x < x_p$ has the same negative influence on the pth quantile of F, since the contamination shifts the pth quantile to the left by $F^{-1}(p-\varepsilon/(1-\varepsilon))$ without considering the value of x.

As an example, we illustrate how to derive the influence function of the interquantile range (IQR_p) (see for example page 111 of Huber, 1981). Let $\mathrm{IQR}_p = x_{1-p} - x_p$,

$$\begin{aligned} \mathrm{IF}(x_0, \mathrm{IQR}_p, F) &= \mathrm{IF}(x_{1-p} - x_p) \\ &= \mathrm{IF}(x_{1-p}) - \mathrm{IF}(x_p) \end{aligned}$$

Then recall the $\mathrm{IF}(x_p)$ which is in (2.35) and (2.37) and also recall that $p \in (0, 0.5)$. Hence, if $x < x_p$ then $x < x_{1-p}$. Therefore,

$$\begin{aligned} \mathrm{IF}(x_0, \mathrm{IQR}_p, F) &= \frac{-p}{f(x_{1-p})} - \frac{p-1}{f(x_p)} \\ &= -pg(1-p) - (p-1)g(p) \end{aligned}$$

When, $x \in [x_p, x_{1-p}]$ then $x \geq x_p$ and $x \leq x_{1-p}$. Therefore,

$$\begin{aligned} \mathrm{IF}(x_0, \mathrm{IQR}_p, F) &= \frac{-p}{f(x_{1-p})} - \frac{p}{f(x_p)} \\ &= -pg(1-p) - pg(p) \end{aligned}$$

When, $x \geq x_{1-p}$ then $x \geq x_p$. Therefore,

$$\begin{aligned}\mathrm{IF}(x_0, \mathrm{IQR}_p, F) &= \frac{p-1}{f(x_{1-p})} - \frac{p}{f(x_p)} \\ &= (p-1)g(1-p) - pg(p)\end{aligned}$$

Finally, we can summarize this as,

$$\mathrm{IF}(x_0, \mathrm{IQR}_p, F) = \begin{cases} \frac{-p}{f(x_{1-p})} - \frac{p-1}{f(x_p)} & x < x_p \\ \frac{-p}{f(x_{1-p})} - \frac{p}{f(x_p)} & x \in [x_p, x_{1-p}] \\ \frac{p-1}{f(x_{1-p})} - \frac{p}{f(x_p)} & x > x_{1-p} \end{cases} .$$

2.3 Asymptotic variances and standard errors

Asymptotic variances

Influence functions also exhibit useful asymptotic properties including an often-convenient means to derive the asymptotic variances of the estimator. Asymptotic variance (ASV) is one of the summary values of the influence function and was earlier called the "expected square" of the influence curve.

As examples, Hampel *et al.* (1986) and Staudte & Sheather (1990) described the relationship between the influence function and the asymptotic variance and how to derive the asymptotic variance using the influence function. We review this material below.

If some distribution G is in a neighbourhood of F then the influence function appears in the first order von Mises expansion of $\mathcal{T}(G)$ which is given by (e.g. p. 53 of Clarke, 2018),

$$\mathcal{T}(G) = \mathcal{T}(F) + \int IF(x_0, \mathcal{T}, F) d(G - F)(x) + R ,$$

where R is the remainder.

For $X \sim F$ and F_n denoting the empirical distribution for n iid random variables distributed F, under some mild regularity conditions such as differentiability of $\mathcal{T}(F)$ and by the Central Limit Theorem we have the following ((see, e.g., page 85 of Hampel *et al.*, 1986) and (see, e.g., page 63 of Staudte & Sheather, 1990)),

For $X \sim F$ and F_n denoting the empirical distribution for n iid random variables distributed F and if $G = F_n$ in (2.3), then for large enough n the expansion of $\mathcal{T}(F_n)$ for F_n in a neighbourhood of F can be written as,

$$\mathcal{T}(F_n) = \mathcal{T}(F) + \int IF(x_0, \mathcal{T}, F)d(F_n - F)(x) + R_n \ .$$

When $n \to \infty$, $R_n \to 0$ in probability and by evaluating the integral over F_n, it resulted in the approximation,

$$\sqrt{n}[\mathcal{T}(F_n) - \mathcal{T}(F)] \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(X_i; \mathcal{T}, F) \ .$$

By the central limit theorem, the right-hand side of (2.3) is asymptotically normal. Therefore,

$$\sqrt{n}[\mathcal{T}(F_n) - \mathcal{T}(F)] \overset{a}{\sim} N(0, \mathrm{ASV}(\mathcal{T}))$$

where $\overset{a}{\sim}$ denotes 'approximately distributed as 'and

$$\mathrm{ASV}[\mathcal{T}(F)] = E\left[\mathrm{IF}(X; \mathcal{T}, F)^2\right] \tag{2.38}$$

is the asymptotic variance of the estimator with functional $\mathcal{T}$. Further details and discussion can be found in Ch. 2 and 3 of Clarke (2018).

As an example, when the $X_1, ..., X_n$ is a simple random sample the influence function for the sample mean where $\mu = \mathcal{M}(F) = \int x dF$ is $\mathrm{IF}(x_0; \mathcal{M}, F) = x_0 - \mu$ as in (2.30). Then (2.3) becomes

$$\overline{X}_n = \mu + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(X_i; \mathcal{M}, F) + R_n \ ,$$

where $R_n = 0$ and

$$\sqrt{n}\left[\overline{X}_n - \mu\right] \overset{a}{\sim} N(0, \mathrm{ASV}(\mathcal{M})) \ ,$$

where $E_F[\mathrm{IF}(X; \mathcal{M}, F)] = 0$, $\mathrm{ASV}[\mathcal{M}(F)] = E\left[\mathrm{IF}(X; \mathcal{M}, F)^2\right] = E_F[(X - \mu)^2] = \sigma^2$ for $E_F(.)$ denoting expectation when $X \sim F$.

When there are two populations, we have to consider the partial influence functions in (2.34) and let F_{n_1} and F_{n_2} denote the empirical distribution functions for iid samples of size n_1 and n_2 from F_1 and F_2 then from (Hampel *et al.*, 1986, pg.196) and Pires & Branco (2002) we have that $\sqrt{n_1 + n_2}[\mathcal{T}(F_{n_1}, F_{n_2}) - \mathcal{T}(F_1, F_2)]$ is asymptotically normal with mean zero and asymptotic variance

$$\mathrm{ASV}[\mathcal{T}(F_1, F_2)] = \frac{1}{\mathrm{w}_1} E_{F_1}[\mathrm{PIF}_1(X; \mathcal{T}, F_1, F_2)^2] + \frac{1}{\mathrm{w}_2} E_{F_2}[\mathrm{PIF}_2(X; \mathcal{T}, F_1, F_2)^2] \tag{2.39}$$

where $\mathrm{w}_i = n_i/(n_1 + n_2)$ $(i = 1, 2)$.

Asymptotic standard deviations

The asymptotic standard deviation of an estimator with functional $\mathcal{T}$, can be found by

$$\text{ASD}[\mathcal{T}(F)] \equiv \sqrt{\text{ASV}[\mathcal{T}(F)]}\ , \tag{2.40}$$

where ASV$[\mathcal{T}(F)]$ is defined as in (2.38).

When there are two populations,

$$\text{ASD}[\mathcal{T}(F_1,F_2)] \equiv \sqrt{\text{ASV}[\mathcal{T}(F_1,F_2)]}\ , \tag{2.41}$$

where ASV$[\mathcal{T}(F_1,F_2)]$ is defined as in (2.39).

2.3.1 Asymptotic Variance based on Delta method

The delta method gives an approximate variance for a nonlinear function of a random variable using Taylor series expansion.

If X_n is a sequence of random variable that satisfies (Verrill, 2003)

$$\sqrt{n}(X_n - \mu) \to N(0,\sigma^2)\ , \tag{2.42}$$

where, μ is mean and then,

$$\sqrt{n}(f(X_n) - f(\mu)) \to N(0, f'(\mu)^2\sigma^2) \tag{2.43}$$

Where, f' is the first derivative of the density function f evaluated at μ (Agresti, 1990).

Therefore, if the $Var(X_n) = \sigma^2$, $Var(f(X_n)) = f'(\mu)^2\sigma^2$.

Let the $X^{(1)},\ldots,X^{(n)}$ be a random sample with $EX_i^{(j)} = \mu_i$ and $Cov(X_i^{(j)}, X_k^{(j)}) = \sigma_{ij}$. Then from multivariate delta method (Papanicolaou, 2009), for for $\mu = (\mu_1,\ldots,\mu_n)$ for which $\theta^2 = \Sigma_i\Sigma_j\sigma_{ij}f_i'(\mu)f_j'(\mu) > 0$,

$$\sqrt{n}(f(\hat{X}^{(1)},\ldots,\hat{X}^{(n)}) - f(\mu_1,\ldots,\mu_p)) \to N(0,\theta^2) \tag{2.44}$$

2.3.2 Asymptotic variance of quantiles

For the quantile estimator with functional x_p and influence function given in (2.35), it can be shown that $E_F[\mathrm{IF}(X;x_p,F)] = 0$ and

$$\mathrm{ASV}(x_p) = E_F[\mathrm{IF}^2(X;x_p,F)] = p(1-p)g^2(p). \tag{2.45}$$

For more details we suggest to refer (David, 1981, Ch.2), (Staudte & Sheather, 1990, page 64) and (DasGupta, 2006a, Ch.7).

Let F be a continuous and positive density f at $\mathcal{G}(p) = x_p = F^{-1}(p)$, then Bahadur (1966) has shown that

$$\sqrt{n}[X_{(np)} - x_p] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}(X_i;x_p,F) + \sqrt{n}R_n$$

where $\mathrm{IF}(X_i;x_p,F)$ is the influence function of the x_p estimator given in (2.35) and (2.37), $R_n = O(n^{-\frac{3}{4}}\log n)$ with probability one and R_n become negligible when $n \to \infty$. Then, using (2.3), the (2.46) can be approximated to,

$$\sqrt{n}[X_{np} - x_p] \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{IF}(X_i;x_p,F). \tag{2.46}$$

From the central limit theorem,

$$\sqrt{n}[X_{np} - x_p] \overset{a}{\sim} N\left(0, \mathrm{ASV}(\mathcal{G})\right) \ ,$$

where $E_F[\mathrm{IF}(X;\mathcal{G},F)] = 0$ and (e.g see pages 63,64 Staudte & Sheather, 1990)

$$\mathrm{ASV}[\mathcal{G}(F)] = \mathrm{ASV}_F[x_p] = E_F[\mathrm{IF}^2(X;x_p,F)] = \frac{p(1-p)}{f^2(x_p)}.$$

Since the quantile density $g(p) = 1/f(x_p)$ as in (2.17), then

$$\mathrm{ASV}_F[x_p] = p(1-p)g^2(p) \ . \tag{2.47}$$

When the statistical functional is applied to the empirical distribution denoted F_n, then the asymptotic variance of $x_p(F_n)$ is then the asymptotic variance of $x_p(F_n)$ is same as in (2.47). For more details, we refer the reader to Bahadur (1966), Staudte & Sheather (1990) and Ch. 7 of DasGupta (2006a).

2.3.3 Asymptotic covariance of quantiles

Similarly, and as also found in the preceding references, the asymptotic covariance between the p-th and q-th quantile estimators is, provided $0 < p < q < 1$,

$$\mathrm{E}\left[\mathrm{IF}(X;\mathcal{G}(\cdot,p),F)\mathrm{IF}(X;\mathcal{G}(\cdot,q),F)\right] = p(1-q)g(p)g(q). \tag{2.48}$$

and when, $0 < q < p < 1$, we have

$$\mathrm{E}\left[\mathrm{IF}(X;\mathcal{G}(\cdot,p),F)\mathrm{IF}(X;\mathcal{G}(\cdot,q),F)\right] = q(1-p)g(p)g(q). \tag{2.49}$$

As an example, we derive the ASV of IQR_p (see for example page 112 of Huber, 1981). Recall the ASV as in (2.38), $\mathrm{ASV}_F[x_p]$ in (2.47) and the asymptotic covariance of quantiles in (2.48). Then,

$$\begin{aligned}
\mathrm{ASV}(\mathrm{IQR}_p) =& E_F[\mathrm{IF}^2(\mathrm{IQR})] \\
=& E_F[(\mathrm{IF}(x_{1-p}) - \mathrm{IF}(x_p))^2] \\
=& E_F[\mathrm{IF}^2(x_{1-p}) + \mathrm{IF}^2(x_p) - 2\mathrm{IF}(x_{1-p})\mathrm{IF}(x_p)] \\
=& E_F[\mathrm{IF}^2(x_{1-p})] + E_F[\mathrm{IF}^2(x_p)] - 2E_F[\mathrm{IF}(x_{1-p})\mathrm{IF}(x_p)] \\
=& p(1-p)g^2(1-p) + p(1-p)g^2(p) - 2p^2 g(p)g(1-p).
\end{aligned}$$

2.4 Estimation

2.4.1 Point estimation

Point estimation when there is a single population

Suppose that for F, there is a parameter of interest, θ (where $\theta \in \Theta$ and Θ is an open convex subset in real line $\mathfrak{R}$), and associated estimator with the weakly continuous (Pg. 41, 42 of Clarke, 2018) statistical functional $\mathcal{T}$ such that $\mathcal{T}(F) = \theta$ and $\mathcal{T}(F_n) = \widehat{\theta}$ then $\mathcal{T}(F_n)$ is the estimator of $\mathcal{T}(F)$, where F_n denotes the empirical distribution function which gives $1/n$ probability at each of the n data points and $\mathcal{T}(F_n)$ converges to $\mathcal{T}(F)$ when n goes to ∞.

Point estimation when there are two populations

Suppose that for F_1 and F_2, there is a parameter of interest, θ where $\theta \in \Theta$ and Θ is an open convex subset in real line $\Re$, and associated estimator with the weakly continuous (Pg. 41, 42 of Clarke, 2018) statistical functional $\mathcal{T}$ such that $\mathcal{T}(F_1, F_2) = \theta$ and $\mathcal{T}(F_{n_1}, F_{n_2}) = \widehat{\theta}$ then $\mathcal{T}(F_{n_1}, F_{n_2})$ is the estimator of $\mathcal{T}(F_1, F_2)$, where F_{n_1} and F_{n_2} denote the empirical distribution functions which gives $1/(n_1 + n_2)$ probability at each of the data points in the combined sample. Then, $\mathcal{T}(F_{n_1}, F_{n_2})$ converges to $\mathcal{T}(F_1, F_2)$ when n_1 and n_2 go to ∞.

2.4.2 Interval estimation

Interval estimation when there is a single population

Let $z_\alpha = \Phi^{-1}(\alpha)$ denote the α quantile of the standard normal distribution. All our $100(1-\alpha)\%$ asymptotic confidence intervals for estimator functional $\mathcal{T}(F)$ will be of the form:

$$\mathcal{T}(F_n) \pm z_{1-\alpha/2}\, \widehat{\mathrm{ASE}}[\mathcal{T}(F_n)] \ , \tag{2.50}$$

where $\widehat{\mathrm{ASE}}[\mathcal{T}(F_n)] = \widehat{\mathrm{ASD}}[\mathcal{T}(F_n)]/\sqrt{n}$ is an estimate of its standard error based on the sample where $\widehat{\mathrm{ASD}}[\mathcal{T}(F_n)]$ is the estimated asymptotic standard deviation. The actual coverage probability of this estimator depends on how quickly the distribution of $\mathcal{T}(F_n)$ approaches normality, as well as the rate of convergence of $\mathcal{T}(F_n)$ to $\mathcal{T}(F)$ and $\widehat{\mathrm{ASE}}[\mathcal{T}(F)]$ to $\mathrm{ASE}[\mathcal{T}(F)]$.

In constructing the interval estimators for the ratios, due to the improved statistical performance such as quicker convergence to normality and symmetrising so that choice of numerator and denominator does not matter, it is common to first construct the interval for the log transformed ratio followed by exponentiation to return to the original ratio scale. Let $\mathcal{W}$ denotes an arbitrary statistical functional and $\mathcal{W}(F) = \ln[\mathcal{T}(F)]$ then, using the Delta Method (e.g. Ch.3 DasGupta, 2006b),

$$\mathrm{ASV}[\mathcal{W}(F)] \doteq \frac{1}{[\mathcal{T}(F)]^2}\, \mathrm{ASV}[\mathcal{T}(F)] \ . \tag{2.51}$$

Then $\widehat{\mathrm{ASE}}[\mathcal{W}(F_n)] = \{\widehat{\mathrm{ASV}}[\mathcal{W}(F_n)]\}^{1/2}/\sqrt{n}$. This enables one to construct the confidence interval for $\ln \mathcal{T}(F)$, which is based on the asymptotic normality of $\ln \mathcal{T}(F_n)$. A confidence interval for $\mathcal{T}(F)$ can then be found by exponentiating the lower and upper bounds.

Interval estimation when there are two populations

The $100(1-\alpha)\%$ confidence intervals for $\mathcal{T}(F_1, F_2)$ will be of the form:

$$\mathcal{T}(F_{n_1}, F_{n_2}) \pm z_{1-\alpha/2}\, \widehat{\mathrm{ASD}}(\mathcal{T}, F_{n_1}, F_{n_2})/\sqrt{n_1+n_2}\ ,$$

where $\mathcal{T}(F_{n_1}, F_{n_2})$ is the estimator of $\mathcal{T}(F_1, F_2)$ and $\widehat{\mathrm{ASD}}(\mathcal{T}, F_{n_1}, F_{n_2})/\sqrt{n_1+n_2}$ is an estimate of its standard deviation (standard error) based on two samples. The actual coverage probability of this estimator depends on how quickly the distribution of $\mathcal{T}(F_{n_1}, F_{n_2})$ approaches normality, as well as the rate of convergence of $\mathcal{T}(F_{n_1}, F_{n_2})$ to $\mathcal{T}(F_1, F_2)$ and $\widehat{\mathrm{ASD}}(\mathcal{T}, F_{n_1}, F_{n_2})$ to $\mathrm{ASD}(\mathcal{T}, F_1, F_2)$.

Let $\mathcal{W}(F_1, F_2) = \ln[\mathcal{T}(F_1, F_2)]$ then, using the Delta Method (Ch.3 DasGupta, 2006b),

$$\mathrm{ASV}[\mathcal{W}(F_1, F_2)] \doteq \frac{1}{[\mathcal{T}(F_1, F_2)]^2}\, \mathrm{ASV}[\mathcal{T}(F_1, F_2)]\ . \tag{2.52}$$

Then $\widehat{\mathrm{ASE}}[\mathcal{W}(F_{n_1}, F_{n_2})] = \{\widehat{\mathrm{ASV}}[\mathcal{W}(F_{n_1}, F_{n_2})]/(n_1+n_2)\}^{1/2}$ enables one to construct the confidence interval for $\mathcal{T}(F_1, F_2)$, which is based on the asymptotic normality of $\mathcal{T}(F_{n_1}, F_{n_2})$.

3. Additional work

In this chapter, we have included some additional works that have not been included in the publications. To avoid reintroducing notations and concepts etc., the reader should first read the corresponding papers that are included in full in the next part of this thesis.

3.1 Additional work related to Paper I

This is supplementary work to Paper I, and so the reader should first read Paper I in Part II.

3.1.1 Partial influence function comparison

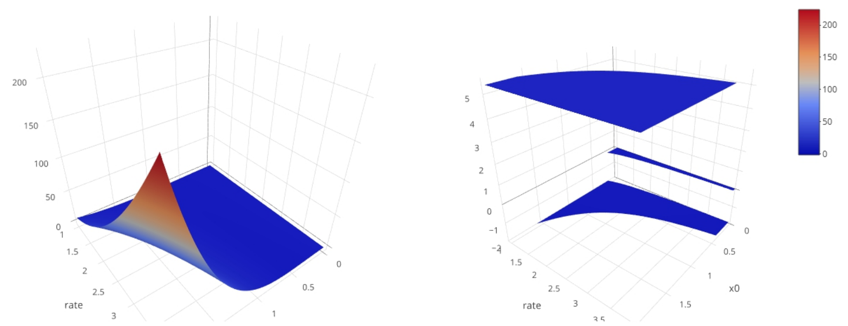


Figure 3.1: PIF_1 comparisons for the ratio of variances (left) and squared ratio of IQRs (right) for ratios for two exponential populations with rates equal to 1 and $p = 0.2$.

Figure 3.1 depicts the PIFs for the ratio of variances (left) and IQRs (right) for $p = 0.2$ for two exponential distributions with rate 1. The PIF for the variance ratio is unbounded and increasing with x_0 and rate. The PIF for the IQR ratio is bounded and in three segments depending on the location of x_0 in relation to $x_{0.2}$ and $x_{0.8}$. Consequently, the IQR ratio will be less influenced by outliers.

3.1.2 Asymptotic variance comparison

As examples, we have selected the $LN(0,1)$, $EXP(1)$ and $Uniform(2,5)$ distributions to compare the asymptotic variances of the ratio of variances and squared ratio of interquantile range estimators.

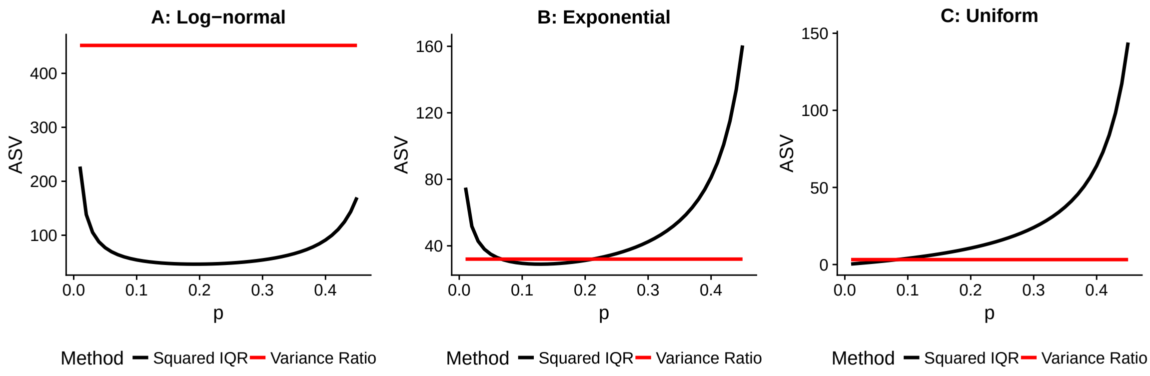


Figure 3.2: ASV comparisons for the $LN(0,1)$, $EXP(1)$ and $Uniform(2,5)$ distributions with assumed equal sample sizes (so $w_1 = w_2 = 1/2$). The distributions are chosen to be equal in each example so that the estimators of $\rho = \rho_p = 1$.

As shown in Figure 3.2, the ASV of the squared IQR ratio (black curve) can vary greatly with p. Here the plots are over the domain $p \in [0.01, 0.45]$ and for choices of p for the log-normal distribution we can see that the ASV is smaller than that for the ratio of variances (red line). For the exponential distribution, a choice of p of around 0.15 will result in a smaller ASV for the squared IQR although if p is either very small or moderately large then ASV for the ratio of variances is smaller. An interesting finding arises for the continuous uniform distribution. The ASV is minimized when p is chosen to be as small as possible. This implies that, in practice, we should choose to select the range (max $-$ min) as the interquantile range to decrease estimator variability.

3.1.3 Extra plot and table for house price data example

In Table 5 of Paper 1, since 30% of intervals gave different conclusions comparing the two methods (ratio of variance and ratio of IQR) of all the pairwise comparisons between suburbs, our next aim is to find the suburbs which show a higher number of different conclusions. Since there are 301 suburbs in the cleaned data set, there are 300 choose 2 or 45,150 unique pairwise comparisons. Table 3.1 shows the names of the suburbs which show a higher number of different conclusions between the two methods and Figure 3.3 depicts the house price distributions of these eleven suburbs. It can be clearly seen that the house price distribution of all the suburbs except Warranwood is positively skewed and some suburb's house price distributions contain outliers and in some cases very extreme values.

Table 3.1: Suburbs which show a higher number of different conclusions between the two methods in the pairwise comparisons

Suburb	No/Yes	Yes/No	Total
Beaumaris	227	7	234
Docklands	9	200	209
Edithvale	13	204	217
Frankston North	0	206	206
Glen Waverley	183	26	209
Hampton Park	10	205	215
Kingsbury	238	1	239
Langwarrin	7	232	239
Oakleigh	203	10	213
Tullamarine	220	10	230
Warranwood	2	211	213

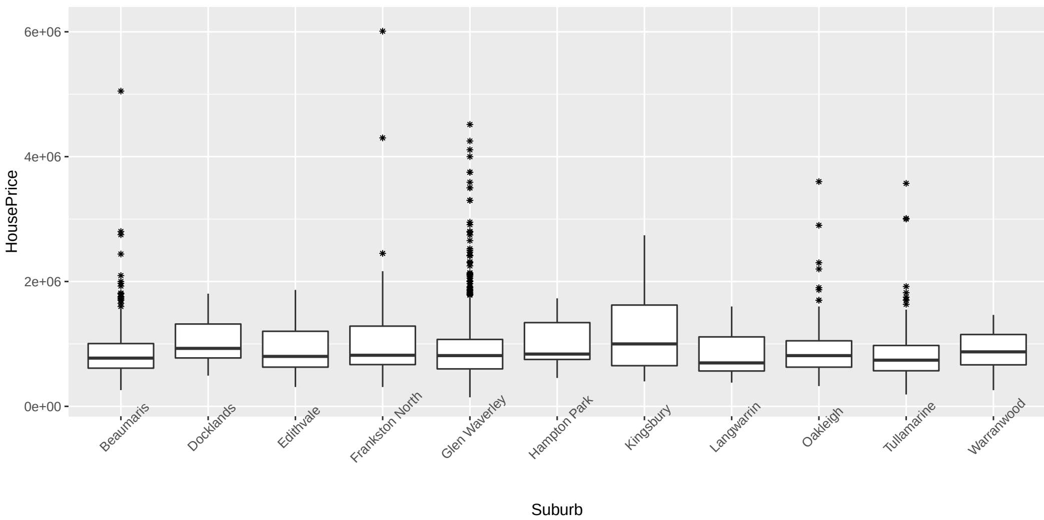


Figure 3.3: Box plots of house price distributions for suburbs which show a higher number of different conclusions between the two methods

3.1.4 Doctor visits data example

The doctor visits data is a sub-sample of 3066 individuals of the AHEAD cohort (born before 1924) for wave 6 (year 2002) from the Health and Retirement Study (HRS) which surveys more than 22,000 Americans over the age of 50 every 2 years. The response variable that we are interested in is the number of doctor visits. We grouped this data into two groups using gender as the grouping variable. The summary statistics of the response variable for the two gender groups can be found in Table 6 of Arachchige *et al.* (2019b).

According to the summary statistics, the doctor visit distributions are positively skewed and there is a large outlier in the female group. The summary statistics suggest a positive skew in the number of female visits to the doctor, even after removing the outliers. Our objective was to compare the variation in the number of doctor visits between males and females. We used the ratio of variance approach and the squared IQR ratio to compare the variation in the number of doctor visits between males and females with and without the outlier.

Table 3.2: 95 % confidence interval lower bounds (LB) and upper bounds (UB) for the doctor visits data.

Confidence Interval Method	x= male, y= female			
	With outlier		Without outlier	
	LB	UB	LB	UB
Ratio of Variance	0.2155	1.8582	0.5367	2.3671
Squared IQR ratio, $p = 0.01$	0.2489	2.5223	0.2814	2.8186
Squared IQR ratio, $p = 0.05$	0.4861	1.2268	0.5356	1.3759
Squared IQR ratio, $p = 0.1$	0.8789	1.1378	0.8799	1.1365
Squared IQR ratio, $p = 0.2$	0.4610	0.9454	0.4610	0.9454
Squared IQR ratio, $p = 0.25$	0.5983	1.1416	0.5985	1.1412

Table 3.2 provides the confidence interval bounds of the 95% confidence intervals using the two methods. It can be clearly seen that there is a large difference between the ratio of variance confidence intervals depending on whether the outlier is included. On the other hand, the confidence interval for the squared IQR ratio is hardly influenced by the outlier. The squared IQR ratio does not measure the same thing for different p. Additionally, the interval for the ratio of variances is wide compared to the interval for the squared IQR ratio with the exception of when $p = 0.01$ is chosen for the latter. This suggests that the IQR approach is a better choice to compare the variation between the two groups for this example.

3.2 Additional work related to paper II

This is supplementary work to Paper II, and so the reader should first read Paper II in Part II.

3.2.1 Confidence interval for median absolute deviation from a target

Bonett & Seier (2003) suggested constructing distribution-free confidence intervals for median absolute deviation from a target by applying the usual confidence interval for the median described on page no. 137 of Snedecor & Cochran (1980) to the transformed values $|Y_i - h|$. Note here that the target, h, is known and fixed so that this differs from the MAD where the target is the median which needs to be estimated.

Snedecor & Cochran (1980) describe a simple way to construct the confidence interval for the population median. Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ be the ordered random sample. Then two of the order statistics become the lower and upper bounds of the confidence interval. The $(1-\alpha)\times 100\%$ confidence interval for position of the population median is,

$$(n+1)/2 \pm z_{(1-\alpha/2)}\sqrt{n}. \tag{3.1}$$

After rounding the lower limit and upper limit to the nearest integer, two order statistics can be selected as the lower and upper bounds for the confidence interval for the population median. When $X_i = |Y_i - M|$ where M is the true population median and Y_i is the actual value, this confidence interval becomes the interval for the median absolute deviation with the exception that M is assumed fixed and not estimated.

In Arachchige & Prendergast (2019), we stated when h is the population median and is known, the simulation results were very good showing that the coverage probabilities are very close to nominal. Here we show the simulation table which is not included in Arachchige & Prendergast (2019). From Table 3.3, the coverage probabilities are close to the nominal coverage (0.95) even for small sample sizes under this assumption of known population median.

Table 3.3: Simulated coverage probabilities (and widths in parentheses) for the 95% confidence interval for the MAD when the target value is the population median and is assumed known.

Sample size	$X \sim$ LN(0,1)	$X \sim$ EXP(1)	$X \sim \chi^2_5$	$X \sim$ PAR(1,7)
50	0.940(0.31)	0.937(0.23)	0.937(1.07)	0.932(0.04)
100	0.944(0.22)	0.940(0.17)	0.937(0.79)	0.946(0.03)
200	0.940(0.15)	0.941(0.12)	0.939(0.56)	0.942(0.02)
500	0.945(0.10)	0.946(0.08)	0.946(0.36)	0.944(0.01)
1000	0.946(0.07)	0.944(0.05)	0.946(0.25)	0.950(0.01)

Table 3.4: Simulated coverage probabilities (and widths in parentheses) for the 95% confidence interval for the MAD when the target value is the estimated median.

Sample size	$X \sim$ LN(0,1)	$X \sim$ EXP(1)	$X \sim \chi^2_5$	$X \sim$ PAR(1,7)
50	0.751(0.31)	0.765(0.24)	0.900(1.09)	0.723(0.04)
100	0.756(0.22)	0.773(0.17)	0.908(0.80)	0.746(0.03)
200	0.755(0.15)	0.777(0.12)	0.912(0.56)	0.737(0.02)
500	0.753(0.10)	0.781(0.08)	0.912(0.36)	0.734(0.01)
1000	0.754(0.07)	0.790(0.05)	0.911(0.25)	0.740(0.01)

Table 3.4, shows the simulated coverage probabilities for the confidence interval for the MAD when the target is the estimated median. The coverage probabilities are typically much lower than the nominal coverage of 0.95 even for large sample sizes under this assumption of estimated population median for these selected distributions. This provides further evidence

of the need for confidence intervals with good coverage such as those we provided in Arachchige & Prendergast (2019).

3.2.2 Confidence interval for ratios of median absolute deviations from a target

Bonett & Seier (2003) also suggested constructing distribution-free confidence intervals for a ratio of median absolute deviations from a target by applying the Price & Bonett (2002) method for transformed values $|Y_{ij} - h|$. The two populations medians were considered as target values. The Price & Bonett (2002) method can be found in detail in Section 2.1 of Arachchige *et al.* (2019a).

Table generated by Excel2LaTeX from sheet 'Results for Bonnet and Series M'

Table 3.5: Simulated coverage probabilities (and widths in parentheses) for the 95% confidence interval for the ratio of MADs when the target values are the two population medians.

Sample sizes (n_1,n_2)	$X \sim$ LN(0,1) $Y \sim$ LN(0,1)	$X \sim$ EXP(1) $Y \sim$ EXP(1)	$X \sim \chi^2_5$ $Y \sim \chi^2_2$	$X \sim$ PAR(1,7) $Y \sim$ PAR(1,3)
50,50	0.906(0.93)	0.892(0.89)	0.900(1.81)	0.889(0.33)
100,100	0.918(0.59)	0.910(0.57)	0.912(1.20)	0.904(0.21)
200,200	0.925(0.39)	0.913(0.38)	0.918(0.81)	0.916(0.14)
200,500	0.878(0.31)	0.871(0.30)	0.859(0.65)	0.861(0.11)
500,500	0.930(0.24)	0.929(0.24)	0.927(0.50)	0.929(0.09)
500,1000	0.901(0.20)	0.896(0.20)	0.879(0.43)	0.895(0.07)
1000,1000	0.934(0.17)	0.935(0.16)	0.929(0.35)	0.930(0.06)

According to the results in Table 3.5, it seems this method works only for moderate to large and equal sample sizes. The coverage probability is bit low for unequal sample sizes.

3.2.3 Mathlevel data example

The Mathlevel data set, available in the `Ecdat` package (Croissant, 2016) in R (Development Core Team, 2018), measures the level of calculus attained for undergraduate students who took advanced micro-economics between August 1983 and August 1986 in the United States. The data set contains records for 609 (male-373, female-236) students and the response variable that we considered is SAT Math score.

Figure 3.4 depicts the math scores for male and female students. It can be clearly seen that there are some outliers for both the male and female students. Table 3.6 shows the estimate and the confidence intervals for the ratio of variances using the F distribution (assuming the data are drawn from normal distributions), the squared ratio of MADs and the

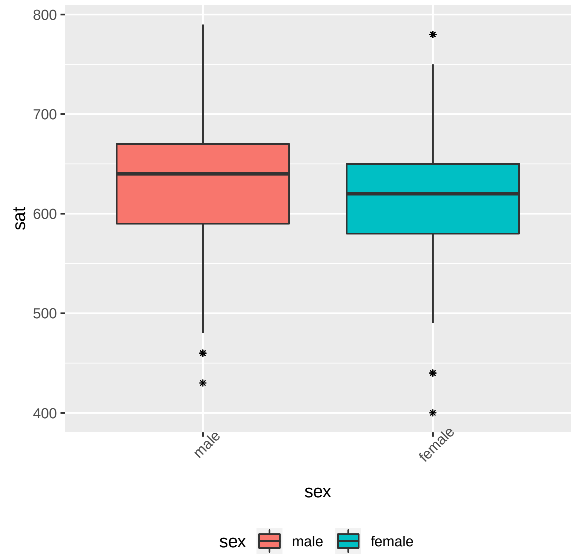


Figure 3.4: Box plot of the SAT Math score for males and females students.

Table 3.6: 95% asymptotic confidence intervals (CI) for the ratio of variances, ratio of MADs (R_M) and difference of MADs (D_M) for SAT Math score for males and females.

Method	Estimate	CI
Ratio of variances	1.2317	(0.9742, 1.5477)
R_M	1.7778	(1.1806, 2.6771)
D_M	10	(3.0825, 16.9175)

difference of MADs. The interval estimator of the ratio of variances includes one, making it difficult to conclude that there are differences in the spread. However, both the interval estimators of the squared ratio of MADs and the difference of MADs indicate a significant difference in the spread of the math scores between males and females.

3.3 Additional work related to paper III

This is supplementary work to Paper III, and so the reader should first read Paper III in Part II.

3.3.1 Comparing two relative spreads using confidence intervals for differences of CV, RCV$_Q$ and RCV$_M$

In Section 4.2 of Arachchige *et al.* (2019b), we introduced the interval estimator for the ratio of two relative spreads to compare the relative spreads of two independent populations. The interval estimators for differences in the two relative spreads can also be used to compare the relative spread of two independent populations. For example, an interval estimator for

$\mathrm{RCV}_{M,1} - \mathrm{RCV}_{M,2}$ is,

$$\widehat{d} \pm z_{1-\alpha/2} \left[\frac{\widehat{\mathrm{ASD}}\left(\mathcal{RCV}_{M,1}, F_{n_1}\right)}{\sqrt{n_1}} + \frac{\widehat{\mathrm{ASD}}\left(\mathcal{RCV}_{M,2}, F_{n_2}\right)}{\sqrt{n_2}} \right], \tag{3.2}$$

where $\widehat{d} = \widehat{\mathrm{rcv}}_{M,1} - \widehat{\mathrm{rcv}}_{M,2}$ and n_1 and n_2 are the sample sizes of the simple random samples from the two populations.

3.3.2 Melbourne house price data example

The house price data example which is described in Section 5.2.2 of Arachchige *et al.* (2019b) is based on the Melbourne house price data set which is available at https://www.kaggle.com/anthonypino/melbourne-housing-market. Figure 2 of Arachchige *et al.* (2019b) depicts the house price distribution of 3 pairs of neighbouring suburbs. Our objective is to check whether there are any differences in the relative spreads of house prices between neighbouring suburbs. In Arachchige *et al.* (2019b), we used interval estimators of the ratios of the relative spreads of CV, RCV_Q and RCV_M and here we use the differences of the relative spreads of CV, RCV_Q and RCV_M to check whether there are any differences in the relative spreads of house prices between neighbouring suburbs.

Table 3.7: 95% confidence interval lower bounds (LB) and upper bounds (UB) for differences of CV, RCV_Q and RCV_M between neighbouring suburbs house prices.

Confidence Interval Method	x=Bundoora y=Kingsbury		x=Beaumaris y=Black Rock		x=Oakleigh y=Oakleigh East	
	LB	UB	LB	UB	LB	UB
$\mathrm{CV}_x - \mathrm{CV}_y$	0.0062	0.2771	-0.1621	0.2478	-0.1789	0.1655
$\mathrm{RCV}_{Q_x} - \mathrm{RCV}_{Q_y}$	-0.5406	0.0283	0.0308	0.3497	-0.3629	0.0354
$\mathrm{RCV}_{M_x} - \mathrm{RCV}_{M_y}$	-0.3152	0.0563	0.0238	0.2546	-0.2624	0.0090

According to the results in Table 3.7, the three measures provide different conclusions. For Bundoora and Kingsbury, the intervals of the differences of CV suggest that there is a difference in the relative spread of house price between these two suburbs while the intervals of the differences of RCV_Q and RCV_M suggest there is no difference in relative spread, but the intervals are wide. For Beaumaris and Black Rock, the differences of CV suggest there is no significant difference in the relative spread of house price between these two suburbs but the other two intervals suggest a significant difference in relative spread. All three measures suggest that there are no significant differences in the relative spread of house price between Oakleigh and Oakleigh East.

3.4 Additional works related to paper IV

This is supplementary work to Paper IV, and so the reader should first read Paper IV in Part II.

3.4.1 Simulation results for the generalized Bowley's coefficient, γ_p

Table 3.8: Simulated coverage probabilities (and widths) for 95% confidence interval estimators for γ_p.

n	Dist.	$\gamma_{p=0.05}$	$\gamma_{p=0.1}$	$\gamma_{p=0.15}$	$\gamma_{p=0.2}$	$\gamma_{p=0.25}$
50	N(2,1)	0.968(0.64)	0.974(0.66)	0.979(0.71)	0.980(0.77)	0.979(0.85)
	LN(0, 1)	0.945(0.49)	0.962(0.58)	0.972(0.66)	0.980(0.74)	0.987(0.83)
	EXP(1)	0.960(0.55)	0.968(0.60)	0.972(0.67)	0.979(0.75)	0.985(0.84)
	Chi(2)	0.956(0.54)	0.967(0.60)	0.974(0.67)	0.979(0.75)	0.981(0.84)
	PAR(1, 7)	0.955(0.57)	0.963(0.61)	0.972(0.69)	0.980(0.76)	0.984(0.86)
100	N(2,1)	0.974(0.46)	0.974(0.46)	0.976(0.49)	0.974(0.54)	0.972(0.59)
	LN(0, 1)	0.955(0.35)	0.964(0.39)	0.972(0.45)	0.971(0.51)	0.976(0.57)
	EXP(1)	0.963(0.37)	0.962(0.41)	0.965(0.45)	0.971(0.51)	0.973(0.57)
	Chi(2)	0.958(0.37)	0.962(0.41)	0.969(0.45)	0.970(0.51)	0.971(0.57)
	PAR(1, 7)	0.959(0.37)	0.962(0.41)	0.965(0.45)	0.972(0.51)	0.974(0.58)
200	N(2,1)	0.973(0.32)	0.974(0.33)	0.973(0.34)	0.972(0.37)	0.967(0.41)
	LN(0, 1)	0.962(0.24)	0.964(0.27)	0.964(0.31)	0.968(0.35)	0.971(0.39)
	EXP(1)	0.958(0.25)	0.963(0.28)	0.963(0.31)	0.964(0.35)	0.966(0.39)
	Chi(2)	0.959(0.25)	0.960(0.28)	0.965(0.31)	0.966(0.35)	0.964(0.39)
	PAR(1, 7)	0.959(0.25)	0.965(0.28)	0.967(0.31)	0.967(0.35)	0.965(0.39)
500	N(2,1)	0.975(0.20)	0.970(0.20)	0.969(0.22)	0.966(0.23)	0.961(0.25)
	LN(0, 1)	0.960(0.15)	0.962(0.17)	0.963(0.19)	0.960(0.21)	0.961(0.24)
	EXP(1)	0.962(0.15)	0.957(0.17)	0.959(0.19)	0.959(0.22)	0.956(0.24)
	Chi(2)	0.959(0.15)	0.960(0.17)	0.956(0.19)	0.958(0.22)	0.956(0.24)
	PAR(1, 7)	0.957(0.15)	0.962(0.17)	0.957(0.19)	0.962(0.21)	0.961(0.24)
1000	N(2,1)	0.972(0.14)	0.964(0.14)	0.963(0.15)	0.960(0.16)	0.955(0.17)
	LN(0, 1)	0.961(0.10)	0.958(0.12)	0.960(0.13)	0.959(0.15)	0.954(0.17)
	EXP(1)	0.956(0.11)	0.955(0.12)	0.959(0.13)	0.956(0.15)	0.955(0.17)
	Chi(2)	0.953(0.11)	0.960(0.12)	0.958(0.13)	0.953(0.15)	0.955(0.17)
	PAR(1, 7)	0.959(0.10)	0.962(0.12)	0.958(0.13)	0.957(0.15)	0.958(0.17)
5000	N(2,1)	0.955(0.06)	0.956(0.06)	0.956(0.06)	0.955(0.07)	0.955(0.08)
	LN(0, 1)	0.952(0.04)	0.956(0.05)	0.953(0.06)	0.955(0.06)	0.952(0.07)
	EXP(1)	0.951(0.05)	0.953(0.05)	0.951(0.06)	0.954(0.07)	0.954(0.07)
	Chi(2)	0.953(0.05)	0.950(0.05)	0.955(0.06)	0.955(0.07)	0.953(0.07)
	PAR(1, 7)	0.950(0.04)	0.953(0.05)	0.953(0.06)	0.954(0.07)	0.955(0.07)
10000	N(2,1)	0.959(0.04)	0.955(0.04)	0.952(0.04)	0.951(0.05)	0.953(0.05)
	LN(0, 1)	0.960(0.03)	0.953(0.04)	0.951(0.04)	0.954(0.05)	0.954(0.05)
	EXP(1)	0.953(0.03)	0.955(0.04)	0.948(0.04)	0.948(0.05)	0.952(0.05)
	Chi(2)	0.955(0.03)	0.956(0.04)	0.951(0.04)	0.950(0.05)	0.951(0.05)
	PAR(1, 7)	0.954(0.03)	0.954(0.04)	0.950(0.04)	0.952(0.05)	0.953(0.05)

Simulated coverages based on 10,000 trials for the interval estimator of γ_p in Paper IV are provided in Table 3.8. The interval estimator of γ_p provides very good coverage compared to the nominal 0.95 and the interval width decreases with increasing sample sizes. Also, $\gamma_{p=0.05}$ provides a lower interval width compared to the γ_p of all the other choices of p

for all n.

3.4.2 Comparison of measures of skewness

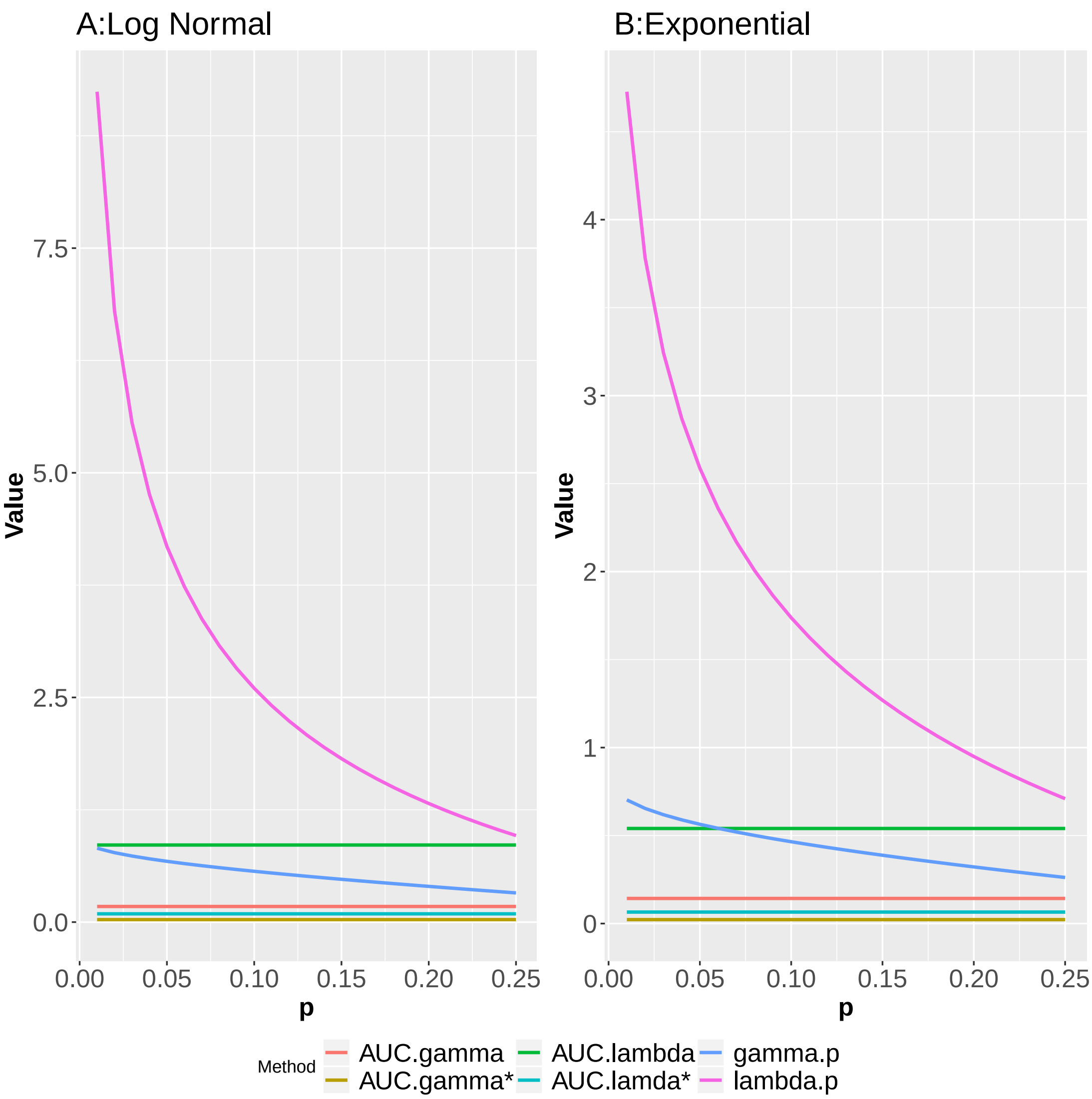


Figure 3.5: Comparison of skewness measures over p for LN(0, 1) and Exp(1) distributions

Figure 3.5 depicts the comparisons of measures of skewness, γ_p (gamma.p), λ_p (lambda.p) with the new measures of skewness, $\bar{\gamma}_p$ (AUC.gamma), $\bar{\gamma}_p^*$ (AUC.gamma), $\bar{\lambda}_p$ (AUC.lambda) and $\bar{\lambda}_p^*$ (AUC.lambda*) over p for the LN(0, 1) and Exp(1) distributions. The γ_p and λ_p measures vary over p while the integrated versions remain constant since they are averaged over all p. The highest variation can be seen in the λ_p over p and the lowest value represents $\bar{\gamma}_p^*$. The values of the $\bar{\gamma}_p$, $\bar{\gamma}_p^*$ and the $\bar{\lambda}_p^*$ are approximately similar and the value of the $\bar{\lambda}_p$ is somewhat higher compared to the others.

3.4.3 Comparison of measures of skewness

We now consider the relative asymptotic variance (rASV) where, e.g., for γ_p, rASV = ASV$/\gamma_p^2$. Figure 3.6 depicts the rASV comparisons of all six skewness measures, γ_p, λ_p, $\bar{\gamma}_p$, $\bar{\gamma}_p^*$, $\bar{\lambda}_p$ and $\overline{\lambda}_p^*$ over $p \in [0.01, 0.25]$ for the LN(0,1) and EXP(1) distributions. For the LN(0,1) distribution, the rASV for the estimator of $\overline{\gamma_p}^*$ (rASV.AUC.gamma) is comparatively higher than the rASV of the other integrated versions. The rASV of both the γ_p (rASV.gamma.p) and λ_p (rASV.lambda.p) estimators are increasing when p increases. The rASV of γ_p also becomes lower with lower values of p. The best choice of p for λ_p to get a minimum rASV is around 0.05 and this choice of p is Groeneveld *et al.* (2009)'s recommendation of $p = 0.05$ for λ_p. When it comes to the EXP(1) distribution, a similar comparison can be seen among the six measures of skewness. However it is clear that for small $p \in [0.01, 0.05]$ rASV of λ_p is lower than the rASV of $\bar{\gamma}_p$ (rASV.AUC.gamma) and $\overline{\lambda}_p^*$ (rASV.AUC.lambda*) and the rASV of the γ_p estimator is less than the rASV of $\overline{\lambda}_p^*$. In addition, selecting a $p \in [0.025, 0.05]$ will be a good choice for λ_p to minimise the rASV.

3.5 R shiny web applications

We created R shiny (Chang *et al.*, 2017) web applications for Publications I, II, III and IV and included the links for the applications in each paper. Here we describe the interface of the R shiny web application related to each publication.

3.5.1 Interface of the R shiny web applications created in Paper I

Figure 3.7 shows the interface of the created R shiny web application which can be accessed via the link given in Arachchige & Prendergast (2019). Here we introduced a wide range of additional distributions in the shiny web application that we did not report in the simulation results given in Arachchige & Prendergast (2019). The user can change the numerator and denominator distributions, parameters, sample sizes, probability and the number of trials to any values they choose. The simulation results can be obtained by clicking on the 'Run Simulation' button and the relevant estimates and the two performance measures, coverage probability (CP) and the average confidence interval width (Width) will be calculated. For clarity, we developed the web application to create two separate tables to compare the performance of the location and scale estimators separately. In addition, anyone can use the " copy" and " Print" buttons to copy or print the results. The first table in Figure 3.7

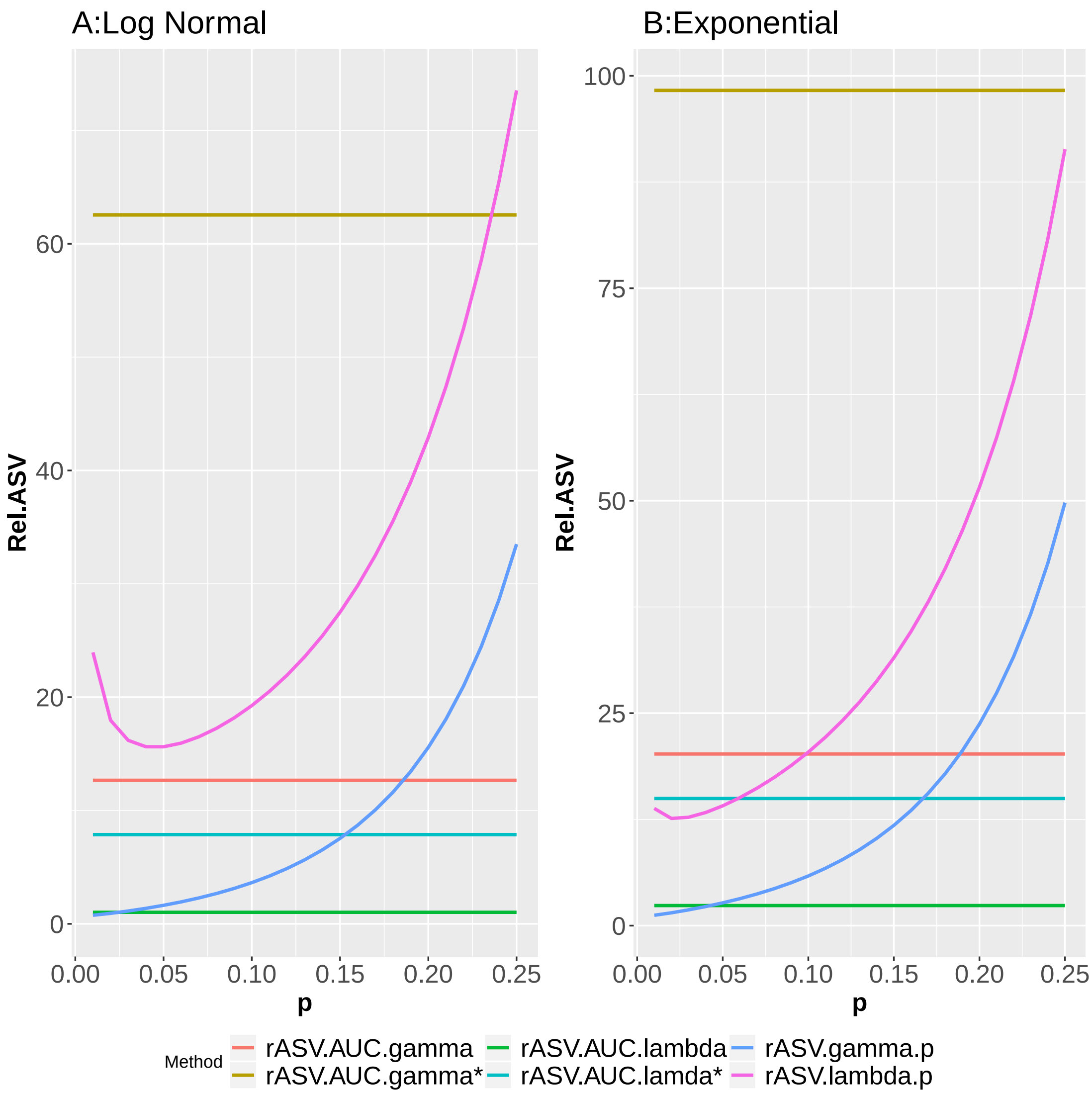


Figure 3.6: Relative asymptotic variance comparisons of skewness measures over p for LN(0, 1) and Exp(1) distributions

illustrates the simulation results for point and interval estimators of t-test (t), Price and Bonnet method (r) Price & Bonett (2002) and ratio of quantiles (rp) when both the numerator and denominator distributions are LN(0,1), sample sizes (n_1, n_2) are 100, 1000 simulation trials and $p = 0.2$ for r_p. The second table in Figure 3.7 details the simulation results for the point and interval estimators of the F-test (F), ratio of variances (R) and squared ratio of IQR (Rp) for the same settings as for the first table.

3.5.2 Interface of the R shiny web applications created in Paper II

Figures 3.8 and 3.9 show the interfaces of the R shiny web applications which are related to the simulations conducted in Arachchige & Prendergast (2019) and are available via following links

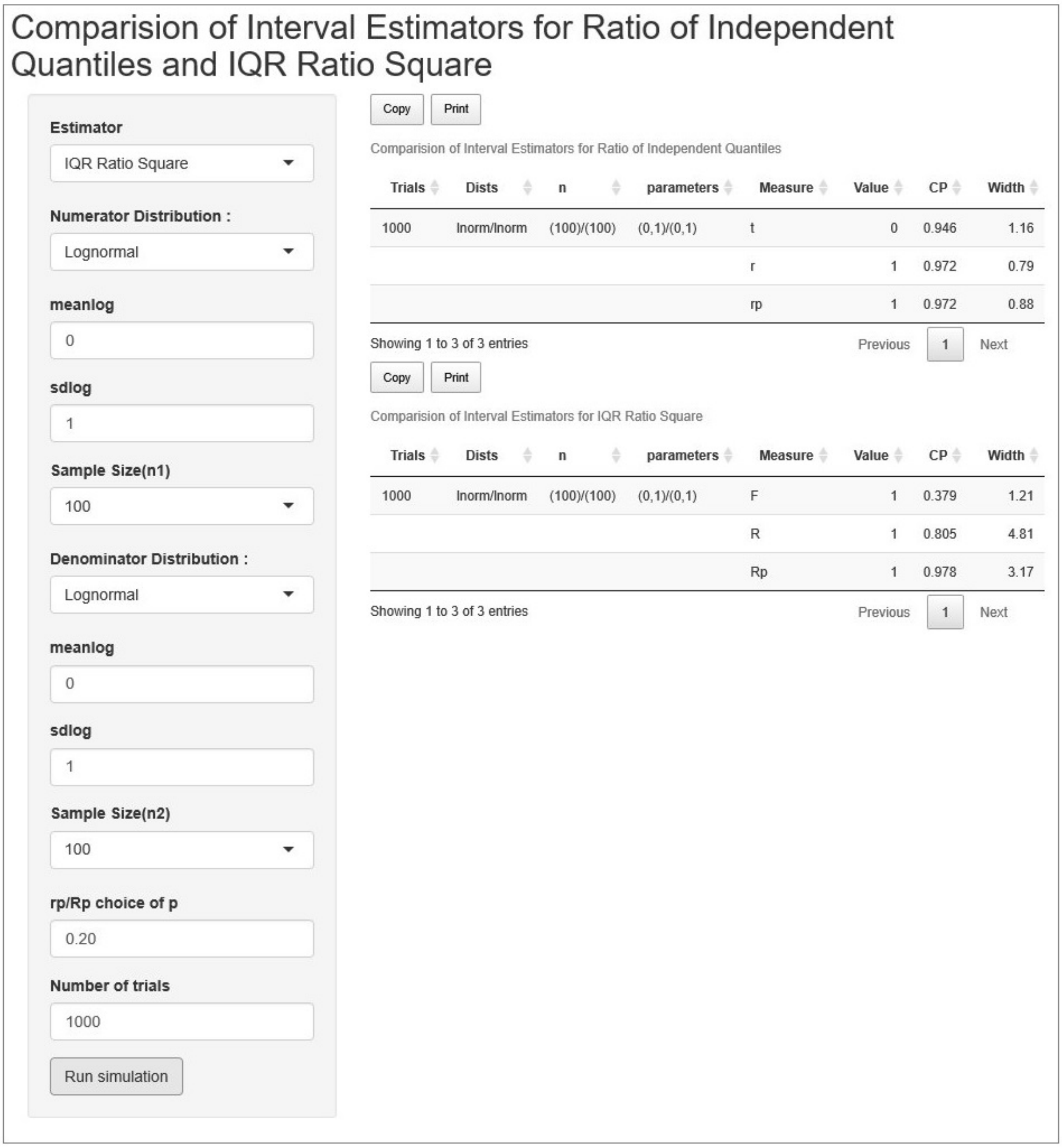


Figure 3.7: The R shiny interface to compare the performance of ratio of variances, squared ratios of IQRs and ratios of squared MADs.

https://lukeprendergast.shinyapps.io/MADcalc/

and

https://lukeprendergast.shinyapps.io/MADRmDm/

Figure 3.8 shows the interface of the created R shiny web application to check the performance of the interval estimator of the MAD which is used to make inferences about spread in a single population. We introduce a wide range of other distributions that we did not report in the simulation results given in Table 1 of Arachchige & Prendergast (2019). The user can change the distribution, parameters, sample size and the number of trials they choose. Simulation results can be obtained by clicking on the ‘Run Simulation

’button and the relevant estimates and the three performance measures, coverage probability (CP), average confidence interval width (Avg.W.) and median confidence interval width (Med.W.) are calculated. Here, we have included median confidence interval width also as a performance measure due to some excessively large confidence interval widths for a small number of intervals that skew the mean and this is noted in Table 1 of Arachchige & Prendergast (2019). Here also, anyone can use the “Copy” and “print” buttons to copy the whole table at once or print the results. The table in Figure 3.8 details the simulation results for the point and interval estimators of MAD for the LN(0,1), EXP(1), χ^2_5 and PAR(1,7) distributions when the sample size (n) is 500 and for 1000 simulation trials.

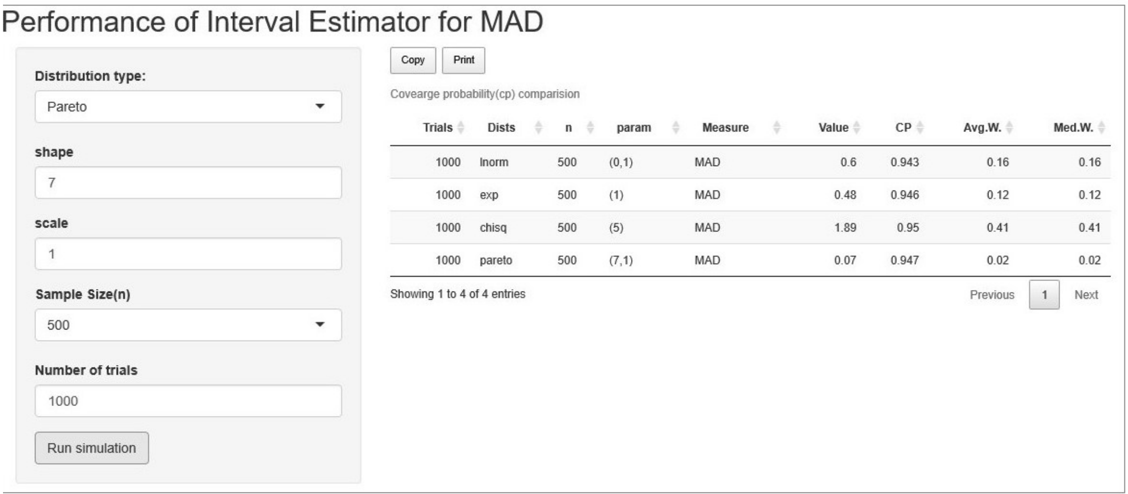


Figure 3.8: The created R shiny interface to check the performance of interval estimators for the MAD.

Figure 3.9 shows the interface of the R shiny web application to compare the performance of the interval estimators of differences and ratios of MADs (Dm and Rm) which are used to compare the spread in two populations. Again, we introduce a wide range of distributions that we did not report in the simulation results given in Table 2 of Arachchige & Prendergast (2019). The user can change the numerator and denominator distributions, parameters, sample sizes and the number of trials they choose. Then, the simulation results can be obtained by clicking on the ‘Run Simulation ’button and the relevant estimates and the three performance measures, coverage probability (CP), average confidence interval width (Avg.W.) and median confidence interval width (Med.W.) are calculated. Here we have included median confidence interval width as a performance measure due to some excessively large confidence interval widths for a small number, between 1% and 2%, of intervals and this is noted in Table 2 of Arachchige & Prendergast (2019). Again, the tables can be copied or printed. The table in Figure 3.9 details the simulation results for the point and interval

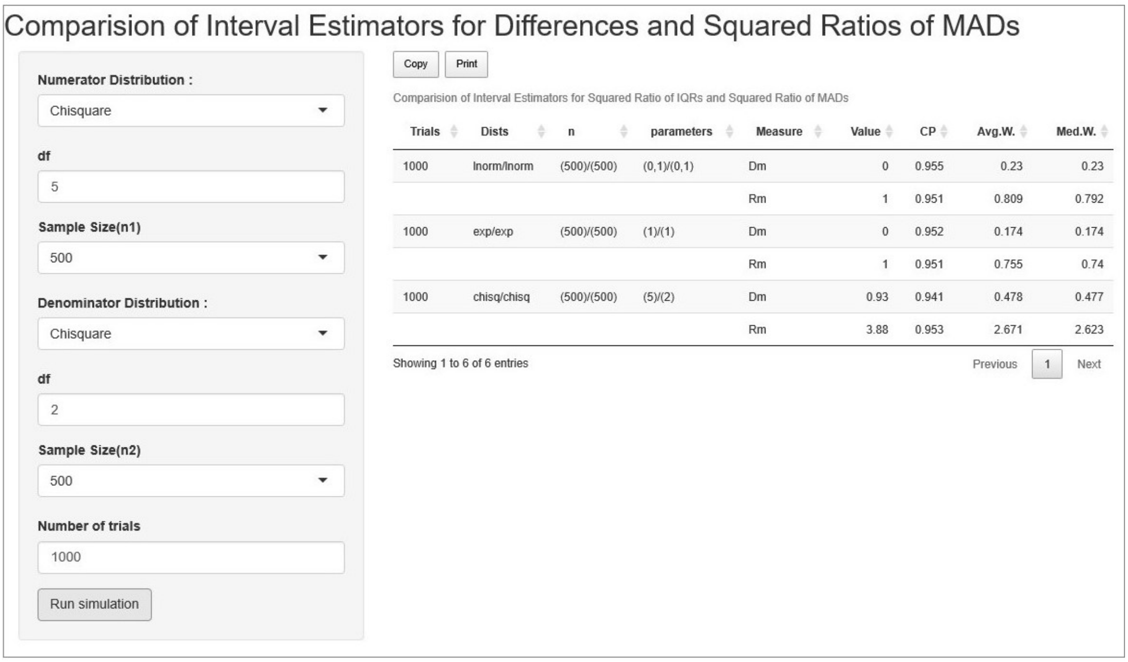


Figure 3.9: The R shiny interface to check the performance of interval estimators of differences of MAD and ratios of MAD

estimators of differences and the ratios of MADs (Dm and Rm) when both the numerator and denominator distributions are LN (0,1) and EXP(1) and when the numerator and denominator distributions and when they are χ^2_5 and χ^2_2 respectively for sample sizes (n_1, n_2) both equal to 500 and for 1000 simulation trials.

3.5.3 Interface of the R shiny web applications related to Paper III

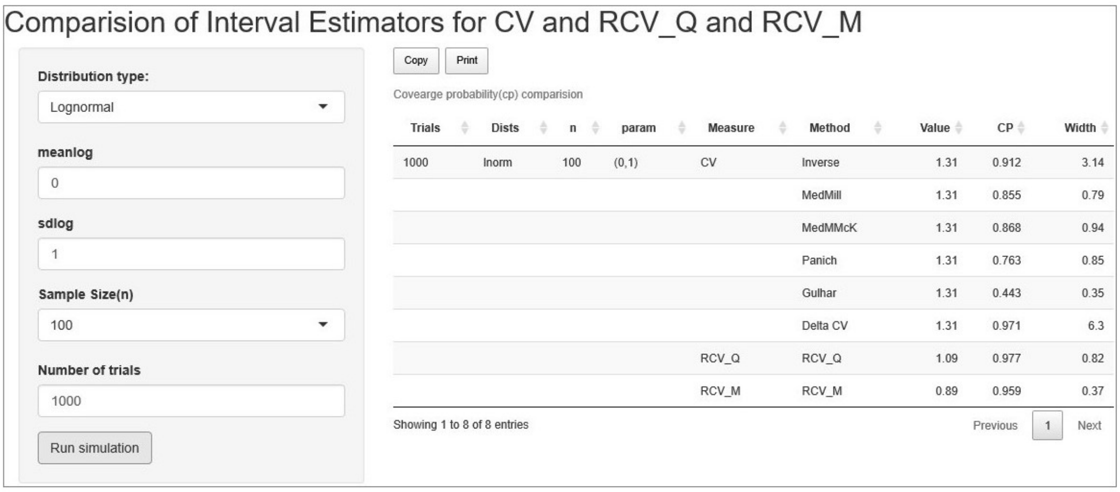


Figure 3.10: The R shiny interface to compare the performance of CV, RCV_Q and RCV_M.

Figure 3.10 shows the interface of the R shiny web application which can be accessed via the link given in Arachchige *et al.* (2019b). The user can change distributions, parameters, sample size and the number of trials as they choose. Then the simulation results can be obtained by clicking on the 'Run Simulation' button and the relevant estimates and the two performance measures, coverage probability (CP) and the average confidence interval width (width) will be calculated for all the confidence interval estimators of CV (various estimators, see the paper for details), RCV_Q and RCV_M. The table in Figure 3.10 provides the simulation results for the point and interval estimators of CV, RCV_Q and RCV_M for the LN(0,1) distributions when the sample size (n) is 100 and for 1000 simulation trials.

3.5.4 R shiny web applications related to paper IV

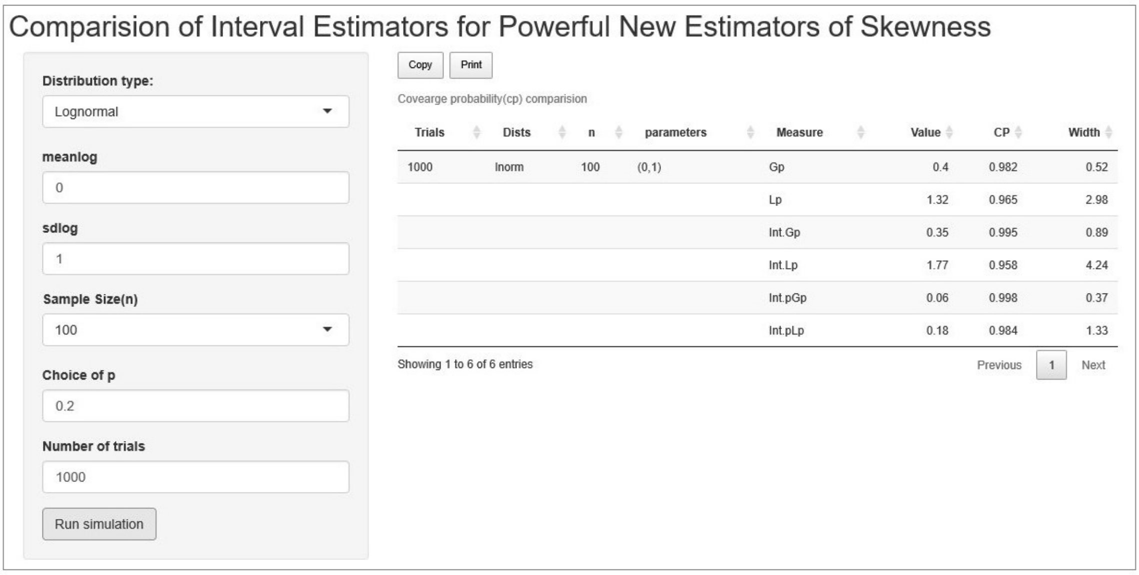


Figure 3.11: The created R shiny interface to compare the performance of interval estimators for γ_p ,λ_p with our new measures $\bar{\gamma}_p$, $\bar{\gamma}_p^*$, $\bar{\lambda}_p$, $\bar{\lambda}_p^*$.

Figure 3.11 shows the interface of the created R shiny web application which related can be accessed via the link

https://lukeprendergast.shinyapps.io/meanskew/

The user can change the distribution, parameters, sample size, probability and the number of trials to which ever they choose. Then the simulation results can be obtained by clicking on the 'Run Simulation 'button and the relevant estimates and the two performance measures, coverage probability (CP) and the average confidence interval width (Width) will be calculated for all the intervals estimators of the measures of skewness, γ_p (G_p), λ_p (L_p) and our new measures $\bar{\gamma}_p$ (int. G_p), $\bar{\gamma}_p^*$ (int. pG_p), $\bar{\lambda}_p$ (int. L_p), $\bar{\lambda}_p^*$ (int. pL_p). The table in Figure 3.11 details the simulation results for the point and interval estimators of CV, RCV_Q and RCV_M for LN(0,1) distributions when the sample size (n) is 100 and for 1000 simulation trials.

4. Summary and future work

This thesis is divided into two parts, as Part I (Chapters 1-3) and Part II. In Chapter 1, previous work related to the research in this thesis was briefly reviewed. In Chapter 2, theory and methods which used for developing new measures were discussed. In Chapter 3, some additional works related to each paper were included, including a description of R Shiny applications that the reader can use. The Appendix consists of the most important R programs related to the all the papers.

Part II consists of four accepted or submitted papers which cover the main contents of this thesis. The objective of this thesis is to introduce quantile-based measures of location, scale, relative spread and skewness with their distribution-free interval estimators to compare two populations or to make inferences about a single population. Paper I and Paper II consist of some research works related to quantile-based measures to compare the location and scale of two populations respectively. Paper III and Paper IV detail research related to the quantile-based measures of relative spread and skewness to make inferences on a population.

In Paper I, distribution free point and interval estimators were introduced for ratios of independent quantiles and squared ratios of IQRs. Robustness properties were investigated using partial influence functions. Asymptotic variance comparisons were conducted, and the respective probabilities related to the minimum ASV in squared ratios of IQRs were shown for selected distributions. Simulation studies were conducted to compare the performance of the interval estimators and the coverage probabilities were very close to the nominal coverage even for moderate sample sizes of 50. Several examples were used as applications to the real-world data and new location and scale estimators provided different conclusions compared to the existing measures of location and scales such as t-test and the F-test. Future work will be to find a new measure of scale which avoids the ”need to choose p” requirement.

In Paper II, distribution-free point and interval estimators for the MAD and difference and ratios of MADs were introduced to make inferences on scale of a single population and to compare the scale of two populations respectively. Robustness properties were investigated using influence functions and partial influence functions. Simulations were conducted to compare the performance of the new estimators. Results suggest that the coverage probabilities are very close to the nominal coverage even for moderate sample sizes of 50. Examples reveal that the difference and ratios of MADs resulted in different conclusions compared to the F- test. Future works will be to find intervals to alternatives to the MAD.

In Paper III, distribution-free point and interval estimators were introduced for two robust versions of the coefficient of variation, the RCV_Q and RCV_M. Parametric and non-parametric bootstrap intervals were introduced for RCV_M and robustness properties of the estimators were investigated using influence functions. The relative asymptotic standard deviation comparisons were made among the estimators. Performances of the new interval estimators were compared using simulations and results suggest that the coverage probabilities are very close to the nominal coverage even for small sample sizes for the intervals of the robust alternatives to the CV when compared to the CV. Examples reveal that different conclusions can be made based on robust and non-robust versions of the CV. Future works may be comparing these new intervals with other existing intervals which were not investigated and finding some other robust alternatives to the CV.

In Paper IV, more powerful robust alternatives were introduced to existing quantile-based measures of skewness. The integrated versions were introduced for two alternatives to the well-known Bowely's skewness coefficient, γ_p and λ_p, to avoid the "need to choose p requirement". The new estimators were tested to determine whether they satisfied the properties that skewness estimators should satisfy. The distribution-free point and interval estimators were constructed for the new measures of skewness. Performances of the new interval estimators were tested using simulations for a wide range of distributions and results suggest that the coverage probabilities are very close to nominal coverage for moderate to large sample sizes. Future work may be to investigate the robustness properties of the new estimators using their influence functions.

The R functions for the two robust versions of the CV which is given in paper III are available in "MKmisc" package in R (Kohl & Kohl, 2019). Therefore, Future work may be to create an R package including the estimators introduced in paper I, paper II and paper Iv or publish the all R scripts on a website or some public repository such as GitHub or GitLab or R-Forge.

A. Example R Programs

A.1 R programs related to the Introduction

A.1.1 R program to compare the SIF and EIF

```
n <- 100
mean <- 0
sigma <- 1
x <- rnorm(n,mean,sigma)
EIF <- rep(0,n)
SIF <- rep(0,n)
M <- mean(x)
S1 <- var(x)
for(i in 1:n){
  EIF[i] <- ((x[i]-M)^2)-S1
  S1.i <- var(x[-i])
  SIF[i] <- (n-1)*(S1-S1.i)
}
EIF
SIF
windows(width = 10, height = 5)
plot(EIF,type="o",col="blue",ann=FALSE)
lines(SIF,type="o",col="red",lty=2)
abline(h=0)
#title(main="Influence Function for the Variance")
title(ylab="Influence Functions")
title(xlab="Index")
g_range <- range(0, EIF, SIF)
legend(1, g_range[2], c("EIF","SIF"), cex=1,
```

```
    col=c("blue","red"),bty="o",pch=1:1,lty=1:2)
```

A.2 R programs related to paper I

A.2.1 Function to calculate an estimate, confidence interval and standard error for the ratio of quantiles, ratio of variance and ratio of squared IQRs

```
VarQuaIQRratio <- function(x, y, p, conf.level = 0.95){
  # Computes the ratio of quantiles, ratio of variances
  # and squared ratio of IQRs between two vectors.
  # Args:
  #   x: One of two numeric vectors whose sample quantile,
  #      sample variance and IQRs are to be calculated.
  #   y: The other numeric vector whose sample quantile,
  #      sample variance and IQR are to be calculated.
  #   conf.level: The relevant confidence level to obtain the
  #               confidence interval for ratio of quantiles,
  #               ratio of variance and squared ratio of IQRs.
  # Returns:
  # The sample ratio of quantiles, ratio of variance and
  # squared ratio of IQRs between x and y, confidence
  # intervals and the Standard errors of the estimates under
  # the specified confidence level.

  n1 <- length(x)
  n2 <- length(y)

  S1 <- sd(x)
  M1 <- mean(x)
  S2 <- sd(y)
  M2 <- mean(y)
  Z1 <- (x - M1) / S1
  Z2 <- (y - M2) / S2

  R <- (S1 / S2) ^ 2
```

```
log.R <- log(R)
Var.R <- R ^ 2 * (mean(Z1 ^ 4 - 1) / n1 +
                    (mean(Z2 ^ 4) - 1) / n2)
Var.log.R <- (1 / R) ^ 2 * Var.R # Using Delta Method
# CI for ratio of variances
CI.log.R <- log.R + c(-1, 1) * qnorm(0.975) * sqrt(Var.log.R)
CI.R <- exp(CI.log.R)

QOR.ln <- function(p) {
  # QOR function for the log-normal
  q.n.0 <- 1 / dnorm(qnorm(p))
  q.n.1 <- qnorm(p) * q.n.0 ^ 2
  q.n.2 <- (1 + 2 * qnorm(p) ^ 2) * q.n.0 ^ 3
  1 / (q.n.0 ^ 2 + 3 * q.n.1 + q.n.2 / q.n.0)
}
# will estimate 1/f(x_p)
qHat <- function(m, p, QOR.FUN = QOR.ln, lambda = NULL,
         hn = 0.15, bw.correct = TRUE, method = "TL", ...){
  m.sorted <- sort(m)
  n <- length(m)
  kernepach <- function(u) 3/4 * (1 - u ^ 2) * (abs(u) <= 1)
  qor <- QOR.FUN(p, ...)
  band.width <- 15 ^ (1/5) * abs(qor) ^ (2/5) / n ^ (1/5)
  if (bw.correct) band.width <- min(band.width, p)
  consts <- kernepach((p - (1:n - 1) / n) / band.width)/
    band.width - kernepach((p - (1:n) / n) /
                           band.width) / band.width
  sum(consts * m.sorted)}

qu.x1 <- qHat(x, p)
qu.x3 <- qHat(x, 1-p)
qu.y1 <- qHat(y, p)
qu.y3 <- qHat(y, 1-p)
ASVar.q1.x <- p * (1 - p) * (qu.x1 ^ 2)
ASVar.q3.x <- p * (1 - p) * (qu.x3 ^ 2)
ASCov.q1q3.x <- p ^ 2 * qu.x1 * qu.x3
```

```
  ASVar.q1.y <- p * (1 - p) * (qu.y1 ^ 2)
  ASVar.q3.y <- p * (1 - p) * (qu.y3 ^ 2)
  ASCov.q1q3.Y <- p ^ 2 * qu.y1 * qu.y3
  ASVar.IQR.x <- ASVar.q1.x + ASVar.q3.x - 2 * ASCov.q1q3.x
  ASVar.IQR.y <- ASVar.q1.y + ASVar.q3.y - 2 * ASCov.q1q3.Y

  Q.x <- quantile(x, p)
  Q.y <- quantile(y, p)
  rp <- Q.x/Q.y
  log.rp <- log(rp)

  Var.rp <- (1 / n1) * (1 / Q.y ^ 2) * ASVar.q1.x
            + (1 / n2) * (rp^2 / Q.y ^ 2) * ASVar.q1.y
  Var.log.rp <- (1 / rp^2) * Var.rp # using Delta Method
  # CI for the ratio of quantiles
  CI.log.rp <- log.rp + c(-1, 1)*qnorm(0.975)*sqrt(Var.log.rp)
  CI.rp <- exp(CI.log.rp)

  IQR.x <- quantile(x, 1-p) - quantile(x, p)
  IQR.y <- quantile(y, 1-p) - quantile(y, p)
  a <- IQR.x / IQR.y
  a <- IQR.x / IQR.y
  Rp <- a ^ 2
  log.a <- log(a)

  Var.a <- (1 / n1) * (1 / IQR.y ^ 2) * ASVar.IQR.x
            + (1 / n2) *  (Rp / IQR.y ^ 2) * ASVar.IQR.y
  Var.log.a <- (1 / Rp) * Var.a # using Delta Method
  # CI for the IQR Ratio square
  CI.log.a <- log.a + c(-1,1) * qnorm(0.975) * sqrt(Var.log.a)
  CI.a <- exp(CI.log.a)
  CI.Rp <- (CI.a) ^ 2

list(Est2 = rp, Est1=R, Est3 = Rp, conf.int2 = CI.rp,
       conf.int2 = CI.R, conf.int3 = CI.Rp, SE.rp=sqrt(Var.rp),
       SE.R = sqrt(Var.R), SE.Rp=sqrt(Var.Rp))}
```

```
x <- rnorm(100)
y <- rnorm(200, sd = 2)
p <- 0.2
QuaIQRratio(x, y, p)
```

A.2.2 Function to calculate the true asymptotic variance of the ratio of variance

```
V <- function(d, lower = 0, upper = Inf,
       subdivisions = 1000L, ...){
  Fx <- function(x, ...){
    x*d(x, ...)
  }
  Fv <- function(x, ...){
    mu <- integrate(Fx, lower = lower, upper = upper,
            subdivisions = subdivisions, ...)$value
    (x - mu)^2*d(x, ...)}
  integrate(Fv, lower = lower, upper = upper,
    subdivisions = subdivisions, ...)$value}

Z4 <- function(d, lower = 0, upper = Inf,
        subdivisions = 1000L, ...){
  Fx <- function(x, ...){
    x*d(x, ...)}
  Fx2 <- function(x, ...){
    x^2*d(x, ...)}
  Fz4 <- function(x, ...){
    mu <- integrate(Fx, lower = lower, upper = upper,
    subdivisions = subdivisions, ...)$value
    sd <- sqrt(integrate(Fx2, lower = lower, upper = upper,
    subdivisions = subdivisions, ...)$value - mu^2)
    ((x - mu)/sd)^4*d(x, ...)
  }
  integrate(Fz4, lower = lower, upper = upper,
    subdivisions = subdivisions, ...)$value}
```

```
ASVrv2 <- function(n1, n2, dlist1 = dlist1, dlist2 = dlist2){
  V1 <- do.call(V, dlist1)
  V2 <- do.call(V, dlist2)
  R  <- V1/V2

  w1 <- n1 / (n1 + n2)
  w2 <- n2 / (n1 + n2)

  E.Z4.x <- do.call(Z4, dlist1)
  E.Z4.y <- do.call(Z4, dlist2)

  A <- E.Z4.x
  B <- E.Z4.y

  ASVar.R <- R ^ 2 * ((A - 1) / w1 + (B - 1) / w2)
  list(Est = R,  ASV.R = ASVar.R )
}
dlist1 <- list(d = dlnorm, lower = 0, upper = Inf,
            meanlog = 0, sdlog = 1)
dlist2 <- list(d = dlnorm, lower = 0, upper = Inf,
            meanlog = 0, sdlog = 1)
ASVrv2(1000, 1200, dlist1 = dlist1, dlist2 = dlist2 )

dlist1 <- list(d = dexp, lower = 0, upper = Inf, rate=1)
dlist2 <- list(d = dexp, lower = 0, upper = Inf, rate=1)
ASVrv2(1000, 1200, dlist1 = dlist1, dlist2 = dlist2 )
```

A.2.3 Function to calculate true asymptotic variance of ratio of quantiles and squared ratio of IQRs

```
ASVQ1 <- function(d, q, p , ...){
  p * (1 - p)  / d(q(p, ...), ...)^2
}

ASVQ3 <- function(d, q, p, ...){
```

```
  p * (1 - p)/ d(q(1-p, ...), ...)^2
}

ASCov <- function(d, q, p, ...){
  p ^ 2 /  (d(q(p, ...), ...)*d(q(1-p, ...), ...))
}

Qu <- function(q, p, ...){
  q(p, ...)
}

InQuRa <- function(q, p, ...){
  q(1-p, ...) - q(p, ...)
}

QunIQRratio <- function(n1, n2, p, d1, q1, d2, q2, ...  ){
  w1 <- n1 / (n1 + n2)
  w2 <- n2 / (n1 + n2)

  ASVar.q1.x <- ASVQ1(d1, q1, p, ...)
  ASVar.q3.x <- ASVQ3(d1, q1, p,...)
  ASCov.q1q3.x <-  ASCov(d1, q1, p, ...)
  ASVar.q1.y <- ASVQ1(d2, q2, p, ...)
  ASVar.q3.y <- ASVQ3(d2, q2, p, ...)
  ASCov.q1q3.Y <-  ASCov(d2, q2, p, ...)
  ASVar.IQR.x <- ASVar.q1.x + ASVar.q3.x - 2 * ASCov.q1q3.x
  ASVar.IQR.y <- ASVar.q1.y + ASVar.q3.y - 2 *ASCov.q1q3.Y

  Q.x <- Qu(q1, p, ...)
  Q.y <- Qu(q2, p, ...)
  rp <- Q.x/Q.y

  ASVar.rp <- (1 / w1) * (1 / Q.y ^ 2) * ASVar.q1.x + (1 / w2) *
    (rp^2 / Q.y ^ 2) * ASVar.q1.y

  IQR.x <- InQuRa(q1, p, ...)
```

```
  IQR.y <- InQuRa(q2, p, ...)
  r <- IQR.x / IQR.y
  R2 <- r ^ 2

  ASVar.r <- (1 / w1) * (1 / IQR.y ^ 2) * ASVar.IQR.x +
    (1 / w2) * (R2 / IQR.y ^ 2) * ASVar.IQR.y
  ASVar.R2 <- (ASVar.r) ^ 2
  list(Est1 = rp, Est2 = R2, ASVar.rp=ASVar.rp, ASV.R2 = ASVar.R2)
}
QunIQRratio(1000, 1200, 0.1, dlnorm, qlnorm, dlnorm, qlnorm)
```

A.3 R programs related to paper II

A.3.1 Function to calculate an estimate, confidence interval and standard error for MAD, difference and squared ratio of MADs

```
library(gld)
CIMADs <- function(x, y, conf.level = 0.95){
  # Computes MAD, the difference and squared ratio
  # of MADs between two vectors.
  # Args:
  #   x: One of two numeric vectors whose MAD
  #      is to be calculated.
  #   y: The other numeric vector whose MAD
  #      is to be calculated.
  #   conf.level: The relevant confidence level to
  #               obtain the confidence interval for
  #               MAD, the difference and squared
  #               ratios of MADs.
  # Returns:
  # The sample MAD of x and difference and squared ratio of
  # MADs between x and y, confidence intervals and the
  # Standard errors of the estimates under the specified
  # confidence level.
  n1 <- length(x)
  n2 <- length(y)
```

```r
MAD <- function(z){
  median(abs(z - median(z)))}

MAD.x <- MAD(x)
MAD.y <- MAD(y)

rm <- MAD.x/MAD.y
Rm <- (rm)^2
log.Rm  <- log(Rm)
Dm <- MAD.x - MAD.y

ASV.mad <- function(z, method){
  lambda <- fit.fkml(z, method = method)$lambda
  m  <- median(z)
  mad <- MAD(z)

  fFinv1 <- dgl(m - mad, lambda1 = lambda)
  fFinv2 <- dgl(m + mad, lambda1 = lambda)
  fFinv3 <- dgl(m, lambda1 = lambda)
  FFinv1 <- pgl(m - mad, lambda1 = lambda)
  FFinv2 <- pgl(m + mad, lambda1 = lambda)

  A <- fFinv1 + fFinv2
  C <- fFinv1 - fFinv2
  B <- C^2 + 4*C*fFinv3*(1 - FFinv2 - FFinv1)

  (1/(4 * A^2))*(1 + B/fFinv3^2) # ASV for the MAD
}
ASV.MAD.x <- ASV.mad(x, method = "TL")
ASV.MAD.y <- ASV.mad(y, method = "TL")

Var.MAD.x <- ASV.MAD.x/n1
Var.MAD.y <- ASV.MAD.y/n2

CI.MAD.x <- MAD.x + c(-1,1) * qnorm(0.975) * sqrt(ASV.MAD.x/n1)
```

```
    Var.Dm <- Var.MAD.x+Var.MAD.y
    Var.rm <- (1 / n1) * (1 / MAD.y ^ 2) * ASV.MAD.x +
           (1 / n2) * (Rm /  MAD.y^ 2) * ASV.MAD.y
    Var.Rm <- 4 * Rm * Var.rm
    Var.log.Rm <- (1 / Rm^2) * Var.Rm

CI.log.Rm <- log.Rm + c(-1,1) * qnorm(0.975) * sqrt(Var.log.Rm)
    CI.Rm <- exp(CI.log.Rm)
    CI.Dm <- Dm + c(-1,1) * qnorm(0.975) * sqrt(Var.Dm)

list(Est1 = MAD.x, Est2 = Rm, Est3 = Dm, conf.int1 = CI.MAD.x,
  conf.int2 = CI.Rm, conf.int3 = CI.Dm, SE.MAD=sqrt(Var.MAD.x),
  SE.Rm=sqrt(Var.Rm), SE.Dm=sqrt(Var.Dm))
}

x <- rnorm(100)
y <- rnorm(200, sd = 2)
CIMADs(x, y)
```

A.4 R programs related to paper III

A.4.1 Function to calculate an estimate, confidence interval and standard error for CV, RCV$_Q$ and RCV$_M$

```
library(gld)
library(stats)

CVRcvrCV <- function(x, conf.level = 0.95){
  n <- length(x)

  S <- sd(x)
  M <- mean(x)
  CV <- S/M

  Mo3.f <- function(d, lower = 0, upper = Inf,
                    subdivisions = 1000L, ...){
```

```
  Fx1 <- function(x, ...){
    x*d(x, ...)}
  Fv1 <- function(x, ...){
    mu1 <- integrate(Fx1, lower = lower, upper = upper,
                     subdivisions = subdivisions, ...)$value
    (x - mu1)^3*d(x, ...)}
  integrate(Fv1, lower = lower, upper = upper,
            subdivisions = subdivisions, ...)$value}

Mo4.f <- function(d, lower = 0, upper = Inf,
                  subdivisions = 1000L, ...){
  Fx2 <- function(x, ...){
    x*d(x, ...)}
  Fv2 <- function(x, ...){
    mu2 <- integrate(Fx2, lower = lower, upper = upper,
                     subdivisions = subdivisions, ...)$value
    (x - mu2)^4*d(x, ...)}
  integrate(Fv2, lower = lower, upper = upper,
            subdivisions = subdivisions, ...)$value}

Mo3 <- do.call(Mo3.f, dlist)
Mo4 <- do.call(Mo4.f, dlist)

ASV.CV <- CV^2 * (CV^2 - 1/4) + (1 /(4 * M^2 * S^2)) *
  Mo4 - (1/ M^3) * Mo3

log.CV <- log(CV)
Var.CV <- ASV.CV/n
Var.log.CV <- (1/CV^2) * Var.CV

# CI for CV
CI.log.CV <- log.CV + c(-1,1) * qnorm(0.975) * sqrt(Var.log.CV)
CI.CV <- exp(CI.log.CV)

QOR.ln <- function(p) {
  # QOR function for the log-normal
```

```
  q.n.0 <- 1 / dnorm(qnorm(p))
  q.n.1 <- qnorm(p) * q.n.0 ^ 2
  q.n.2 <- (1 + 2 * qnorm(p) ^ 2) * q.n.0 ^ 3
  1 / (q.n.0 ^ 2 + 3 * q.n.1 + q.n.2 / q.n.0)}

# will estimate 1/f(x_p)
qHat<- function(m, p, QOR.FUN = QOR.ln, lambda = NULL, hn = 0.15,
                bw.correct = TRUE, method = "TL", ...){
  m.sorted <- sort(m)
  n <- length(m)
  kernepach <- function(u) 3 / 4 * (1 - u ^ 2) * (abs(u) <= 1)
  qor <- QOR.FUN(p, ...)
  band.width <- 15 ^ (1 / 5) * abs(qor) ^ (2 / 5) / n ^ (1 / 5)
  if (bw.correct) band.width <- min(band.width, p)
  consts <- kernepach((p - (1:n - 1) / n) / band.width)/
    band.width - kernepach((p - (1:n) / n) /
                           band.width) / band.width
  sum(consts * m.sorted)}

qu1 <- qHat(x, 0.25)
qu2 <- qHat(x, 0.5)
qu3 <- qHat(x, 0.75)
ASVar.q1 <- 0.25 * (1 - 0.25) * (qu1 ^ 2)
ASVar.q3 <- 0.75 * (1 - 0.75) * (qu3 ^ 2)
ASCov.q1q3 <- 0.25 * (1- 0.75) * qu1 * qu3
ASVar.IQR <- ASVar.q1 + ASVar.q3 - 2 * ASCov.q1q3
ASVar.q2 <- 0.5 * (1 - 0.5) * (qu2 ^ 2)
ASCov.q3q2 <- 0.5 * (1 - 0.75) * qu3 * qu2
ASCov.q1q2 <- 0.25 * (1 - 0.5) * qu1 * qu2

md <- median(x)
IQR <- IQR(x, na.rm = FALSE, type = 8)
RCV.Q <- 0.75 * (IQR / md)
log.RCV.Q <- log(RCV.Q)


Var.RCV.Q <- (1 / n) * (RCV.Q ^ 2) * (ASVar.IQR / IQR ^ 2 +
```

```
                      ASVar.q2 / md ^ 2 - 2 * (ASCov.q3q2
                      - ASCov.q1q2)/(IQR*md))
Var.log.RCV.Q <- (1/RCV.Q^2) * Var.RCV.Q # Using Delta Method

# CI interval for Rcv
CI.log.RCV.Q <- log.RCV.Q + c(-1,1) * qnorm(0.975) *
  sqrt(Var.log.RCV.Q)
CI.RCV.Q <- exp(CI.log.RCV.Q)

mad <- mad(x, center = median(x), constant = 1.4826, na.rm = FALSE,
    low = FALSE, high = FALSE)

RCV.M <- mad/md
log.RCV.M <- log(RCV.M)

ASV <- function(x, method){
  lambda <- fit.fkml(x, method = method)$lambda
  md <- median(x)
  mad <- mad(x)

  fFinv1 <- dgl(md - mad, lambda1 = lambda)
  fFinv2 <- dgl(md + mad, lambda1 = lambda)
  fFinv3 <- dgl(md, lambda1 = lambda)
  FFinv1 <- pgl(md - mad, lambda1 = lambda)
  FFinv2 <- pgl(md + mad, lambda1 = lambda)

  A <- fFinv1 + fFinv2
  C <- fFinv1 - fFinv2
  B <- C^2 + 4*C*fFinv3*(1 - FFinv2 - FFinv1)

  rho1 <- 1/(4*fFinv3^2) # ASV for the median
  rho2 <- (1/(4 *A^2))*(1 + B/fFinv3^2) # ASV for the MAD
  rho12 <- (1/(4*A*fFinv3))*(1 - 4*FFinv1 + C/fFinv3)
  #ASCOV between median and MAD

  list(rho1 = rho1, rho2 = rho2, rho12 = rho12)}
```

```
  res <- ASV(x, method = "TL")
  ASV.md <- res$rho1
  ASV.MAD <- res$rho2
  ASCov.md.MAD <- res$rho12

  Var.RCV.M <- (1/n) * RCV.M^2 * ((ASV.MAD/mad^2) +
        (ASV.md/md^2) - 2 * (ASCov.md.MAD/(mad * md)))
  Var.log.RCV.M <- (1/RCV.M^2) * Var.RCV.M

  # CI interval for rCV
  CI.log.RCV.M <- log.RCV.M + c(-1,1) * qnorm(0.975) *
    sqrt(Var.log.RCV.M)
  CI.RCV.M <- exp(CI.log.RCV.M)

  list(Est1 = CV, Est2 =RCV.Q, Est3 =RCV.M,
       conf.int1 = CI.CV , conf.int2 = CI.RCV.Q,
       conf.int3 = CI.RCV.M, SE.CV = sqrt(Var.CV),
       SE.RCV.Q = sqrt(Var.RCV.Q), SE.RCV.M = sqrt(Var.RCV.M))}

lamda <- 1
x <- rexp(100, rate=lamda)
dlist <- list(d = dexp, lower = 0, upper = Inf, rate=lamda)

CVRcvrCV(x)
```

A.4.2 Function to calculate true asymptotic variance of CV, RCV_Q and RCV_M

```
moment1 <- function(d, lower = 0, upper = Inf,
            subdivisions = 1000L, ...){
  e1.f <- function(x, ...){
    x*d(x, ...)
  }
  mom1 <- integrate(e1.f, lower = lower, upper = upper,
  subdivisions = subdivisions,  ...)$value
```

```
}
moment2 <- function(d, lower = 0, upper = Inf,
            subdivisions = 1000L, ...){
  e2.f <- function(x, ...){
    x^2*d(x, ...)
  }
  mom2 <- integrate(e2.f, lower = lower, upper = upper,
           subdivisions = subdivisions, ...)$value
}
moment3 <- function(d, lower = 0, upper = Inf,
            subdivisions = 1000L, ...){
  e3.f <- function(x, ...){
    x^3*d(x, ...)
  }
  mom3 <- integrate(e3.f, lower = lower, upper = upper,
           subdivisions = subdivisions, ...)$value
}
moment4 <- function(d, lower = 0, upper = Inf,
            subdivisions = 1000L, ...){
  e4.f <- function(x, ...){
    x^4*d(x, ...)
  }
  mom4 <- integrate(e4.f, lower = lower, upper = upper,
           subdivisions = subdivisions, ...)$value
}

Variance <- function(d, ...){
  mu <- moment1(d, ...)
  e2 <- moment2(d, ...)
  e2 - mu^2
}

cv.f <- function( d, ...){
  mu <- moment1(d, ...)
  V <- Variance(d, ...)
  sqrt(V)/mu
```

```
}

#Lets take E(x - mu)^3 = E(X^3) - 3* mu * E(X^2) + 2 * mu^3 = Mo3
Mo3 <- function( d, ...){
  mu <- moment1(d, ...)
  e2 <- moment2(d, ...)
  e3 <- moment3(d, ...)
  e3 - 3 * mu * e2 + 2 * mu^3
}

#Lets take E(x-mu)^4 = E(X^4) - 4* mu * E(x^3)
#           + 6 * mu^2 * E(X^2) - 3 * mu^4 = Mo4
Mo4 <- function( d, ...){
  mu <- moment1(d, ...)
  e2 <- moment2(d, ...)
  e3 <- moment3(d, ...)
  e4 <- moment4(d, ...)
  e4 - 4 * mu * e3 + 6 * mu^2 * e2 - 3 * mu^4
}

# ASV(CV) = CV^2 * (CV^2 - 1/4) +(1 / (4 * mu^2 * V))
#             * E(x - mu)^4 - (1/mu^3) * E(x-mu)^3
ASV.CV <- function(d, ...){
  cv <- cv.f(d, ...)
  mu <- moment1(d, ...)
  V <- Variance(d, ...)
  M3 <- Mo3(d, ...)
  M4 <- Mo4(d, ...)
  cv^2 * (cv^2 - (1/4)) + (1/(4 * mu^2 * V)) * M4 - (1/mu^3) * M3
}

ASV.CV(dexp, rate = 1)

ASVQ1.f <- function(d, q, a = 0.25 , ...){
  a * (1 - a)  / d(q(a, ...), ...)^2
}
```

```r
ASVQ3.f <- function(d, q, c = 0.75, ...){
  c * (1 - c)/ d(q(c, ...), ...)^2
}
ASCovQ1Q3.f <- function( d, q, a=0.25, c=0.75, ...){
  a ^ 2 /  (d(q(a, ...), ...) * d(q(c, ...), ...))
}
ASV.IQR.f <- function(d, q, ...){
  ASVQ1 <- ASVQ1.f(d, q, ...)
  ASVQ3 <- ASVQ3.f(d, q, ...)
  ASCovQ1Q3 <- ASCovQ1Q3.f(d, q, ...)
  ASVQ1 + ASVQ3 - 2 * ASCovQ1Q3
}
rcv.f <- function(q, a = 0.25, b = 0.5, c = 0.75, ...){
  0.75 * (q(c, ...) - q(a, ...))/q(b, ...)
}
ASVQ2.f <- function(d, q, b = 0.5 , ...){
  b * (1 - b)  / d(q(b, ...), ...)^2
}
ASCovQ1Q2.f <- function(d, q, a=0.25, b=0.5, ...){
  a * (1 - b) / (d(q(a, ...), ...)*d(q(b, ...), ...))
}
ASCovQ3Q2.f <- function(d, q, b=0.5, c=0.75, ...){
  b * (1 - c) /  (d(q(b, ...), ...) * d(q(c, ...), ...))
}
IQR.f <- function(q, a = 0.25, c = 0.75, ...){
  q(c, ...) - q(a, ...)
}
Q2.f <- function(q, b = 0.5, ...){
  q(b, ...)
}

ASV.Rcv_Q <- function(d, q, ...){
  ASV.IQR <- ASV.IQR.f(d, q, ...)
  rcv <- rcv.f(q, ...)
  IQR <- IQR.f(q, ...)
  Q2 <- Q2.f(q, ...)
```

```
  ASVQ2 <- ASVQ2.f(d, q, ...)
  ASCovQ3Q2 <- ASCovQ3Q2.f(d, q, ...)
  ASCovQ1Q2 <- ASCovQ1Q2.f(d, q, ...)
  rcv^2 * ((1 / IQR^2) * ASV.IQR + (1 / Q2^2) * ASVQ2
  - 2 * (ASCovQ3Q2 - ASCovQ1Q2)*(1 / (Q2 * IQR)))
}

ASV.Rcv_Q(dexp, qexp, rate = 1)

mad1 <- function(d, q, param){
  m1 <- do.call(q, c(0.5, param))
  abs.x.m1 <- function(x, d, param, m1){
    do.call(d, c(x = x + m1, param))
     + do.call(d, c(x = - x + m1, param))
  }
  abs.x.m1.vec <- Vectorize(abs.x.m1, "x")
  f <- function(x, d, param, m1){
    integrate(abs.x.m1.vec, lower = 0, upper = x, d = d,
     param = param, m1 = m1)$value - 0.5
  }
  upper <- abs(do.call(q, c(0.75, param)) + m1)
  uniroot(f, interval = c(0, upper), d = d,
   param = param, m1 = m1)$root
}

ASV.Rcv_M <- function(d, q, p, param){
  m <- do.call(q, c(p = 0.5, param))
  mad <- mad1(d, q, param)
  fFinv1 <- do.call(d, c(m - mad, param))
  fFinv2 <- do.call(d, c(m + mad, param))
  fFinv3 <- do.call(d, c(m, param))
  FFinv1 <- do.call(p, c(m - mad, param))
  FFinv2 <- do.call(p, c(m + mad, param))

  A <- fFinv1 + fFinv2
  C <- fFinv1 - fFinv2
```

```
  B <- C^2 + 4*C*fFinv3*(1 - FFinv2 - FFinv1)

  ASV.m <- 1/(4*fFinv3^2) # ASV for the median
  ASV.MAD <- (1/(4 *A^2))*(1 + B/fFinv3^2) # ASV for the MAD
  #ASCOV between median and MAD
  ASCov.m.MAD <- (1/(4*A*fFinv3))*(1 - 4*FFinv1 + C/fFinv3)
  (1.4826*mad/m)^2 * ((ASV.MAD/mad^2) +
   (ASV.m/m^2) - 2 * (ASCov.m.MAD/(mad * m)))
}
ASV.Rcv_M(dexp, qexp, pexp, list(rate=1))
```

A.5 R programs related to paper IV

A.5.1 Function to calculate an estimate, confidence interval and standard error for γ_p, λ_p, and AUC measures

```
library(gdata)
Skew <- function(x, p, conf.level = 0.95){
n <- length(x)
QOR.ln <- function(p) {
  # QOR function for the log-normal
  q.n.0 <- 1 / dnorm(qnorm(p))
  q.n.1 <- qnorm(p) * q.n.0 ^ 2
  q.n.2 <- (1 + 2 * qnorm(p) ^ 2) * q.n.0 ^ 3
  1 / (q.n.0 ^ 2 + 3 * q.n.1 + q.n.2 / q.n.0)
}

# will estimate 1/f(x_p)
qHat<- function(m, p, QOR.FUN = QOR.ln, lambda = NULL,
       hn = 0.15, bw.correct = TRUE, method = "TL", ...){
  m.sorted <- sort(m)
  n <- length(m)
  kernepach <- function(u) 3 / 4 * (1 - u ^ 2) * (abs(u) <= 1)
  qor <- QOR.FUN(p, ...)
  band.width <- 15 ^ (1 / 5) * abs(qor) ^ (2 / 5) / n ^ (1 / 5)
  if (bw.correct) band.width <- min(band.width, p)
```

```
  consts <- kernepach((p - (1:n - 1) / n) / band.width)/
  band.width - kernepach((p - (1:n) / n) / band.width) /
  band.width
  sum(consts * m.sorted)
}

qu1 <- qHat(x, p)
qu2 <- qHat(x, 0.5)
qu3 <- qHat(x, 1-p)
ASVar.q1 <- p * (1 - p) * (qu1 ^ 2)
ASVar.q2 <- 0.5 * (1 - 0.5) * (qu2 ^ 2)
ASVar.q3 <- p * (1 - p) * (qu3 ^ 2)
ASCov.q1q3 <- p ^ 2 * qu1 * qu3
ASCov.q3q2 <- 0.5 * p * qu3 * qu2
ASCov.q1q2 <- p * (1 - 0.5) * qu1 * qu2

ASVar.iqr1 <- ASVar.q1 + ASVar.q3 - 2 * ASCov.q1q3
ASVar.iqr2 <- ASVar.q1 + ASVar.q2 - 2 * ASCov.q1q2
ASV.a <- ASVar.q3 + ASVar.q1 + 4 * ASVar.q2 + 2 * ASCov.q1q3
         - 4 * ASCov.q1q2 - 4 * ASCov.q3q2
ASCov.aiqr1 <- ASVar.q3 - ASVar.q1 - 2 *   ASCov.q3q2
               + 2 * ASCov.q1q2
ASCov.aiqr2 <- ASCov.q3q2 -  ASCov.q1q3 + 3 * ASCov.q1q2
               - ASVar.q1 - 2 * ASVar.q2

q1 <- quantile(x, p, type=8)
q2 <- quantile(x, 0.5, type=8)
q3 <- quantile(x, 1-p, type=8)
iqr1 <- quantile(x, 1-p, type=8) - quantile(x, p, type=8)
iqr2 <- quantile(x, 0.5, type=8) - quantile(x, p, type=8)
a <- q3 + q1 - 2 * q2

gammap <- a / iqr1
log.gammap <- log(gammap)
ASV.gammap <- (gammap) ^ 2  * ((ASV.a /a ^ 2 )
              + (ASVar.iqr1 / iqr1 ^ 2)
```

```
            - (2 * ASCov.aiqr1) / (iqr1 * a))
Var.gammap <- ASV.gammap/n
# using Delata Method
Var.log.gammap <- (1 / gammap ^ 2) * Var.gammap

# CI for Gammap
CI.log.gammap <- log.gammap + c(-1, 1) * qnorm(0.975)
                 * sqrt(Var.log.gammap)
CI.gammap <- exp(CI.log.gammap)

lamdap <- a / iqr2
log.lamdap <- log(lamdap)

ASV.lamdap <- (lamdap) ^ 2  * ((ASV.a /a ^ 2 )
            + (ASVar.iqr2 /iqr2 ^ 2)
            - (2 * ASCov.aiqr2) / (iqr2 * a))
Var.lamdap <- ASV.lamdap/n
# using Delata Method
Var.log.lamdap <- (1 / lamdap ^ 2) * Var.lamdap

CI.log.lamdap <- log.lamdap + c(-1, 1) * qnorm(0.975)
                 * sqrt(Var.log.lamdap)
CI.lamdap <- exp(CI.log.lamdap)

J  <- 100
us <- (1:J - 0.5)/J
Q3 <- quantile(x, 1-us/2)
Q1 <- quantile(x, us/2)
Q2 <- quantile(x, 0.5)
IQR1 <- Q3 - Q1
IQR2 <- Q2 - Q1
A <- Q3 + Q1 - 2 * Q2
Gammap <- A / IQR1
Int.Gammap <- sum(Gammap)/J
log.Int.Gammap <- log(Int.Gammap)
```

```
Qhat <- function(x, v, bw.correct = TRUE, QOR.FUN = QOR.ln, ...){
  v <- c(us/2, 1 - us/2)
  qor <- QOR.FUN(v, ...)
  bw <- 15^(1/5)*abs(qor)^(2/5)/n^(1/5)
  if (bw.correct) bw[v <= bw] <- v[v <= bw]

  kernepach <- function(v) 3/4*(1 - v^2)*(abs(v) <= 1)
  m1 <- matrix(v, nrow = 2*J, ncol = n, byrow = FALSE)
  m2 <- matrix(1:n, nrow = 2*J, ncol = n, byrow = TRUE)

  consts <- kernepach((m1 - (m2 - 1)/n)*(1/bw))*(1/bw) -
    kernepach((m1 - m2/n)*(1/bw))*(1/bw)

  x.sorted <- sort(x)
  q.hat <- c(consts%*%x.sorted)

  q.hat.1 <- q.hat[1:(length(q.hat)/2)]
  q.hat.2 <- q.hat[-(1:length(q.hat)/2)]

  list(q.hat=q.hat, q.hat.1=q.hat.1, q.hat.2=q.hat.2)
}

q2.hat <- qHat(x, 0.5)
res <- Qhat(x)
q.hat  <- res$q.hat
q.hat.1 <- res$q.hat.1
q.hat.2 <- res$q.hat.2

B <- ((us/2) %*% t(1-us/2))
lowerTriangle(B) <- upperTriangle(B, byrow=TRUE)
B

Cov.1mp.1mq <- (1/n) * B * (q.hat.2 %*% t(q.hat.2))
Cov.1mp.q <- (1/n) * ((us/2) %*% t(us/2))
             * (q.hat.2 %*% t(q.hat.1))
Cov.p.1mq <- (1/n) * ((us/2) %*% t(us/2))
```

```
             * (q.hat.1 %*% t(q.hat.2))
Cov.p.q <-  (1/n) * B * (q.hat.1 %*% t(q.hat.1))
Cov.1mp <-  (1/n) * (us/2) * 0.5 * (q.hat.2 %*%
            t(rep(q2.hat, length(us/2))))
Cov.p <- (1/n) * (us/2) * 0.5 * (q.hat.1 %*%
          t(rep(q2.hat, length(us/2))))
Cov.1mq <-  (1/n) * (rep(0.5, length(us/2)) %*% t(us/2))
            * (rep(q2.hat, length(us/2)) %*% t(q.hat.2))
Cov.q <- (1/n) * (rep(0.5, length(us/2)) %*% t(us/2))
          * (rep(q2.hat, length(us/2)) %*% t(q.hat.1))
Var <- (1/n) * 0.5^2 * (rep(q2.hat, length(us/2))
        %*% t(rep(q2.hat, length(us/2))))

Cov.Ap.Aq <- Cov.1mp.1mq + Cov.1mp.q + Cov.p.1mq + Cov.p.q
             - 2 * Cov.1mp - 2 * Cov.p - 2 * Cov.1mq
             - 2 * Cov.q + 4 * Var
Cov.Ap.IQR1q <- Cov.1mp.1mq - Cov.1mp.q + Cov.p.1mq
                - Cov.p.q - 2 * Cov.1mq + 2 *  Cov.q
Cov.IQR1p.Aq <- Cov.1mp.1mq + Cov.1mp.q - Cov.p.1mq
                - Cov.p.q - 2 * Cov.1mp + 2 * Cov.p
Cov.IQR1p.IQR1q <- Cov.1mp.1mq - Cov.1mp.q - Cov.p.1mq + Cov.p.q

rc <- matrix(Gammap, ncol = J, nrow = J, byrow = FALSE)

Cov.Gp.Gq <- ((1/IQR1) %*% t(1/IQR1)) * (Cov.Ap.Aq
             -  Cov.Ap.IQR1q * t(rc) -  Cov.IQR1p.Aq * rc
             + (Gammap %*% t(Gammap)) * Cov.IQR1p.IQR1q)
sum(Cov.Gp.Gq)

Var.int.Gammap <- (1/J^2) *  sum(Cov.Gp.Gq)
log.Var.int.Gammap <- (1/Int.Gammap^2) * Var.int.Gammap

# CI for integrated Gamma_P
CI.log.Int.Gammap <- log.Int.Gammap + c(-1, 1) * qnorm(0.975)
                     * sqrt(log.Var.int.Gammap)
CI.Int.Gammap  <- exp(CI.log.Int.Gammap)
```

```
Lambdap <- A / IQR2
Int.Lambdap <- sum(Lambdap)/J
log.Int.Lambdap <- log(Int.Lambdap)

Cov.Ap.IQR2q <- Cov.1mp - Cov.1mp.q + Cov.p
                - Cov.p.q + 2 *  Cov.q  - 2 * Var
Cov.IQR2p.Aq <- Cov.1mq + Cov.q - Cov.p.1mq - Cov.p.q
                + 2 * Cov.p - 2 * Var
Cov.IQR2p.IQR2q <- Cov.p.q - Cov.p - Cov.q + Var

lc <- matrix(Lambdap, ncol = J, nrow = J, byrow = FALSE)

Cov.Lp.Lq <- ((1/IQR2) %*% t(1/IQR2)) * (Cov.Ap.Aq
             -  Cov.Ap.IQR2q * t(lc) -  Cov.IQR2p.Aq * lc
             + (Lambdap %*% t(Lambdap)) * Cov.IQR2p.IQR2q)
sum(Cov.Lp.Lq)

Var.int.Lambdap <- (1/J^2) *  sum(Cov.Lp.Lq)
Var.log.int.Lambdap <- (1/Int.Lambdap^2) * Var.int.Lambdap

CI.log.Int.Lambdap <- log.Int.Lambdap + c(-1, 1) *
           qnorm(0.975) * sqrt(Var.log.int.Lambdap)
CI.Int.Lambdap <- exp(CI.log.Int.Lambdap)

p.Gammap <- (us/2) * (A / IQR1)
Int.p.Gammap <- sum(p.Gammap)/J
log.Int.p.Gammap <- log(Int.p.Gammap)

Cov.p.Gp.p.Gq <- ((us/2) %*% t(us/2)) *  Cov.Gp.Gq
sum(Cov.p.Gp.p.Gq)

Var.int.p.Gammap <- (1/J^2) *  sum(Cov.p.Gp.p.Gq)
Var.log.int.p.Gammap <- (1/Int.p.Gammap^2) * Var.int.p.Gammap

# CI for integrated Gamma_P
```

```r
CI.log.Int.p.Gammap <- log.Int.p.Gammap + c(-1, 1)
        * qnorm(0.975) * sqrt(Var.log.int.p.Gammap)
CI.Int.p.Gammap <- exp(CI.log.Int.p.Gammap)

p.Lambdap <- (us/2) * Lambdap
Int.p.Lambdap <- sum(p.Lambdap)/J
log.Int.p.Lambdap <- log(Int.p.Lambdap)

Cov.p.Lp.p.Lq <- ((us/2) %*% t(us/2)) * Cov.Lp.Lq
sum(Cov.p.Lp.p.Lq)

Var.int.p.Lambdap <- (1/J^2) *  sum(Cov.p.Lp.p.Lq)
Var.log.int.p.Lambdap <- (1/Int.p.Lambdap^2) * Var.int.p.Lambdap

# CI for integrated Gamma_P
CI.log.Int.p.Lambdap <- log.Int.p.Lambdap + c(-1, 1) *
              qnorm(0.975) * sqrt(Var.log.int.p.Lambdap)
CI.Int.p.Lambdap <- exp(CI.log.Int.p.Lambdap)

list(Est1 = gammap, Est2 =lamdap, Est3 =Int.Gammap,
     Est4 =Int.Lambdap, Est5 =Int.p.Gammap,
     Est6 =Int.p.Lambdap, conf.int1 = CI.gammap ,
     conf.int2 = CI.lamdap, conf.int3 = CI.Int.Gammap,
     conf.int4 = CI.Int.Lambdap, conf.int5 = CI.Int.p.Gammap,
     conf.int6 = CI.Int.p.Lambdap, SE.gammap = sqrt(Var.gammap),
     SE.lamdap=sqrt(Var.lamdap),
     SE.int.Gammap = sqrt(Var.int.Gammap),
     SE.int.Lambdap = sqrt(Var.int.Lambdap),
     SE.int.p.Gammap = sqrt(Var.int.p.Gammap),
     SE.int.p.Lambdap =sqrt(Var.int.p.Lambdap))
}

lamda <- 1
x <- rexp(100, rate=lamda)
Skew(x, 0.2)
```

A.5.2 Function to calculate true asymptotic variance of γ_p, λ_p and AUC measures

```
ASV.Skew <- function(d, q, p, param){
ASVar.q1 <- p * (1 - p)/do.call(d, c(do.call(q,
            c(p, param)), param))^2
ASVar.q2 <- 0.5 * (1 - 0.5) / do.call(d, c(do.call(q,
            c(0.5, param)), param))^2
ASVar.q3 <- p * (1 - p) / do.call(d, c(do.call(q,
            c(1-p, param)), param))^2
ASCov.q1q3 <- p ^ 2 / do.call(d, c(do.call(q, c(p, param)),
     param))* do.call(d, c(do.call(q, c(p, param)), param))
ASCov.q3q2 <- 0.5 * p / do.call(d, c(do.call(q, c(0.5, param)),
        param)) * do.call(d, c(do.call(q, c(p, param)), param))
ASCov.q1q2 <- p * (1 - 0.5) /do.call(d, c(do.call(q, c(p, param)),
       param)) * do.call(d, c(do.call(q, c(0.5, param)), param))

ASVar.IQR1 <- ASVar.q1 + ASVar.q3 - 2 * ASCov.q1q3
ASV.A <- ASVar.q3 + ASVar.q1 + 4 * ASVar.q2 + 2 * ASCov.q1q3
         - 4 * ASCov.q1q2 - 4 * ASCov.q3q2
ASCov.AIQR1 <- ASVar.q3 - ASVar.q1 - 2 * ASCov.q3q2
               + 2 * ASCov.q1q2

q1 <- do.call(q, c(p, param))
q2 <- do.call(q, c(0.5, param))
q3 <- do.call(q, c(1-p, param))
IQR1 <- q3 - q1
A <- q3 + q1 - 2 * q2
Gammap <- A / IQR1

ASV.Gammap <- (Gammap) ^ 2  * ((ASV.A / A ^ 2)
               + (ASVar.IQR1 / IQR1 ^ 2)
               - (2 * ASCov.AIQR1) / (IQR1 * A))
ASVar.IQR2 <- ASVar.q1 + ASVar.q2 - 2 * ASCov.q1q2
ASV.A <- ASVar.q3 + ASVar.q1 + 4 * ASVar.q2 + 2 * ASCov.q1q3
         - 4 * ASCov.q1q2 - 4 * ASCov.q3q2
```

```r
  ASCov.AIQR2 <- ASCov.q3q2 - ASCov.q1q3 + 3 * ASCov.q1q2
                 - ASVar.q1 - 2 * ASVar.q2

  IQR2 <- q2 - q1
  Lamdap <- A / IQR2

  ASV.Lamdap <- (Lamdap) ^ 2  * ((ASV.A /A ^ 2 )
                 + (ASVar.IQR2 /IQR2 ^ 2)
                 - (2 * ASCov.AIQR2) / (IQR2 * A))

   J <- 100
  us <- (1:J - 0.5)/J

  S <- matrix(0, J, J)
  S1 <- matrix(0, J, J)
  S2 <- matrix(0, J, J)
  S3 <- matrix(0, J, J)

  for(j in 1:J){
    for(k in 1:J){
      if (us[j]/2 < us[k]/2){
        A <- (us[j]/2) * (1-us[k]/2)
      }
      else A <- (1-us[j]/2) * (us[k]/2)

Cov.1mp.1mq <- A / (do.call(d, c(do.call(q, c(1-us[j]/2,
                param)), param)) * do.call(d, c(do.call(q,
                c(1-us[k]/2, param)), param)))
Cov.1mp.q <- (us[j]/2) * (us[k]/2) / (do.call(d, c(do.call(q,
              c(1-us[j]/2, param)), param)) * do.call(d,
              c(do.call(q, c(us[k]/2, param)), param)))
Cov.p.1mq <- (us[j]/2) * (us[k]/2) / (do.call(d, c(do.call(q,
              c(us[j]/2, param)), param)) * do.call(d,
              c(do.call(q, c(1-us[k]/2, param)), param)))
Cov.p.q <-  A / (do.call(d, c(do.call(q, c(us[j]/2, param)),
            param)) * do.call(d, c(do.call(q,
```

```
             c(us[k]/2, param)), param)))
Cov.1mp <-  0.5 * (us[j]/2) / (do.call(d, c(do.call(q,
             c(1-us[j]/2, param)), param)) * do.call(d,
             c(do.call(q, c(0.5, param)), param)))
Cov.p <-  0.5 * (us[j]/2) / (do.call(d, c(do.call(q,
           c(us[j]/2, param)), param)) * do.call(d,
           c(do.call(q, c(0.5, param)), param)))
Cov.1mq <-  0.5 * (us[k]/2) / (do.call(d, c(do.call(q,
             c(1-us[k]/2, param)), param)) * do.call(d,
             c(do.call(q, c(0.5, param)), param)))
Cov.q <- 0.5 * (us[k]/2) / (do.call(d, c(do.call(q,
          c(us[k]/2, param)), param)) * do.call(d,
          c(do.call(q, c(0.5, param)), param)))
Var <- (0.5)^2 / do.call(d, c(do.call(q,
        c(0.5, param)), param))^2

Cov.Ap.Aq <- Cov.1mp.1mq + Cov.1mp.q + Cov.p.1mq + Cov.p.q
              - 2 * Cov.1mp - 2 * Cov.p - 2 * Cov.1mq
              - 2 * Cov.q + 4 * Var
Cov.Ap.IQR1q <- Cov.1mp.1mq - Cov.1mp.q + Cov.p.1mq - Cov.p.q
                 - 2 * Cov.1mq + 2 * Cov.q
Cov.IQR1p.Aq <- Cov.1mp.1mq + Cov.1mp.q - Cov.p.1mq - Cov.p.q
                 - 2 * Cov.1mp + 2 * Cov.p
Cov.IQR1p.IQR1q <- Cov.1mp.1mq - Cov.1mp.q - Cov.p.1mq + Cov.p.q

IQR.p <-  do.call(q, c(1-us[j]/2, param))
           - do.call(q, c(us[j]/2, param))
IQR.q <-  do.call(q, c(1-us[k]/2, param))
           - do.call(q, c(us[k]/2, param))

Gamma.p <- (do.call(q, c(1-us[j]/2, param))
            + do.call(q, c(us[j]/2, param))
            - 2 * do.call(q, c(0.5, param)))/
            (do.call(q, c(1-us[j]/2, param))
            - do.call(q, c(us[j]/2, param)))
Gamma.q <- (do.call(q, c(1-us[k]/2, param))
```

```
          + do.call(q, c(us[k]/2, param))
          - 2 * do.call(q, c(0.5, param)))/
          (do.call(q, c(1-us[k]/2, param))
          - do.call(q, c(us[k]/2, param)))

Cov.Ap.IQR2q <- Cov.1mp - Cov.1mp.q + Cov.p
                - Cov.p.q + 2 *  Cov.q  - 2 * Var
Cov.IQR2p.Aq <- Cov.1mq + Cov.q - Cov.p.1mq - Cov.p.q
                + 2 * Cov.p - 2 * Var
Cov.IQR2p.IQR2q <- Cov.p.q - Cov.p - Cov.q + Var

IQR2.p <- do.call(q, c(0.5, param))
         - do.call(q, c(us[j]/2, param))
IQR2.q <- do.call(q, c(0.5, param))
         - do.call(q, c(us[k]/2, param))

Lamda.p <- (do.call(q, c(1-us[j]/2, param))
          + do.call(q, c(us[j]/2, param))
          - 2 * do.call(q, c(0.5, param)))/
          (do.call(q, c(0.5, param))
          - do.call(q, c(us[j]/2, param)))
Lamda.q <- (do.call(q, c(1-us[k]/2, param))
          + do.call(q, c(us[k]/2, param))
          - 2 * do.call(q, c(0.5, param)))/
          (do.call(q, c(0.5, param))
          - do.call(q, c(us[k]/2, param)))

S[j, k] <- (1/IQR.p) * (1/IQR.q) * (Cov.Ap.Aq
          -  Gamma.q * Cov.Ap.IQR1q - Gamma.p *
          Cov.IQR1p.Aq + Gamma.p * Gamma.q * Cov.IQR1p.IQR1q)
S1[j, k] <- (us[j]/2) * (us[k]/2) * S[j, k]
S2[j, k] <- (1/IQR2.p) * (1/IQR2.q) * (Cov.Ap.Aq
           -  Lamda.q * Cov.Ap.IQR2q
           - Lamda.p * Cov.IQR2p.Aq
           + Lamda.p * Lamda.q * Cov.IQR2p.IQR2q)
S3[j, k] <- (us[j]/2) * (us[k]/2) * S2[j, k]
```

```
    }
  }
  ASVar.int.Gammap <- (1/J^2) * sum(S)
  ASVar.int.p.Gammap <- (1/J^2) * sum(S1)
  ASVar.int.Lamdap <- (1/J^2) * sum(S2)
  ASVar.int.p.Lamdap <- (1/J^2) * sum(S3)

  list (ASV.Gp=ASV.Gammap, ASV.Lp=ASV.Lamdap,
  ASV.int.Gp= ASVar.int.Gammap,
  ASV.int.pGp=ASVar.int.p.Gammap,
  ASV.int.Lp=ASVar.int.Lamdap,
  ASV.int.pLp=ASVar.int.p.Lamdap)
}
ASV.Skew(dlnorm, qlnorm, p=0.2, list(meanlog=1, sdlog=1))
```

Bibliography

AGRESTI, ALAN. 1990. *Categorical data analysis*. Vol. 482. John Wiley & Sons.

AHN, K. 1995. On the use of coefficient of variation for uncertainty analysis in fault tree analysis. *Reliability Engineering & System Safety*, **47**(3), 229–230.

ALBATINEH, AN, KIBRIA, BMG, WILCOX, ML, & ZOGHEIB, B. 2014. Confidence interval estimation for the population coefficient of variation using ranked set sampling: a simulation study. *Journal of Applied Statistics*, **41**(4), 733–751.

AMIRI, S, & ZWANZIG, S. 2010. An improvement of the nonparametric bootstrap test for the comparison of the coefficient of variations. *Communications in StatisticsâSimulation and Computation®*, **39**(9), 1726–1734.

ANANTHAKRISHNAN, R, & SOMAN, MK. 1989. Statistical distribution of daily rainfall and its association with the coefficient of variation of rainfall series. *International Journal of Climatology*, **9**(5), 485–500.

ANSARI, AR, BRADLEY, RA, *et al.* 1960. Rank-sum tests for dispersions. *The Annals of Mathematical Statistics*, **31**(4), 1174–1189.

ARACHCHIGE, CNPG, & PRENDERGAST, LA. 2019. Confidence intervals for median absolute deviations. *arXiv preprint arXiv:1910.00229*.

ARACHCHIGE, CNPG, CAIRNS, M, & PRENDERGAST, LA. 2019a. Interval estimators for ratios of independent quantiles and interquantile ranges. *Communications in Statistics-Simulation and Computation*, 1–17.

ARACHCHIGE, CNPG, PRENDERGAST, LA, & STAUDTE, RG. 2019b. Robust analogues to the coefficient of variation. *arXiv preprint arXiv:1907.01110*.

ARKIN, H, & COLTON, RR. 1935. Statistical methods. barnes and noble. *Inc., New York*, **50**.

ARNOLD, BC, & GROENEVELD, RA. 1995. Measuring skewness with respect to the mode. *The American Statistician*, **49**(1), 34–38.

ATKINSON, AB. 1970. On the measurement of inequality. *Journal of Economic Theory*, **2**(3), 244–263.

BAHADUR, RR. 1966. A note on quantiles in large samples. *The Annals of Mathematical Statistics*, **37**(3), 577–580.

BALAKRISHNAN, N, & MA, CW. 1990. A comparative study of various tests for the equality of two opulation variances. *Journal of Statistical Computation and Simulation*, **35**(1-2), 41–89.

BANIK, S, & KIBRIA, BMG. 2011. Estimating the population coefficient of variation by confidence intervals. *Communications in Statistics-Simulation and Computation*, **40**(8), 1236–1261.

BENJAMINI, Y, & KRIEGER, AM. 1996. Concepts and measures for skewness with data-analytic implications. *Canadian Journal of Statistics*, **24**(1), 131–140.

BISSELL, AF, & FERGUSON, RA. 1975. The jackknife–toy, tool or two-edged weapon? *Journal of the Royal Statistical Society: Series D (The Statistician)*, **24**(2), 79–100.

BOFINGER, E. 1975. Estimation of a density function using order statistics 1. *Australian Journal of Statistics*, **17**(1), 1–7.

BONETT, DG. 2006. Confidence interval for a coefficient of quartile variation. *Computational Statistics & Data Analysis*, **50**(11), 2953–2957.

BONETT, DG, & SEIER, E. 2003. Confidence intervals for mean absolute deviations. *The American Statistician*, **57**(4), 233–236.

BONETT, DG, & SEIER, E. 2005. Confidence interval for a coefficient of dispersion in nonnormal distributions. *Biometrical Journal*, **47**(1), 144–148.

BOSHNAKOV, GN. 2003. Confidence characteristics of distributions. *Statistics & Probability Letters*, **63**(4), 353–360.

BOSHNAKOV, GN. 2007. Some measures for asymmetry of distributions. *Statistics & Probability Letters*, **77**(11), 1111–1116.

BOWLEY, AL. 1920. *Elements of statistics*. Vol. 2. PS King.

BOX, GEP, & ANDERSEN, SL. 1955. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society: Series B (Methodological)*, **17**(1), 1–26.

BOYLE, GW, & RAO, RKS. 1988. The mean-generalized coefficient of variation selection rule and expected utility maximization. *Southern Economic Journal*, 1–8.

BRIEF, RP, & OWEN, J. 1969. A note on earnings risk and the coefficient of variation. *The Journal of Finance*, **24**(5), 901–904.

BROWN, MB, & FORSYTHE, AB. 1974. Robust tests for the equality of variances. *Journal of the American Statistical Association*, **69**(346), 364–367.

BRYS, G, HUBERT, M, & STRUYF, A. 2003. A comparison of some new measures of skewness. *Pages 98–113 of: Developments in robust statistics*. Springer.

BRYS, G, HUBERT, M, & STRUYF, A. 2004. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, **13**(4), 996–1017.

BULENT, A, & HAMZA, G. 2018. Bootstrap confidence intervals for the coefficient of quartile variation. *Communications in Statistics - Simulation and Computation*, **0**(0), 1–9.

CABRAS, S, MOSTALLINO, G, & RACUGNO, W. 2006. A nonparametric bootstrap test for the equality of coefficients of variation. *Communications in Statistics-Simulation and Computation*, **35**(3), 715–726.

CAMPBELL, NA. 1978. The influence function as an aid in outlier detection in discriminant analysis. *Applied Statistics*, 251–258.

CHANG, W, CHENG, J, ALLAIRE, JJ, XIE, Y, & MCPHERSON, J. 2017. *shiny: Web application framework for r*. R package version 1.0.5.

CHARLIER, CVL. 1905. *Uber das fehlergesetz, uppsala*.

CHATURVEDI, A, & RANI, U. 1996. Fixed-width confidence interval estimation of the inverse coefficient of variation in a normal population. *Microelectronics Reliability*, **36**(9), 1305–1308.

CHEN, J, & FLEISHER, BM. 1996. Regional income inequality and economic growth in china. *Journal of Comparative Economics*, **22**(2), 141–164.

CHENG, M-Y, SUN, S, *et al.* 2006. Bandwidth selection for kernel quantile estimation. *Journal of Chinese Statistical Association, 44*, 271–295.

CHU, JT, *et al.* 1957. Some uses of quasi-ranges. *The Annals of Mathematical Statistics*, **28**(1), 173–180.

CHUNG, N, ZHANG, XD, KREAMER, A, LOCCO, L, KUAN, P-F, BARTZ, S, LINSLEY, PS, FERRER, M, & STRULOVICI, B. 2008. Median absolute deviation to improve hit selection for genome-scale rnai screens. *Journal of Biomolecular Screening*, **13**(2), 149–158.

CLARKE, BR. 2018. *Robustness theory and application*. John Wiley & Sons.

COLE, TJ, BELLIZZI, MC, FLEGAL, KM, & DIETZ, WH. 2000. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ*, **320**(7244), 1240.

CONOVER, WJ, & CONOVER, WJ. 1980. Practical nonparametric statistics.

CONOVER, WJ, JOHNSON, ME, & JOHNSON, MM. 1981. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, **23**(4), 351–361.

CRITCHLEY, F, & JONES, MC. 2008. Asymmetry and gradient asymmetry functions: Density-based skewness and kurtosis. *Scandinavian Journal of Statistics*, **35**(3), 415–437.

CRITCHLEY, F, & VITIELLO, C. 1991. The influence of observations on misclassification probability estimates in linear discriminant analysis. *Biometrika*, **78**(3), 677–690.

CROISSANT, Y. 2016. *Ecdat: data sets for econometrics. r package version 0.3-1.*

CURTO, JD, & PINTO, JC. 2009. The coefficient of variation asymptotic distribution in the case of non-iid random variables. *Journal of Applied Statistics*, **36**(1), 21–32.

D'ALVISE, R, MOLONEY, C, & RAMPONI, G. 1999. Statistical methods for the evaluation of the coefficient of variation in sar images. *Pages 1253–1255 of: Ieee 1999 international geoscience and remote sensing symposium. igarss'99 (cat. no. 99ch36293)*, vol. 2. IEEE.

DASGUPTA, A. 2006a. *Asymptotic theory of statistics and probability*. Springer.

DASGUPTA, A. 2006b. *Asymptotic Theory of Statistics and Probability*. New York, NY: Springer.

DAVID, FN, & JOHNSON, NL. 1956. Some tests of significance with ordered variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **18**(1), 1–31.

DAVID, HA. 1981. *Order statistics*. John Wiley & Sons.

DAVID, HA. 1998. Early sample measures of variability. *Statistical Science*, 368–377.

DE, P, GHOSH, JB, & WELLS, CE. 1996. Scheduling to minimize the coefficient of variation. *International Journal of Production Economics*, **44**(3), 249–253.

DEVELOPMENT CORE TEAM, R. 2018. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

DRAPER, NR, & HUNTER, WG. 1969. Transformations: some examples revisited. *Technometrics*, **11**(1), 23–40.

EDGEWORTH, FY. 1908. The law of error. part i. *Transactions of the Cambridge Philosophical Society*, **20**, 36.

EDWARD MILLER, G. 1991. Asymptotic test statistics for coefficients of variation. *Communications in Statistics-Theory and Methods*, **20**(10), 3351–3363.

EFRON, B, & TIBSHIRANI, RJ. 1994. *An introduction to the bootstrap*. CRC Press.

EISENBERG, DT. 2016. Telomere length measurement validity: the coefficient of variation is invalid and cannot be used to compare quantitative polymerase chain reaction and southern blot telomere length measurement techniques. *International Journal of Epidemiology*, **45**(4), 1295–1298.

EPANECHNIKOV, VA. 1969. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, **14**(1), 153–158.

FALK, M. 1986. On the estimation of the quantile density function. *Statistics & Probability Letters*, **4**(2), 69–73.

FALK, M. 1997. Asymptotic independence of median and mad. *Statistics & Probability Letters*, **34**(4), 341–345.

FIELLER, EC. 1932. A numerical test of the adequacy of at mckay's approximation. *Journal of the Royal Statistical Society*, **95**(4), 699–702.

FISHER, RA. 1922. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **222**(594-604), 309–368.

FLIGNER, MA, & KILLEEN, TJ. 1976. Distribution-free two-sample tests for scale. *Journal of the American Statistical Association*, **71**(353), 210–213.

GASTWIRTH, JL. 1982. Statistical properties of a measure of tax assessment uniformity. *Journal of Statistical Planning and Inference*, **6**(1), 1–12.

GASTWIRTH, JL. 1988. *Statistical reasoning in law and public policy: Tort law, evidence and health*. Vol. 2. Elsevier.

GAUSS, CF. 1816. *Demonstratio nova altera theorematis omnem functionem algebraicam*. Apvd Henricvm Dieterich.

GAYEN, AK. 1950. The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika*, **37**(3/4), 236–255.

GBÜNEBERG, H, BAINS, GS, BEBBY, BJ, BILES, L, SMITH, CAB, WEISS, BA, *et al.* 1966. A search for genetic effects of high natural radioactivity in south india. *A Search for Genetic Effects of High Natural Radioactivity in South India.*

GEARY, RC. 1947. Testing for normality. *Biometrika*, **34**(3/4), 209–242.

GENG, S, WANG, WJ, & MILLER, C. 1979. Small sample size comparisons of tests for homogeneity of variances by monte-carlo: Small sample size comparisons of tests. *Communications in Statistics-Simulation and Computation*, **8**(4), 379–389.

GONG, J, & LI, Y. 1999. Relationship between the estimated weibull modulus and the coefficient of variation of the measured strength for ceramics. *Journal of the American Ceramic Society*, **82**(2), 449–452.

GROENEVELD, RA. 1991. An influence function approach to describing the skewness of a distribution. *The American Statistician*, **45**(2), 97–102.

GROENEVELD, RA. 2011a. Influence functions for the coefficient of variation, its inverse, and cv comparisons. *Communications in Statistics - Theory and Methods*, **40**(23), 4139–4150.

GROENEVELD, RA. 2011b. Influence functions for the coefficient of variation, its inverse, and cv comparisons. *Communications in Statistics-Theory and Methods*, **40**(23), 4139–4150.

GROENEVELD, RA, & MEEDEN, G. 1984. Measuring skewness and kurtosis. *The Statistician*, 391–399.

GROENEVELD, RA, MEEDEN, G, *et al.* 2009. An improved skewness measure. *Metron*, **67**(3), 325.

GULHAR, M, KIBRIA, G, ALBATINEH, A, & AHMED, NU. 2012. A comparison of some confidence intervals for estimating the population coefficient of variation: a simulation study. *Sort*, **36**(1).

HALL, IJ. 1972. Some comparisons of tests for equality of variances. *Journal of Statistical Computation and Simulation*, **1**(2), 133–194.

HALL, P, & PADMANABHAN, AR. 1997. Adaptive inference for the two-sample scale problem. *Technometrics*, **39**(4), 412–422.

HALL, P, & WILSON, SR. 1991. Two guidelines for bootstrap hypothesis testing. *Biometrics*, 757–762.

HAMER, AJ, STRACHAN, JR, BLACK, MM, IBBOTSON, C, & ELSON, RA. 1995. A new method of comparative bone strength measurement. *Journal of Medical Engineering & Technology*, **19**(1), 1–5.

HAMPEL, FR. 1968. Contribution to the theory of robust estimation. *Ph. D. Thesis, University of California, Berkeley*.

HAMPEL, FR. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**(346), 383–393.

HAMPEL, FR, RONCHETTI, EM, ROUSSEEUW, PJ, & STAHEL, WA. 1986. *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

HENDRICKS, WA, & ROBEY, KW. 1936. The sampling distribution of the coefficient of variation. *The Annals of Mathematical Statistics*, **7**(3), 129–132.

HETTMANSPERGER, TP. 1984. Two-sample inference based on one-sample sign statistics. *Applied Statistics*, 45–51.

HILLIER, FS, & SO, KC. 1991. The effect of the coefficient of variation of operation times on the allocation of storage space in production line systems. *IIE Transactions*, **23**(2), 198–206.

HINKLEY, DV. 1975. On power transformations to symmetry. *Biometrika*, **62**(1), 101–111.

HODGES JR, JOSEPH L, & LEHMANN, EL. 1963. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 598–611.

HOLLANDER, M, & WOLFE, DA. 1999. Nonparametric statistical methods.

HOSKEN, DJ, BUSS, DL, & HODGSON, DJ. 2018. Beware the f test (or, how to compare variances). *Animal Behaviour*, **136**, 119–126.

HÖSSJER, O, *et al.* 1996. Incomplete generalized l-statistics. *The Annals of Statistics*, **24**(6), 2631–2654.

HUBER, PJ. 1981. *Robust statistics*. Wiley.

HYNDMAN, RJ, & FAN, Y. 1996. Sample quantiles in statistical packages. *The American Statistician*, **50**(4), 361–365.

IGLEWICZ, B. 1967. Some properties of the sample coefficient of variation.

IGLEWICZ, B, & MYERS, RH. 1970. Comparisons of approximations to the percentage points of the sample coeffcient of variation. *Technometrics*, **12**(1), 166–169.

JONES, MC. 1992. Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics*, **44**(4), 721–727.

JUREČKOVÁ, J, PICEK, J, & SCHINDLER, M. 2019. *Robust statistical methods with r*. Chapman and Hall/CRC.

KAPLAN, EL, & MEIER, P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457–481.

KIBRIA, BMG. 2006. Modified confidence intervals for the mean of the asymmetric distribution. *Pakistan Journal Of Statistics-All Series-*, **22**(2), 109.

KLOTZ, J, *et al.* 1962. Nonparametric tests for scale. *The Annals of Mathematical Statistics*, **33**, 498–512.

KOHL, MATTHIAS, & KOHL, MAINTAINER MATTHIAS. 2019. Package âmkmiscâ.

KOOPMANS, LH, OWEN, DB, & ROSENBLATT, JI. 1964. Confidence intervals for the coefficient of variation for the normal and log normal distributions. *Biometrika*, **51**(1/2), 25–32.

LAX, DA. 1985. Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, **80**(391), 736–741.

LEHMANN, EL, & D'ABRERA, HJ. 1975. *Nonparametrics: statistical methods based on ranks.* Holden-Day.

LEVENE, H. 1961. Robust tests for equality of variances. *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*, 279–292.

LEYS, C, LEY, C, KLEIN, O, BERNARD, P, & LICATA, L. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, **49**(4), 764–766.

LIM, T-S, & LOH, W-Y. 1996. A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, **22**(3), 287–301.

LOHRDING, RK. 1975. A two sample test of equality of coefficients of variation or relative errors. *Journal of Statistical Computation and Simulation*, **4**(1), 31–36.

LOVITT, WV, & HOLTZCLAW, HF. 1929. *Statistics: By william vernon lovitt and henry f. holtzclaw*. Prentice-Hall, Incorporated.

MA, S, & ZHANG, X. 1991. Some rules of variation coefficient of annual rainfall in xinjiang. *ACTA Meteorol SINICA*, **49**(1), 39–45.

MACGILLIVRAY, HL, *et al.* 1986. Skewness and asymmetry: measures and orderings. *The Annals of Statistics*, **14**(3), 994–1011.

MAHMOUDVAND, R, & HASSANI, H. 2009. Two new confidence intervals for the coefficient of variation in a normal distribution. *Journal of Applied Statistics*, **36**(4), 429–442.

MANIYA, GM. 1961. Remarks on nonparametric estimates of a bivariate probability density. *Soobscenija Akademii Nauk Gruzinskoj SSR*, **27**(4), 385–400.

MAROZZI, M. 2011. Levene type tests for the ratio of two scales. *Journal of Statistical Computation and Simulation*, **81**(7), 815–826.

MAROZZI, M. 2012. A combined test for differences in scale based on the interquantile range. *Statistical Papers*, **53**(1), 61–72.

MCGIBNEY, G, & SMITH, MR. 1993. An unbiased signal-to-noise ratio measure for magnetic resonance images. *Medical Physics*, **20**(4), 1077–1078.

MCKAY, AT. 1932. Distribution of the coefficient of variation and the extended" t" distribution. *Journal of the Royal Statistical Society*, **95**(4), 695–698.

MCKEAN, JW, & SCHRADER, RM. 1984. A comparison of methods for studentizing the sample median. *Communications in Statistics-Simulation and Computation*, **13**(6), 751–773.

MEEDEN, G. 1993. Noninformative nonparametric bayesian estimation of quantiles. *Statistics & Probability Letters*, **16**(2), 103–109.

MILLER, EG, & KARSON, MJ. 1977. Testing equality of two coefficients of variation. *Pages 278–283 of: American statistical association: Proceedings of the business and economics section, part i*, vol. 95.

MILLER, RG. 1968. Jackknifing variances. *The Annals of Mathematical Statistics*, **39**(2), 567–582.

MILLER, RG. 1974. The jackknife-a review. *Biometrika*, **61**(1), 1–15.

MILLER JR, RUPERT G. 1997. *Beyond anova: basics of applied statistics*. Chapman and Hall/CRC.

MISHRA, SK, *et al.* 2008. A new method of robust linear regression analysis: some monte carlo experiments. *Journal of Applied Economic Sciences (JAES)*, **3**(03), 261–269.

MOOD, AM. 1950. Introduction to the theory of statistics.

MOOD, AM. 1954. On the asymptotic efficiency of certain nonparametric two-sample tests. *The Annals of Mathematical Statistics*, 514–522.

MOORE, DS, & YACKEL, JW. 1977. Consistency properties of nearest neighbor density function estimators. *The Annals of Statistics*, 143–154.

MOSES, LE. 1963. Rank tests of dispersion. *The Annals of Mathematical Statistics*, 973–983.

NADARAYA, EA. 1964. Estimation of a bivariate probability density. *Akademiia Nauk Guzinskoi SSR Soobscheniia*, **36**, 267–268.

NG, CK. 2006. Performance of three methods of interval estimation of the coefficient of variation. *InterStat*.

NGATCHOU-WANDJI, J. 2006. On testing for the nullity of some skewness coefficients. *International Statistical Review*, **74**(1), 47–65.

NORRIS, N. 1938. Some efficient measures of relative dispersion. *The Annals of Mathematical Statistics*, **9**(3), 214–220.

O'BRIEN, RALPH G. 1979. A general anova method for robust tests of additive models for variances. *Journal of the American Statistical Association*, **74**(368), 877–880.

O'BRIEN, RALPH G. 1981. A simple test for variance effects in experimental designs. *Psychological Bulletin*, **89**(3), 570.

OJA, H. 1981. On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics*, 154–168.

PAN, G. 1999. On a levene type test for equality of two variances. *Journal of Statistical Computation and Simulation*, **63**(1), 59–71.

PANG, W-K, LEUNG, P-K, HUANG, W-K, & LIU, W. 2005. On interval estimation of the coefficient of variation for the three-parameter weibull, lognormal and gamma distribution: A simulation-based approach. *European Journal of Operational Research*, **164**(2), 367–377.

PANICHKITKOSOLKUL, W. 2009. Improved confidence intervals for a coefficient of variation of a normal distribution. *Thailand Statistician*, **7**(2), 193–199.

PAPANICOLAOU, ALEX. 2009. Taylor approximation and the delta method. *April*, **28**, 2009.

PARZEN, E. 1962. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**(3), 1065–1076.

PARZEN, E. 1977. *Nonparametric statistical data science: A unified approach based on density estimation and testing for'white noise'.* Tech. rept. STATE UNIV OF NEW YORK AT BUFFALO AMHERST STATISTICAL SCIENCE DIV.

PARZEN, E. 1979. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, **74**(365), 105–121.

PEARSON, ES. 1931. The analysis of variance in cases of non-normal variation. *Biometrika*, **23**(1/2), 114–133.

PEARSON, ES. 1932. Comparison of at mckay's approximation with experimental sampling results. *Journal of the Royal Statistical Society*, **95**(4), 703–704.

PEARSON, K. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, **185**, 71–110.

PEARSON, K. 1895. X. contributions to the mathematical theory of evolution.âii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*, 343–414.

PEARSON, K. 1896. Vii. mathematical contributions to the theory of evolution.âiii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 253–318.

PIRES, AM, & BRANCO, JA. 2002. Partial influence functions. *Journal of Multivariate Analysis*, **83**(2), 451–468.

PRATT, JW. 1964. Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, **59**(307), 665–680.

PRENDERGAST, LA, & STAUDTE, RG. 2016a. Exploiting the quantile optimality ratio in finding confidence intervals for quantiles. *Stat*, **5**, 70–81.

PRENDERGAST, LA, & STAUDTE, RG. 2016b. Quantile versions of the Lorenz curve. *Electronic Journal of Statistics*, **10**(2), 1896–1926.

PRENDERGAST, LA, & STAUDTE, RG. 2017a. A Simple and Effective Inequality Measure. *The American Statistician*, **Early view**.

PRENDERGAST, LA, & STAUDTE, RG. 2017b. When large n is not enough-Distribution-free interval estimators for ratios of quantiles. *The Journal of Economic Inequality*, **15**(3), 277–293.

PRICE, RM, & BONETT, DG. 2002. Distribution-free confidence intervals for difference and ratio of medians. *Journal of Statistical Computation and Simulation*, **72**(2), 119–124.

PYKE, R. 1965. Spacings. *Journal of the Royal Statistical Society. Series B (Methodological)*, 395–449.

PYKE, R, *et al.* 1972. Spacings revisited. *In: Proceedings of the sixth berkeley symposium on mathematical statistics and probability, volume 1: Theory of statistics*. The Regents of the University of California.

RAGHAVACHARI, M. 1965. The two-sample scale problem when locations are unknown. *The Annals of Mathematical Statistics*, **36**(4), 1236–1242.

RAJPUT, DS, SINGH, PK, & BHATTACHARYA, M. 2011. Feature selection with efficient initialization of clusters centers for high dimensional data clustering. *Pages 293–297 of: 2011 international conference on communication systems and network technologies*. IEEE.

RAO, KA, & BHATT, ARS. 1995. Tests for coefficient of variation. *Journal of the Indian Society of Agricultural Statistics*, **47**(3), 225–229.

REED, GF, LYNN, F, & MEADE, BD. 2002. Use of coefficient of variation in assessing variability of quantitative assays. *Clinical and Diagnostic Laboratory Immunology*, **9**(6), 1235–1239.

REIMANN, C, FILZMOSER, P, GARRETT, RG, & DUTTER, R. 2008. *Statistical data analysis explained: applied environmental statistics with r*.

REISS, ROLF-DIETER. 1989. *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer science & business media.

RIEDER, HELMUT. 2012. *Robust asymptotic statistics*. Vol. 1. Springer Science & Business Media.

ROSENBLATT, M. 1956. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 832–837.

ROUSSEEUW, PJ, & CROUX, C. 1993. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**(424), 1273–1283.

ROUSSEEUW, PJ, & RONCHETTI, E. 1981. Influence curves of general statistics. *Journal of Computational and Applied Mathematics*, **7**(3), 161–166.

RUBIN, DB. 1981. The bayesian bootstrap. *The Annals of Statistics*, 130–134.

SACHS, L. *Applied statistics. 1982.*

SHAPIRO, HM. 2005. *Practical flow cytometry*. John Wiley & Sons.

SHARMA, KK, & KRISHNA, H. 1994. Asymptotic sampling distribution of inverse coefficient-of-variation and its applications. *IEEE Transactions on Reliability*, **43**(4), 630–633.

SHEATHER, SJ, & MARRON, JS. 1990. Kernel quantile estimators. *Journal of the American Statistical Association*, **85**(410), 410–416.

SHI, W, & GOLAM K, BM. 2007. On some confidence intervals for estimating the mean of a skewed population. *International Journal of Mathematical Education in Science and Technology*, **38**(3), 412–421.

SHOEMAKER, LH. 1995. Tests for differences in dispersion based on quantiles. *The American Statistician*, **49**(2), 179–182.

SHOEMAKER, LH. 1999. Interquartile tests for dispersion in skewed distributions. *Communications in Statistics-Simulation and Computation*, **28**(1), 189–205.

SHOEMAKER, LH. 2003. Fixing the f test for equal variances. *The American Statistician*, **57**(2), 105–114.

SHOEMAKER, LH, & HETTMANSPERGER, TP. 1982. Robust estimates and tests for the one-and two-sample scale models. *Biometrika*, **69**, 47–53.

SIEGEL, S, & TUKEY, JW. 1960. A nonparametric sum of ranks procedure for relative spread in unpaired samples. *Journal of the American Statistical Association*, **55**(291), 429–445.

SINGH, N, SOMAN, MK, & KUMAR, KK. 1987. Use of co-efficient of variation in determining rainfall probabilities in a humid region. *Mausam*, **38**, 261–264.

SNEDECOR, GW, & COCHRAN, WG. 1980. *Statistical methods*.

SORENSON, H. 1936. Psychology in education.

STAUDTE, RG. 2014. Inference for quantile measures of skewness. *Test*, **23**(4), 751–768.

STAUDTE, RG, & SHEATHER, SJ. 1990. *Robust estimation and testing*. John Wiley & Sons.

SU, JQ, & WEI, LJ. 1993. Nonparametric estimation for the difference or ratio of median failure times. *Biometrics*, 603–607.

SU, S. 2009. Confidence intervals for quantiles using generalized lambda distributions. *Computational Statistics & Data Analysis*, **53**(9), 3324–3333.

SUBRAHMANYA N, K, & ARUNA RAO, K. 2003. Tests of coefficients of variation of normal population. *Communications in Statistics-Simulation and Computation*, **32**(3), 641–661.

SUKHATME, BV. 1957. On certain two-sample nonparametric tests for variances. *The Annals of Mathematical Statistics*, **28**(1), 188–194.

SUKHATME, BV. 1958. A two sample distribution free test for comparing variances. *Biometrika*, **45**(3/4), 544–548.

SUKHATME, BV, *et al.* 1958. Testing the hypothesis that two populations differ only in location. *The Annals of Mathematical Statistics*, **29**(1), 60–78.

TUKEY, JW. 1965. Which part of the sample contains the information? *Proceedings of the National Academy of Sciences*, **53**(1), 127–134.

UMPHREY, GJ. 1983. A comment on mckay's approximation for the coefficient of variation. *Communications in Statistics-Simulation and Computation*, **12**(5), 629–635.

VAN VALEN, L. 2005. The statistics of variation. *Pages 29–47 of: Variation*. Elsevier.

VAN ZWET, WR. 1964. Convex transformations of random variables, math. *Centrum, amsterdam*, **20**.

VANGEL, MG. 1996. Confidence intervals for a normal coefficient of variation. *The American Statistician*, **50**(1), 21–26.

VARMUZA, K, & FILZMOSER, P. 2009. *Introduction to multivariate statistical analysis in chemometrics*. CRC press.

VASTRAD, C, *et al.* 2013. Identification of outliers in oxazolines and oxazoles high dimension molecular descriptor dataset using principal component outlier detection algorithm and comparative numerical study of other robust estimators. *arXiv preprint arXiv:1312.2861*.

VERRILL, S. 2003. Confidence bounds for normal and lognormal distribution coefficients of variation. *Res. Pap. FPL-RP-609. Madison, WI: US Department of Agriculture, Forest Service, Forest Products Laboratory: 13 pages*, **609**.

WANG, J-L, & HETTMANSPERGER, TP. 1990. Two-sample inference for median survival times based on one-sample procedures for censored survival data. *Journal of the American Statistical Association*, **85**(410), 529–536.

WEI, LJ, & GAIL, MH. 1983. Nonparametric estimation for a scale-change with censored observations. *Journal of the American Statistical Association*, **78**(382), 382–388.

WEINRAUB, HJ, & KUHLMAN, BR. 1994. The effect of common stock beta variability on the variability of the portfolio beta. *Journal of Financial and Strategic decisions*, **7**(2).

WELLMANN, J, & GATHER, U. 2003. Identification of outliers in a one-way random effects model. *Statistical Papers*, **44**(3), 335–348.

WELSH, AH. 1988. Asymptotically efficient estimation of the sparsity function at a point. *Statistics & Probability Letters*, **6**(6), 427–432.

WILK, MB, & GNANADESIKAN, R. 1968. Probability plotting methods for the analysis of data. *Biometrika*, **55**(1), 1–17.

WORTHINGTON, AC, & HIGGS, H. 2003. Risk, return and portfolio diversification in major painting markets: The application of conventional financial analysis to unconventional investments: Discussion paper no 148.

WU, C, ZHAO, Y, & WANG, Z. 2002. The median absolute deviations and their applications to shewhart control charts. *Communications in Statistics-Simulation and Computation*, **31**(3), 425–442.

YULE, GU. 1912. *An introduction to the theory of statistics*. London: Griffin.

ZAR, JH. 1984. Biostatistical analysis, englewood cliffs. *NY: Prentice-Hall*, 360.

ZWILLINGER, D, & KOKOSKA, S. 1999. *Crc standard probability and statistics tables and formulae*. Crc Press.

Part II

Papers

Paper I

Interval estimators for ratios of independent quantiles and interquantile ranges

Arachchige, Chandima NPG, Cairns, Maxwell, & Prendergast, Luke A. 2019a. Communications in statistics-simulation and computation, 1-17.

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Interval estimators for ratios of independent quantiles and interquantile ranges

Chandima N. P. G. Arachchige, Maxwell Cairns & Luke A. Prendergast

To cite this article: Chandima N. P. G. Arachchige, Maxwell Cairns & Luke A. Prendergast (2019): Interval estimators for ratios of independent quantiles and interquantile ranges, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2019.1634814

To link to this article: https://doi.org/10.1080/03610918.2019.1634814

Published online: 05 Jul 2019.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1634814

Interval estimators for ratios of independent quantiles and interquantile ranges

Chandima N. P. G. Arachchige, Maxwell Cairns, and Luke A. Prendergast

Department of Mathematics and Statistics, La Trobe University, Melbourne, Victoria, Australia

ABSTRACT
Recent research has shown that interval estimators with good coverage properties are achievable for some functions of quantiles, even when sample sizes are not large. Motivated by this, we consider interval estimators for the ratios of independent quantiles and interquantile ranges that will be useful when comparing location and scale for two samples. Simulations show that the intervals have excellent coverage properties for a wide range of distributions, including those that are heavily skewed. Examples are also considered that highlight the usefulness of using these approaches to compare location and scale.

ARTICLE HISTORY
Received 17 December 2018
Accepted 17 June 2019

KEYWORDS
Asymptotic variance; Coverage probability; Partial influence functions

1. Introduction

It is common to use the t-test to compare the means of two independent populations and under the assumption of normality of those populations. However, when the distributions are skewed, medians may be a more appropriate measure of location. Non-parametric alternatives to the t-test, such as tests for the difference, or ratio, of two medians are available (e.g. Price and Bonett 2002). Similarly, when comparing the spread of two populations, normality is again often assumed and the standard F-test employed based on the ratio of two independent sample variances. However, it has long been known that the F-test can be unreliable when normality is violated (see, e.g. Brown and Forsythe, 1974). For a recent discussion see Hosken, Buss, and Hodgson (2018) who advise "do not use F tests to compare variances". As a non-parametric alternative, Shoemaker (1999) introduced a test using differences in *interquantile ranges*. They found that the test is reliable for many distributions, including those that are heavily skewed.

In this paper, we propose interval estimators for ratios of quantiles and interquantile ranges, which are scale-free and thus easily interpretable. We begin by detailing some related existing methods in Sec. 2, which also allows us to introduce notations used throughout. In Sec. 3, we obtain partial influence functions for ratios of independent quantile and interquantile range estimators. Partial influence functions, which can be used to study robustness properties, can also be used to obtain asymptotic variances of

CONTACT Luke A. Prendergast luke.prendergast@latrobe.edu.au Department of Mathematics and Statistics, La Trobe University, Melbourne, Victoria 3086, Australia.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

estimators. It is the latter which is our main focus and the asymptotic variances are used to construct interval estimators in Sec. 4. We conduct simulations and then provide some examples in Sec. 5.

2. Notations and related existing methods

Let F_1 denote the distribution function for random variable X and f_1 denote the density. For a $p \in [0, 1]$, let the pth quantile be $x_p = F_1^{-1}(p) = \inf\{x : F_1(x) \geq p\}$. Also, let $g_1(p) = 1/f_1(x_p)$ denote the *quantile density function* (Tukey 1965; Parzen 1979) and its reciprocal, which we denote $q_1(p) = f_1(x_p)$, is the *density quantile function*. Similarly, let F_2 denote the distribution function for random variable Y with $y_p = F_2^{-1}(p) = \inf\{y : F_2(y) \geq p\}, g_2(p) = 1/f_2(y_p)$ and $q_2(p) = f_2(y_p)$. Also let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ denote simple random samples of size n_1 and n_2 from F_1 and F_2 respectively.

2.1. The price and Bonett method

Price and Bonett (2002) proposed an asymptotic confidence interval for a ratio of medians, which does not require identically shaped distributions. Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n_1)}$ and $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n_2)}$ be the ordered random samples. Let $\hat{\eta}_1$ and $\hat{\eta}_2$ be the usual sample medians obtained from each sample which are estimators of η_1 and η_2. Let $X^*_{(i)} = \ln X_{(i)}$ and $Y^*_{(i)} = \ln Y_{(i)}$, assuming $X_{(i)}$ and $Y_{(i)}$ are both non-negative and let $\hat{\eta}^*_1$ and $\hat{\eta}^*_2$ denote the sample medians of these log-transformed samples. An asymptotic distribution-free $(1-\alpha) \times 100\%$ confidence interval for η_1/η_2 is

$$\left(\frac{\hat{\eta}_1}{\hat{\eta}_2}\right) \exp\left\{\pm z_{1-\alpha/2}\sqrt{\mathrm{Var}\left(\hat{\eta}^*_1\right) + \mathrm{Var}\left(\hat{\eta}^*_2\right)}\right\}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ percentile of the standard normal distribution and $\mathrm{Var}(\hat{\eta}^*_j)$ $(j = 1,\ 2)$ is the variance of $\hat{\eta}^*_j$. The Price and Bonett (2001) modification of the McKean-Schrader estimator McKean and Schrader (1984) is used where

$$\mathrm{Var}\left(\hat{\eta}^*_1\right) = \left(\frac{\left(X^*_{(n_1-c_1+1)} - X^*_{(c_1)}\right)}{2z_1}\right)^2$$

where $c_1 = (n_1 + 1)/2 - n_1^{1/2}$, rounded to the nearest integer, $z_1 = \Phi^{-1}(1-p_1/2)$ and $p_1 = \sum_{i=0}^{c_1-1}[n_1!/i!(n_1 - i)!](0.5)^{(n_1-i)}$. $\mathrm{Var}(\hat{\eta}^*_2)$ is similarly defined.

2.2. Shoemaker's test

For $p \in (0, 0.5)$, the interquantile range is denoted $\mathrm{IQR}_p(X) = x_{1-p} - x_p$ which is the usual *interquartile range* when $p = 0.25$. Let $\hat{x}_p$ denote the estimator of x_p. Shoemaker (1995) uses the influence function (Hampel 1974) to calculate the asymptotic variance of the $\mathrm{IQR}_p(X)$ estimators. We will provide more detail on the influence function in the next section. Using our notations for the density quantile function at the start of this section, the asymptotic variance for $\mathrm{IQR}_p(X)$ estimator is $\omega_1^2 = p\{q_1(p) + q_1(1-p) - p[q_1(p) + q_1(1-p)]^2\}/[q_1^2(p)q_1^2(1-p)]$. Hence, for ω_2^2 denoting the asymptotic variance for the estimator of $\mathrm{IQR}_p(Y)$, the Shoemaker (1999) test statistic is

$$Z = \frac{(\hat{x}_{1-p} - \hat{x}_p) - (\hat{y}_{1-p} - \hat{y}_p)}{\sqrt{\omega_1^2/n_1 + \omega_2^2/n_2}} \quad . \tag{1}$$

Z is asymptotically $N(0,1)$ distributed when $\mathrm{IQR}_p(X) = \mathrm{IQR}_p(Y)$ provided that f_1 and f_2 are positive and continuous for the quantiles used in the interquantile ranges. As the estimator of the denominator for $q_1(p)$, and then similarly $q_1(1-p), q_2(p)$ and $q_2(1-p)$, Shoemaker (1999) uses $n(\hat{x}_p, h_n)/(2n_1h_n)$ where $n(\hat{x}_p, h_n)$ is the number of observations falling in the interval $\hat{x}_p \pm h_n$ with bandwidth $h_n = 1.3s/n_1^{1/5}$. Simulation results for a wide variety of distributions validate the use of this test, and certainly superiority over the F-test in the presence of skew. It is suggested that one should choose p between 0.1 and 0.25 for improved power.

3. Ratios of quantiles and interquantile ranges

In this section, we introduce the ratio estimators and ultimately derive their asymptotic variances.

3.1. The ratio estimators

We continue with the notations already introduced and assume $p \in (0,1)$. We define the population ratio of quantiles r_p and associated estimator to be

$$r_p = \left(\frac{x_p}{y_p}\right) \text{ and } \hat{r}_p = \left(\frac{\hat{x}_p}{\hat{y}_p}\right) \quad . \tag{2}$$

Remark 1. As with ratios of means, care should be taken when the numerator and denominator can be of opposite sign, due to difficulty in interpreting negative ratios, or when the denominator is small in magnitude comparative to the numerator. In such case, the ratio can be extremely sensitive to small changes in the quantiles. An example of where such care should be taken is when X and Y represent changes (e.g. change in a particular measurement following intervention) which can be either positive or negative and in such situations a value close to zero may be meaningful, yet the ratio may not be. For example, if $x_{0.25} = 0.2$ and $y_{0.25} = 0.05$ where $x_{0.25} - y_{0.25} = 0.15$ represents a negligible, i.e. clinically insignificant, difference. Then the ratio $r_{0.25} = 4$ may falsely imply a large difference between the two that may be mistaken for being clinically significant. One can make this ratio arbitrarily large by making only small changes to the value of $y_{0.25}$.

Assuming $p \in (0, 0.5)$, we define the population squared ratio of IQR_ps and associated estimator to be

$$R_p = \left[\frac{\mathrm{IQR}_p(X)}{\mathrm{IQR}_p(Y)}\right]^2 \text{ and } \hat{R}_p = \left(\frac{\hat{x}_{1-p} - \hat{x}_p}{\hat{y}_{1-p} - \hat{y}_p}\right)^2 \quad . \tag{3}$$

We focus on the squared ratio of IQRs since it is analogous to the ratio of variances although it is simple to obtain estimators for the ratio of IQRs by a square-root transformation. A nice property of R_p is that it is equivalent to the ratio of variances for many distributions, as shown below.

Lemma 1. *Let $G(\mu, \sigma)$ be the distribution function of a location-scale family with location and scale parameters μ and σ. If $X \sim G(\mu_1, \sigma_1)$ and $Y \sim G(\mu_2, \sigma_2)$, $R_p = \mathrm{Var}(X)/\mathrm{Var}(Y)$ for any $p \in (0, 1/2)$.*

Proof. The proof is obvious when noting that for $Z_1 \sim G(0, 1)$ and $Z_2 \sim G(0, 1)$, we can write $X = \sigma_1 Z_1 + \mu_1$ and $Y = \sigma_2 Z_2 + \mu_2$ so that the quantile functions for X and Y may each be written $\sigma_1 Q(p) + \mu_1$ and $\sigma_2 Q(p) + \mu_2$ where $Q(p)$ is the quantile function for $G(0,\ 1)$. Hence, $R_p = \sigma_1^2/\sigma_2^2$. For $Z \sim \mathrm{G}(0, 1), \mathrm{Var}(X) = \sigma_1^2 \mathrm{Var}(Z)$ and $\mathrm{Var}(Y) = \sigma_2^2 \mathrm{Var}(Z)$ so that $R_p = \mathrm{Var}(X)/\mathrm{Var}(Y)$. □

From Lemma 1, R_p is equal to the ratio of variances when the distributions are from the same location-scale family. This means that an estimator of R_p is a direct competitor to the ratio of variances for such distributions.

Remark 2. For an appropriately chosen value of p, the ratio R_{p} can be particularly advantageous when dealing with asymmetric distributions with noticeable skew. For example, consider the case of the lognormal distribution with $X \sim \mathrm{LN}\,(0, 1.25)$ and $Y \sim \mathrm{LN}\,(0, 1)$. While a plot of the densities (not shown) reveals a difference between the two, that difference is not so great as is implied by the large ratio of variances equal to 3.85. This is due to the variance in the numerator being 'drawn' toward a more extreme right-skew that outweighs the low probability mass in the extreme right tail when compared to the distribution of Y. However, by choosing $p = 0.25$ so that we are comparing the range of the middle 50% of the distributions for X and Y, the squared ratio of IQRs is much smaller at 1.70. For choices $p = 0.1$ and $p = 0.05$, the squared ratio of IQRs is 2.05 and 2.38 respectively. Still somewhat smaller, these imply that the ratio of variances is potentially misleading in this example.

3.2. Influence functions and partial influence functions

Let F denote a distribution function and, for $\epsilon \in [0, 1]$, define the contamination distribution to be $F_\epsilon = (1-\epsilon)F + \epsilon\Delta_{x_0}$ where Δ_{x_0} has all of its mass at the contaminant x_0. Suppose that for F there is a parameter of interest, θ, and associated estimator with the statistical functional T such that $T(F) = \theta$ and $T(F_n) = \hat{\theta}$. For example, for the mean parameter μ, we have $\mu = T(F) = \int x dF$. The influence function (IF, Hampel 1974) is then defined to be

$$\mathrm{IF}(x_0; T, F) \equiv \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \frac{\partial}{\partial \epsilon} T(F_\epsilon)\bigg|_{\epsilon=0}$$

which is the rate of change in T, at F, when a small amount of contamination is introduced. Influence functions are therefore useful tools to understand the behavior of estimators in the presence of certain observations types, including outliers.

Let f denote the probability density function of F and let Q_p denote the functional for the pth quantile where $Q_p(F) = x_p$. The influence function of the pth quantile is well known (e.g., p.59 of Staudte and Sheather 1990) to be

$$\mathrm{IF}\left(x_0; Q_p, F\right) = [p - I(x_p \geq x_0)]g(p) \tag{4}$$

where $g(p) = 1/f(x_p)$ is the quantile density defined earlier.

Influence functions also exhibit useful asymptotic properties including an often convenient means to derive asymptotic variances such as those computed for the IQR by Shoemaker (1995). For $X \sim F$ and F_n denoting the empirical distribution for n iid random variables distributed as F, under some mild regularity conditions such as differentiability of $T(F)$ and by the Central Limit Theorem we have (see, e.g., page 63 of Staudte and Sheather 1990),

$$\sqrt{n}\left[T(F_n) - T(F)\right] \overset{a}{\sim} N\left(0, \mathrm{ASV}(T)\right) \tag{5}$$

where $\overset{a}{\sim}$ denotes 'approximately distributed as' and $\mathrm{ASV}(T) = E[\mathrm{IF}(X; T, F)^2]$ is the asymptotic variance (ASV) of the estimator with functional T.

For the quantile estimator with functional Q_p and influence function given in (4), it can be shown that $E_F[\mathrm{IF}(X; Q_p, F)] = 0$ and

$$\mathrm{ASV}(Q_p) = E_F\left[\mathrm{IF}^2\left(X; Q_p, F\right)\right] = p(1-p)g^2(p) \ . \tag{6}$$

In our context, we have two populations and therefore consider partial influence functions (PIF, Pires and Branco 2002). We have two PIFs, where contamination is introduced to each of the populations while the other population remains uncontaminated. The first PIF of the estimator with functional T at (F_1, F_2) is

$$\mathrm{PIF}_1(x_0; T, F_1, F_2) = \lim_{\epsilon \to 0}\left[\frac{T\left[(1-\epsilon)F_1 + \epsilon\Delta_{x_0}, F_2\right] - T(F_1, F_2)}{\epsilon}\right] \ ,$$

with $\mathrm{PIF}_2(x_0; T, F_1, F_2)$ defined similarly. Let F_{n_1} and F_{n_2} denote empirical distribution functions for iid samples of size n_1 and n_2 from F_1 and F_2 then, from Pires and Branco (2002), we have that $\sqrt{n_1 + n_2}[T(F_{n_1}, F_{n_2}) - T(F_1, F_2)]$ is asymptotically normal with mean zero and asymptotic variance

$$\mathrm{ASV}(T) = \frac{1}{w_1}E_{F_1}\left[\mathrm{PIF}_1(X; T, F_1, F_2)^2\right] + \frac{1}{w_2}E_{F_2}\left[\mathrm{PIF}_2(X; T, F_1, F_2)^2\right] \tag{7}$$

where $w_i = n_i/(n_1 + n_2)$ $(i = 1, 2)$ and $E_F(.)$ denotes expectation at distribution F.

Partial influence functions for the ratio of quantiles

Let r_p be the functional for ratio of quantiles so that

$$r_p(F_1, F_2) = \left[\frac{Q_p(F_1)}{Q_p(F_2)}\right] = r_p.$$

Recall $Q_p(F_1) = x_p$ and $Q_p(F_2) = y_p$ to distinguish between the populations.

Theorem 1. *For* $\mathrm{IF}(x_0; Q_p, F)$ *as defined in (4), the partial influence functions of* r_p *for contamination introduced to each of* F_1 *and* F_2 *are*

$$\mathrm{PIF}_1(x_0; r_p, F_1, F_2) = \left[\frac{\mathrm{IF}\left(x_0; Q_p, F_1\right)}{y_p}\right], \mathrm{PIF}_2(x_0; r_p, F_1, F_2) = -r_p\left[\frac{\mathrm{IF}(x_0; Q_p, F_2)}{y_p}\right].$$

The proof of Theorem 1 is in Appendix A.1.

Partial influence functions for the ratio of variances

Let the mean and variance of the distribution described by F be μ and σ^2. For $\mathcal{T}$ denoting the functional for the mean estimator, we have $\mathcal{T}(F) = \int xdF = \mu$. Let $\mathcal{V}$ be the functional for the variance estimator where $\mathcal{V}(F) = \int [x - T(F)]^2 dF = \sigma^2$. Then $\mathcal{T}(F_\epsilon) = \int xd[(1-\epsilon)F + \epsilon\Delta_{x_0}] = (1-\epsilon)\mu + \epsilon x_0$ and $\mathcal{V}(F_\epsilon) = \sigma^2 + \epsilon[(x_0-\mu)^2 - \sigma^2] - \epsilon^2(x_0-\mu)^2$. Consequently, $\mathrm{IF}(x_0; \mathcal{V}, F) = (x_0-\mu)^2 - \sigma^2$ is the influence function for the variance estimator.

Let $\mathcal{T}(F_j) = \mu_j, \mathcal{V}(F_j) = \sigma_j^2$ $(j = 1,\ 2)$. For $\mathcal{R}$ denoting the functional for the ratio of variances, we have $\mathcal{R}(F_1, F_2) = \mathcal{V}(F_1)/\mathcal{V}(F_2) = \sigma_1^2/\sigma_2^2 = \rho$. Then, for $z_j = (x_0-\mu_j)/\sigma_j$ $(j = 1,\ 2)$, the PIFs for $\mathcal{R}$ are

$$\mathrm{PIF}_1(x_0; \mathcal{R}, F_1, F_2) = \rho\left(z_1^2 - 1\right), \mathrm{PIF}_2(x_0; \mathcal{R}, F_1, F_2) = -\rho\left(z_2^2 - 1\right) \tag{8}$$

As expected, the PIFs are unbounded in x_0 indicating that outliers can exert unlimited influence.

Partial influence functions for the squared IQR ratio

Let $\mathcal{R}_p$ be the functional for the squared ratio of IQRs so that

$$\mathcal{R}_p(F_1, F_2) = \left[\frac{\mathcal{Q}_{1-p}(F_1) - \mathcal{Q}_p(F_1)}{\mathcal{Q}_{1-p}(F_2) - \mathcal{Q}_p(F_2)}\right]^2 = \rho_p.$$

Theorem 2. *For* $\mathrm{IF}(x_0; \mathcal{Q}_p, F)$ *is defined in (4), the partial influence functions of* $\mathcal{R}_p$ *for contamination introduced to each of* F_1 *and* F_2 *are*

$$\mathrm{PIF}_1\left(x_0; \mathcal{R}_p, F_1, F_2\right) = \frac{2\rho_p}{x_{1-p} - x_p}\left[\mathrm{IF}\left(x_0; \mathcal{Q}_{1-p}, F_1\right) - \mathrm{IF}\left(x_0; \mathcal{Q}_p, F_1\right)\right],$$
$$\mathrm{PIF}_2\left(x_0; \mathcal{R}_p, F_1, F_2\right) = -\frac{2\rho_p}{y_{1-p} - y_p}\left[\mathrm{IF}\left(x_0; \mathcal{Q}_{1-p}, F_2\right) - \mathrm{IF}\left(x_0; Q_p, F_2\right)\right].$$

The proof of Theorem 2 is in Appendix A.2.

3.2.1. Asymptotic variances

Recall that μ_i and σ_i denote the mean and standard deviation of F_i $(i = 1, 2)$ and that $\rho = \sigma_1^2/\sigma_2^2$. Then from (7) and (8), it is straight forward to show that the ASV for the ratio of variances estimator is

$$\mathrm{ASV}(\mathcal{R}; n_1, n_2) = \rho^2\left\{\frac{1}{w_1}\left[E_{F_1}\left(Z_1^4\right) - 1\right] + \frac{1}{w_2}\left[E_{F_2}\left(Z_2^4\right) - 1\right]\right\} \tag{9}$$

where $Z_i = (X-\mu_i)/\sigma_i$ so that $E_{F_i}(Z_i^4)$ is the scaled fourth central moment of F_i $(i = 1,\ 2)$. Recall that $g_1(p) = 1/f_1(x_p)$ and $g_2(p) = 1/f_2(y_p)$ are the quantile density functions. We now provide the ASV for the ratio of quantiles.

Theorem 3. *The asymptotic variances of* $\sqrt{n_1 + n_2}\, r_p(F_{n_1}, F_{n_2})$ *and* $\sqrt{n_1 + n_2}\, \mathcal{R}_p(F_{n_1}, F_{n_2})$ *are*

$$\mathrm{ASV}(r_p; n_1, n_2) = p(1-p) r_p^2\left\{\frac{g_1^2(p)}{w_1 x_p^2} + \frac{g_2^2(p)}{w_2 y_p^2}\right\}.$$

Table 1. Choice of p to minimize ASV for the squared IQR ratio.

Distribution	p	Distribution	p	Distribution	p
Exp(λ)	0.128	Beta(0.1,0.1)	0	Gamma(1)	0.128
Unif(a, b)	0	Beta(0.5,0.5)	0	Gamma(2)	0.110
Log Normal(0,1)	0.193	Beta(1)	0	Gamma(10)	0.081
Log Normal(1)	0.193	Beta(10)	0.055	PAR(1)	0.282
N(μ, σ^2)	0.069	Weibull(0.5)	0.181	PAR(2)	0.224
Chi-Squared(1)	0.127	Weibull(1)	0.128	PAR(3)	0.198
Chi-Squared(2)	0.128	Weibull(2)	0.069	PAR(5)	0.173
Chi-Squared(25)	0.079	Weibull(10)	0.081	PAR(7)	0.161

and

$$\begin{aligned}\mathrm{ASV}\left(\mathcal{R}_p; n_1, n_2\right) = 4p\rho_p^2\Bigg\{ &\frac{g_1^2(p) + g_1^2(1-p) - p\left[g_1(p) + g_1(1-p)\right]^2}{w_1(x_{1-p} - x_p)^2} \\ &+ \frac{g_2^2(p) + g_2^2(1-p) - p\left[g_2(p) + g_2(1-p)\right]^2}{w_2(y_{1-p} - y_p)^2}\Bigg\}.\end{aligned}$$

The proof of Theorem 3 is in Appendix A.3.

Corollary 1. *Suppose that X and Y are both random variables from the same location-scale family such that the density of X may be written $f(x; \mu_1, \sigma_1)$ and the density of Y $f(y; \mu_2, \sigma_2)$ where μ_1, μ_2 and σ_1, σ_2 are the respective location and scale parameters. Let q_{1-p} and q_p denote the $(1-p)$th and pth quantiles of the distribution with density $f(\cdot; 0, 1)$ and $g_0(1-p) = 1/f(q_{1-p}; 0, 1)$ and $g_0(p) = 1/f(q_p; 0, 1)$ the respective quantile densities. Then*

$$\mathrm{ASV}\left(\mathcal{R}_p; n_1, n_2\right) = 4p\frac{\sigma_1^4}{\sigma_2^4}\left\{\frac{g_0^2(p) + g_0^2(1-p) - p\left[g_0(p) + g_0(1-p)\right]^2}{w_1(1 - w_1)(q_{1-p} - q_p)^2}\right\}.$$

Proof. Since X and Y are from the same location-scale family, then $x_{1-p} - x_p = \eta_1(q_{1-p} - q_p), y_{1-p} - y_p = \eta_2(q_{1-p} - q_p)$ and

$$f(x; \mu_1, \eta_1) = \frac{1}{\sigma_1} f\left(\frac{\sigma_1 x - \mu_1}{\sigma_1}; 0, 1\right), f(y; \mu_2, \sigma_2) = \frac{1}{\sigma_2} f\left(\frac{\sigma_2 y - \mu_2}{\sigma_2}; 0, 1\right).$$

Using these results $g_1(p) = g_0(p)\sigma_1$ and $g_2(p) = g_0(p)\sigma_2$. The result follows after some simplification and noting that $w_2 = 1 - w_1$. □

Remark 3. Since the ASV in Corollary 1 depends on location and scale only through σ_1^4/σ_2^4 which is a common factor to all terms, then the choice of p that minimizes the ASV is independent of the location and scale parameters.

We now explore the choices of p that result in the minimum ASV of the squared IQR ratio estimator for several distributions. As shown in Table 1, the choice of p that minimizes the ASV varies for different distributions. The choice of p that minimizes the ASV for the exponential, uniform and normal distributions does not depend on the parameters of these distributions (see Remark 3 which is a consequence of Corollary 1). For distributions considered with the exception of small shape parameter for the Pareto type II (PAR), choosing a $p < 0.25$ gives a smaller ASV than if one were to use the ratio of interquartile ranges. This agrees with the observations of Shoemaker (1999). Our interest is mainly in applications to skewed data and we therefore favor $p = 0.2$ which would give good results for log normal and Pareto-type distributions.

4. Interval estimators

The quantiles are estimated using a weighted average of adjacent order statistics. We use the type 8 weights recommended by Hyndman and Fan (1996) which is approximately median-unbiased and is available in the R quantile function. To estimate the quantile density, g_1 and then similarly for g_2, we use the kernel density estimator

$$\hat{g}_1(p, b) = \sum_{i=1}^{n_1} X_{(i)} \left\{ k_b\left(p - \frac{i-1}{n_1}\right) - k_b\left(p - \frac{i}{n_1}\right) \right\}$$

with kernel function k_b, for which we use the Epanechnikov (1969) kernel, and bandwidth b. It has recently been shown that excellent confidence interval coverage for estimators of functions of quantiles can be obtained for sample sizes even as low as 30 (see Prendergast and Staudte 2016b, 2017, 2018). These works use the Quantile Optimality Ratio (QOR, Prendergast and Staudte 2016a) to choose the optimal b for estimating the quantile densities. We therefore use the QOR in selecting our b although other choices of b are also possible.

4.1. Approximate variances of the estimators

Let $\hat{\rho} = S_1^2/S_2^2$ be the estimator of σ_1^2/σ_2^2 where $S_i^2 = \mathcal{V}(F_{n_i})$ $(i = 1,\ 2)$ are the sample variance estimators. Let $\{X_i\}_{i=1}^{n_1}$ denote the simple random sample for the first sample with sample mean $\bar{X} = \mathcal{T}(F_{n_1})$ and $\{Y_i\}_{i=1}^{n_2}$ denote the simple random sample for the second with $\bar{Y} = \mathcal{T}(F_{n_2})$. From (9),

$$\mathrm{Var}(\hat{\rho}) \approx \frac{\hat{\rho}^2}{n_1 + n_2}\left[\frac{1}{w_1}\left(\overline{Z_1^4} - 1\right) + \frac{1}{w_2}\left(\overline{Z_2^4} - 1\right)\right]$$

where

$$\overline{Z_1^4} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\frac{X_i - \bar{X}}{S_1}\right)^4, \overline{Z_2^4} = \frac{1}{n_2}\sum_{i=1}^{n_2}\left(\frac{Y_i - \bar{Y}}{S_2}\right)^4.$$

Let $\mathrm{Var}(\hat{r}_p)$ denote the variance of the ratio of quantiles estimator. Then, from Theorem 3

$$\mathrm{Var}(\hat{r}_p) \approx \frac{p(1-p)\hat{r}_p^2}{n_1 + n_2}\left\{\frac{\hat{g}_1^2(p)}{w_1\hat{x}_p^2} + \frac{\hat{g}_2^2(p)}{w_2\hat{y}_p^2}\right\},$$

where $\hat{g}_i(p)$ $(i = 1,\ 2)$ is the estimated quantile density using the QOR method. Similarly, let $\mathrm{Var}(\hat{\rho}_p)$ denote the variance of the squared ratio of IQRs estimator. Then, from Theorem 3

Table 2. Simulated coverage probabilities (and widths) for the 95% confidence interval estimators for the interval based on the Price and Bonnet (rows labeled PB) method and the interval in (10) (r_p) with $p = 0.25, 0.5, 0.75$. True values of the ratios are included in the top rows.

Sample size (n_1,n_2)		$X \sim$ LN $(0,1)$ $Y \sim$ LN $(0,1)$	$X \sim$ EXP (1) $Y \sim$ EXP (1)	$X \sim \chi_5^2$ $Y \sim \chi_2^2$	$X \sim$ PAR$(1,7)$ $Y \sim$ PAR$(1,3)$
	PB =	**1.00**	**1.00**	**3.14**	**0.40**
	$r_{0.25} =$	**1.00**	**1.00**	**4.65**	**0.42**
	$r_{0.5} =$	**1.00**	**1.00**	**3.14**	**0.40**
	$r_{0.75} =$	**1.00**	**1.00**	**2.39**	**0.37**
50,50	PB	0.961(1.12)	0.965(1.31)	0.964(3.29)	0.963(0.57)
	$r_{0.25}$	0.963(1.18)	0.959(1.71)	0.957(6.05)	0.962(0.75)
	$r_{0.5}$	0.972(1.17)	0.967(1.30)	0.966(3.20)	0.973(0.60)
	$r_{0.75}$	0.977(1.38)	0.970(1.16)	0.965(2.24)	0.971(0.55)
100,100	PB	0.960(0.75)	0.962(0.88)	0.962(2.21)	0.958(0.38)
	$r_{0.25}$	0.965(0.82)	0.961(1.19)	0.958(4.31)	0.959(0.51)
	$r_{0.5}$	0.970(0.77)	0.960(0.87)	0.958(2.19)	0.966(0.39)
	$r_{0.75}$	0.972(0.89)	0.962(0.77)	0.965(1.52)	0.969(0.35)
200,200	PB	0.952(0.51)	0.953(0.59)	0.950(1.50)	0.954(0.26)
	$r_{0.25}$	0.967(0.58)	0.957(0.82)	0.957(3.04)	0.962(0.36)
	$r_{0.5}$	0.962(0.53)	0.961(0.60)	0.957(1.52)	0.960(0.26)
	$r_{0.75}$	0.967(0.59)	0.960(0.53)	0.959(1.04)	0.966(0.24)
200,500	PB	0.948(0.42)	0.951(0.49)	0.953(1.11)	0.947(0.21)
	$r_{0.25}$	0.965(0.48)	0.953(0.68)	0.959(2.24)	0.958(0.30)
	$r_{0.5}$	0.961(0.44)	0.958(0.50)	0.960(1.12)	0.961(0.22)
	$r_{0.75}$	0.965(0.49)	0.958(0.43)	0.961(0.78)	0.961(0.19)
500,500	PB	0.952(0.32)	0.947(0.36)	0.950(0.93)	0.949(0.16)
	$r_{0.25}$	0.962(0.35)	0.957(0.51)	0.957(1.90)	0.955(0.22)
	$r_{0.5}$	0.958(0.33)	0.959(0.37)	0.957(0.95)	0.957(0.16)
	$r_{0.75}$	0.963(0.36)	0.954(0.32)	0.955(0.64)	0.960(0.14)
500,1000	PB	0.946(0.27)	0.947(0.31)	0.946(0.73)	0.948(0.13)
	$r_{0.25}$	0.960(0.31)	0.953(0.44)	0.957(1.49)	0.953(0.19)
	$r_{0.5}$	0.960(0.28)	0.953(0.32)	0.955(0.74)	0.959(0.14)
	$r_{0.75}$	0.961(0.31)	0.959(0.28)	0.954(0.51)	0.960(0.12)
1000,1000	PB	0.945(0.22)	0.948(0.25)	0.945(0.65)	0.946(0.11)
	$r_{0.25}$	0.959(0.25)	0.958(0.36)	0.952(1.33)	0.954(0.15)
	$r_{0.5}$	0.955(0.23)	0.952(0.26)	0.957(0.66)	0.954(0.11)
	$r_{0.75}$	0.958(0.25)	0.957(0.23)	0.954(0.45)	0.957(0.10)

$$\mathrm{Var}(\hat{\rho}_p) \approx \frac{4p\hat{R}_p^2}{n_1+n_2}\left\{\frac{\hat{g}_1^2(p)+\hat{g}_1^2(1-p)-p\left[\hat{g}_1(p)+\hat{g}_1(1-p)\right]^2}{w_1\left(\hat{x}_{1-p}-\hat{x}_p\right)^2} + \frac{\hat{g}_2^2(p)+\hat{g}_2^2(1-p)-p\left[\hat{g}_2(p)+\hat{g}_2(1-p)\right]^2}{w_2\left(\hat{y}_{1-p}-\hat{y}_p\right)^2}\right\}.$$

4.2. Asymptotic confidence intervals

In constructing our interval estimators for the ratios, we use the log transformation and exponentiate to return to the ratio scale. For a random variable $W > 0$, using the Delta method it follows that $\mathrm{Var}[\ln(W)] \approx \mathrm{Var}(W)/W^2$. Hence, approximate $(1-\alpha)100\%$ confidence intervals for the ratio of quantiles, ratio of variances and squared ratio of IQRs are

$$\exp\left[\ln(\hat{r}_p) \pm z_{1-\alpha/2}\frac{1}{\hat{r}_p}\sqrt{\mathrm{Var}(\hat{r}_p)}\right] \tag{10}$$

Table 3. Simulated coverage probabilities (and widths) for the 95% confidence interval estimators for the interval based on the F-test (rows labeled F) and the intervals in (11) for the ratio of variances (R) and squared ratio of IQRs (R_p) with several choices of p. (*median widths reported due to excessively large average widths after back exponentiation). True values of the ratios are included in the top rows.

Sample size (n_1,n_2)		$X \sim$ LN(0,1) $Y \sim$ LN(0,1)	$X \sim$ EXP(1) $Y \sim$ EXP(1)	$X \sim \chi^2_5$ $Y \sim \chi^2_2$	$X \sim$ PAR(1,7) $Y \sim$ PAR(1,3)
	$F =$	**1.00**	**1.00**	**2.50**	**0.05**
	$R =$	**1.00**	**1.00**	**2.50**	**0.05**
	$R_{0.05} =$	**1.00**	**1.00**	**2.84**	**0.10**
	$R_{0.10} =$	**1.00**	**1.00**	**3.01**	**0.11**
	$R_{0.20} =$	**1.00**	**1.00**	**3.18**	**0.13**
50,50	F	0.445(2.01)	0.705(1.39)	0.756(3.47)	0.405(0.14)
	R	0.778(6.11)	0.867(2.27)	0.869(4.91)	0.714(0.35)
	$R_{0.05}$	0.975(15.25*)	0.971(7.67)	0.969(18.14)	0.977(0.74*)
	$R_{0.10}$	0.978(20.64)	0.967(4.60)	0.968(11.44)	0.977(2.75)
	$R_{0.20}$	0.978(7.97)	0.971(4.24)	0.971(11.91)	0.974(0.93)
100,100	F	0.389(1.15)	0.689(0.88)	0.741(2.19)	0.341(0.08)
	R	0.829(4.14)	0.896(1.59)	0.903(3.38)	0.740(0.23)
	$R_{0.05}$	0.977(9.57)	0.970(2.79)	0.970(6.75)	0.976(1.03)
	$R_{0.10}$	0.975(4.14)	0.970(2.18)	0.965(5.81)	0.975(0.42)
	$R_{0.20}$	0.975(3.14)	0.962(2.17)	0.967(6.21)	0.970(0.38)
200,200	F	0.348(0.70)	0.686(0.58)	0.746(1.47)	0.296(0.04)
	R	0.861(2.96)	0.915(1.10)	0.926(2.40)	0.766(0.15)
	$R_{0.05}$	0.978(2.99)	0.968(1.54)	0.965(3.83)	0.975(0.27)
	$R_{0.10}$	0.973(2.07)	0.966(1.30)	0.965(3.50)	0.968(0.22)
	$R_{0.20}$	0.973(1.74)	0.961(1.31)	0.963(3.78)	0.965(0.21)
200,500	F	0.340(0.54)	0.678(0.48)	0.770(1.20)	0.314(0.03)
	R	0.872(2.44)	0.930(0.91)	0.937(1.89)	0.831(0.12)
	$R_{0.05}$	0.971(2.12)	0.965(1.18)	0.962(2.74)	0.972(0.17)
	$R_{0.10}$	0.969(1.56)	0.963(1.02)	0.962(2.60)	0.968(0.15)
	$R_{0.20}$	0.963(1.35)	0.960(1.03)	0.961(2.88)	0.963(0.16)
500,500	F	0.324(0.39)	0.679(0.36)	0.741(0.90)	0.257(0.02)
	R	0.897(1.87)	0.935(0.70)	0.939(1.53)	0.775(0.10)
	$R_{0.05}$	0.972(1.40)	0.963(0.84)	0.962(2.13)	0.967(0.13)
	$R_{0.10}$	0.965(1.08)	0.962(0.74)	0.960(2.01)	0.964(0.11)
	$R_{0.20}$	0.964(0.96)	0.959(0.76)	0.960(2.20)	0.960(0.12)
500,1000	F	0.307(0.33)	0.681(0.31)	0.765(0.77)	0.272(0.02)
	R	0.906(1.63)	0.940(0.60)	0.947(1.26)	0.830(0.08)
	$R_{0.05}$	0.966(1.14)	0.962(0.71)	0.962(1.70)	0.970(0.10)
	$R_{0.10}$	0.963(0.90)	0.956(0.63)	0.962(1.64)	0.964(0.09)
	$R_{0.20}$	0.962(0.81)	0.961(0.64)	0.954(1.81)	0.958(0.10)
1000,1000	F	0.302(0.26)	0.675(0.25)	0.738(0.63)	0.215(0.02)
	R	0.918(1.34)	0.940(0.49)	0.941(1.08)	0.793(0.07)
	$R_{0.05}$	0.967(0.89)	0.963(0.57)	0.961(1.43)	0.968(0.08)
	$R_{0.10}$	0.963(0.71)	0.959(0.51)	0.958(1.38)	0.960(0.08)
	$R_{0.20}$	0.959(0.65)	0.955(0.52)	0.956(1.51)	0.961(0.08)

$$\exp\left[\ln(\hat{\rho}) \pm z_{1-\alpha/2}\frac{1}{\hat{\rho}}\sqrt{\mathrm{Var}(\hat{\rho})}\right] \mathit{and} \exp\left[\ln\left(\hat{\rho}_p\right) \pm z_{1-\alpha/2}\frac{1}{\hat{\rho}_p}\sqrt{\mathrm{Var}\left(\hat{\rho}_p\right)}\right] \tag{11}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)\times 100$ percentile of the standard normal distribution and where for the variances we use the approximations from Sec. 4.1.

5. Simulations and examples

5.1. Simulations

A simulation study was conducted to compare the performance among estimators by considering coverage probability and the average confidence interval width as the performance measures. We have selected the lognormal, exponential, chi-square and Pareto distributions with different parameter choices with the sample sizes $n = \{50, 100, 200, 500, 1000\}$ and 10,000 simulation trials.

Simulated coverages for the Price and Bonett method and interval estimator in (10) are provided in Table 2. The Price and Bonett (PB) method for the ratio medians provides very good coverage compared to the nominal 0.95 and the interval width decreases with increasing sample sizes. Similar results can be seen for the ratio of quantiles interval estimator when we choose $p = 0.5$ for the ratio of medians. Coverages suggest that the use of $r_{0.5}$ provides slightly more conservative coverage but with similar interval width. For ratios of the quartiles ($p = 0.25$ and $p = 0.75$) coverages are again very good with none reported below the nominal 0.95 and most less than 0.97. The highest coverages were reported for the smaller sample size setting where $n_1 = n_2 = 50$.

Simulated coverages for the F-test and the interval estimators in (11) are provided in Table 3. The F-test approach refers to the standard method for getting an interval for the ratio of variances under the assumption that the data has been sampled from normal distributions. Consequently, the coverages for the intervals based on the F-test (rows labeled *F*) are poor due to the violation of underlying normality. The interval for ratio of variances using the asymptotic interval provides reasonable coverage for some of the distributions but not when the sample sizes are small to moderate where the intervals appear to be too narrow. On the other hand, the coverages for the squared IQR ratio interval is very good for all distributions, including for the smaller sample sizes. For the distributions we have considered here, the squared ratio of IQRs is preferred due to superior coverage. We have seen this across a broad range of distributions distributions and the reader can verify this by using our web application detailed next.

5.1.1. A shiny web application for the performance comparisons of the intervals

For further comparisons, we have developed a Shiny (Chang et al., 2017) web application that readers can use to run the simulations with different parameter choices. This can be found at https://lukeprendergast.shinyapps.io/IQR_ratio/.

The user can change the distribution, parameters, sample size, probability and the number of trials according to their choices. Once the desired options are selected, the 'Run Simulation' button can be pressed and the relevant estimates, coverage probability (cp) and the average width of the confidence interval (w) will be calculated according to their input choices.

5.2. Examples

As examples, we have selected two different data sets in different contexts.

5.2.1. Prostate data

The prostate data set, which we obtained from the depthTools package (Lopez-Pintado and Torrente 2013) in R, is a normalized subset of gene expression data of the 100 most variable genes for 25 randomly selected tumor and 25 randomly selected normal

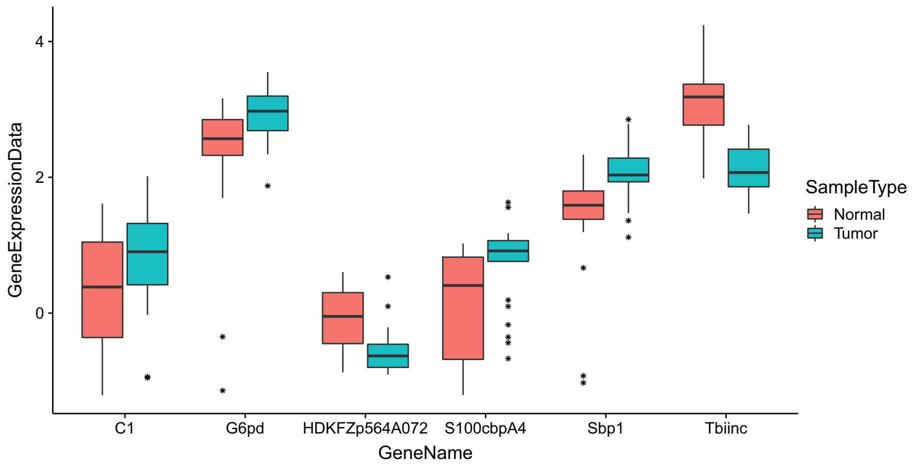


Figure 1. Box plots of the gene expression data for tumoral and normal samples of selected six genes.

prostate samples from Singh et al. (2002). In Remark 1 we cautioned the use of ratios of quantiles when quantiles may be of different sign such as is the case here. We therefore restrict our attention to looking for differences in spread between the tumor and normal samples.

Since the sample sizes are comparatively small, we have chosen $p = 0.1, 0.2, 0.25$ to construct the confidence interval for the ratio of IQRs. We found that there are six genes, which lead to very different conclusions depending on whether the ratio of variances or ratio of IQRs is used. These genes, including their abbreviations where applicable, are Carboxylesterase 1 (C1), Glucose-6-phosphate dehydrogenase (G6pd), HDKFZp564A072, S100 calcium-binding protein A4 (S100cbpA4), Selenium binding protein 1 (Sbp1) and Thymosin beta, identified in neuroblastoma cells (Tbiinc).

Box plots of the genes separated according to groups are shown in Figure 1. There is at least one outlier or extreme value in at least one of the two groups in all genes except for Tbiinc. Ignoring outliers, the boxplots suggest differences in spread for C1, HDKFZp564A072 and S100cbpA4.

In Table 4 we provide the estimated asymptotic 95% confidence intervals for the ratio of variances and squared ratios of IQRs from (11) for the six genes. While the boxplots indicate a difference in spread for C1, HDKFZp564A072 and S100cbpA4, this is not convincingly reflected in the ratio of variances intervals most likely due to the presence of outliers. However, the interval estimators for the squared ratio of IQRs capture the differences. For the G6pd, Sbpl and Tibiinc genes, the conclusions are reversed where the ratio of variances suggest significant differences in spread while this is not the case for those based on the IQRs.

5.2.2. Melbourne house price data

Since house prices are usually highly skewed, the sample mean is often not indicative of a typical house price. Therefore, the median is the most popular measure used

Table 4. Estimates (Est.) and 95% asymptotic confidence intervals (CI) for the ratio of variances (column labeled R) and squared ratios of IQRs (R_p) with $p = 0.1, 0.2, 0.25$ for the six selected genes.

Gene		R	$R_{0.1}$	$R_{0.2}$	$R_{0.25}$
C1	**Est.**	**1.531**	**2.548**	**4.520**	**2.420**
	CI	(0.777, 3.016)	(1.206, 5.384)	(2.124, 9.618)	(1.080, 5.426)
G6pd	**Est.**	**6.496**	**2.269**	**1.680**	**1.071**
	CI	(2.085, 20.243)	(0.363, 14.163)	(0.295, 9.564)	(0.140, 8.179)
HDKFZp-	**Est.**	**1.930**	**3.714**	**2.864**	**4.870**
564A072	**CI**	(0.847, 4.397)	(1.533, 8.997)	(1.506, 5.448)	(2.357, 10.065)
S100-	**Est.**	**1.748**	**1.987**	**3.338**	**24.257**
cbpA4	**CI**	(0.950, 3.217)	(1.260, 3.136)	(1.431, 7.786)	(3.211, 183.277)
Sbp1	**Est.**	**3.459**	**1.042**	**0.674**	**1.431**
	CI	(1.193, 10.029)	(0.190, 5.728)	(0.084, 5.387)	(0.172, 11.929)
Tbiinc	**Est.**	**2.227**	**1.787**	**1.177**	**1.185**
	CI	(1.163, 4.265)	(0.868, 3.678)	(0.531, 2.611)	(0.533, 2.634)

understanding house price markets. Similarly, variances, and ratios of variances, may be difficult to interpret for skewed data (see, e.g., Remark 2 for a brief example), and IQRs can be more informative when seeking to understand house price spread. Motivated by this, we now apply our intervals to Melbourne house clearance data from January 2016 obtained from the website https://www.kaggle.com/anthonypino/melbourne-housing-market. When describing house prices, it is also common to focus on quartiles (see, e.g., Taylor and Watling 2011) so in what follows we choose $p = 0.25$ and $p = 0.75$.

This data set contains prices of three types of houses (house, unit, townhouse) within different suburbs in Melbourne, Australia. Restricting to suburbs with more than 10 houses sold left data for 301 suburbs. Our focus will be on comparing house prices between suburbs. To get an understanding of how often different findings could result depending on whether variances or IQRs were used, we obtained the intervals for every pairing of suburbs which resulted in 45,150 confidence intervals for each ratio.

Table 5 represents the proportions of times that a different conclusion is reached (assuming conclusions are reached based on whether the ratio intervals include one or not) when comparing spread using either variances or IQRs. We were surprised that over 30% of the time there was a difference depending on which interval was used. This helps to highlight why one should choose which ratio is best suited to their purpose carefully.

To illustrate, Figure 2 depicts the house price distribution of a selected three pairs of neighboring suburbs. As expected, the house price distributions are positively skewed and it can be seen that there are at least a few outliers in all suburbs except for Kingsbury. When considering the middle 50% of house prices, there are noticeable differences in variation between each of the neighboring suburbs in all three pairings.

From Table 6, the Price and Bonnet and asymptotic methods for the ratio of medians are consistent in their findings. A significant difference in median house price is detected between Beaumaris and Black Rock, and the intervals for the other pairings, while including the ratio one, also suggest that there may be differences. The ratio of variance intervals are very wide making it difficult to determine whether there are real differences in spread, despite two of the estimated ratios being substantially less than one. However, the interval estimators for the ratio of interquartle ranges to detect differences in spread and this agrees with other premise that there were notable differences in the spread for the middle 50% of house prices. Putting this together and thinking

Table 5. Proportion of comparisons giving different conclusions based on ratio of variances/ratio of IQRs ($p = 0.25$). Here we count the number of times that the intervals differ in terms of whether they include one.

Type of Conclusion	Count	Percentage(%)
R does not include one and $R_{0.25}$ includes one	7914	17.53
R includes one and $R_{0.25}$ does not include one	5708	12.64
Both intervals include one or do not include one	31528	69.83
Total	45150	100.00

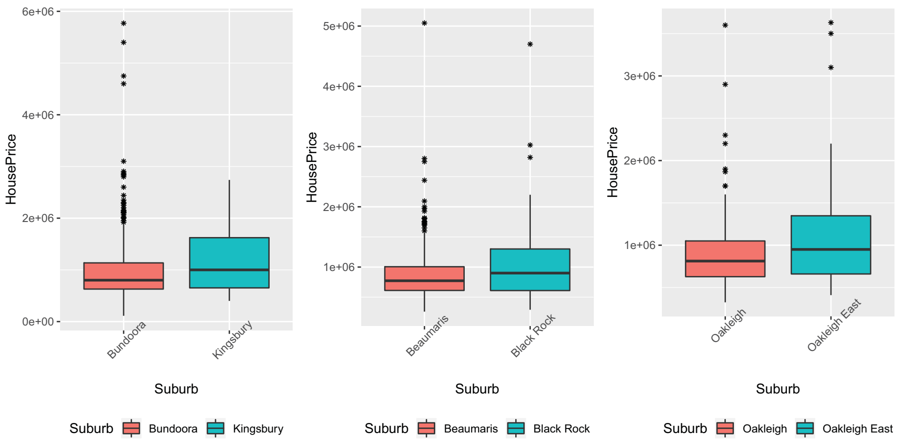


Figure 2. House price comparisons of selected three pair of neighboring suburbs.

Table 6. 95% confidence intervals (CI) for the Price and Bonnet method (row labeled PB), asymptotic interval for the ratio of quantiles (r), ratio of variances R and ratio of interquartile ranges $R_{0.25}$ for selected three pairs of neighboring suburbs.

	Bundoora/Kingsbury		Beaumaris/Black Rock		Oakleigh/Oakleigh East	
Ratio	Estimate	CI	Estimate	CI	Estimate	CI
PB	0.801	(0.628, 1.020)	0.858	(0.741, 0.994)	0.855	(0.707, 1.035)
$r_{0.25}$	0.965	(0.797, 1.168)	1.003	(0.858, 1.172)	0.952	(0.809, 1.119)
$r_{0.5}$	0.801	(0.597, 1.074)	0.858	(0.745, 0.989)	0.855	(0.713, 1.026)
$r_{0.75}$	0.670	(0.537, 0.911)	0.773	(0.655, 0.913)	0.779	(0.625, 0.971)
R	1.194	(0.555, 1.803)	0.650	(0.424, 2.361)	0.698	(0.436, 2.295)
$R_{0.25}$	0.272	(0.115, 0.643)	0.325	(0.166, 0.640)	0.377	(0.147, 0.967)

about what it means for a potential home buyer, as an example we consider the Beaumaris and Black Rock neighboring suburbs. A typical (median) house in Beaumaris was significantly cheaper (ratio of medians = 0.858, 95% CI [0.745, 0.989]) and the spread of prices notably smaller for the middle 50% of houses. This reduced spread is also reflected in the approximate equivalent price at the 25th percentile ($\hat{r}_{0.25} = 1.003$, 95% CI [0.86, 1.17]).

6. Summary and discussion

We have introduced interval estimators for ratios of quantiles and interquantile ranges. The intervals have very good coverage, even for samples as small as 50 for a wide range of distributions. Our examples highlight that very different conclusions can be arrived at when using ratios of interquantile ranges instead of ratios of variances. Future work will consider how to best choose p or the creation of a combined interval that does not require p to be chosen as was done recently by Marozzi (2012) for hypothesis tests of variation using IQRs.

References

Brown, M. B., and A. B. Forsythe. 1974. Robust tests for the equality of variances. *Journal of the American Statistical Association* 69 (346):364–7. doi:10.2307/2285659.

Chang, W., J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson. 2017. Shiny: Web application framework for R. R package version 1.0.5.

Epanechnikov, V. A. 1969. Nonparametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* 14 (1):153–8. doi:10.1137/1114019.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69 (346):383–93. doi:10.1080/01621459.1974.10482962.

Hosken, D. J., D. L. Buss, and D. J. Hodgson. 2018. Beware the f test (or, how to compare variances). *Animal Behaviour* 136:119–26. doi:10.1016/j.anbehav.2017.12.014.

Hyndman, R. J., and Y. Fan. 1996. Sample quantiles in statistical packages. *The American Statistician* 50 (4):361–5. doi:10.2307/2684934.

Lopez-Pintado, S., and A. Torrente. 2013. Depthtools: Depth tools package. R package version 0.4.

Marozzi, M. 2012. A combined test for differences in scale based on the interquantile range. *Statistical Papers* 53 (1):61–72. doi:10.1007/s00362-010-0308-7.

McKean, J. W., and R. M. Schrader. 1984. A comparison of methods for studentizing the sample median. *Communications in Statistics-Simulation and Computation* 13 (6):751–73. doi:10.1080/03610918408812413.

Parzen, E. 1979. Nonparametric statistical data modeling. *Journal of the American Statistical Association* 74 (365):105–21. doi:10.1080/01621459.1979.10481621.

Pires, A. M., and J. A. Branco. 2002. Partial influence functions. Journal of Multivariate Analysis. 83 (2):451–68. doi:10.1006/jmva.2001.2055.

Prendergast, L. A., and R. G. Staudte. 2016a. Exploiting the quantile optimality ratio in finding confidence intervals for quantiles. *STAT* 5 (1):70–81. doi:10.1002/sta4.105.

Prendergast, L. A., and R. G. Staudte. 2016b. Quantile versions of the Lorenz curve. *Electronic Journal of Statistics* 10 (2):1896–926. doi:10.1214/16-EJS1154.

Prendergast, L. A., and R. G. Staudte. 2017. When large n is not enough-distribution-free interval estimators for ratios of quantiles. The Journal of Economic Inequality 15 (3):277–93. doi:10.1007/s10888-017-9347-9.

Prendergast, L. A., and R. G. Staudte. 2018. A simple and effective inequality measure. *The American Statistician* 72 (4):328–43. doi:10.1080/00031305.2017.1366366.

Price, R. M., and D. G. Bonett. 2001. Estimating the variance of the sample median. *Journal of Statistical Computation and Simulation* 68 (3):295–305. doi:10.1080/00949650108812071.

Price, R. M., and D. G. Bonett. 2002. Distribution-free confidence intervals for difference and ratio of medians. *Journal of Statistical Computation and Simulation* 72 (2):119–24. doi:10.1080/00949650212140.

Shoemaker, L. H. 1995. Tests for differences in dispersion based on quantiles. *The American Statistician* 49 (2):179–82. doi:10.2307/2684634.

Shoemaker, L. H. 1999. Interquartile tests for dispersion in skewed distributions. *Communications in Statistics - Simulation and Computation* 28 (1):189–205. doi:10.1080/03610919908813543.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1 (2):203–9. doi:10.1016/S1535-6108(02)00030-2.

Staudte, R. G., and S. J. Sheather. 1990. *Robust estimation and testing*. New York: John Wiley & Sons.
Taylor, E., and R. Watling. 2011. Long run patterns of housing prices in Melbourne. Paper presented at the Proceedings of the State of Australian Cities Conference.
Tukey, J. W. 1965. Which part of the sample contains the information? *Proceedings of the National Academy of Sciences* 53 (1):127–34. doi:10.1073/pnas.53.1.127.

A. Appendix

A.1 Proof of Theorem 1

A power series expansion for $\mathcal{Q}_p(F_\epsilon)$ can be written as $\mathcal{Q}_p(F) + \epsilon \mathrm{IF}(x_0; \mathcal{Q}_p, F) + O(\epsilon^2)$. Setting $F_\epsilon = (1-\epsilon)F_1 + \epsilon\Delta_{x_0}$ where $\mathcal{Q}_p(F_1) = x_p$ and for simplicity, write $\mathrm{IF}_{1,p} = \mathrm{IF}(x_0; \mathcal{Q}_p, F_1)$ and recall $\mathcal{Q}_p(F_2) = y_p$. Then the first PIF is

$$\mathrm{PIF}_1(x_0; r_p, F_1, F_2) = \lim_{\epsilon \downarrow 0} \left\{ \frac{x_p + \epsilon \mathrm{IF}_{1,p} + O(\epsilon^2) - x_p}{\epsilon y_p} \right\} = \frac{\mathrm{IF}_{1,p}}{y_p}.$$

Setting $F_\epsilon = (1-\epsilon)F_2 + \epsilon\Delta_{x_0}$ and letting $\mathrm{IF}_{2,p} = \mathrm{IF}(x_0; \mathcal{Q}_p, F_2)$ the second PIF is

$$\begin{aligned}
\mathrm{PIF}_2(x_0; r_p, F_1, F_2) &= \lim_{\epsilon \downarrow 0} \left[\frac{x_p / \left(y_p + \epsilon \mathrm{IF}_{2,p} + O(\epsilon^2) \right) - x_p / y_p}{\epsilon} \right] \\
&= \lim_{\epsilon \downarrow 0} \left\{ \frac{x_p \left[y_p - \left(y_p + \epsilon \mathrm{IF}_{2,p} + O(\epsilon^2) \right) \right]}{\epsilon y_p \left(y_p + \epsilon \mathrm{IF}_{2,p} + O(\epsilon^2) \right)} \right\} \\
&= \lim_{\epsilon \downarrow 0} \left\{ \frac{-x_p \left[\epsilon \mathrm{IF}_{2,p} + O(\epsilon^2) \right)]}{\epsilon y_p \left[y_p + \epsilon \mathrm{IF}_{2,p} + O(\epsilon^2) \right]} \right\}.
\end{aligned}$$

The proof concludes after canceling the ϵ terms and taking the limit.

A.2. Proof of Theorem 2

A power series expansion for $\mathcal{Q}_{1-p}(F_\epsilon) - \mathcal{Q}_p(F_\epsilon)$ can be written as

$$\mathcal{Q}_{1-p}(F) - \mathcal{Q}_p(F) + \epsilon \left[\mathrm{IF}\left(x_0; \mathcal{Q}_{1-p}, F\right) - \mathrm{IF}\left(x_0; \mathcal{Q}_p, F\right) \right] + O(\epsilon^2).$$

Setting $F_\epsilon = (1-\epsilon)F_1 + \epsilon\Delta_{x_0}$ where $\mathcal{Q}_p(F_1) = x_p$, we have $[\mathcal{Q}_{1-p}(F_\epsilon) - \mathcal{Q}_p(F_\epsilon)]^2$ can be written

$$(x_{1-p} - x_p)^2 + 2\epsilon (x_{1-p} - x_p) \left[\mathrm{IF}\left(x_0; \mathcal{Q}_{1-p}, F_1\right) - \mathrm{IF}\left(x_0; \mathcal{Q}_p, F_1\right) \right] + O(\epsilon^2). \tag{12}$$

For simplicity, write $\mathrm{PIF}_1 = \mathrm{PIF}_1(x_0; \mathcal{R}_p, F_1, F_2)$ and $\mathrm{IF}_{1,p} = \mathrm{IF}(x_0; \mathcal{Q}_p, F_1)$ and recall $\mathcal{Q}_p(F_2) = y_p$. Since $\rho_p = (x_{1-p} - x_p)^2 / (y_{1-p} - y_p)^2$, the first PIF is

$$\begin{aligned}
\mathrm{PIF}_1 &= \lim_{\epsilon \downarrow 0} \left\{ \frac{(x_{1-p} - x_p)^2 + 2\epsilon (x_{1-p} - x_p) [\mathrm{IF}_{1,1-p} - \mathrm{IF}_{1,p}] + O(\epsilon^2) - (x_{1-p} - x_p)^2}{\epsilon (y_p - y_{1-p})^2} \right\} \\
&= \frac{2\rho_p}{(x_{1-p} - x_p)} [\mathrm{IF}_{1,1-p} - \mathrm{IF}_{1,p}].
\end{aligned}$$

Let $\mathcal{IQR}_p(F) = \mathcal{Q}_{1-p}(F) - \mathcal{Q}_p(F)$ be the functional for the IQR at p and for the second PIF set $F_\epsilon = (1-\epsilon)F_2 + \epsilon\Delta_{x_0}$. Then

$$\begin{aligned}
\mathrm{PIF}_2 &= \lim_{\epsilon\downarrow 0}\left\{\frac{(x_{1-p}-x_p)^2\left[\mathcal{IQR}^2(F_\epsilon)\right]^{-1}-(x_{1-p}-x_p)^2/(y_{1-p}-y_p)^2}{\epsilon}\right\}\\
&= \lim_{\epsilon\downarrow 0}\left\{\frac{(x_{1-p}-x_p)^2(y_{1-p}-y_p)^2-(x_{1-p}-x_p)^2\mathcal{IQR}^2(F_\epsilon)}{\epsilon(y_{1-p}-y_p)^2\mathcal{IQR}^2(F_\epsilon)}\right\}\\
&= \lim_{\epsilon\downarrow 0}\left\{\frac{-2\epsilon(x_{1-p}-x_p)^2(y_{1-p}-y_p)[\mathrm{IF}_{2,1-p}-\mathrm{IF}_{2,p}]+O(\epsilon^2)}{\epsilon(y_{1-p}-y_p)^2\mathcal{IQR}^2(F_\epsilon)}\right\}
\end{aligned}$$

when using (12) but evaluated at F_2 and letting $\mathrm{IF}_{2,p}=\mathrm{IF}(x_0;\mathcal{Q}_p,F_2)$. The proof concludes after canceling the ϵ terms and taking the limit.

A.3. Proof of Theorem 3

Note $\mathrm{IF}(x_0;\mathcal{Q}_p,F_1)^2=[p^2+(1-2p)I(x_p\geq x_0)]g_1^2(p)$. Then, from above and Theorem 1 and noting that, for example, $E_{F_1}[I(x_p\geq X)]=p$, for $X\sim F_1$,

$$E_{F_1}\left[\mathrm{PIF}_1(x_0;r_p,F_1,F_2)^2\right]=\frac{r_p^2}{x_p^2}E\left[\mathrm{IF}\left(X;\mathcal{Q}_p,F_1\right)^2\right]=\frac{p(1-p)r_p^2}{x_p^2}\left\{g_1^2(p)\right\}.$$

$E_{F_2}[\mathrm{PIF}_2(x_0;r_p,F_1,F_2)^2]$ is derived similarly and the asymptotic variance follows by applying (7).

For the ratio of IQRs, first note $\mathrm{IF}(x_0;\mathcal{Q}_{1-p},F_1)^2=[(1-p)^2-(1-2p)I(x_{1-p}\geq x_0)]g_1^2(1-p)$. $\mathrm{IF}(x_0;\mathcal{Q}_p,F_1)\mathrm{IF}(x_0;\mathcal{Q}_{1-p},F_1)=p[(1-p)-I(x_{1-p}\geq x_0)+I(x_p\geq x_0)]g_1(p)g_1(1-p)$ since $p<1-p$. For simplicity let $\mathrm{IF}_p(X)=\mathrm{IF}(X;\mathcal{Q}_p,F_1)$. Then, from above and Theorem 2 and noting that, for example, $E_{F_1}[I(x_p\geq X)]=p$, for $X\sim F_1$,

$$\begin{aligned}
E_{F_1}\left[\mathrm{PIF}_1^2\right] &= \frac{4\rho_p^2}{(x_{1-p}-x_p)^2}\left\{E\left[\mathrm{IF}_{1-p}^2(X)\right]+E\left[\mathrm{IF}_p^2(X)\right]-2E\left[\mathrm{IF}_{1-p}(X)\mathrm{IF}_p(X)\right]\right\}\\
&= \frac{4p\rho_p^2}{(x_{1-p}-x_p)^2}\left\{g_1^2(p)+g_1^2(1-p)-p\left[g_1(p)+g_1(1-p)\right]^2\right\}.
\end{aligned}$$

$E_{F_2}[\mathrm{PIF}_2^2]$ is derived similarly and the asymptotic variance follows by applying (7).

Paper II

Confidence intervals for median absolute deviations

Arachchige, Chandima NPG, & Prendergast, Luke A. 2019. arxiv preprint arxiv:1910.00229.

Submitted to Journal of Statistics in Biopharmaceutical Research.

Confidence intervals for median absolute deviations

Chandima N. P. G. Arachchige
Department of Mathematics and Statistics, La Trobe University
18201070@students.latrobe.edu.au

Luke A. Prendergast
Department of Mathematics and Statistics, La Trobe University
luke.prendergast@latrobe.edu.au

March 19, 2020

Abstract

The median absolute deviation (MAD) is a robust measure of scale that is simple to implement and easy to interpret. Motivated by this, we introduce interval estimators of the MAD to make reliable inferences for dispersion for a single population and ratios and differences of MADs for comparing two populations. Our simulation results show that the coverage probabilities of the intervals are very close to the nominal coverage for a variety of distributions. We have used partial influence functions to investigate the robustness properties of the difference and ratios of independent MADs.

Keywords: asymptotic variance, partial influence functions, robust

1 Introduction

The median absolute deviation is a robust measure of dispersion (MAD, see e.g. Hampel, 1974; Hampel *et al.* , 1986). Defined as the median of the absolute residuals from the median, the MAD is a suitable scale measure to accompany the median. Hampel (1974) referred to the MAD as the "median deviation" and it had first received attention even as early as Gauss (1816), and later rediscovered by Hampel (1968). The MAD is the *most robust* estimator of scale as measured by robustness measures such as the break-down point and gross error sensitivity (Hampel, 1974). The breakdown point of an estimator is the proportion of contamination that the estimator can handle before providing unreliable results and for the MAD this is equal to 1/2 (the maximum). The MAD estimator has what is known as a bounded influence function so that the amount of influence any observational type can exert on the estimator is limited. More will be said on the influence function later.

Arachchige *et al.* (2019a) showed that excellent coverages for interval estimators of ratios of interquantile ranges can be achieved. This makes these intervals more suitable than those for ratio of variances when normality cannot be assumed. Then, Arachchige *et al.* (2019b) considered interval estimators for robust versions of the coefficient of variation, one of which uses the MAD in place of the standard deviation (and the median to replace the mean). Motivated by these good coverage properties, we consider interval estimators for the MAD and for ratios and differences of independent MADs as robust alternatives to intervals based on sample variances. To the best of our knowledge, and not to confuse the MAD with the *mean* absolute deviation for which interval estimators with good coverage have been introduced by Bonett & Seier (2003), no one has introduced these interval estimators for the MAD. The very good coverage properties, that we will highlight later, ensure inferences about dispersion based on the MAD are possible.

In Section 2 we provide some necessary notations before considering influence functions for ratios of MADs. In Section 3 we consider confidence intervals for MADs, differences of MADs and ratios of MADs with coverage properties explored via simulations in Section 4. Examples are also considered in Section 4 and we conclude in Section 5.

2 Notations and influence functions

Let X denote a random variable and F its distribution function. Then Hampel (1974) defined the median absolute deviation (MAD) as

$$\mathrm{MAD}(X) = \mathrm{med} \mid X - M \mid \ , \tag{2.1}$$

where 'med' denotes the median and $M = \mathrm{med}(X) = F^{-1}(0.5)$ is the population median. Let $X_1, \ldots, X_n$ denote a random sample of n observations. Then the MAD estimate is simply the median of the absolute residuals from the sample median. That is, for m denoting the sample median, $\widehat{\mathrm{MAD}}$ is the sample median of the $|X_1 - m|, \ldots, |X_n - m|$. While inference, for a single MAD may be of interest, it is often the case that comparison of dispersion measures, such as the MAD, is needed to compare two populations.

Consider two independent random variables $X \sim F_1$ and $Y \sim F_2$ and let us consider $\mathrm{MAD}(X)$ and $\mathrm{MAD}(Y)$. Then, the population squared ratio of MADs, which we denote as R_M, and associated estimator can be define as

$$R_M = \left[\frac{\mathrm{MAD}(X)}{\mathrm{MAD}(Y)}\right]^2 \ \text{ and } \ \widehat{R}_M = \left[\frac{\widehat{\mathrm{MAD}}(X)}{\widehat{\mathrm{MAD}}(Y)}\right]^2 . \tag{2.2}$$

Here we have suggested the squared ratio of MADs since it is the analogue to the ratio of variances and, in fact, equal to ratio of variances for some distributions (e.g. normal). However, the ratio of MADs may also be used. Another possibility is the difference of MADs, D_M, where

$$D_M = \mathrm{MAD}(X) - \mathrm{MAD}(Y) \ \text{ and } \ \widehat{D}_M = \widehat{\mathrm{MAD}}(X) - \widehat{\mathrm{MAD}}(Y) \ . \tag{2.3}$$

2.1 Influence function and partial influence functions

Define the contamination distribution to be $F_\epsilon = (1-\epsilon)F + \epsilon\Delta_x$, where $\epsilon \in [0, 1]$ is the proportion of contamination and Δ_x has all of its mass at the contaminant x. Consider an estimator functional $\mathcal{T}$ such that $\mathcal{T}(F) = \theta$ and $\mathcal{T}(F_n) = \widehat{\theta}$ where F_n denotes the empirical distribution function for sample of n observations. The relative influence on $\mathcal{T}(F)$ of ϵ proportion of contaminated observations at x is given by, $[\mathcal{T}(F_\epsilon)-\mathcal{T}(F)]/\epsilon$, where $\mathcal{T}(F_\epsilon)=(1-\epsilon)\mathcal{T}(F)+\epsilon\Delta_x$. Then, the influence function (IF Hampel, 1974) is defined as,

$$\mathrm{IF}(x;\mathcal{T},F)=\lim_{\epsilon\downarrow 0}\frac{\mathcal{T}(F_\epsilon)-\mathcal{T}(F)}{\epsilon}\equiv\frac{\partial}{\partial\epsilon}\mathcal{T}(F_\epsilon)\Big|_{\epsilon=0}.$$

When more than one population exists, the IF is determined by contaminating one population while the other population remains uncontaminated. Pires & Branco (2002) defines this notion as "partial IFs" (PIFs) and in our context with two populations we have two PIFs. The first PIF of the estimator with functional $\mathcal{T}$ at $(F_1,\ F_2)$ is

$$\mathrm{PIF}_1(x;\mathcal{T},F_1,F_2)=\lim_{\epsilon\to 0}\left[\frac{\mathcal{T}[(1-\epsilon)F_1+\epsilon\Delta_{x_0},F_2]-\mathcal{T}(F_1,F_2)}{\epsilon}\right] \tag{2.4}$$

and with $\mathrm{PIF}_2(x;\mathcal{T},F_1,F_2)$ defined similarly.

Now, consider the functional for the standardized MAD denoted by $\mathcal{MAD}$ so that $\mathcal{MAD}(F)=\mathrm{MAD}_X$. Hampel (1974) gives the influence function for the MAD when F is the normal distribution and further details can be found on page 107 of Hampel *et al.* (1986). Let $f=F'$ denote the density function then, assuming $f(M)$ and $2[f(M+\mathrm{MAD}_X)+f(M-\mathrm{MAD}_X)]$ are nonzero, a general form of the IF for the MAD exists; e.g. see Theorem 1.5.7 (page 22 and 23) of see page 137 of Huber (1981) or page 16 of Andersen (2008). This is given as

$$\mathrm{IF}(x;\,\mathcal{MAD},F)=\frac{[\mathrm{sign}(x-M)-\mathrm{MAD}_X]-\dfrac{f(M+\mathrm{MAD}_X)-f(M-\mathrm{MAD}_X)}{f(M)}\mathrm{sign}(x-M)}{2[f(M+\mathrm{MAD}_X)+f(M-\mathrm{MAD}_X)]}. \tag{2.5}$$

2.1.1 Partial influence functions of the difference and squared ratio of MADs

Let $\mathcal{D}_M$ be the functional for the difference of MADs so that,

$$\mathcal{D}_M(F_1,F_2)=\mathcal{MAD}(F_1)-\mathcal{MAD}(F_2)$$

then the PIFs are $\mathrm{PIF}_1(x;\mathcal{D}_M,F_1,F_2)=\mathrm{IF}(x;\mathcal{MAD},F_1)$ and $\mathrm{PIF}_2(x;\mathcal{D}_M,F_1,F_2)=-\mathrm{IF}(x;\mathcal{MAD},F_2)$. These are trivial and previous studies on robustness of the MAD may be considered for this context. We therefore do not explore the difference PIFs further.

Let $\mathcal{R}_M$ be the functional for the squared ratio of MADs so that,

$$\mathcal{R}_M(F_1,F_2)=\left[\frac{\mathcal{MAD}(F_1)}{\mathcal{MAD}(F_2)}\right]^2 .$$

Then the PIFs for the squared ratio of MADs are given below.

Theorem 2.1. *For $PIF(x;T,F_1,F_2)$ as defined in (2.4), the PIFs of $\mathcal{R}_M$ are*

$$PIF_1(x;\mathcal{R}_M,F_1,F_2)=\frac{2\,\mathcal{R}_M(F_1,F_2)}{\mathcal{MAD}(F_1)}IF(x;\mathcal{MAD},F_1),$$
$$PIF_2(x;\mathcal{R}_M,F_1,F_2)=-\frac{2\,\mathcal{R}_M(F_1,F_2)}{\mathcal{MAD}(F_2)}IF(x;\mathcal{MAD},F_2).$$

The proof of Theorem 2.1 is in Appendix A.2 and we consider some examples of the first PIF next.

2.1.2 Partial influence functions comparison

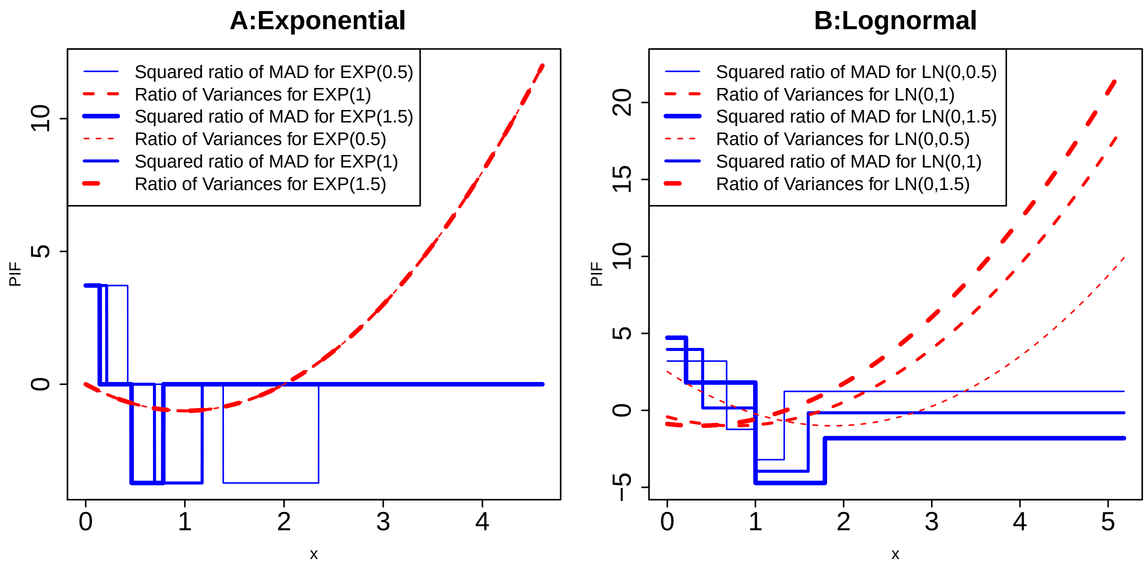


Figure 1: PIF_1 comparisons for (A) two exponential populations both with rates 0.5, 1 and 1.5 and (B) two log-normal populations both with μ=0 and σ=0.5,1,1.5.

Figure 1 depicts the PIFs of the first population for the squared ratio of MADs and the ratio of variances (see Arachchige *et al.* , 2019a, for these). In Plot A we consider the ratio of variances and squared ratio of MADs for two exponential distributions, both with rates equal to 0.5, or 1 or 1.5. Similarly, in Plot B we do this for two log normal distributions both with μ=0 and $\sigma = 0.5$ or 1 or 1.5. Since the numerator and denominator distributions are the same, both are estimators of one and therefore the PIFs are comparable. As expected, the PIFs of the ratio of variances is unbounded indicating that outliers can exert large influence on the estimator. The PIFs of the squared ratio of MAD is bounded and the influence of any large outliers is limited, and far less than for the ratio of variances. For the exponential distribution, the PIFs of ratio of variances do not depend on the rate parameter. However, for the log-normal distribution the PIF for the ratio of variances increases quickly with increasing σ.

3 Asymptotic confidence intervals

In their discussion of intervals for the mean absolute deviations, Bonett & Seier (2003) provide suggestions for median absolution deviations from a fixed point, h. They suggest using intervals for the median and where the data used is the transformed $|X_i - h|$s. When h is the population median, i.e. $h = M$, and this median is known, simulations (not shown) result in good coverage that is close to nominal. However, when M is not known and needs to be estimated, this approach typically results in coverage that is too low (e.g. less than 0.8 for a nominal 0.95). In this section we therefore provide confidence intervals that have good coverage properties, as shown by our simulations that follow.

Asymptotic normality and associated variance of the MAD can be found in Falk (1997) who provide the asymptotic joint normality between the median and MAD estimators. We again let $\text{MAD}_X = \mathcal{MAD}(F)$ and also let $\mathcal{MAD}(F_n) = \widehat{\text{MAD}}_X$. Then, if F is continuous near, and differentiable at, the median M, $M - \text{MAD}_X$ and $M + \text{MAD}_X$ with $f(M) > 0$ and $B_1 = f(M - \text{MAD}_X) + f(M + \text{MAD}_X) > 0$, we have

$$\sqrt{n}\left(\widehat{\text{MAD}}_X - \text{MAD}_X\right) \overset{\text{approx.}}{\sim} N(0, \text{ASV}) \ ,$$

where '$\overset{\text{approx.}}{\sim}$' denotes 'approximately distributed'. The asymptotic variance of the MAD estimator is

$$\text{ASV} = \text{ASV}(\mathcal{MAD}; F) = \frac{1}{4B_1^2}\left[1 + \frac{B_2}{[f(M)]^2}\right] \ , \tag{3.1}$$

where B_1 is given above and $B_2 = B_3^2 + 4B_3 f(M)\left[1 - F(M + \text{MAD}_X) - F(M - \text{MAD}_X)\right]$ with $B_3 = f(M - \text{MAD}_X) - f(M + \text{MAD}_X)$.

We used the ASV in (3.1) and the Delta method (see e.g., chapter 3 of DasGupta, 2006) to derive the asymptotic variance of the ratios of MADs. The asymptotic variance of $\sqrt{n_1 + n_2}\mathcal{R}_M(F_{n_1}, F_{n_2})$ is

$$\text{ASV}(\mathcal{R}_M; n_1, n_2) = 4\mathcal{R}_M^2(F_1, F_2)\left[\frac{\text{ASV}(\mathcal{MAD}, F_1)}{w_1\, \mathcal{MAD}^2(F_1)} + \frac{\text{ASV}(\mathcal{MAD}, F_2)}{w_2\, \mathcal{MAD}^2(F_2)}\right] \tag{3.2}$$

where $w_i = n_i/(n_1 + n_2)$ for $i = 1, 2$.

Since the two populations are independent, deriving the asymptotic variance of the difference of MAD is straightforward.

$$\text{ASV}(\mathcal{D}_M; n_1, n_2) = \text{ASV}(\mathcal{MAD}, F_1) + \text{ASV}(\mathcal{MAD}, F_2). \tag{3.3}$$

Throughout, let $\widehat{\text{ASD}}(\cdot) = \sqrt{\widehat{\text{ASV}}(\cdot)}$ denote the estimated asymptotic standard deviation estimate. Note that the ASV depends on both f and F, the density and distribution functions. There are several options to estimate these, but we choose to use the very flexible Generalized Lambda Distribution (GLD) which, for the FKML parameterization (Freimer *et al.* , 1988), is defined in terms of its quantile function, $Q(p)$,

$$Q(p) = \lambda_1 + \lambda_2^{-1}\left\{\lambda_3^{-1}(p^{\lambda_3} - 1) - \lambda_4^{-1}[(1-p)^{\lambda_4} - 1]\right\} \ ,$$

where λ_1, λ_2, λ_3 and λ_4 are the location, inverse scale and two shape parameters respectively. To estimate the GLD parameters we use a recent approach introduced by Dedduwakumara *et al.* (2019a) which is computationally efficient making it useful for our simulations that follow. However, other estimators can also be used. We then use these parameter estimates with the density and distribution functions for the GLD in R `gld` package (King *et al.* , 2016).

Based on asymptotic normality of the MAD (e.g. Falk, 1997), an asymptotic $(1-\alpha)\%$ confidence interval for MAD is given as

$$[L, U]_{\text{MAD}} = \left[\widehat{\text{MAD}}_X \pm z_{1-\alpha/2}\,\frac{\widehat{\text{ASD}}(\mathcal{MAD}, F_n)}{\sqrt{n}}\right], \tag{3.4}$$

where the $z_{1-\alpha/2}$ is the $(1-\alpha/2)\times 100$ percentile of the standard normal distribution.

When constructing the interval estimator for the squared ratio of MADs, we first derive the confidence interval for the log transformed ratio and then exponentiate to return to the ratio scale. Let $\mathcal{W}(F_1, F_2) = \ln[\mathcal{R}_M(F_1, F_2)]$ then, using the Delta method, it is straightforward to show that $\text{ASV}(\mathcal{W}, F_1, F_2) \doteq \text{ASV}(\mathcal{R}_M, F_1, F_2)/[\mathcal{R}_M(F_1, F_2)]^2$. Then a $(1-\alpha)\%$ confidence interval estimator for R_M is given as

$$[L, U]_{R_M} = \exp\left[\ln(\widehat{R}_M) \pm z_{1-\alpha/2}\,\frac{\widehat{\text{ASD}}(\mathcal{R}_M, F_{n_1}, F_{n_2})}{\widehat{R}_M\sqrt{n_1+n_2}}\right], \tag{3.5}$$

where $\widehat{R}_M$ is the squared ratio of MADs estimator and the ASV is in (3.2).

Finally, a $(1-\alpha)\%$ confidence interval for the difference in MADs is simply

$$[L, U]_{D_M} = \widehat{D}_M \pm z_{1-\alpha/2}\,\frac{\widehat{\text{ASD}}(\mathcal{D}_M, F_{n_1}, F_{n_2})}{\sqrt{n_1+n_2}}\ , \tag{3.6}$$

where $\widehat{D}_M$ is the difference of MADs estimator and the ASV can be found in (3.3).

4 Simulations and Examples

We begin by conducting simulations to assess the coverage properties of the interval estimations for data generated from several distributions. As pointed out earlier, we have used a new estimator of the GLD parameters provided by Dedduwakumara *et al.* (2019b) since it exhibits very good performance and is very efficient making it useful for our simulations. In Appendix A.2, we provide R code for the interval estimators using readily available estimators for the GLD from the `gld` package (King *et al.* , 2016). In that code we have opted for Titterington's method (Titterington, 1985) since it to has good performance, albeit is more time consuming.

4.1 Simulations

To investigate the performance of the MAD, squared ratio of MADs and difference of MADs intervals we consider simulated coverage probability and the average confidence interval width as performance measures. We have selected the log normal (LN), exponential (EXP), chi-square (χ^2_5) and Pareto (PAR) distributions with different sample sizes of $n = 50, 100, 200, 500, 1000$. Each simulation consists of 10,000 trials.

Simulated coverages and widths for the interval estimator of MADs, from (3.4), are provided in Table 1 for several distributions. The coverage probabilities are all close to the nominal level of 0.95, even for $n = 50$ where coverages were approximately in the vicinity of 0.93-0.94.

Table 1: Simulated coverage probabilities (and widths in parentheses) for the 95% confidence interval for the MAD (* denotes median width reported due to excessively large widths for a small number of intervals that skew the mean).

Sample size	$X \sim$ LN(0,1)	$X \sim$ EXP(1)	$X \sim \chi^2_5$	$X \sim$ PAR(1,7)
True MAD =	0.599	0.481	1.895	0.075
50	0.938 (1.43)	0.936 (1.93)	0.927 (1.25*)	0.939 (0.34)
100	0.940 (0.37)	0.939 (0.29)	0.938 (0.91)	0.939 (0.05)
200	0.938 (0.26)	0.947 (0.20)	0.942 (0.65)	0.944 (0.03)
500	0.945 (0.16)	0.948 (0.12)	0.947 (0.41)	0.949 (0.02)
1000	0.946 (0.12)	0.951 (0.09)	0.944 (0.29)	0.947 (0.01)

Table 2: Simulated coverage probabilities (and widths in parentheses) for the 95% confidence interval for the squared ratio of MADs (R_M) and difference of MADs (D_M) (* Median width reported due to excessively large widths for a small number, between 1% and 2%, of intervals).

Sample sizes (n_1,n_2)	Measure	$X \sim$ LN(0,1) $Y \sim$ LN(0,1)	$X \sim$ EXP(1) $Y \sim$ EXP(1)	$X \sim \chi^2_5$ $Y \sim \chi^2_2$	$X \sim$ PAR(1,7) $Y \sim$ PAR(1,3)
	True R_M =	1	1	3.876	0.148
	True D_M =	0	0	0.932	-0.119
50,50	R_M	0.958 (3.71*)	0.971 (4.03*)	0.955 (12.14*)	0.978 (0.91*)
	D_M	0.967 (2.55)	0.972 (3.49)	0.956 (1.54*)	0.967 (1.17)
100,100	R_M	0.949 (2.23)	0.958 (1.87*)	0.954 (6.48*)	0.960 (0.33*)
	D_M	0.954 (0.52)	0.958 (0.42)	0.952 (1.08)	0.951 (0.16)
200,200	R_M	0.953 (1.37)	0.946 (1.28)	0.950 (4.51)	0.952 (0.22)
	D_M	0.945 (0.37)	0.950 (0.28)	0.950 (0.76)	0.947 (0.10)
200,500	R_M	0.946 (1.09)	0.951 (1.02)	0.950 (3.47)	0.952 (0.17)
	D_M	0.945 (0.31)	0.951 (0.23)	0.946 (0.69)	0.956 (0.07)
500,500	R_M	0.946 (0.81)	0.952 (0.75)	0.949 (2.69)	0.950 (0.12)
	D_M	0.948 (0.23)	0.953 (0.17)	0.950 (0.48)	0.947 (0.06)
500,1000	R_M	0.947 (0.69)	0.952 (0.64)	0.948 (2.23)	0.951 (0.10)
	D_M	0.947 (0.20)	0.949 (0.15)	0.949 (0.45)	0.948 (0.04)
1000,1000	R_M	0.947 (0.56)	0.949 (0.52)	0.949 (1.87)	0.950 (0.09)
	D_M	0.944 (0.16)	0.950 (0.12)	0.952 (0.34)	0.948 (0.04)

Coverages become closer to the nominal level as the sample size increases and, as expected the interval widths decrease with increasing sample size.

Simulated coverages for interval estimators of squared ratio of MADs and difference of MADs are provided in Table 2 for several distributions. Results show excellent coverages compared to the coverages of F-test (the coverage probabilities for interval estimator of the F-test can be found in Table 3 of Arachchige *et al.* , 2019a) which are poor due to the violation of underlying normality assumptions). Coverages are very close to the nominal 0.95 for both the squared MAD ratio and difference of MAD for all the selected distributions, including smaller sample sizes. There are some slightly conservative coverages only for $n = 50$ and for other sample sizes the coverages become very close. For smaller sample sizes a very small number of the intervals were very wide (between 1% and 2%) so we report the median width instead.

4.2 Prostate data example

The prostate data set, which is available in the `depthTools` package (Lopez-Pintado & Torrente, 2013), is a normalized subset of the Singh *et al.* (2002) prostate data set. The data consists of gene expressions for the 100 most variable genes for 25 normal and 25 tumoral samples.

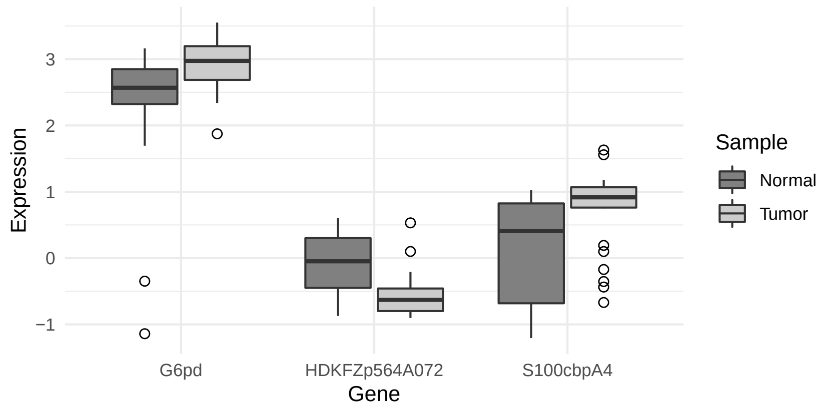


Figure 2: Box plots of three interesting genes selected from the prostate data set.

We selected three genes that are interesting when comparing intervals for ratios of variances and those based on the MADs. These three genes are three of the six that were considered by Arachchige *et al.* (2019a). The genes and their abbreviations we consider are Glucose-6-phosphate dehydrogenase (G6pd), HDKFZp564A072 and calcium-binding protein A4 (S100cbpA4). Box plots of the genes are provided in Figure 2 where we note that, ignoring extreme outliers, the spread for the bulk of the data looks similar for G6pd and very different for HDKFZp564A072 and S100cbpA4.

In Table 3 we provide the point estimate and asymptotic 95% confidence intervals for the ratio of variances (from the F-test assuming underlying normality), the squared ratios of MADs and difference of MADs for the three selected genes. When ignoring two outliers for G6pd the spread looks similar, however the interval for the ratio of variances suggests a large difference in variance between the two. This is not the case for the MAD intervals where the point and

Table 3: 95% asymptotic confidence intervals (CI) for the ratios of variances resulting from the F-test (F), the ratio of MADs (R_M) and difference of MADs (D_M) for the three selected genes.

	F		R_M		D_M	
Gene	Est.	CI	Est.	CI	Est.	CI
G6pd	6.496	(2.863, 14.742)	1.000	(0.268, 3.734)	0.000	(-0.185, 0.185)
HDKFZp564A072	1.930	(0.850, 4.379)	5.013	(1.211, 20.761)	0.213	(0.035, 0.391)
S100cbpA4	1.748	(0.770, 3.968)	8.725	(1.440, 52.856)	0.301	(-0.013, 0.615)

intervals estimates suggest little difference. For HDKFZp564A072 and S100cbpA4 the intervals tell a different story. The ratio of variance intervals do not find a significant difference, while the MAD intervals do, or in the case of the difference very close to. We favor the findings from the MAD due to the obvious difference in spread for the bulk of the data as depicted in the box plots. This difference in findings is likely due to the group with smaller spread for most data, have extreme outliers that increases the sample variance so that it is similar to the sample variance for the other group. The MADs are not affected by these outliers. Arachchige *et al.* (2019a) provide similar contrasting results when comparing an asymptotic interval for the ratio of variances and intervals based on the interquantile range.

5 Summary and discussion

The MAD is a robust estimator of scale exhibiting good robustness properties. We have considered interval estimators for the MAD, ratios of MADs and differences of MADs. Simulation results for the interval estimators showed excellent coverages even for small sample sizes such as $n = 50$ for all distributions we considered. Our example reveals that different conclusions can be made by using ratios of MADs and differences of MADs compared to intervals for the ratio of variances which is influenced by outliers. Future extensions to this work would be to consider intervals for alternatives to the MAD (e.g. see Rousseeuw & Croux, 1993).

A Appendix

A.1 Proof of Theorem 2.1

Proof. A power series expansion of $\mathcal{MAD}(F_\epsilon)$ can be written as

$$\mathcal{MAD}(F_\epsilon) = \mathcal{MAD}(F) + \epsilon \mathrm{IF}(x; \mathcal{MAD}, F) + O(\epsilon^2) \ .$$

Let $F_\epsilon = (1-\epsilon)F_1 + \epsilon\Delta_x$, then we have

$$[\mathcal{MAD}(F_\epsilon)]^2 = \mathcal{MAD}^2(F_1) + 2\epsilon\mathcal{MAD}(F_1)\mathrm{IF}(x; \mathcal{MAD}, F_1) + O(\epsilon^2) \ .$$

Therefore, the first PIF is

$$\text{PIF}_1(x;\mathcal{R}_M,F_1,F_2)=\lim_{\epsilon\downarrow 0}\left\{\frac{\mathcal{MAD}^2(F_1)+2\epsilon\mathcal{MAD}(F_1)\text{IF}(x;\mathcal{MAD},F_1)+O(\epsilon^2)-\mathcal{MAD}^2(F_1)}{\epsilon\mathcal{MAD}^2(F_2)}\right\}$$

For the second PIF set $F_\epsilon=(1-\epsilon)F_2+\epsilon\Delta_x$. Then

$$\begin{aligned}\text{PIF}_2(x;\mathcal{R}_M,F_1,F_2)&=\lim_{\epsilon\downarrow 0}\left\{\frac{\mathcal{MAD}^2(F_1)\left[\mathcal{MAD}^2(F_\epsilon)\right]^{-1}-\mathcal{MAD}^2(F_1)/\mathcal{MAD}^2(F_2)}{\epsilon}\right\}\\&=\lim_{\epsilon\downarrow 0}\left\{\frac{\mathcal{MAD}^2(F_1)\mathcal{MAD}^2(F_2)-\mathcal{MAD}^2(F_1)\mathcal{MAD}^2(F_\epsilon)}{\epsilon\mathcal{MAD}^2(F_2)\mathcal{MAD}^2(F_\epsilon)}\right\}\\&=\lim_{\epsilon\downarrow 0}\left\{\frac{-2\epsilon\mathcal{MAD}^2(F_1)\mathcal{MAD}(F_2)\text{IF}(x;\mathcal{MAD},F_2)+O(\epsilon^2)}{\epsilon\mathcal{MAD}^2(F_2)\mathcal{MAD}^2(F_\epsilon)}\right\}\end{aligned}$$

Recall the $\text{IF}(x;\mathcal{MAD},F)$ in (2.5) and evaluated at F_1 and F_2. Finally, the PIF_1 and PIF_2 can be obtained by taking the limit by noting that $\lim_{\epsilon\downarrow 0}[O(\epsilon^2)/\epsilon]=0$. □

A.2 R code for interval estimators

```
# This codes uses the gld R package for estimation of the GLD since it is
# readily available in R.
library(gld)
library(stats)

mad <- mad(x, center = median(x), constant = 1, na.rm = FALSE,
      low = FALSE, high = FALSE)

asv.mad <- function(x, method = "TM"){
  lambda <- fit.fkml(x, method = method)$lambda
  m  <- median(x)
  mad.x <- mad(x)
  fFinv <- dgl(c(m - mad.x, m + mad.x, m), lambda1 = lambda)
  FFinv <- pgl(c(m - mad.x, m + mad.x), lambda1 = lambda)
  A <- fFinv[1] + fFinv[2]
  C <- fFinv[1] - fFinv[2]
  B <- C^2 + 4*C*fFinv[3]*(1 - FFinv[2] - FFinv[1])
  (1/(4 * A^2))*(1 + B/fFinv[3]^2)
}

ci.mad <- function(x, y = NULL, gld.est = "TM",
          two.samp.diff = TRUE, conf.level = 0.95){
  alpha <- 1 - conf.level
```

```
  z <- qnorm(1 - alpha/2)
  x <- x[!is.na(x)]
  est <- mad.x <- mad(x)
  n.x <- length(x)
  asv.x <- asv.mad(x, method = gld.est)
  if(is.null(y)){
    ci <- mad.x + c(-z, z)*sqrt(asv.x/n.x)
  } else{
    y <- y[!is.na(y)]
    mad.y <- mad(y)
    n.y <- length(y)
    asv.y <- asv.mad(y, method = gld.est)
    if(two.samp.diff){
      est <- mad.x - mad.y
      ci <- est + c(-z, z)*sqrt(asv.x/n.x + asv.y/n.y)
    } else{
      est <- (mad.x/mad.y)^2
      log.est <- log(est)
      var.est <- 4 * est * ((1/mad.y^2)*asv.x/n.x + (est/mad.y^2)*asv.y/n.y)
      Var.log.est <- (1 / est^2) * var.est
      ci <- exp(log.est + c(-z, z) * sqrt(Var.log.est))
    }
  }
  list(Estimate = est, conf.int = ci)
}

x <- rlnorm(100)
y <- rlnorm(200, meanlog = 1.2)

ci.mad(x) # single sample
ci.mad(x, y) # two sample difference
ci.mad(x, y, two.samp.diff = FALSE) # two sample squared ratio
```

References

Andersen, R. 2008. *Modern methods for robust regression.* Sage.

Arachchige, C. NPG, Cairns, M., & Prendergast, L. A. 2019a. Interval estimators for ratios of independent quantiles and interquantile ranges. *Commun. Stat. B-Simul*, 1–17.

Arachchige, C. NPG, Prendergast, L. A., & Staudte, R. G. 2019b. Robust analogues to the coefficient of variation. *arXiv preprint arXiv:1907.01110.*

Bonett, D. G., & Seier, E. 2003. Confidence intervals for mean absolute deviations. *Am. Stat.*, **57**(4), 233–236.

DasGupta, A. 2006. *Asymptotic Theory of Statistics and Probability.* New York, NY: Springer.

Dedduwakumara, D. S., Prendergast, L. A., & Staudte, R. G. 2019a. An efficient estimator of the parameters of the generalized lambda distribution. *arXiv preprint arXiv:1907.06336.*

Dedduwakumara, D. S., Prendergast, L. A., & Staudte, R. G. 2019b. A simple and efficient method for finding the closest generalized lambda distribution to a specific model. *Cogent Mathematics & Statistics*, **6**, 1–11.

Falk, Michael. 1997. Asymptotic independence of median and MAD. *Stat. Probabil. Lett.*, **34**(4), 341–345.

Freimer, M., Kollia, G., Mudholkar, G. S., & Lin, C. T. 1988. A study of the generalized tukey lambda family. *Commun. Stat. Theory Methods*, **17**(10), 3547–3567.

Gauss, Carl F. 1816. *Demonstratio nova altera theorematis omnem functionem algebraicam.* apvd Henricvm Dieterich.

Hampel, F. R. 1968. Contribution to the theory of robust estimation. *Ph. D. Thesis, University of California, Berkeley.*

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.*, **69**, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. 1986. *Robust statistics: the approach based on influence functions.* John Wiley & Sons.

Huber, P. J. 1981. *Robust statistics.* Wiley.

King, R., Dean, B., & Klinke, S. 2016. *gld: Estimation and Use of the Generalised (Tukey) Lambda Distribution.* R package version 2.4.1.

Lopez-Pintado, S., & Torrente, A. 2013. *depthtools: Depth tools package.* R package version 0.4.

Pires, A. M., & Branco, J. A. 2002. Partial influence functions. *J. Multivariate Anal.*, **83**(2), 451–468.

Rousseeuw, P. J., & Croux, C. 1993. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.*, **88**(424), 1273–1283.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., *et al.* . 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**(2), 203–209.

Paper III

Robust analogues to the coefficient of variation

Arachchige, C. NPG, Prendergast, L. A., & Staudte, R. G. 2019b. arXiv preprint arXiv:1907.01110.

Submitted to Journal of Applied Statistics.

Robust analogues to the Coefficient of Variation

Chandima N. P. G. Arachchige
Department of Mathematics and Statistics, La Trobe University
18201070@students.latrobe.edu.au

Luke A. Prendergast
Department of Mathematics and Statistics, La Trobe University
luke.prendergast@latrobe.edu.au

Robert G. Staudte
Department of Mathematics and Statistics, La Trobe University
r.staudte@latrobe.edu.au

March 19, 2020

Abstract

The coefficient of variation (CV) is commonly used to measure relative dispersion. However, since it is based on the sample mean and standard deviation, outliers can adversely affect the CV. Additionally, for skewed distributions the mean and standard deviation do not have natural interpretations and, consequently, neither does the CV. Here we investigate the extent to which quantile-based measures of relative dispersion can provide appropriate summary information as an alternative to the CV. In particular, we investigate two measures, the first being the interquartile range (in lieu of the standard deviation), divided by the median (in lieu of the mean), and the second being the median absolute deviation (MAD), divided by the median, as robust estimators of relative dispersion. In addition to comparing the influence functions of the competing estimators and their asymptotic biases and variances, we compare interval estimators using simulation studies to assess coverage.

Keywords: influence function, median absolute deviation, quantile density

1 Introduction

The coefficient of variation (CV), defined to be the ratio of the standard deviation to the mean, is the most commonly used method of measuring relative dispersion. It has applications in many areas, including engineering, physics, chemistry, medicine, economics and finance, to name just a few. For example, in analytical chemistry the CV is widely used to express the precision and repeatability of an assay (Reed *et al.* , 2002). In finance the coefficient of variation is often considered useful in measuring relative risk (Miller & Karson, 1977) where a test of the equality of the CVs for two stocks can be performed to compare risk. In economics, the CV is a summary statistic of inequality (e.g. Atkinson, 1970; Chen & Fleisher, 1996). Other examples use the CV to assess the homogeneity of bone test samples (Hamer *et al.* , 1995), assessing strength of ceramics (Gong & Li, 1999) and as a summary statistic to describe the development of age- and sex-specific cut off points for body-mass indexing in overweight children (Cole *et al.* , 2000).

The lack of robustness to outliers of moment-based measures such as the mean and standard deviation has long been known. Almost a century ago Lovitt & Holtzclaw (1929) proposed a measure called the "coefficient of variability "based on the upper and lower quartiles (Q_3 and Q_1). Promoted as an alternative to the CV, it was defined to be $(Q_3 - Q_1)/(Q_3 + Q_1)$. Bonett (2006) have since called this measure the "coefficient of quartile variation " and introduced an interval estimator which exhibited good coverage even for small samples. This measure was recently re-investigated by Bulent & Hamza (2018) and they have constructed bootstrap confidence intervals that typically provide conservative coverage. Another alternative measure is to take the ratio of the mean absolute deviation from the median divided by the median. This measure has applications in tax assessments (Gastwirth, 1982) and confidence intervals have been considered by Bonett & Seier (2005). The mean absolute deviation is still non-robust to outliers, and robustness can be improved (see e.g. Shapiro, 2005; Reimann *et al.* , 2008; Varmuza & Filzmoser, 2009) by instead using the interquartile range (IQR) or the *median* absolute deviation (MAD).

For decades, interval estimation for the CV has attracted the attention of many researchers. For example, Gulhar *et al.* (2012) compared no less than 15 parametric and non-parameic confidence interval estimators of the population CV. To the best of our knowledge interval estimators have not been introduced for the coefficient of variation based on the IQR and MAD. Therefore, given the obvious need for interval estimators that has attracted the interest for many others, one aim of this paper is to provide reliable interval estimators. We are motivated to do so by noting the excellent coverage achieved for measures based on ratios of quantiles, even for small samples (Prendergast & Staudte, 2016b, 2017a,b; Arachchige *et al.* , 2019).

2 Notations and some selected methods

Let $X_1, X_2, ..., X_n$ be an independent and identically distributed sample of size n from a distribution with distribution function F. Then the sample mean estimator is $\overline{X} = n^{-1}\sum_i^n X_i$ and sample variance estimator is $S^2 = \sum_i^n (X_i - \overline{X})^2/(n-1)$. The sample coefficient of variation estimator is then $\widehat{\mathrm{CV}} = S/\overline{X}$. Next let $\mathcal{F}$ be the class of all right-continuous cdfs on the positive axis; that is each $F \in \mathcal{F}$ satisfies $F(0) = 0$. For a sample denoted $x_1, \ldots, x_n$, the statistics $\overline{x}$, s, and $\widehat{\mathrm{cv}} = s/\overline{x}$ are the observed values of the $\overline{X}$, S and $\widehat{\mathrm{CV}}$ estimators above, and are therefore estimates of the unknown population parameters $\mu = \mathrm{E}_F[X]$, $\sigma = \sqrt{\mathrm{E}_F[(X-\mu)^2]}$ and $\mathrm{CV} = \sigma/\mu$, assuming the first two moments of F exist.

For each such $F \in \mathcal{F}$ define the associated left-continuous *quantile function* of F by $Q(u) \equiv \inf\{x: \; F(x) \geq u\}$, for $0 < u < 1$. When the population F is understood to be fixed but unknown, we sometimes simply write $x_u = Q(u)$ and write the corresponding estimators of these population quantiles as $\widehat{x}_u$. We restrict attention to the quartiles $x_{0.25}$, $x_{0.5}$ and $x_{0.75}$, the sample estimates of which we denote q_1, m and q_3 for convenience.

2.1 Selected interval estimators of the CV

We begin by describing the inverse method (Sharma & Krishna, 1994) for obtaining an interval estimator for the CV since it is perhaps the most naturally arising interval involving only basic principles. As additional methods for comparison later, we have chosen four of the 15 considered in Gulhar *et al.* (2012) that exhibited comparatively good performance in terms of coverage.

While parametric interval estimators for the CV have typically been developed assuming an underlying normal distribution, such as those that we present below, for large sample sizes, they can also perform well (Gulhar *et al.* , 2012) when there are deviations from normality due to the Central Limit Theorem.

The inverse method

Using the above notation, for suitably large n, $\overline{x}/s$ is approximately $N(0, 1/n)$ distributed. An approximate $(1-\alpha/2) \times 100\%$ confidence interval for μ/σ is therefore $\overline{x}/s \pm z_{1-\alpha/2}/\sqrt{n}$. Noting that μ/σ is simply the inverse of the population CV, an approximate 95% confidence interval for the CV can therefore be obtained by inverting this interval for μ/σ, giving (Sharma & Krishna, 1994)

$$\left\{ \left[\frac{1}{\widehat{\text{cv}}} + z_{1-\alpha/2} \left(\frac{1}{n^{1/2}} \right) \right]^{-1}, \left[\frac{1}{\widehat{\text{cv}}} - z_{1-\alpha/2} \left(\frac{1}{n^{1/2}} \right) \right]^{-1} \right\} . \tag{2.1}$$

Robustness of this interval estimator was recently re-investigated by Groeneveld (2011).

The median-modified Miller interval (Med Mill)

The CV estimator has an approximate asymptotic normal distribution with mean CV and variance $(n-1)^{-1}\text{CV}^2(0.5+\text{CV}^2)$ leading to an asymptotic interval proposed by Miller (1991). In noting that the mean is a poor summary statistic of central location for skewed distributions, Gulhar *et al.* (2012) proposed a median modification where the sample median replaces the sample mean in s. Let $\tilde{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2}$ and $\widetilde{\text{cv}} = \tilde{s}/\overline{x}$, the interval estimator is

$$\left\{ \widetilde{\text{cv}} - z_{1-\alpha/2}\sqrt{(n-1)^{-1}\widetilde{\text{cv}}^2 \left(0.5 + \widetilde{\text{cv}}^2\right)}, \; \widetilde{\text{cv}} + z_{1-\alpha/2}\sqrt{(n-1)^{-1}\widetilde{\text{cv}}^2 \left(0.5 + \widetilde{\text{cv}}^2\right)} \right\} . \tag{2.2}$$

While simulations conducted by Gulhar *et al.* (2012) using data sampled from a chi-square and gamma distribution showed typically good results for the Miller (1991) interval, coverage was often better, if not at least similar, when using the median modification. With our interest mainly in skewed distributions, we focus on the median modified interval in (2.2).

Median modification of the modified McKay (Med MMcK)

Gulhar *et al.* (2012) also introduced a median modification to the *modified McKay interval* (McKay, 1932; Vangel, 1996). The median-modified interval is

$$\left\{ \widetilde{\text{cv}} \sqrt{\left(\frac{\chi^2_{n-1,1-\alpha/2}+2}{n} - 1\right) \widetilde{\text{cv}}^2 + \frac{\chi^2_{n-1,1-\alpha/2}}{n-1}},\ \widetilde{\text{cv}} \sqrt{\left(\frac{\chi^2_{n-1,\alpha/2}+2}{n} - 1\right) \widetilde{\text{cv}}^2 + \frac{\chi^2_{n-1,\alpha/2}}{n-1}} \right\}, \tag{2.3}$$

where $\chi^2_{n-1,\alpha}$ is the 100α-th percentile of a chi-square distribution with $(n-1)$ degrees of freedom. We focus on this median modified interval based on the results in Gulhar *et al.* (2012).

The Panich method

Panichkitkosolkul (2009) has further modified the Modified McKay (Vangel, 1996) interval by replacing the sample CV with the maximum likelihood estimator for a normal distribution, $\tilde{k} = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}/(\sqrt{n}\overline{x})$. The interval is

$$\left\{ \tilde{k} \sqrt{\left(\frac{\chi^2_{n-1,1-\alpha/2}+2}{n} - 1\right) \tilde{k}^2 + \frac{\chi^2_{n-1,1-\alpha/2}}{n-1}},\ \tilde{k} \sqrt{\left(\frac{\chi^2_{n-1,\alpha/2}+2}{n} - 1\right) \tilde{k}^2 + \frac{\chi^2_{n-1,\alpha/2}}{n-1}} \right\}. \tag{2.4}$$

The Gulhar method

Using the fact that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ when data is sampled from the normal distribution, Gulhar *et al.* (2012) proposed the interval,

$$\left(\frac{\sqrt{(n-1)}\,\widehat{\text{cv}}}{\sqrt{\chi^2_{n-1,1-\alpha/2}}},\ \frac{\sqrt{(n-1)}\,\widehat{\text{cv}}}{\sqrt{\chi^2_{n-1,\alpha/2}}} \right), \tag{2.5}$$

which compared favorably to the median-modified intervals for larger CV values. We therefore use this interval as one of the competitors.

2.2 Two robust versions of the CV

We now consider two robust alternatives for the CV that are based on quantiles. The denominator for the measures is the median, a preferred measure of centrality than the mean for skewed distributions.

2.2.1 A version based on the IQR

An option for the numerator is to use the interquartile range (IQR). Shapiro (2005) gives this alternative as

$$\text{RCV}_Q = 0.75 \times \frac{\text{IQR}}{m}\ , \tag{2.6}$$

where the multiplicative factor 0.75 makes RCV_Q comparable to the CV for a normal distribution. More precisely, anyone can use 0.741 as the factor. To the best of our knowledge there has been no research into interval estimators of the RCV_Q and this will be one of our foci shortly.

2.2.2 A version based on the median absolute deviation

The median absolute deviation (Hampel, 1974, MAD) is defined to be

$$\mathrm{MAD} = \mathrm{med} \mid x_i - m \mid \ , \tag{2.7}$$

where, for 'med'denoting median and i=1,...,n. Using the MAD for relative dispersion has been recently proposed (e.g. Reimann *et al.* , 2008; Varmuza & Filzmoser, 2009) giving

$$\mathrm{RCV}_M = 1.4826 \times \frac{\mathrm{MAD}}{m} \ . \tag{2.8}$$

The multiplier $1.4826 = 1/\Phi^{-1}(3/4)$, where Φ^{-1} denotes the quantile function for the $N(0,1)$ distribution, is used to achieve equivalence between $1.4826 \times \mathrm{MAD}/m$ and the standard deviation at the normal model. $1.4826 \times \mathrm{MAD}/m$ is commonly called the *standardized MAD*.

3 Some comparisons between the measures

The question of interest is, can we do just as well (or better) in assessing the relative dispersion by replacing the population concepts μ and σ by the median $m = x_{0.5}$ and interquartile range $\mathrm{IQR} = q_3 - q_1$ or the MAD?

Table 1: A comparison of the CV, RCV_Q and RCV_M for several distributions. LN refers to the log-normal distribution, $\mathrm{WEI}(\lambda, \alpha)$ and $\mathrm{PAR}(\lambda, \alpha)$ to the Weibull and Pareto Type II distributions with scale parameter λ and shape parameter α.

Distribution	CV	0.75* IQR/m	1.4826*MAD/m
Normal(μ,σ^2)	$\frac{\sigma}{\mu}$	$\frac{3}{4}\frac{\sigma}{\mu}\left[\Phi^{-1}(0.75) - \Phi^{-1}(0.25)\right]$	$\frac{\sigma}{\mu}$
EXP(λ)	1	1.189	1.030
Uniform(a,b)	$\frac{1}{\sqrt{3}} \cdot \frac{(b-a)}{(b+a)}$	$\frac{3}{4} \cdot \frac{(b-a)}{(b+a)}$	$\frac{1}{\Phi^{-1}(3/4)} \cdot \frac{(b-a)}{(b+a)}$
WEI(λ, 1)	1	1.189	1.029
WEI(λ, 2)	0.523	0.578	0.565
WEI(λ, 5)	0.229	0.232	0.229
χ^2_2	1	1.189	1.030
χ^2_5	0.632	0.681	0.646
$\chi^2_{\nu\to\infty}$	$\to 0$	$\to 0$	$\to 0$
LN(μ, 1)	1.311	1.090	0.888
LN(μ, 2)	7.321	2.695	1.333
PAR(λ, 2.5)	2.236	1.453	1.120
PAR(λ, 5)	1.291	1.313	1.077

In Table 1 we compare the CV, RCV_Q and RCV_M for several distributions. In most cases, the results show an approximate equivalence between the three measures when the underlying population is normal and closer agreement between the two for many other distributions. Hereafter our main interest is comparing the concepts CV, RCV_Q and RCV_M and the natural estimators of them.

3.1 Properties

An essential property of a measure of relative dispersion is scale invariance. The CV is well-established, so competing measures should give roughly the same values when the underlying distribution is uni-modal and skewed to the right, As we have seen by examples, the plug-in estimator $s/\bar{x}$ of CV suffers from over-sensitivity to outliers. Table 2 provides a rough summary of results in this work.

Table 2: Desirable properties of measures of dispersion and their estimators. Here '+', '0' and '−' indicate the property always, sometimes or never holds.

Property	CV	RCV_Q	RCV_M
P1: Scale invariant	+	+	+
P2: Simple to understand	+	+	0
P3: Widely accepted and used	+	0	0
P4: Defined for all F	0[1]	+	+
P5: Bounded influence function	−	+	+

Property	$\widehat{\mathrm{CV}}$	$\widehat{\mathrm{RCV}}_Q$	$\widehat{\mathrm{RCV}}_M$
P6: Consistency	0[2]	+	+
P7: Asymptotic normality	0	+	+
P8: Standard error formula available	+	+	+
P9: Unaffected by 1% moderate outliers	0	+	+
P10: Unaffected by 1% extreme outliers	−	+	+
P11: Reliable coverage of confidence intervals	−	+	+

In the next section, we briefly describe the methodology required to find standard errors and confidence intervals for CV, RCV_Q and RCV_M. We also investigate the robustness properties of the point estimators using theoretical methods and simulation studies and we illustrate our methods on a real data set. Finally, a summary and discussion of further possible work is in Section 6.

3.2 Influence functions

Consider a distribution function F and suppose that a parameter of interest from F is θ. Let $\mathcal{T}$ be a statistical function for estimator of θ such that $\mathcal{T}(F) = \theta$ and $\mathcal{T}(F_n) = \widehat{\theta}$, for F_n denoting

[1]The CV is only defined if F has a finite variance, but this is usually satisfied for diameter distribution models.

[2]Consistency and asymptotic normality require the existence of certain moments for F.

an empirical distribution function for sample of n observations from F, denotes an estimate of θ. Now, for $0 \le \epsilon \le 1$, define the 'contamination'distribution (F_ϵ) to have positive probability ϵ on x (the contamination point) and $1-\epsilon$ on the distribution F such that $F_\epsilon = (1-\epsilon)F + \epsilon\Delta_x$ where Δ_x denotes the distribution function that puts all of its mass at the point x. The influence of the contamination on the estimator with functional $\mathcal{T}$, relative to proportion of contamination, is $[\mathcal{T}(F_\epsilon) - \mathcal{T}(F)]/\epsilon$. The influence function (Hampel, 1974) is then defined for each x as

$$\text{IF}(x; \mathcal{T}, F) = \lim_{\epsilon \downarrow 0} \frac{\mathcal{T}(F_\epsilon) - \mathcal{T}(F)}{\epsilon} \equiv \frac{\partial}{\partial \epsilon} \mathcal{T}(F_\epsilon)\Big|_{\epsilon=0}.$$

A convenient way to appreciate the usefulness of the influence function in studying estimators is to consider the power series expansion $\mathcal{T}(F_\epsilon) = T(F) + \epsilon \text{IF}(x; \mathcal{T}, F) + O(\epsilon^2)$. So that, ignoring the error term $O(\epsilon^2)$ which is negligible for small ϵ, increasing $|\text{IF}(x; \mathcal{T}, F)|$ results in increasing influence of contamination on the estimator. Consequently, the influence function provides a very useful tool in the study of robustness of estimators.

One can show that (e.g., Hampel *et al.* , 1986; Staudte & Sheather, 1990) for $X \sim F$, the mean and variance at F of the random influence function are $\text{E}_F[\text{IF}(X;\, \mathcal{T}, F)] = 0$ and $\text{Var}_F[\text{IF}(X;\, \mathcal{T}, F)] = \text{E}_F[\text{IF}^2(X;\, \mathcal{T}, F)]$. A reason for finding this last variance is that it arises in the asymptotic variance of the functional of $\mathcal{T}(F_n)$; that is,

$$n\, \text{Var}[\mathcal{T}(F_n)] \to \text{ASV}\,(\mathcal{T}, F) = \text{E}_F[\text{IF}^2(X;\, \mathcal{T}, F)]\ . \tag{3.1}$$

3.2.1 Influence function of the CV

Let $\mathcal{M}$ and $\mathcal{V}$ denote the functional for the usual mean and variance estimators such that, at F, $\mathcal{M}(F) = \int x dF = \mu$ and $\mathcal{V}(F) = \int [x - \mathcal{M}(F)]^2\, df = \sigma^2$. The respective influence functions are $\text{IF}(x;\, \mathcal{M}, F) = x - \mu$ and $\text{IF}(x;\, \mathcal{V}, F) = (x-\mu)^2 - \sigma^2$. For convenience in notation, let $\mathcal{CV}$ also denote the functional for the CV. Groeneveld (2011) derives the influence function as

$$\text{IF}(x; \mathcal{CV}, F) = \text{CV}\left[\frac{\text{IF}(x;\, \mathcal{V}, F)}{2\sigma^2} - \frac{\text{IF}(x;\, \mathcal{M}, F)}{\mu}\right]. \tag{3.2}$$

3.2.2 Influence function of the IQR-based RCV

The influence function of the pth quantile $x_p = \mathcal{G}(F; p) = F^{-1}(p)$ is well-known (Staudte & Sheather, 1990, p.59) to be $\text{IF}[x;\, \mathcal{G}(\,\cdot, p), F] = \{p - I[x_p \ge x]\}\, g(p)$, where $\mathcal{G}'(F; p) = g(p) = 1/f(x_p)$ is the *quantile density* of $\mathcal{G}$ at p. The influence function of the ratio of two quantiles $\rho_{p,q}(F) = x_p/x_q = \mathcal{G}(\,\cdot, p)/\mathcal{G}(\,\cdot, q)$ is then found to be Prendergast & Staudte (2017a):

$$\text{IF}(x;\, \rho_{p,q}, F) = \rho_{p,q} \left\{ \frac{\text{IF}[x;\, \mathcal{G}(\,\cdot, p), F]}{x_p} - \frac{\text{IF}[x;\, \mathcal{G}(\,\cdot, q), F]}{x_q} \right\}\ . \tag{3.3}$$

It then follows that the influence function of $\mathcal{RCV}_Q(F) = 0.75\, \text{IQR}/m$ in terms of (3.3) is

$$\text{IF}(x;\, \mathcal{RCV}_Q, F) = 0.75\, \left[\text{IF}(x;\, \rho_{3/4,1/2}, F) - \text{IF}(x;\, \rho_{1/4,1/2}, F)\right]\ . \tag{3.4}$$

The IFs for the median and IQR are both bounded, but both have very different breakdown points.

3.2.3 Influence function of the MAD-based RCV

Let $\mathcal{MAD}$ denote the functional for the standardized MAD. The influence function for the MAD estimator was described by Hampel (1974) and its form for the standardized MAD for the standard normal distribution is (see, e.g., page 107 of Hampel *et al.* , 1986)

$$\mathrm{IF}(x;\,\mathcal{MAD},\Phi) = \frac{1}{4\Phi^{-1}(0.75)\phi\left[\Phi^{-1}(0.75)\right]}\mathrm{sign}\left[|x| - \Phi^{-1}(0.75)\right] \ . \tag{3.5}$$

It is not suitable for us to study the influence function for $\mathcal{RCV}_M$ at the standard normal model since the median is equal to zero. However, the influence function for the standardized MAD for an arbitrary mean, μ, for the normal distribution is simply (3.5) shifted to be centred at μ and therefore equal to $\mathrm{IF}(x;\,\mathcal{MAD},\Phi_\mu) = \mathrm{IF}(x-\mu;\,\mathcal{MAD},\Phi)$ where we let Φ_μ denote the distribution function for the $N(\mu,1)$ distribution.

Let $\mathcal{RCV}_M$ be the statistical functional for the MAD-based RCV such that $\mathcal{RCV}_M(F) = \mathcal{MAD}(F)/\mathcal{G}(F,1/2) = \mathrm{RCV}_M$. Hence, using the Product Rule and the Chain Rule, the influence function for the RCV_M estimator is

$$\begin{aligned}\mathrm{IF}(x;\mathcal{RCV}_M,F) &= \frac{\partial}{\partial\epsilon}\mathcal{RCV}_M(F_\epsilon^{(x)})\big|_{\epsilon=0}\\ &= \frac{\mathrm{IF}(x;\,\mathcal{MAD},\Phi_\mu)}{m} - \mathrm{RCV}_M\frac{\mathrm{IF}(x;\,\mathcal{G}(\,\cdot,1/2),F)}{m} \ .\end{aligned} \tag{3.6}$$

The general form of the influence for the MAD can be found in, for example, page 137 of Huber (1981), page 16 of Andersen (2008) and page 37 of Wilcox (2011) and this will be used to plot the influence functions for the non-Gaussian examples that follow.

3.2.4 Example influence function comparisons

To compute the true value for the MAD for the distributions being considered for influence function comparisons, and also when required later, we used the R function we have provided in Section B. Readers can use this code to compute the true MAD for any distributions.

In Plot A of Figure 1 we plot the influence functions for the three measures. The influence functions for the two robust measures are almost identical. In fact, it is know that the influence functions for the IQR and MAD are the same for the normal distribution (see page 110 of Hampel *et al.* , 1986) so that the measures share the same robustness properties for this model. The differences in Figure 1 are due to the multiplier 0.75 for the IQR based measure chosen to give approximate equivalence, instead of exact, for the normal. However, this does not generalize to all distributions. As expected, the influence function for the CV is unbounded, meaning that outliers are expected to have uncapped influence on the estimator as they move further from the population mean. On the other hand, the influence functions for the robust measures are bounded. Extreme outliers are expected to have no more influence on the estimators when compared to, say, those closer to the 25% and 75% percentiles. However, the discontinuities at the median and the 25% and 75% percentiles, suggest that the estimators are more sensitive locally in these areas.

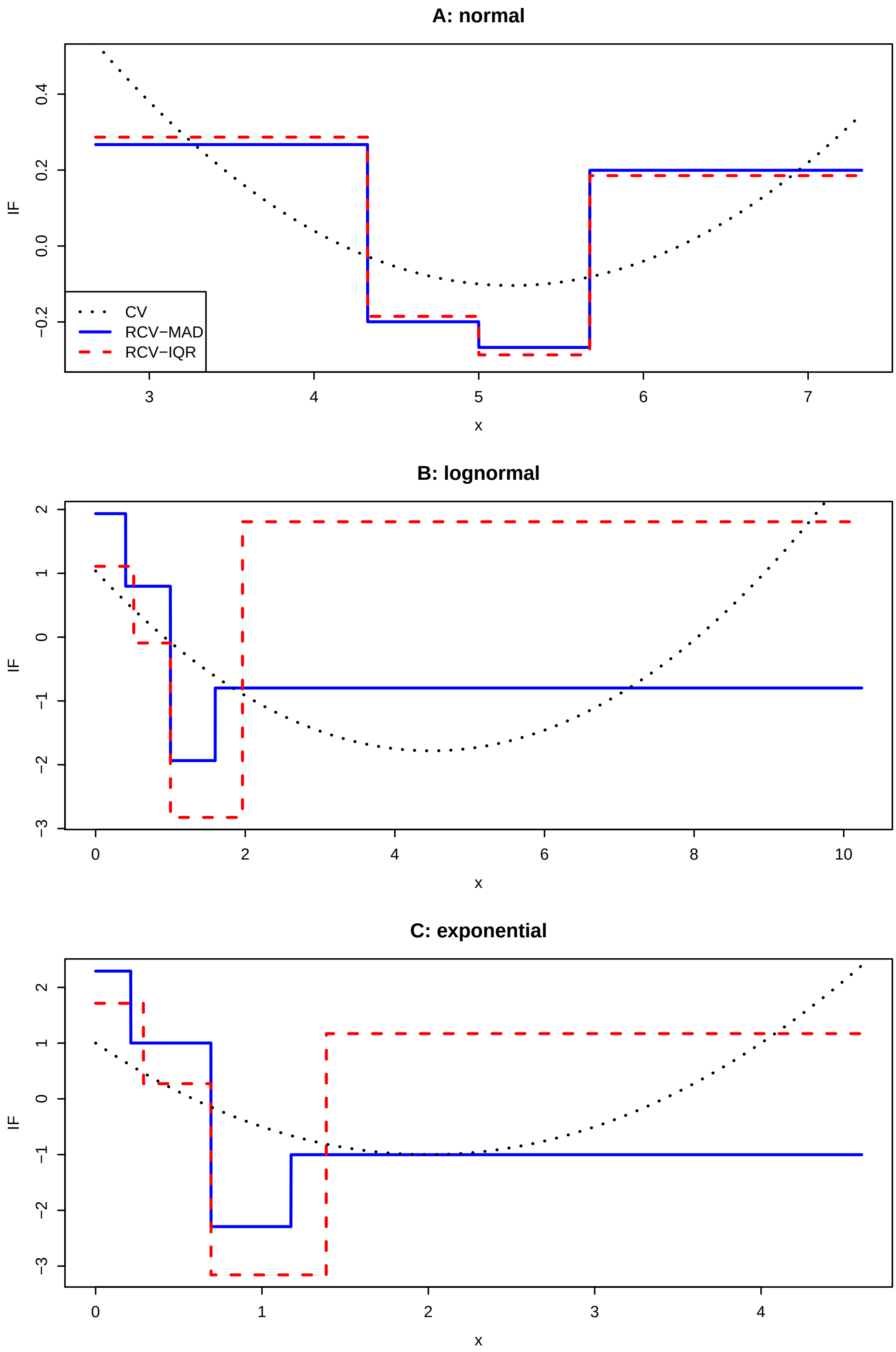


Figure 1: Influence function comparisons between the three measures: CV (black, solid), RCV_M (blue, dash) and RCV_Q (red, dots) for (A) the normal, (B) log-normal and (C) exponential distributions.

3.3 Asymptotic variances and standard deviations

In this section, we further compare the estimators by deriving their asymptotic variances. As discussed in Section 3.2, for an estimator with functional $\mathcal{T}$, the asymptotic standard deviation can be found by $\mathrm{ASD}(\mathcal{T}, F) \equiv \sqrt{\mathrm{ASV}(\mathcal{T}, F)} = \sqrt{\{\mathrm{E}_F[\mathrm{IF}^2(X;\, \mathcal{T}, F)]\}}$. We now derive the ASVs for the estimators before comparing their relative asymptotic standard deviations.

3.3.1 Asymptotic Variance of the CV estimator

Recall $\mu = \mathcal{M}(F)$ is the mean for distribution F and let $\mu_k = \mathrm{E}_F[\{X - \mathcal{M}(F)\}^k]$ denotes the kth central moment of $X \sim F$ where $\mu_2 = \sigma^2 = \mathcal{V}(F)$ denotes the variance. The influence function for the mean is $\mathrm{IF}(x; \mathcal{M}, F) = x - \mu$ and $E\left[\mathrm{IF}(X; \mathcal{M}, F)^2\right] = \sigma^2 = \mathrm{ASV}(\mathcal{M}, F)$, the asymptotic variance of the mean estimator. Similarly, $\mathrm{IF}(x; \mathcal{V}, F) = (x - \mu)^2 - \sigma^2$ and $E\left[\mathrm{IF}(X; \mathcal{V}, F)^2\right] = \mu_4 - \sigma^4 = \mathrm{ASV}(\mathcal{V}, F)$. Before deriving the ASV for the CV estimator, we note that $E\left[\mathrm{IF}(X; \mathcal{M}, F)\mathrm{IF}(X; \mathcal{V}, F)\right]$, which is the asymptotic covariance between the mean and variance estimators, is equal to $\mu_3 - \sigma^2$. Now, from (3.2),

$$\begin{aligned}
\mathrm{E}\left[\mathrm{IF}(X; \mathcal{CV}, F)^2\right] =& [\mathcal{CV}(F)]^2 \Bigg\{ \frac{\mathrm{ASV}(\mathcal{V}, F)}{4\sigma^4} + \frac{\mathrm{ASV}(\mathcal{M}, F)}{\mu^2} \\
& - \frac{E\left[\mathrm{IF}(X; \mathcal{V}, F)\mathrm{IF}(X; \mathcal{M}, F)\right]}{\sigma^2 \mu} \Bigg\} \\
\mathrm{ASV}(\mathcal{CV}, F) =& \mathrm{CV}^2 \left(\frac{\mu_4 - \sigma^4}{4\sigma^4} + \frac{\sigma^2}{\mu^2} - \frac{\mu_3}{\sigma^2 \mu} \right) ,
\end{aligned} \tag{3.7}$$

assuming that the fourth moment exists.

Note that for $X \sim F$, $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$ so that $\mathrm{ASV}(\mathcal{CV}, F) = \mathrm{CV}^2\left(1/2 + \mathrm{CV}^2\right)$ which is the asymptotic variance used by Miller (1991) in the construction of the asymptotic interval for the CV detailed in Section 2.1.

3.3.2 Asymptotic Variance of the RCV_Q estimator

The asymptotic variance of the estimator of x_p, the p-th quantile, is well known to be (eg. Ch.2 of David, 1981; DasGupta, 2006, Ch.3) $\mathrm{ASV}(\mathcal{G}, F; p) = p(1-p)g^2(p)$ where, as denoted earlier, $g(p) = 1/f(x_p)$ and f is the density function. This can be verified also using $\mathrm{E}\left[\mathrm{IF}(X; \mathcal{G}(\cdot, p), F)^2\right]$. Similarly, and as also found in the preceding references, the asymptotic covariance between the p-th and q-th quantile estimators is, $\mathrm{E}\left[\mathrm{IF}(X; \mathcal{G}(\cdot, p), F)\mathrm{IF}(X; \mathcal{G}(\cdot, q), F)\right] = p(1-q)g(p)g(q)$, provided $0 < p < q < 1$.

Asymptotic variance for $\mathrm{RCV}_Q = 0.75\,\mathrm{IQR}/m$ is obtained by a straightforward but lengthy derivation of $\mathrm{E}\left[\mathrm{IF}(X; \mathcal{RCV}_Q, F)^2\right]$ with $\mathrm{IF}(X; \mathcal{RCV}_Q, F)$ defined in (3.4) (or by using the Delta method). After simplifying, it is

Theorem 3.1. *The asymptotic variance for the estimator of RCV_Q is*

$$ASV(\mathcal{RCV}_Q, F) = \frac{RCV_Q^2}{4}\left\{\frac{3\left[g^2(3/4)+g^2(1/4)\right]-2\,g(3/4)g(1/4)}{4\times IQR^2} + \frac{g^2(1/2)}{m^2} - \frac{g(1/2)\left[g(3/4)-g(1/4)\right]}{m\times IQR}\right\}.$$

The proof of Theorem 3.1 is in Section A.

3.3.3 Asymptotic Variance of the RCV$_M$ estimator

Falk (1997) proves the asymptotic joint normality of the $m(F_n)$ and $\mathcal{MAD}(F_n)$ estimators. Let $f = F'$ be the density function associated with F. If F is continuous near and differentiable at $F^{-1}(1/2)$, $F^{-1}(1/2) - \text{MAD}$ and $F^{-1}(1/2) + \text{MAD}$ with $f(F^{-1}(1/2)) > 0$ and $C1 = f(F^{-1}(1/2) - \text{MAD}) + f(F^{-1}(1/2) + \text{MAD}) > 0$, then

$$\sqrt{n}[m(F_n) - F^{-1}(1/2),\, \mathcal{MAD}(F_n) - \mathcal{MAD}(F)]^\top \overset{\text{approx.}}{\sim} N(\mathbf{0}, \mathbf{\Sigma})\ ,$$

where '$\overset{\text{approx.}}{\sim}$'denotes 'approximately distributed as for suitably large n', $\mathbf{0}$ is a column vector zeroes and $\mathbf{\Sigma}$ is a two-dimensional covariance matrix with $\text{vec}(\mathbf{\Sigma}) = [\rho_1, \rho_{12}, \rho_{12}, \rho_2]$. Hence, ρ_1, ρ_2 are the asymptotic variances of the median and MAD estimators respectively and ρ_{12} is the asymptotic covariance between the two. They are (e.g. Falk, 1997),

$$\rho_1 = \frac{1}{4f^2(F^{-1}(1/2))}\ ,\ \rho_2 = \frac{1}{4C_1^2}\left[1 + \frac{C_2}{[f(F^{-1}(1/2)]^2}\right]$$

$$\text{and}\ \ \rho_{12} = \frac{1}{4C_1 f(F^{-1}(1/2))}\left[1 - 4F(F^{-1}(1/2) - \text{MAD}) + \frac{C_3}{f(F^{-1}(1/2))}\right]$$

where $C_3 = f(F^{-1}(1/2) - \text{MAD}) - f(F^{-1}(1/2) + \text{MAD})$ and $C_2 = C_3^2 + 4C_3 f(F^{-1}(1/2))(1 - F(F^{-1}(1/2) + \text{MAD}) - F(F^{-1}(1/2) - \text{MAD}))$.

Using the above results and the Delta method (see e.g. DasGupta, 2006), we derived the asymptotic variance of the RCV$_M$ as given below,

$$\text{ASV}(\mathcal{RCV}_M, F) = \text{RCV}_M^2\left(\frac{\rho_1}{m^2} + \frac{\rho_2}{\text{MAD}^2} - \frac{2\rho_{12}}{m\times\text{MAD}}\right)\ . \tag{3.8}$$

3.3.4 Relative asymptotic standard deviation comparisons

As an example, the asymptotic standard deviation (ASD) for the RCV$_M$ estimator is given as $\text{ASD}(\mathcal{RCV}_M, F) = \sqrt{\text{ASV}(\mathcal{RCV}_M, F)}$ and the ASDs for the other estimators are determined similarly. Later, we will construct approximate confidence intervals for the measures and therefore it make sense that we use the ASE for comparisons here. Since the CV, RCV$_Q$ and RCV$_M$ represent different values we use the relative (to the population parameter) ASD (RASE) to compare the estimators. For example, for the RCV$_M$ estimator this is defined to be $\text{rASD}(\mathcal{RCV}_M, F) = \text{ASD}(\mathcal{RCV}_M, F)/\mathcal{RCV}_M(F)$.

Table 3: Relative ASD (rASD) comparisons for the estimators of CV, RCV_Q and RCV_M for the $\mathrm{N}(5, \sigma^2)$, $\mathrm{LN}(0, \sigma)$, $\mathrm{EXP}(\lambda)$ and $\mathrm{PAR}(\alpha)$ distributions.

Distribution		rASD for the CV estimator	rASD for the RCV_Q estimator	rASD for the RCV_M estimator
$\mathrm{N}(5, \sigma^2)$	$\sigma = 0.50$	0.714	1.173	1.173
	$\sigma = 1$	0.735	1.193	1.193
	$\sigma = 1.5$	0.768	1.225	1.225
	$\sigma = 2$	0.812	1.270	1.270
	$\sigma = 2.5$	0.866	1.324	1.324
	$\sigma = 3$	0.927	1.388	1.388
$\mathrm{LN}(0, \sigma)$	$\sigma = 0.10$	0.721	1.172	1.164
	$\sigma = 0.25$	0.801	1.199	1.149
	$\sigma = 0.5$	1.151	1.294	1.098
	$\sigma = 0.75$	2.075	1.438	1.017
	$\sigma = 1$	4.674	1.621	0.914
	$\sigma = 1.5$	49.298	2.062	0.669
$\mathrm{EXP}(\lambda)$	λ	1	1.594	0.950
$\mathrm{PAR}(\alpha)$	$\alpha = 0.50$	Undefined	3.223	0.419
	$\alpha = 1$	Undefined	2.236	0.664
	$\alpha = 1.5$	Undefined	1.976	0.735
	$\alpha = 2$	Undefined	1.862	0.785
	$\alpha = 2.5$	Undefined	1.799	0.816
	$\alpha = 3$	Undefined	1.760	0.837
	$\alpha = 4$	54.482	1.714	0.864
	$\alpha = 4.5$	5.619	1.699	0.873
	$\alpha = 5$	3.724	1.687	0.880
	$\alpha = 5.5$	2.937	1.678	0.887
	$\alpha = 6$	2.500	1.670	0.892
	$\alpha = 6.5$	2.221	1.664	0.897

To compare the rASD for the estimators of CV, RCV_Q and RCV_M, we have selected normal and lognormal distributions, both with varying σ, exponential and the Pareto type II distribution with varying shape. From Table 3, the rASD for RCV_Q and RCV_M are a little higher than the rASD of CV for the normal distribution. However, RCV_Q and RCV_M estimators compare favorably to the CV for skewed distributions such as the lognormal and Pareto. The p^{th} central moment of Pareto type II distribution exists only if $\alpha > p$ so that the rASD for the CV estimator is undefined for $\alpha < 4$ since it requires the fourth central moment. When comparing RCV_q and RCV_M, the RCV_M estimator is the better performer with smaller (or equal to in the case of the normal) rASD.

4 Inference

We want to compare point and interval estimators of $\mathrm{CV} = \sigma/\mu$, $\mathrm{RCV}_Q = 0.75\,\mathrm{IQR}/x_{0.5}$ and $\mathrm{RCV}_M = 1.4826\,\mathrm{MAD}/x_{0.5}$. First, we introduce asymptotic Wald-type intervals using the asymptotic standard errors from earlier. With recent results highlighting very good coverage for estimators based on ratios of quantiles even for small samples (Prendergast & Staudte, 2016b, 2017a,b; Arachchige *et al.* , 2019), we are confident of similarly good coverage for RCV_Q. We also propose an asymptotic interval for RCV_M as well as bootstrap intervals.

We estimate the pth quantile $x_p = G(p) = F^{-1}(p)$ by the Hyndman & Fan (1996) quantile estimator $\widehat{x}_p = \widehat{G}(p)$, which is a linear combination of two adjacent order statistics. It is readily available as the Type 8 quantile estimator on the R software (Development Core Team, 2018).

4.1 Asymptotic confidence intervals

Let $z_\alpha = \Phi^{-1}(\alpha)$ denote the α quantile of the standard normal distribution. All our $100(1-\alpha)\%$ confidence intervals for measures of relative spread $\mathcal{T}(F)$ will be of the form:

$$\mathcal{T}(F_n) \pm z_{1-\alpha/2}\,\widehat{\mathrm{ASD}}(\mathcal{T}, F_n)/\sqrt{n}\ , \tag{4.1}$$

where $\mathcal{T}(F_n)$ is the estimator of $\mathcal{T}(F)$ and $\widehat{\mathrm{ASD}}(\mathcal{T}, F_n)/\sqrt{n}$ is an estimate of its standard deviation (standard error) based on the sample. The actual coverage probability of this estimator depends on how quickly the distribution of $\mathcal{T}(F_n)$ approaches normality, as well as the rate of convergence of $\mathcal{T}(F_n)$ to $\mathcal{T}(F)$ and $\widehat{\mathrm{ASD}}(\mathcal{T}, F_n)$ to $\mathrm{ASD}(\mathcal{T}, F)$.

In constructing the interval estimators for the ratios, due to improved statistical performance such as quicker convergence to normality, it is common to first construct the interval for the log-transformed ratio followed by exponentiation to return to the original ratio scale. Let $W(F) = \ln[\mathcal{T}(F)]$ then, using the Delta Method (e.g. Ch.3 of DasGupta, 2006),

$$\mathrm{ASV}(W, F) \doteq \frac{1}{[\mathcal{T}(F)]^2}\,\mathrm{ASV}(\mathcal{T}, F)\ . \tag{4.2}$$

Then $\widehat{\mathrm{ASD}}(W, F_n) = \{\widehat{\mathrm{ASV}}(W, F_n)\}^{1/2}$, where $\widehat{\mathrm{ASV}}(W, F_n))$ is an estimate of the asymptotic variance, enables one to construct the confidence interval for $W(F)$, which is based on the asymptotic normality of $W(F_n)$, before exponentiating to the original scale.

4.1.1 Confidence interval for CV

A $(1-\alpha) \times 100\%$ confidence interval for the CV, which is based on the asymptotic normality of $\widehat{\mathrm{CV}}$ when the first four moments of F exist is

$$[L, U]_{\mathrm{CV}} \equiv \exp\left[\ln\left(\widehat{\mathrm{cv}}\right) \pm z_{1-\alpha/2}\,\frac{\widehat{\mathrm{ASD}}(\mathcal{CV}, F_n)}{\widehat{\mathrm{cv}}\sqrt{n}}\right] \tag{4.3}$$

and later we define this confidence interval method as “Delta CV ”in our simulation study. The ASV for the CV estimator is given in (3.7) and to obtain our asymptotic standard error we replace the population CV, σ and μ with $\widehat{\mathrm{cv}}$, sample standard deviation s and sample mean $\overline{x}$ respectively. To estimate μ_j (the jth central moment) we use $n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^j$.

4.1.2 Confidence interval for RCV$_Q$

A large-sample confidence interval for $\mathrm{RCV}_Q = 0.75\,\mathrm{IQR}/m$ is in terms of the estimate $\widehat{\mathrm{rcv}}_Q = 0.75(\widehat{x}_{0.75} - \widehat{x}_{0.25})/\widehat{x}_{0.5}$

$$[L, U]_{\mathrm{RCV}_Q} = \exp\left[\ln\left(\widehat{\mathrm{rcv}}_Q\right) \pm z_{1-\alpha/2}\,\frac{\widehat{\mathrm{ASD}}(\mathcal{RCV}_Q, F_n)}{\widehat{\mathrm{rcv}}_Q\sqrt{n}}\right]\,. \tag{4.4}$$

The $\mathrm{ASV}(\mathcal{RCV}_Q, F))$ is given in Theorem 3.1 and to obtain $\widehat{\mathrm{ASD}}(\mathcal{RCV}_Q, F_n) = \sqrt{\widehat{\mathrm{ASV}}(\mathcal{RCV}_Q, F_n)}$, one needs to replace each x_p by $\widehat{x}_p$ and each $g(p)$ by $\widehat{g}(p)$. For $\widehat{g}(p)$, we use a kernel density estimator with the Epanechnikov (1969) kernel and optimal bandwidth using the quantile optimality ratio of Prendergast & Staudte (2016a).

4.1.3 Confidence interval for RCV$_M$

A large-sample confidence interval for $\mathrm{RCV}_M = 1.4826\,\mathrm{MAD}/m$ is in terms of $\widehat{\mathrm{rcv}}_M = 1.4826\,\widehat{\mathrm{MAD}}/\widehat{x}_{0.5}$,

$$[L, U]_{\mathrm{RCV}_M} = \exp\left[\ln\left(\widehat{\mathrm{rcv}}_M\right) \pm z_{1-\alpha/2}\,\frac{\widehat{\mathrm{ASD}}(\mathcal{RCV}_M, F_n)}{\widehat{\mathrm{rcv}}_M\sqrt{n}}\right]\,. \tag{4.5}$$

Estimation of the MAD is trivial, requiring only routine coding if functionality is not already available (i.e. it is simply the median of the ordered absolute differences of the x_is from the sample median). We also need to estimate ρ_1, ρ_2 and ρ_{12} in (3.8) and a simple approach using readily available software is use the FKML parameterization (Freimer *et al.* , 1988) of the Generalized Lambda Distribution (GLD). Defined in terms of its quantile function

$$Q(p) = \lambda_1 + \frac{1}{\lambda_2}\left(\frac{p^{\lambda_3} - 1}{\lambda_3} - \frac{(1-p)^{\lambda_4} - 1}{\lambda_4}\right),$$

where λ_i $(i = 1, \ldots, 4)$ are location, inverse scale and two shape parameters, the GLD can approximate a very wide range of probability distributions (e.g. Karian & Dudewicz, 2000; Dedduwakumara *et al.* , 2019). To do so we use the method of moments estimators and density and quantile functions for the GLD in R `gld` package (King *et al.* , 2016). It is then simple to estimate ρ_1, ρ_2 and ρ_{12} using the quantile and density functions with the estimated GLD parameters and the estimated MAD.

Additional to the asymptotic interval above, we also consider two bootstrap confidence intervals.

Non-parametric bootstrap

A non-parametric bootstrap re-samples n observations with replacement from the sample and estimates the MAD. This is repeated B times and let $\widehat{\mathrm{MAD}}^i$ $(i = 1, \ldots, B)$ denote the ith estimated MAD. The lower and upper bounds for the 95% bootstrap interval is then the 0.025 and 0.975 quantiles of the estimated $\widehat{\mathrm{MAD}}^i$s.

Parametric bootstrap

The parametric bootstrap interval is obtained in the same way as the non-parametric bootstrap with the exception that the sampling is done from a nominated, or estimated, density function. In this case, we use the estimated density from the FKML GLD as described above for the asymptotic interval. This is called the Generalized Bootstrap by Dudewicz (1992) who also uses the GLD, albeit with a different parameterization, as one example.

4.2 Confidence intervals for comparing two relative spreads

When data from two independent groups are available, it is straightforward to obtain interval estimators for the comparison of relative spread for each group. Given that empirical evidence suggests excellent coverage can be achieved in the single sample case by using a log transformation, we propose to use the log ratio of two independent relative spread estimators with a back exponentiation to the ratio scale. For example, an interval estimator for $\mathrm{RCV}_{M,1}/\mathrm{RCV}_{M,2}$ where $\mathrm{RCV}_{M,1}$ and $\mathrm{RCV}_{M,2}$ are the relative MAD-based spread for independent populations, is, where for simplicity $\widehat{r} = \widehat{\mathrm{rcv}}_{M,1}/\widehat{\mathrm{rcv}}_{M,2}$,

$$\exp\left[\ln(\widehat{r}) \pm z_{1-\alpha/2}\left\{\frac{\widehat{\mathrm{ASD}}\,(\mathcal{RCV}_{M,1}, F_n)}{\widehat{\mathrm{rcv}}_{M,1}\sqrt{n_1}} + \frac{\widehat{\mathrm{ASD}}\,(\mathcal{RCV}_{M,2}, F_n)}{\widehat{\mathrm{rcv}}_{M,2}\sqrt{n_2}}\right\}\right] , \tag{4.6}$$

where n_1 and n_2 are the sample sizes for simple random samples from the populations and where the estimates and asymptotic standard errors can be found as above for the single sample setting.

5 Simulations and Examples

5.1 Simulations

Firstly, a simulation study was conducted to compare the performance of the interval estimator of RCV_Q and asymptotic CV interval given in 4.1 with the methods given in Section 2.1 using coverage probability and width as performance measures. We have selected normal (N), log normal (LN), exponential (EXP), chi-square (χ^2) and Pareto (PAR) distributions with different parameter choices and with sample sizes $n = \{50, 100, 200, 500, 1000\}$. 10,000 simulation trials were used.

Table 4: Simulated coverage probabilities (and widths) for 95% confidence interval estimators for RCV_Q, Delta CV and the intervals for CV described in Section 2.1. (* median widths reported due to excessively large average widths after back-exponentiation.)

Sample Size(n)	Distribution	Panich	Med Mill	Med MMcK	Gulhar Method	Inverse Method	Delta CV	RCV_Q
50	N(5, 1)	0.927(0.08)	0.937(0.08)	0.941(0.08)	0.943(0.08)	0.838(0.06)	0.929(0.08)	0.979(0.16)
	LN(0, 1)	0.688(0.97)	0.817(1.03)	0.803(1.07)	0.508(0.48)	0.808(4.85)	0.997(7.81*)	0.983(1.30)
	EXP(1)	0.965(0.78)	0.978(0.73)	0.981(0.88)	0.887(0.40)	0.992(3.54)	0.997(0.68)	0.985(1.30)
	Chi(5)	0.954(0.34)	0.971(0.35)	0.966(0.36)	0.918(0.26)	0.999(0.76)	0.959(0.33)	0.977(0.58)
	PAR(1, 4)	0.746(1.12)	0.866(1.19)	0.836(1.22)	0.552(0.52)	0.720(2.97)	1.000(3.57E+9*)	0.985(1.70)
100	N(5, 1)	0.938(0.06)	0.949(0.06)	0.948(0.06)	0.943(0.06)	0.900(0.05)	0.938(0.06)	0.978(0.11)
	LN(0, 1)	0.755(0.85)	0.842(0.77)	0.867(0.96)	0.453(0.35)	0.926(2.69)	0.980(5.64)	0.975(0.82)
	EXP(1)	0.979(0.55)	0.988(0.52)	0.991(0.62)	0.863(0.28)	1.000(2.06)	0.983(0.43)	0.971(0.84)
	Chi(5)	0.966(0.24)	0.961(0.34)	0.975(0.26)	0.909(0.18)	1.000(0.59)	0.953(0.22)	0.971(0.39)
	PAR(1, 4)	0.812(0.99)	0.887(0.88)	0.914(1.11)	0.471(0.37)	0.890(1.79)	1.000(1.19E+6*)	0.978(1.07)
200	N(5, 1)	0.947(0.04)	0.946(0.04)	0.945(0.04)	0.940(0.04)	0.955(0.04)	0.942(0.04)	0.979(0.08)
	LN(0, 1)	0.783(0.67)	0.828(0.56)	0.892(0.75)	0.404(0.25)	0.979(2.46)	0.970(2.30)	0.967(0.55)
	EXP(1)	0.987(0.39)	0.988(0.37)	0.997(0.43)	0.850(0.20)	1.000(1.44)	0.974(0.29)	0.966(0.57)
	Chi(5)	0.976(0.17)	0.967(0.17)	0.978(0.18)	0.911(0.13)	1.000(0.47)	0.955(0.15)	0.968(0.27)
	PAR(1, 4)	0.822(0.78)	0.871(0.65)	0.929(0.87)	0.422(0.27)	0.970(4.20)	0.999(1.16E+4*)	0.969(0.71)
500	N(5, 1)	0.944(0.03)	0.949(0.03)	0.950(0.03)	0.944(0.02)	0.987(0.03)	0.950(0.03)	0.967(0.05)
	LN(0, 1)	0.792(0.44)	0.782(0.36)	0.923(0.49)	0.360(0.16)	0.998(2.21)	0.965(1.21)	0.961(0.33)
	EXP(1)	0.991(0.25)	0.960(0.23)	0.994(0.27)	0.841(0.12)	1.000(1.00)	0.959(0.18)	0.960(0.35)
	Chi(5)	0.976(0.11)	0.956(0.11)	0.966(0.11)	0.914(0.08)	1.000(0.36)	0.951(0.09)	0.960(0.17)
	PAR(1, 4)	0.833(0.52)	0.828(0.42)	0.952(0.58)	0.368(0.17)	0.995(1.85)	0.999(291.02*)	0.963(0.43)
1000	N(5, 1)	0.952(0.02)	0.949(0.02)	0.951(0.02)	0.943(0.02)	0.997(0.03)	0.954(0.02)	0.960(0.03)
	LN(0, 1)	0.751(0.31)	0.739(0.26)	0.874(0.35)	0.336(0.11)	0.999(1.51)	0.959(0.81)	0.959(0.23)
	EXP(1)	0.992(0.18)	0.884(0.16)	0.964(0.19)	0.834(0.09)	1.000(0.79)	0.955(0.13)	0.958(0.24)
	Chi(5)	0.979(0.08)	0.928(0.08)	0.950(0.08)	0.906(0.06)	1.000(0.29)	0.949(0.07)	0.958(0.12)
	PAR(1, 4)	0.797(0.37)	0.794(0.30)	0.923(0.41)	0.339(0.12)	0.998(1.65)	0.998(52.74*)	0.956(0.30)

Table 5: Simulated Coverage probabilities (and widths) for 95% bootstrap (non-parametric and parametric) confidence interval estimators for RCV_M

Sample size(n)	Distribution	Method: Non-parametric	Parametric	Asymptotic
50	N(5, 1)	0.9740(0.141)	0.9616(0.131)	0.9525(0.134)
	LN(0, 1)	0.9772(0.479)	0.9839(0.441)	0.9665(0.524)
	EXP(1)	0.9758(0.565)	0.9893(0.508)	0.9719(0.601)
	Chi(5)	0.9763(0.421)	0.9840(0.394)	0.9557(0.413)
	PAR(1, 4)	0.9777(0.549)	0.9874(0.493)	0.9751(0.619)
100	N(5, 1)	0.9759(0.099)	0.9795(0.093)	0.9493(0.094)
	LN(0, 1)	0.9749(0.337)	0.9859(0.327)	0.9673(0.370)
	EXP(1)	0.9762(0.402)	0.9946(0.374)	0.9648(0.411)
	Chi(5)	0.9738(0.296)	0.9776(0.284)	0.9588(0.291)
	PAR(1, 4)	0.9748(0.389)	0.9933(0.362)	0.9697(0.414)
200	N(5, 1)	0.9725(0.069)	0.9826(0.066)	0.9520(0.066)
	LN(0, 1)	0.9724(0.235)	0.9688(0.236)	0.9726(0.265)
	EXP(1)	0.9720(0.282)	0.9965(0.270)	0.9591(0.287)
	Chi(5)	0.9704(0.207)	0.9848(0.201)	0.9576(0.205)
	PAR(1, 4)	0.9729(0.272)	0.9903(0.261)	0.9681(0.283)
500	N(5, 1)	0.9644(0.043)	0.9851(0.042)	0.9505(0.042)
	LN(0, 1)	0.9668(0.147)	0.9257(0.150)	0.9757(0.169)
	EXP(1)	0.9624(0.177)	0.9962(0.173)	0.9564(0.180)
	Chi(5)	0.9678(0.129)	0.9877(0.127)	0.9574(0.129)
	PAR(1, 4)	0.9681(0.171)	0.9570(0.167)	0.9635(0.176)
1000	N(5, 1)	0.9582(0.030)	0.9861(0.029)	0.9495(0.030)
	LN(0, 1)	0.9616(0.103)	0.8247(0.106)	0.9793(0.120)
	EXP(1)	0.9612(0.124)	0.9757(0.123)	0.9569(0.128)
	Chi(5)	0.9640(0.091)	0.9834(0.090)	0.9571(0.092)
	PAR(1, 4)	0.9606(0.119)	0.8029(0.118)	0.9621(0.124)

In Table 4 we provide the simulation results for the CV and RCV_Q intervals. For simplicity, the RCV_M results follow in Table 5 where the bootstrap and asymptotic intervals are compared. From Table 4, the Panich, Med Mill and Gulhar interval estimators for the CV perform really well for the normal distribution and when the sample size increases coverage reach to the nominal coverage. However, coverages was typically below nominal for skewed distributions pointing to unreliable performance of the estimators. The Delta CV interval of (4.1.1) provides improved coverage and close to nominal when the sample size increases, with the exception for the PAR(5,1) distribution for which the CV is undefined. The interval estimator for RCV_Q was conservative being slightly above nominal for these simulations. The asymptotic interval for RCV_M (Table 5) provide excellent coverage, even for $n = 50$ and all distributions considered. With notable narrower intervals and very good coverage, the use of RCV_M and associated asymptotic interval estimators using estimated GLD functions are practically enticing. However, there does not appear to be a benefit for using a bootstrap approach where coverage was typically more conservative.

5.1.1 A Shiny web application for the performance comparisons of the intervals

For further comparisons, we have developed a Shiny (Chang *et al.* , 2017) web application that readers can use to run the simulations with different parameter choices. This can be found at `https://lukeprendergast.shinyapps.io/Robust_CV/`. The user can change the distribution, parameters, sample size, probability and the number of trials according to their choices. Once the desired options are selected, the 'Run Simulation 'button can be pressed and the relevant estimates, coverage probability (cp) and the average width of the confidence interval (w) will be calculated according to their input choices. In addition to that in the bottom right hand corner of the web page it will shows the time taken to run the each simulation.

5.2 Examples

We have selected two different data sets, which are named as doctor visits data and Melbourne house price data to apply our findings to real world data.

5.2.1 Doctor visits data

We selected the doctor visits data set used in Heritier *et al.* (2009) to apply our findings to a real world problem. The doctor visits data is a subsample of 3066 individuals of the AHEAD cohort (born before 1924) for wave 6 (year 2002) from the Health and Retirement Study (HRS) which surveys more than 22,000 Americans over the age of 50 every 2 years. We grouped this data in to two groups by taking the gender as the grouping variable. The response variable that we were interested is the number of doctor visits. Table 6 provides summary statistics of the response variable for the two gender groups.

From Table 6, the summary statistics suggest that the doctor visits distributions are positively skewed which is common for count variables. There is also a large outlier in the female group with a number of doctor visits equal to 750. We removed the outlier form the data set and again calculated the descriptive statistics for female group as shown in the 3^{rd} column of the above Table 6. The mean for the female group reduces after the removal of the outlier and the summary statistics still suggest positive skew.

Our objective was to compare the relative spread of the number of doctor visits between males and females. We used CV, RCV_Q and RCV_M to compare the relative spread of the number of doctor visits between males and females with and without an outlier.

Table 7 provides the confidence interval bounds of the 95 percent confidence intervals for

Table 6: Summary Statistics of number of doctor visits between Male and Female

Summary Statistic	Male	Female	Female (without outlier)
Sample Size	987	2079	2078
Minimum	0	0	0
1st Quartile	4	4	4
Median	8	8	8
Mean	12.08	12.8	12.45
3rd Quartile	14	15	15
Maximum	300	750	365

Table 7: 95 % confidence interval lower bounds (LB) and upper bounds (UB) for the number of doctor visits.

Sample	CV	RCV_Q	RCV_M
Male	(1.283, 2.016)	(0.837, 1.050)	(0.681, 0.807)
Female	(1.298, 2.801)	(0.943, 1.128)	(0.700, 0.786)
Female, outlier excluded	(1.237, 1.746)	(0.943, 1.128)	(0.699, 0.786)

the three measures. The confidence interval for CV is greatly influenced by whether or not the outlier in the female data is included. This is not the case for the interval for quantile-based measures. Additionally, in comparison, the interval CV is wide compared to the intervals for RCV_Q and RCV_M.

5.2.2 Melbourne house price data

The median is the most popular summary measure used to describe housing markets. Motivated by this, we applied our measures to Melbourne house clearance data from January 2016 which is available at https://www.kaggle.com/anthonypino/melbourne-housing-market. This data set contains suburb-wise prices for three types of houses (house, unit, townhouse). There is data for 369 suburbs and we removed the suburbs, which contain less than 10 houses sold leaving 301 suburbs.

We selected three pairs of suburbs which were considered by (Arachchige *et al.* , 2019) to calculate the interval estimators for ratios CV, RCV_Q and RCV_M to assess differences in relative spread of house prices.

Figure 2 depicts there are outliers for all suburbs except for Kingsbury. Additionally, there are differences in spread for the house price distributions between each neighboring suburb.

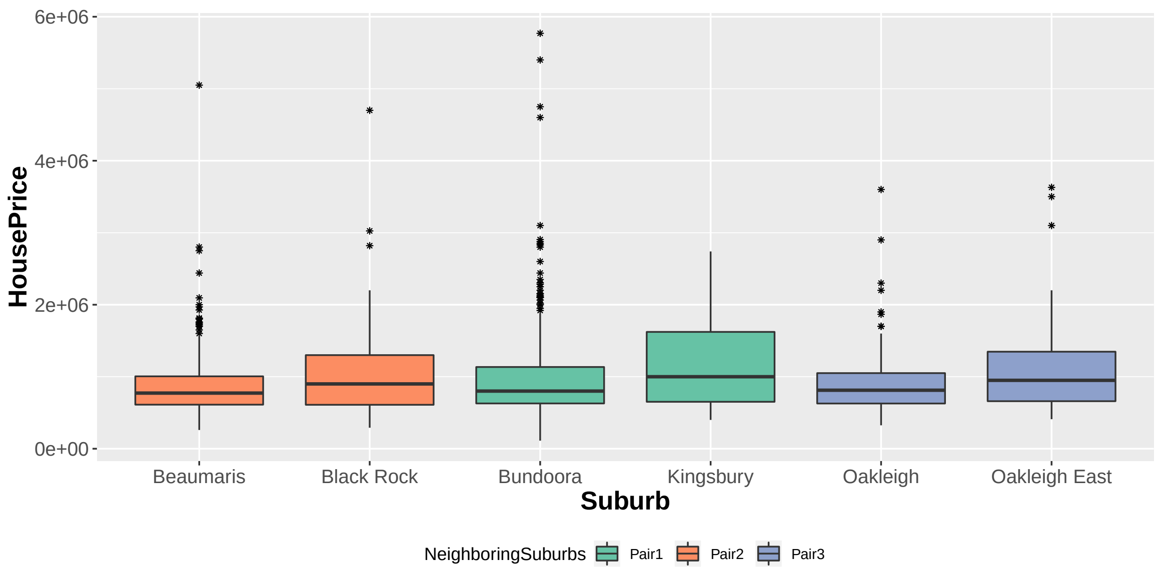


Figure 2: House price comparisons of selected three pair of neighboring suburbs

Table 8: 95 % confidence interval lower bounds (LB) and upper bounds (UB) for ratios of CV, RCV_Q and RCV_M between neighboring suburbs house prices.

Confidence Interval Method	x =Bundoora y =Kingsbury		x =Black Rock y =Beaumaris		x =Oakleigh y =Oakleigh East	
	LB	UB	LB	UB	LB	UB
$\mathrm{CV}_x/\mathrm{CV}_y$	1.0156	1.6079	0.6525	1.3225	0.7219	1.3519
$\mathrm{RCV}_{Q_x}/\mathrm{RCV}_{Q_y}$	0.4336	0.9736	0.4844	0.9243	0.4607	1.0914
$\mathrm{RCV}_{M_x}/\mathrm{RCV}_{M_y}$	0.5392	1.0808	0.5751	0.9366	0.5286	1.0218

Ratios of the measure are reported in Table 8 to see whether there is a difference in relative spread between suburbs. Comparing Bundoora and Kingsbury, the measures provide different insights. While the box plot suggests greater spread in Kingsbury, the ratio of CVs suggests otherwise having been highly influence by outliers in Bundoora. The ratios of RCV_Q and RCV_M suggest greater relative spread in Kingsbury which is in better agreement with what is shown in the box plots. For Beaumaris and Black Rock, a significant difference is not found for the CVs and the interval is wide. However, the other intervals suggest a significant difference. All three measures suggest there is not a significant difference in relative spread of house price between Oakleigh and Oakleigh East, although the intervals do tend to suggest that there is for RCV_Q and RCV_M. Overall, the intervals are narrower for the quantile-based measures having not been so greatly influence by outliers.

6 Summary and discussion

We have proposed interval estimators for alternative robust measures of relative spread to the coefficient of variation. RCV_Q, a scalar multiple of the interquartile range divided by median, is simple and the associated confidence intervals have very good coverage over a diverse range of distribution types. Similarly, RCV_M where the MAD is used instead of the interquartile range, interval also have excellent coverage and typically has smaller variability than the estimator for RCV_Q making it a preferred candidate to be used instead of the CV. While we also considered bootstrap interval estimators for RCV_M, the asymptotic Wald-type interval based on the approximate variances, and covariance between, the MAD and median achieved excellent coverage even for sample sizes as small as 50. These robust intervals compare very favorably to the CV where coverage is typically poor when the data is not sampled from a normal distribution. Our examples highlighted that they can provide very different insights into relative spread when compared to the CV, and the use of quantile-based measures is more easily justified when data is skewed due to difficulty interpreting the mean and variance.

A Proof of Theorem 3.1

Recall $\text{IF}(x;\, \rho_{p,q}, F)$ and $\text{IF}(x;\, \mathcal{RCV}_Q, F)$ in (3.3) and (3.4) respectively. For simplicity let $\text{IF}(x;\, \rho_{p,q}, F) = \text{IF}_{\rho_{p,q}}$, $\text{IF}(x;\, \mathcal{RCV}_Q, F) = \text{IF}_{\mathcal{RCV}_Q}$, $\text{IF}[\mathcal{G}(\,\cdot, p)] = \text{IF}_{\mathcal{G},p}$ and $\text{ASV}\,(\mathcal{G}, F; p) = \text{ASV}_{\mathcal{G},p}$. Then

$$\text{E}(\text{IF}^2_{\mathcal{RCV}_Q}) = 0.75^2\left[\text{E}\left(\text{IF}^2_{\rho_{3/4,1/2}}\right) + \text{E}\left(\text{IF}^2_{\rho_{1/4,1/2}}\right) - 2\text{E}\left(\text{IF}_{\rho_{3/4,1/2}}\text{IF}_{\rho_{1/4,1/2}}\right)\right]. \tag{A.1}$$

It can be shown,

$$\begin{aligned}
\text{E}\left(\text{IF}^2_{\rho_{3/4,1/2}}\right) &= \rho^2_{3/4,1/2}\,\text{E}\left[\left(\frac{\text{IF}_{\mathcal{G},3/4}}{x_{3/4}} - \frac{\text{IF}_{\mathcal{G},1/2}}{x_{1/2}}\right)\right]^2 \\
&= \frac{x^2_{3/4}}{x^2_{1/2}}\left[\frac{\text{E}\left(\text{IF}^2_{\mathcal{G},3/4}\right)}{x^2_{3/4}} + \frac{\text{E}\left(\text{IF}^2_{\mathcal{G},1/2}\right)}{x^2_{1/2}} - \frac{2\text{E}\left(\text{IF}_{\mathcal{G},3/4}\text{IF}_{\mathcal{G},1/2}\right)}{x_{3/4}x_{1/2}}\right] \\
&= \frac{1}{x^2_{1/2}}\left[\text{ASV}_{\mathcal{G},3/4} + \frac{x^2_{3/4}\text{ASV}_{\mathcal{G},1/2}}{x^2_{1/2}} - \frac{2x_{3/4}\text{E}\left(\text{IF}_{\mathcal{G},3/4}\;\text{IF}_{\mathcal{G},1/2}\right)}{x_{1/2}}\right].
\end{aligned} \tag{A.2}$$

Similarly,

$$\text{E}\left(\text{IF}^2_{\rho_{1/4,1/2}}\right) = \frac{1}{x^2_{1/2}}\left[\text{ASV}_{\mathcal{G},1/4} + \frac{x^2_{1/4}\text{ASV}_{\mathcal{G},1/2}}{x^2_{1/2}} - \frac{2x_{1/4}\text{E}\left(\text{IF}_{\mathcal{G},1/4}\;\text{IF}_{\mathcal{G},1/2}\right)}{x_{1/2}}\right] \tag{A.3}$$

and

$$
\begin{aligned}
\mathrm{E}\left(\mathrm{IF}_{\rho_{3/4,1/2}}\,\mathrm{IF}_{\rho_{1/4,1/2}}\right) =& \rho_{3/4,1/2}\times\rho_{1/4,1/2}\,\mathrm{E}\Bigg[\left(\frac{\mathrm{IF}_{\mathcal{G},3/4}}{x_{3/4}}-\frac{\mathrm{IF}_{\mathcal{G},1/2}}{x_{1/2}}\right)\\
&\times\left(\frac{\mathrm{IF}_{\mathcal{G},1/4}}{x_{1/4}}-\frac{\mathrm{IF}_{\mathcal{G},1/2}}{x_{1/2}}\right)\Bigg]\\
=&\frac{1}{x_{1/2}^2}\Bigg[\mathrm{E}\left(\mathrm{IF}_{\mathcal{G},3/4}\,\mathrm{IF}_{\mathcal{G},1/4}\right)-\frac{x_{1/4}\mathrm{E}\left(\mathrm{IF}_{\mathcal{G},3/4}\,\mathrm{IF}_{\mathcal{G},1/2}\right)}{x_{1/2}}\\
&-\frac{x_{3/4}\mathrm{E}\left(\mathrm{IF}_{\mathcal{G},1/4}\,\mathrm{IF}_{\mathcal{G},1/2}\right)}{x_{1/2}}+\frac{x_{3/4}x_{1/4}\mathrm{E}\left(\mathrm{IF}^2_{\mathcal{G},1/2}\right)}{x_{1/2}^2}\Bigg]. \qquad \text{(A.4)}
\end{aligned}
$$

Substituting the above (A.2), (A.3),(A.4) in (A.1) and using $\mathrm{ASV}\left(\mathcal{G},F;p\right)=p(1-p)g^2(p)$ gives

$$
\begin{aligned}
\mathrm{E}[\mathrm{IF}^2_{\mathcal{RCV}_Q}] =& \frac{0.75^2(x_{3/4}-x_{1/4})^2}{x_{1/2}^2}\Bigg\{\frac{\mathrm{ASV}_{\mathcal{G},3/4}+\mathrm{ASV}_{\mathcal{G},1/4}-2\mathrm{E}\left(\mathrm{IF}_{\mathcal{G},3/4}\,\mathrm{IF}_{\mathcal{G},1/4}\right)}{(x_{3/4}-x_{1/4})^2}\\
&+\frac{\mathrm{ASV}_{\mathcal{G},1/2}}{x_{1/2}^2}-\frac{2\left[\mathrm{E}\left(\mathrm{IF}_{\mathcal{G},3/4}\,\mathrm{IF}_{\mathcal{G},1/2}\right)-\mathrm{E}\left(\mathrm{IF}_{\mathcal{G},1/4}\,\mathrm{IF}_{\mathcal{G},1/2}\right)\right]}{x_{1/2}(x_{3/4}-x_{1/4})}\Bigg\}\\
=&\frac{\mathcal{RCV}_Q^2}{4}\Bigg\{\frac{3\left[g^2(3/4)+g^2(1/4)\right]-2\,g(1/4)g(3/4)}{4\times IQR^2}\\
&+\frac{g^2(1/2)}{m^2}-\frac{g(1/2)\left[g(3/4)-g(1/4)\right]}{m\times IQR}\Bigg\}. \qquad \text{(A.5)}
\end{aligned}
$$

B Computing the true MAD

Computing the true value of MAD is not a trivial task. We provide an R function below that can be uses to compute true value of the MAD for a user-specified distribution. While we have our own code, that ”” package can also be used.

```
mad <- function(dist, param){
  # Computes the true value of the MAD for a specific
  # distribution with desired parameter choices.
  #
  # Args:
  #    dist: The distribution whose MAD
  #          is to be calculated.
  #param: The parameter choices of the selected
  #            distribution whose MAD is to be calculated.
  #
  # Returns:
```

```
  # The true value of the MAD for a specific
  # distribution with desired parameter choices.

  qf <- paste0("q", dist)
   m <- do.call(qf, c(p = 0.5, param)) # find median
  abs.x.m <- function(x, dist, param, m){
    df <- paste0("d", dist)
    do.call(df, c(x = x + m, param))
            + do.call(df, c(x = - x + m, param))
  }
  abs.x.m.vec <- Vectorize(abs.x.m, "x")

  f <- function(x, dist, param, m){
    integrate(abs.x.m.vec, lower = 0, upper = x,
         dist = dist, param = param, m = m)$value - 0.5
  }
  upper <- abs(do.call(qf, c(p = 0.75, param)) + m)
  uniroot(f, interval = c(0, upper), dist = dist,
           param = param, m = m)$root
}
mad("lnorm", list(meanlog=0, sdlog=1))
mad("exp", list(rate=1))
```

References

Andersen, R. 2008. *Modern methods for robust regression.* Sage.

Arachchige, C. NPG, Cairns, M., & Prendergast, L. A. 2019. Interval estimators for ratios of independent quantiles and interquantile ranges. *Commun. Stat. B-Simul. (accepted, June).*

Atkinson, A. B. 1970. On the measurement of inequality. *J. Econ. Theor.*, **2**(3), 244–263.

Bonett, D. G. 2006. Confidence interval for a coefficient of quartile variation. *Comput. Stat. Data An.*, **50**(11), 2953–2957.

Bonett, D. G., & Seier, E. 2005. Confidence interval for a coefficient of dispersion in nonnormal distributions. *Biometrical J.*, **47**(1), 144–148.

Bulent, A., & Hamza, G. 2018. Bootstrap confidence intervals for the coefficient of quartile variation. *Commun. Stat. B-Simul.*, **In Press**, 1–9.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y, & McPherson, J. 2017. *shiny: Web application framework for r.* R package version 1.0.5.

Chen, J., & Fleisher, B. M. 1996. Regional income inequality and economic growth in china. *J. Comp. Econ.*, **22**(2), 141–164.

Cole, T. J., Bellizzi, M. C., Flegal, K. M., & Dietz, W. H. 2000. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ Brit. Med. J.*, **320**(7244), 1240.

DasGupta, A. 2006. *Asymptotic Theory of Statistics and Probability.* New York, NY: Springer.

David, H.A. 1981. *Order Statistics.* New York: John Wiley & Sons.

Dedduwakumara, D. S., Prendergast, L. A., & Staudte, R. G. 2019. A simple and efficient method for finding the closest generalized lambda distribution to a specific model. *Cogent Math. Stat.*, **6**, 1–11.

Development Core Team, R. 2018. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Dudewicz, E. J. 1992. The generalized bootstrap. *Pages 31–37 of: Bootstrapping and related techniques.* Springer.

Epanechnikov, V. A. 1969. Nonparametric estimation of a multivariate probability density. *Theor. Probab. Appl.*, **14**, 153–158.

Falk, M. 1997. Asymptotic independence of median and mad. *Stat. Probabil. Lett.*, **34**(4), 341–345.

Freimer, M., Mudholkar, G. S., Kollia, G., & Lin, C. T. 1988. A study of the generalized Tukey lambda family. *Comm. Stat. A-Theor.*, **17**, 3547–3567.

Gastwirth, J. L. 1982. Statistical properties of a measure of tax assessment uniformity. *J. Stat. Plan. Infer.*, **6**(1), 1–12.

Gong, J., & Li, Y. 1999. Relationship between the estimated weibull modulus and the coefficient of variation of the measured strength for ceramics. *J. Am. Ceram. Soc.*, **82**(2), 449–452.

Groeneveld, R. A. 2011. Influence functions for the coefficient of variation, its inverse, and cv comparisons. *Comm. Stat. A-Theor.*, **40**(23), 4139–4150.

Gulhar, M., Kibria, G., Albatineh, A., & Ahmed, N. U. 2012. A comparison of some confidence intervals for estimating the population coefficient of variation: a simulation study. *Sort*, **36**(1).

Hamer, A. J., Strachan, J. R., Black, M. M., Ibbotson, C., & Elson, R. A. 1995. A new method of comparative bone strength measurement. *J. Med. Eng. Technol.*, **19**(1), 1–5.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.*, **69**, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons.

Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M.-P. 2009. *Robust methods in biostatistics*. Vol. 825. John Wiley & Sons.

Huber, P. J. 1981. *Robust statistics*. Wiley.

Hyndman, R. J., & Fan, Y. 1996. Sample quantiles in statistical packages. *Am. Stat.*, **50**, 361–365.

Karian, Z. A., & Dudewicz, E. J. 2000. *Fitting Statistical Distributions: the generalized lambda distribution and generalized bootstrap methods*. Chapman and Hall.

King, R., Dean, B., & Klinke, S. 2016. *gld: Estimation and Use of the Generalised (Tukey) Lambda Distribution*. R package version 2.4.1.

Lovitt, W. V., & Holtzclaw, H. F. 1929. *Statistics*. Prentice-Hall, Incorporated.

McKay, A. T. 1932. Distribution of the coefficient of variation and the extended" t" distribution. *J. R. Stat. Soc.*, **95**(4), 695–698.

Miller, E. G. 1991. Asymptotic test statistics for coefficients of variation. *Comm. Stat. A-Theor.*, **20**(10), 3351–3363.

Miller, E. G., & Karson, M. J. 1977. Testing equality of two coefficients of variation. *Pages 278–283 of: Am. Stat. Assoc.: Proceedings of the Bus. Econ. Section, Part I*, vol. 95.

Panichkitkosolkul, W. 2009. Improved confidence intervals for a coefficient of variation of a normal distribution. *Thailand Statistician*, **7**(2), 193–199.

Prendergast, L. A., & Staudte, R. G. 2016a. Exploiting the Quantile Optimality Ratio in finding Confidence Intervals for Quantiles. *Stat*, **5**, 70–81.

Prendergast, L. A., & Staudte, R. G. 2016b. Quantile versions of the Lorenz curve. *Electron. J. Stat.*, **10**(2), 1896–1926.

Prendergast, L.A., & Staudte, R.G. 2017a. When large n is not enough—distribution-free interval estimators for ratios of quantiles. *J. Econ. Inequal.*, 1–17. doi:10.1007/s10888-017-9347-9.

Prendergast, L.A., & Staudte, R.G. 2017b. A simple and effective inequality measure. *Am. Stat.*

Reed, G. F., Lynn, F., & Meade, B. D. 2002. Use of coefficient of variation in assessing variability of quantitative assays. *Clin. Diagn. Lab. Immun.*, **9**(6), 1235–1239.

Reimann, C., Filzmoser, P., Garrett, R. G., & Dutter, R. 2008. *Statistical data analysis explained: applied environmental statistics with r.*

Shapiro, H. M. 2005. *Practical flow cytometry.* John Wiley & Sons.

Sharma, K. K., & Krishna, H. 1994. Asymptotic sampling distribution of inverse coefficient-of-variation and its applications. *IEEE T. Reliab.*, **43**(4), 630–633.

Staudte, R. G., & Sheather, S. J. 1990. *Robust Estimation and Testing.* New York: Wiley.

Vangel, M. G. 1996. Confidence intervals for a normal coefficient of variation. *Am. Stat.*, **50**(1), 21–26.

Varmuza, K., & Filzmoser, P. 2009. *Introduction to multivariate statistical analysis in chemometrics.* CRC press.

Wilcox, R. R. 2011. *Introduction to robust estimation and hypothesis testing.* Academic press.

Paper IV

Mean skewness measures

Arachchige, Chandima NPG & Prendergast, Luke A 2019b. *Mean skewness measures* arXiv:submit/2971320

Mean skewness measures

Chandima N.P.G. Arachchige[*], Luke A. Prendergast[†]
Department of Mathematics and Statistics, La Trobe University
Melbourne, Victoria, Australia, 3086

March 19, 2020

Abstract

Skewness measures can be used to measure the level of asymmetry of a distribution. Given the prevalence of statistical methods that assume underlying symmetry, and also the desire for symmetry in order to make meaningful judgements for common summary measures (e.g. the sample mean), reliably quantifying asymmetry is an important problem. There are several measures, among them generalizations of Bowley's well known skewness coefficient, that use sample quartiles and other quantile-based measures. The main drawbacks of many measures is that they are either limited to quartiles and do not take into account more extreme tail behavior, or that they require one to choose other quantiles (i.e. choose a value for p different from 0.25) in place of the quartiles. Our objective is to (i) average the skewness measures over all p and (ii) provide interval estimators for the new measure with good coverage properties. Our simulation results show that the interval estimators perform very well for all distributions considered.

Keywords: Bowley's coefficient of skewness, quantile-based skewness

1 Introduction

Let Q_1, Q_2 and Q_3 denote the quartiles of a population distribution, so that Q_2 is the median, then the well-known Bowley's coefficient (Yule, 1912; Bowley, 1920) given as $B_1 = (Q_3 + Q_1 - 2Q_2)/(Q_3 - Q_1)$ is a robust measure of skewness. Note that when the distribution is symmetric, then $B_1 = 0$ since $Q_3 - Q_2 = Q_2 - Q_1$. The magnitude of B_1 grows as the difference between $Q_3 - Q_2$ and $Q_2 - Q_1$ increasing implies increasing skewness. A more general case of the Bowley's coefficient (David & Johnson, 1956) can venture further into the tails than when using the first and third quartiles. This measure has been considered further by Hinkley (1975), Groeneveld & Meeden (1984) and Staudte (2014) who provided distribution free confidence intervals for the measure. Groeneveld *et al.* (2009) introduced an improved version for right skewed distributions for which good point and interval estimators can be easily obtained. This measure is appropriate only when the direction of the skewness is known, although in practice simple data visualisations can be used to decide. However, as stated by Groeneveld *et al.* (2009), the measure is easier to interpret and typically more sensitive to skewness.

The generalised Bowley's and the Groeneveld *et al.* (2009) measures require one to choose the extremity of the quantiles used. To overcome this, Groeneveld & Meeden (1984) integrated both the numerator and denominator of the measure over p. Motivated by this, we introduce

[*]18201070@students.latrobe.edu.au
[†]Corresponding author:luke.prendergast@latrobe.edu.au

integrated versions of the measures which have simple interpretations and for which interval estimators with good coverage properties are available.

In Section 2 we introduce notations and existing measures of skewness using quantiles. In Section 3 we consider integration of the measures over p before comparing with other measures in Section 4. Point and interval estimators are provided in Section 5 with simulations assessing coverage and applications to some examples following in Section 6. We then conclude the work in Section 7.

2 Notations and some selected methods

Let F denote the distribution function for random variable X and f denote the density function. For a $p \in [0,\ 1]$, let the pth quantile be $x_p = G(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$ so that, for example, $x_{0.5} = Q_2$ is the population median and $x_{0.25} = Q_1$ and $x_{0.75} = Q_3$ the other quartiles. Let $g(p) = 1/f\,(x_p)$ denote the *quantile density function* (Tukey, 1965; Parzen, 1979) and its reciprocal, which we denote $q(p) = f\,(x_p)$, is the *density quantile function.* Also let $X, \ldots, X_n$ denote a simple random sample of size n from F. Throughout let $\widehat{x}_p$ denote the estimator of x_p where we use the Hyndman & Fan (1996) quantile estimator which can be found as the Type 8 quantile estimator in R software (Development Core Team, 2018).

2.1 Generalized skewness coefficients

Using the notations above, the generalized Bowley's coefficient is defined as

$$\gamma_p = \frac{x_{1-p} + x_p - 2x_{0.5}}{x_{1-p} - x_p} \tag{1}$$

for $p \in (0, 0.5)$ (Hinkley, 1975; Groeneveld & Meeden, 1984). For later use we define the pth *interquantile skewness* to be $S_p = x_{1-p} + x_p - 2x_{0.5}$ and the pth *interquantile range* as $R_{1,p} = x_{1-p} - x_p$ so that $\gamma_p = S_p/R_{1,p}$. We denote the estimator of γ_p as

$$g_p = \frac{s_p}{r_{1,p}}$$

where $s_p = \widehat{x}_{1-p} + \widehat{x}_p - 2\widehat{x}_{0.5}$ and $r_{1,p} = \widehat{x}_{1-p} - \widehat{x}_p$.

Groeneveld *et al.* (2009) introduced a variation of γ_p that is simple to interpret and often more sensitive to skewness. For right-skewed distributions, this measure is defined as

$$\lambda_p = \frac{x_{1-p} + x_p - 2x_{0.5}}{x_{0.5} - x_p} \tag{2}$$

for $p \in (0, 0.5)$. Let $R_{2,p} = x_{0.5} - x_p$ so that $\lambda_p = S_p/R_{2,p}$ with estimator

$$l_p = \frac{s_p}{r_{2,p}}$$

where $r_{2,p} = \widehat{x}_{0.5} - \widehat{x}_p$. For left-skewed distributions, the measure can be adapted to $S_p/(x_{1-p} - x_{0.5})$. For simplicity we will focus on the use of λ_p as defined in (2) noting that findings will similarly hold when re-defining for left-skewed distributions.

To overcome the need for choosing a p for γ_p, Groeneveld & Meeden (1984) integrated both the numerator and denominator with respect to p finding

$$b_3 = \frac{\int_0^{0.5} S_p dp}{\int_0^{0.5} R_{1,p} dp} = \frac{\mu - x_{0.5}}{E|X - x_{0.5}|} \tag{3}$$

where μ is the mean for distribution F.

3 New skewness measures

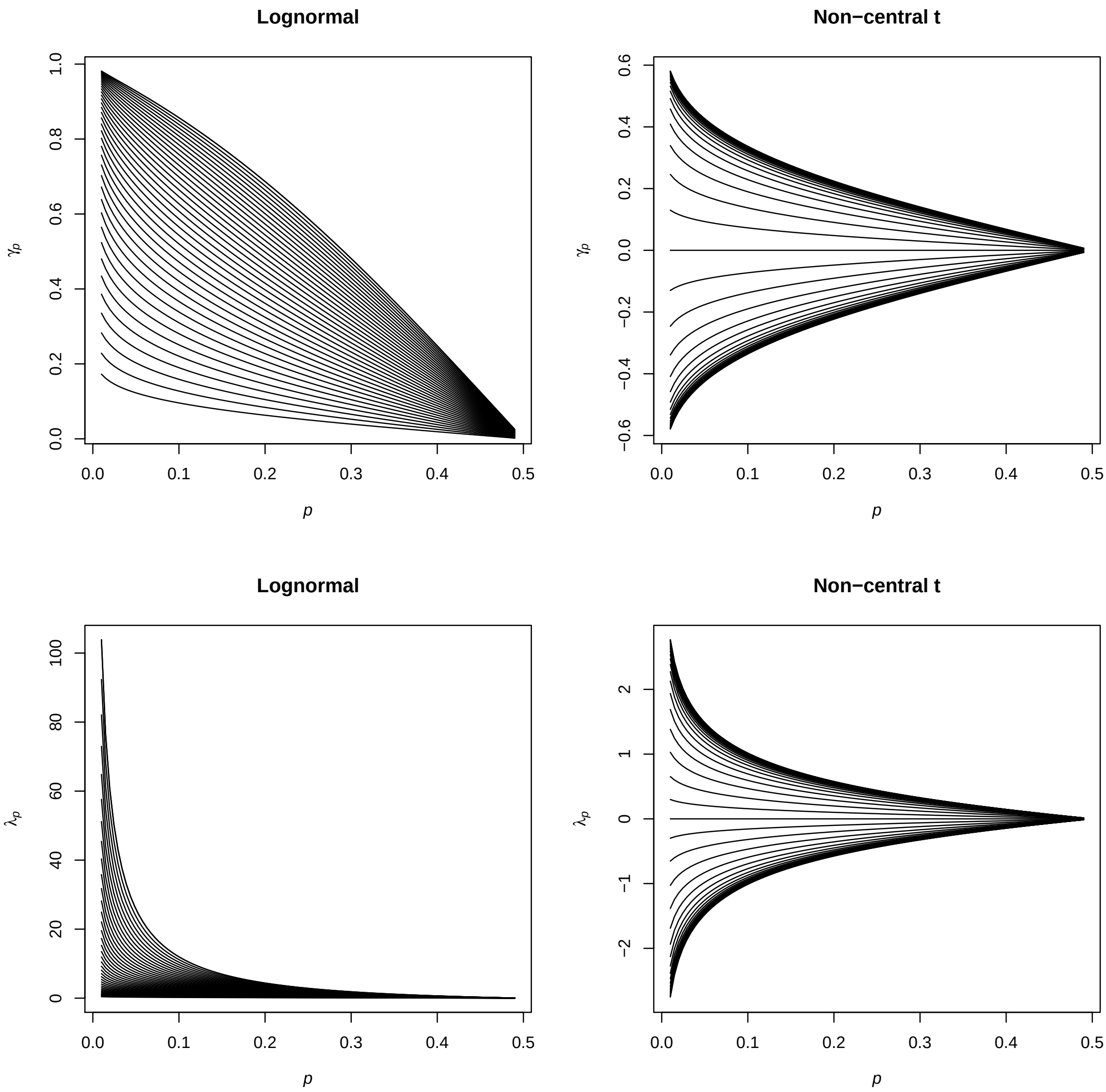


Figure 1: Examples of the curves of γ_p (top row) and λ_p (bottom row) for the lognormal distribution, $\mathrm{LN}(0, \sigma)$, with varying σ from 0.01 to 2.0 and the non-central t distribution, $t_\nu(\mathrm{ncp})$ with $\nu = 5$ and non-centrality parameter (ncp) varying from -10 to 10.

Note that γ_p and λ_p may be thought of as sensitivity curves over the domain of $p \in [0, 0.5]$. In Figure 1 we plot the curves of γ_p (top row) and λ_p (bottom row) for two distributions: the lognormal distribution, $\mathrm{LN}(0, \sigma)$, with varying σ from 0.01 to 2.0 and also the non-central t distribution, $t_\nu(\mathrm{ncp})$ with $\nu = 5$ and non-centrality parameter (ncp) varying from -10 to 10. For the lognormal, the curves are plotted for each choice of σ, where smaller σ are associated with the curves with smaller skew (lower vertical axis values). For the non-central t, the curves with negative skewness are for ncp < 0 (left skew), for ncp $= 0$ the curve is constant at zero (symmetry) and for ncp> 0 the curves are for positive skew. Skewness increases with increasing ncp. In all cases skewness is maximised when the most extreme quantiles are used, i.e. smallest p, and this is also when the distance between the curves is maximised suggesting greater sensitivity in detecting skewness for smallest p. In practice, however, such extreme quantiles are difficult to estimate and so a not-so-small p would be chosen.

3.1 Area under the skewness curve and mean skewness

One way to avoid choosing p is to calculate the Area Under the sensitivity Curve (AUC) by integrating γ_p and λ_p over $p \in [0, 0.5]$. That is, we define

$$\mathrm{AUC}_\gamma = \int_0^{0.5} \gamma_p \, dp = \int_0^{0.5} \frac{S_p}{R_{1,p}} \, dp \text{ and } \mathrm{AUC}_\lambda = \int_0^{0.5} \lambda_p \, dp = \int_0^{0.5} \frac{S_p}{R_{2,p}} \, dp. \tag{4}$$

An interpretation of the AUC above exists in the form of mean skewness. Let $U \sim \mathrm{Uniform}(0, 1/2)$, then

$$E(\gamma_U) = \int_0^{0.5} \frac{1}{2} \frac{S_u}{R_{1,u}} du = \frac{1}{2} \mathrm{AUC}_\gamma \tag{5}$$

so that the expected value for a point randomly chosen on the sensitivity curve is equal to one half of the AUC. This similarly true for the λ_p measures where $E(\lambda_U) = \mathrm{AUC}_\lambda/2$.

Remark 1. If one wanted the AUC and expected sensitivity above to be equal, then we could re-define

$$\tilde{\gamma}_u = \frac{x_{1-u/2} + x_{u/2} - 2x_{0.5}}{x_{1-u/2} - x_{u/2}} = \frac{S_{u/2}}{R_{1,u/2}} \tag{6}$$

for $u \in [0, 1]$ so that $\tilde{\gamma}_u = \gamma_{u/2}$. Then $\tilde{\lambda}_u$ could be similarly defined with $\tilde{\lambda}_u = \lambda_{u/2}$. We would then have that, for $U \sim \mathrm{Uniform}(0, 1)$, $E(\tilde{\gamma}_U) = \mathrm{AUC}_{\tilde{\gamma}}$ and $E\left(\tilde{\lambda}_U\right) = \mathrm{AUC}_{\tilde{\lambda}}$.

3.2 Weighting with respect to p

Given that large values of γ_p and λ_p can result when p is small, we could give less emphasis to the extremes by using $\gamma_p^* = p\gamma_p$ and $\lambda_p^* = p\lambda_p$.

In Figure 2 we plot the curves for $p\gamma_p$ (top row) and $p\lambda_p$ (bottom row). Note that less weighting is now given to the extremes such that greater emphasis is placed on a choice of p associated with greater density. For p_γ, that choice of p is between 0.2 and 0.3 so that the peak in skew is approximately for when the measure is based on quartiles. For $p\lambda_p$, the choice of p is between approximately 0.05 and 0.1 depending on the σ chosen. This is in contrast to the other measures (see Figure 1), where peak skew occurred at the smallest p (i.e. for the most extreme quantiles). Define the integrated γ_p^*, λ_p^* to be,

$$\mathrm{AUC}_{\gamma^*} = \int_0^{0.5} \gamma_p^* \, dp = \int_0^{0.5} p \left(\frac{S_p}{R_{1,p}} \right) dp \text{ and } \mathrm{AUC}_{\lambda^*} = \int_0^{0.5} \lambda_p^* \, dp = \int_0^{0.5} p \left(\frac{S_p}{R_{2,p}} \right) dp \tag{7}$$

where, as before, these are one half of the mean skew over p.

As example comparisons, all measures are depicted in Figure 3 for the lognormal with varying σ and the non-central t with varying ncp.

4 Properties and comparisons with other measures

4.1 Properties

Oja (1981) defined four desirable properties which are desirable for skewness measures. Let β be a skewness measure where, for distribution function F, $\beta(F)$ is the measure of skewness for the distribution F. Further, for $X \sim F$ denoting a random variable, for convenience we also let $\beta(X) = \beta(F)$. These four properties are:

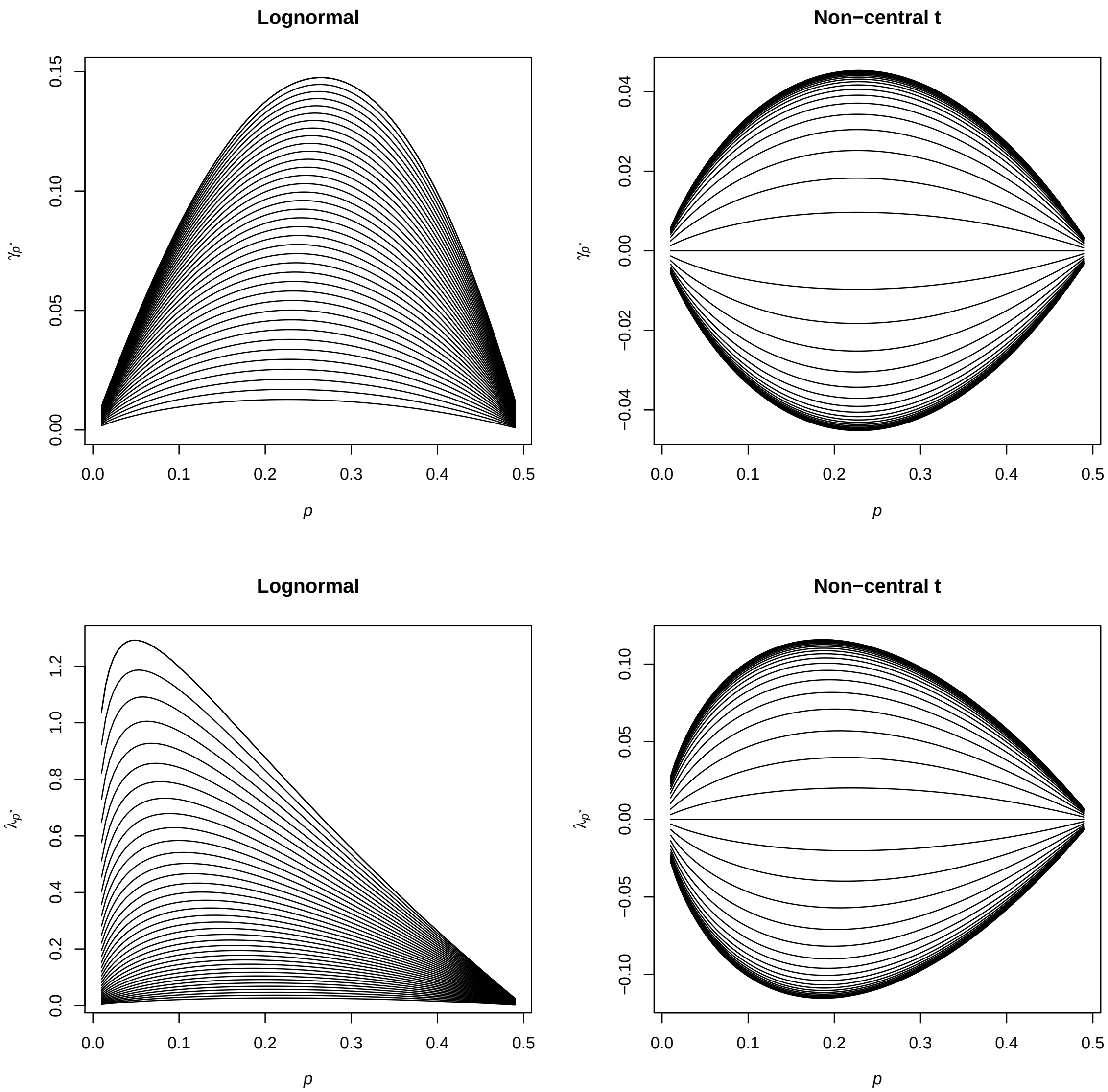


Figure 2: Examples of the curves of $p\gamma_p$ (top row) and $p\lambda_p$ (bottom row) for the lognormal distribution, LN(μ, σ), with varying σ from 0.01 to 2.0 and the non-central t distribution, t_ν(ncp) with $\nu = 5$ and non-centrality parameter (ncp) varying from -10 to 10.

P1. $\beta(cX + d) = \beta(X)$ for constants $c > 0$ and $-\infty < d < \infty$.

P2. $\beta(F) = 0$ for symmetric F.

P3. $\beta(-F) = -\beta(F)$.

P4. If $F <_c G$ then $\beta(F) \leq \beta(G)$.

The notation '$<_c$', used by Groeneveld & Meeden (1984) and Groeneveld *et al.* (2009), is read as 'F c-precedes G' meaning that distribution F is at least as skewed to the right as distribution G. Groeneveld & Meeden (1984) has shown that γ_p satisfies Properties P1 - P4 while Groeneveld *et al.* (2009) has shown that P1, P2 and P4 hold for λ_p. In both cases, for

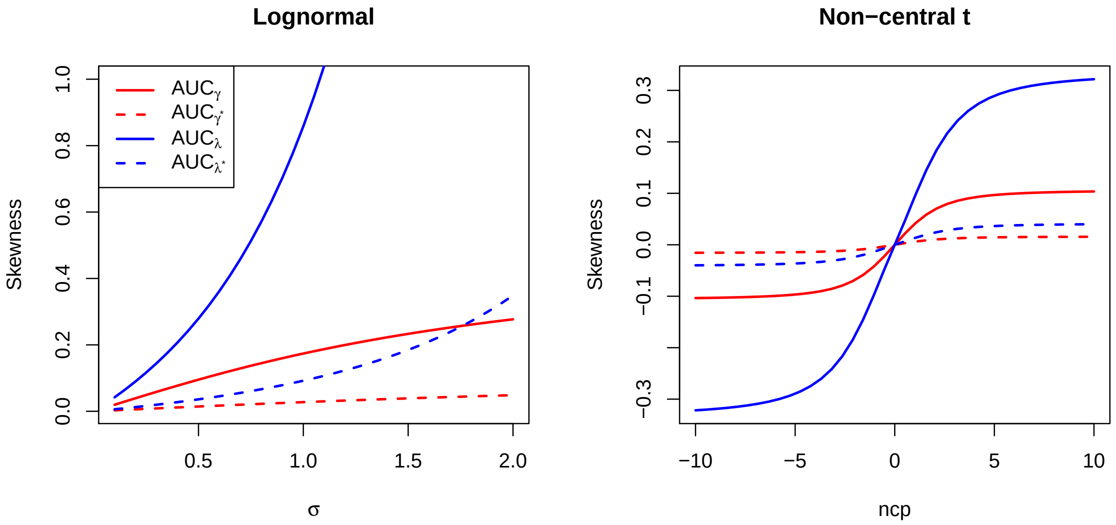


Figure 3: Examples of the AUC measures for the lognormal distribution, LN(μ, σ), with varying σ from 0.01 to 2.0 and the non-central t distribution, $t_\nu(\text{ncp})$ with $\nu = 5$ and non-centrality parameter (ncp) varying from -10 to 10.

Table 1: Desirable properties of measures of skewness and their estimators. Here '+' and '−' indicate the property is satisfied and not satisfied respectively for eahc of the measures.

Property	γ_p	λ_p	AUC$_\gamma$	AUC$_\lambda$	AUC$_{\gamma^*}$	AUC$_{\lambda^*}$
P1	+	+	+	+	+	+
P2	+	+	+	+	+	+
P3	+	−	+	−	+	−
P4	+	+	+	+	+	+

P4 to hold it is required that $G^{-1}(F(x))$ is convex. Given this, it is straightforward then to show that they also hold for the AUC measures and the properties are summarised in Table 1.

4.2 Comparisons of skewness for parametric families

We have carried out a comparison of the skewness measures γ_p, λ_p with our AUC measures over a wide range of distributions with different parameter choices.

Table 2: Comparison of measures of skewness for wide range of distributions

Distribution	$\gamma_{p=0.05}$	$\gamma_{p=0.1}$	$\gamma_{p=0.15}$	$\gamma_{p=0.2}$	$\gamma_{p=0.25}$	$\lambda_{p=0.05}$	$\lambda_{p=0.1}$	$\lambda_{p=0.15}$	$\lambda_{p=0.2}$	$\lambda_{p=0.25}$	AUC_{γ}	AUC_{λ}	AUC_{γ^*}	AUC_{λ^*}
LN(0, 1)	0.676	0.565	0.476	0.398	0.325	4.180	2.602	1.819	1.320	0.963	0.175	0.858	0.028	0.092
LN(1, 2)	0.928	0.857	0.776	0.687	0.588	25.835	11.976	6.948	4.383	2.853	0.277	5.704	0.049	0.349
Exp(λ)	0.564	0.465	0.388	0.322	0.262	2.587	1.738	1.269	0.950	0.710	0.144	0.540	0.022	0.065
χ^2_2	0.564	0.465	0.388	0.322	0.262	2.587	1.738	1.269	0.950	0.710	0.144	0.540	0.022	0.065
χ^2_5	0.354	0.281	0.230	0.188	0.151	1.096	0.782	0.596	0.462	0.356	0.087	0.242	0.013	0.032
χ^2_{25}	0.156	0.122	0.099	0.080	0.064	0.369	0.277	0.219	0.175	0.138	0.038	0.085	0.006	0.012
PAR(1,4)	0.680	0.568	0.477	0.398	0.325	4.250	2.625	1.827	1.322	0.963	0.175	0.872	0.028	0.092
PAR(1,7)	0.633	0.525	0.440	0.366	0.298	3.446	2.210	1.571	1.154	0.850	0.162	0.709	0.025	0.080
PAR(1,100)	0.325	0.469	0.392	0.325	0.264	0.963	1.768	1.289	0.963	0.719	0.145	0.551	0.023	0.066
Beta(2,5)	0.223	0.431	0.145	0.118	0.095	0.574	0.177	0.338	0.268	0.211	0.055	0.132	0.008	0.018
Beta(5,10)	0.106	0.083	0.068	0.055	0.044	0.238	0.182	0.145	0.117	0.092	0.026	0.056	0.004	0.008
WEI(0.5)	0.893	0.823	0.746	0.661	0.568	16.776	9.273	5.869	3.899	2.624	0.268	3.180	0.047	0.279
WEI(1)	0.564	0.465	0.388	0.322	0.262	2.587	1.738	1.269	0.950	0.710	0.144	0.540	0.022	0.065
WEI(2)	0.194	0.148	0.118	0.095	0.076	0.482	0.348	0.269	0.211	0.164	0.046	0.108	0.007	0.015
WEI(10)	-0.099	-0.148	-0.121	-0.099	-0.080	-0.180	-0.257	-0.215	-0.180	-0.147	-0.046	-0.080	-0.007	-0.013
Gamma(2)	0.397	0.929	0.260	0.213	0.172	1.317	0.929	0.702	0.541	0.414	0.098	0.287	0.015	0.037
Gamma(5)	0.248	0.195	0.158	0.129	0.104	0.660	0.484	0.377	0.296	0.231	0.061	0.149	0.009	0.020
Gamma(10)	0.174	0.136	0.111	0.090	0.072	0.423	0.316	0.249	0.198	0.156	0.042	0.097	0.006	0.014
F(1, 6)	0.829	0.735	0.645	0.556	0.466	9.717	5.552	3.640	2.503	1.742	0.232	1.929	0.039	0.178
F(2, 8)	0.398	0.568	0.477	0.398	0.325	1.322	2.625	1.827	1.322	0.963	0.175	0.872	0.028	0.092

Table 2 represents the values for γ_p, λ_p for $p = \{0.05, 0.1, 0.15, 0.2, 0.25\}$ and AUC_γ, AUC_{γ^*}, AUC_λ and AUC_{λ^*} for several distributions. Since the quantile function for the exponential distribution is a multiple of $1/\lambda$ where λ is the rate parameter, skewness does not depend on the rate and so we provide the results for a general λ. Consequently the skewness measures for the χ^2_2 distribution (the exponential distribution with rate 1/2) are also equal to these values. Other examples include the Pareto Type II distribution (PAR) with varying shape parameter where skewness decreases with increasing shape and similarly with the Gamma distribution. Increases and decreases among the skewness measures agree within and among distributions.

5 Estimation and inference

In this section we discuss estimation of the AUCs and provide confidence intervals.

5.1 Estimation

To estimate the pth quantile, x_p, we use the Hyndman & Fan (1996) quantile estimator, which we denote $\widehat{x}_p$, and which is a linear combination of two adjacent order statistics. It is readily available as the Type 8 quantile estimator in the R software package. We let g_p, l_p, g_p^*, l_p^* be the estimates of the skewness measures γ_p, γ_p^*, λ_p and λ_p^*.

For an arbitrary F, closed-form expressions are not available for the AUCs of the skewness measures. Recent research integrating ratios of functions of quantiles over p (e.g. Prendergast & Staudte, 2016, 2018), used summation approximations over a finite number of different ps. Approximate standard errors and subsequent confidence intervals were also found for the measures considered resulting in good coverage. We therefore consider this approach.

Let $p_j = 0.5(j - 1/2)/J$ for $j = 1, 2, \ldots, J$ so that we estimate the AUC for γ_p as

$$\widehat{\mathrm{AUC}}_\gamma \equiv \frac{1}{J}\sum_{j=1}^{J} \widehat{\gamma}_{p_j} \ . \tag{8}$$

In the context of their estimators, Prendergast & Staudte (2016, 2018), showed that $J = 100$ provides an excellent approximation to the integral, including for the standard errors that follow. We too therefore choose $J = 100$. We define $\widehat{\mathrm{AUC}}_\lambda$, $\widehat{\mathrm{AUC}}^*_\gamma$ and $\widehat{\mathrm{AUC}}^*_\lambda$ similarly. If the mean skewness measure over the curve is desired, then the AUC estimate simply needs to be halved.

5.2 Asymptotic variances

In this section we provide estimates of the asymptotic variances for the γ_p, λ_p and the AUC estimators. Staudte (2014) has already derived the asymptotic variance of γ_p using the Delta method (e.g. Ch.3 of DasGupta, 2006). It is

$$\sigma^2_{1p} = n\mathrm{Var}[g_p] \doteq {\gamma_p}^2 \left\{ \frac{n\mathrm{Var}[s_p]}{S_p^2} + \frac{n\mathrm{Var}[r_{1,p}]}{R^2_{1,p}} - \frac{2n\mathrm{Cov}[s_p, r_{1,p}]}{S_p R_{1,p}} \right\} \ , \tag{9}$$

where the estimators s_p and r_{1p} and population values S_p and $R_{1,p}$ defined as in Section 2.

We similarly derived the asymptotic variance of estimator for λ_p finding

$$\sigma^2_{2p} = n\mathrm{Var}[l_p] \doteq {\lambda_p}^2 \left\{ \frac{n\mathrm{Var}[s_p]}{S_p^2} + \frac{n\mathrm{Var}[r_{2,p}]}{R^2_{2,p}} - \frac{2n\mathrm{Cov}[s_p, r_{2p}]}{S_p R_{2,p}} \right\} . \tag{10}$$

We have also derived asymptotic co-variances needed in the variances for the AUC measures where

$$\begin{aligned}
\sigma_{1,pq} = n\mathrm{Cov}(g_p, g_q) &\doteq \frac{n}{r_{1,p}r_{1,q}}\left[\mathrm{Cov}(s_p, s_q) - \gamma_q \mathrm{Cov}(s_p, r_{1,q}) - \gamma_p \mathrm{Cov}(r_{1,p}, s_q)\right. \\
&\quad \left. + \gamma_p\gamma_q \mathrm{Cov}(r_{1,p}, r_{1,q})\right], \\
\sigma_{2,pq} \equiv n\mathrm{Cov}(l_p, l_q) &\doteq \frac{n}{r_{2,p}r_{2,q}}\left[\mathrm{Cov}(s_p, s_q) - \lambda_q \mathrm{Cov}(s_p, r_{2,q}) - \lambda_p \mathrm{Cov}(r_{2,p}, s_q)\right. \\
&\quad \left. + \lambda_p\lambda_q \mathrm{Cov}(r_{2,p}, r_{2,q})\right]
\end{aligned}$$

where setting $p = q$ gives the asymptotic variances for the estimators of γ_p and λ_p. The formulas for the co-variances in the above are given in Appendix A. Then the asymptotic variance of our AUC estimators are given as

$$n\mathrm{Var}\left(\widehat{\mathrm{AUC}}_\gamma\right) \doteq \frac{1}{J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}\sigma_{1,p_j p_k}, \quad n\mathrm{Var}\left(\widehat{\mathrm{AUC}}_\lambda\right) \doteq \frac{1}{J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}\sigma_{2,p_j p_k}. \tag{11}$$

We also let $n\mathrm{Var}\left(\widehat{\mathrm{AUC}}^*_\lambda\right)$ and $n\mathrm{Var}\left(\widehat{\mathrm{AUC}}^*_\lambda\right)$ denote the asymptotic variances for the AUC estimators of g_p and l_p. We do not show them here, since each can be obtained by, for example, multiplying $\sigma_{1,p_j p_k}$ by $p_j p_k$ in the above asymptotic variance expressions.

5.3 Interval estimators for the AUCs and differences of AUCs

Let V denote the estimate of $\mathrm{Var}\,(l_p)$, V_γ denote the estimate of $\mathrm{Var}\left(\widehat{\mathrm{AUC}}_\gamma\right)$ and similarly V_λ the estimate of $\mathrm{Var}\left(\widehat{\mathrm{AUC}}_\lambda\right)$. To obtain these, we need estimates of σ_{1p} σ_{2p}, $\sigma_{1,pq}$ and $\sigma_{2,pq}$. These need estimates of the quantile density functions $1/f(x_p)$ and $1/f(x_q)$ where f is the density function. To estimate the quantile density functions, we use a kernel density estimator studied by, e.g. Falk (1986) and Welsh (1988) with bandwidth determined by the *quantile optimality ratio (QOR)* of Prendergast & Staudte (2016a). The bandwidth based on the QOR typically resulted in slightly conservative intervals for quantiles and so are favored by us for our simulations. Code is available on request, or if desired standard bandwidths and density estimators for f could be used although our preference is to estimate the quantile density directly rather than thake the inverse of an estimated f.

Let $z_\alpha = \Phi^{-1}(\alpha)$ denote the α quantile of the standard normal distribution. All our $100(1-\alpha)\%$ confidence intervals for measures of skewness will be of the form, e.g. for AUC_γ,

$$\widehat{\mathrm{AUC}}_\gamma \pm z_{1-\alpha/2}\,\mathrm{SE}_\gamma \ , \tag{12}$$

where $\mathrm{SE}_\gamma = \sqrt{V_\gamma}$. If an interval for mean skewness over the interval $p \in [0, 0.5]$ was desired (which is half the AUC), then all that is required is to halve the lower and upper bounds of the AUC interval.

When there are two independent groups, we can construct interval estimators to compare the differences in skewness. E.g, an interval estimator for $\mathrm{AUC}_{\gamma,1} - \mathrm{AUC}_{\gamma,2}$ is,

$$\widehat{\mathrm{AUC}}_{\gamma,1} - \widehat{\mathrm{AUC}}_{\gamma,2} \pm z_{1-\alpha/2}\ \mathrm{SE}_{\gamma,1,2} \ . \tag{13}$$

where $\mathrm{SE}_{\gamma,1,2} = \sqrt{V_{\gamma,1} + V_{\gamma,2}}$ and the $V_{\gamma,i}$s are the variances of the respective AUCs.

6 Simulations and Examples

We now consider simulations to assess coverage of the interval estimators before considering two examples.

6.1 Simulations

A simulation study was conducted to compare the performance of interval estimator of λ_p with our new measures AUC_γ, AUC_{γ^*}, AUC_λ and AUC_{λ^*} by considering coverage probability (cp) and the average confidence interval width (w) as the performance measures. We have selected normal, log normal, exponential, chi-square and Pareto distributions with different parameter choices and the sample sizes $n = \{50, 100, 200, 500, 1000\}$. We used 10,000 simulation trials to our simulation study since the standard error of the estimated coverage probability for the nominal 0.95 level is less than 0.005 for 10,000 simulation trials (Staudte, 2014).

Table 3: Simulated coverage probabilities (and widths) for 95% confidence interval estimators for λ_p.

n	Dist.	$\lambda_{p=0.05}$	$\lambda_{p=0.1}$	$\lambda_{p=0.15}$	$\lambda_{p=0.2}$	$\lambda_{p=0.25}$
50	N(2,1)	0.964(1.35)	0.964(1.43)	0.962(1.54)	0.961(1.69)	0.950(1.92)
	LN(0, 1)	0.955(12.91)	0.960(7.35)	0.963(5.66)	0.961(4.82)	0.955(4.47)
	EXP(1)	0.959(6.23)	0.960(4.58)	0.958(3.95)	0.955(3.66)	0.952(3.56)
	χ^2_2	0.960(6.21)	0.950(4.57)	0.955(3.99)	0.952(3.66)	0.953(3.60)
	PAR(1, 7)	0.959(9.29)	0.960(5.91)	0.961(4.79)	0.957(4.25)	0.954(4.06)
100	N(2,1)	0.966(0.94)	0.967(0.96)	0.964(1.02)	0.965(1.12)	0.962(1.25)
	LN(0, 1)	0.963(7.78)	0.964(4.58)	0.968(3.53)	0.965(3.02)	0.958(2.72)
	EXP(1)	0.963(4.02)	0.962(2.98)	0.960(2.54)	0.960(2.33)	0.955(2.24)
	χ^2_2	0.959(3.98)	0.960(2.95)	0.960(2.54)	0.959(2.34)	0.953(2.24)
	PAR(1, 7)	0.960(5.73)	0.962(3.72)	0.962(3.06)	0.961(2.69)	0.954(2.49)
200	N(2,1)	0.971(0.65)	0.969(0.66)	0.967(0.70)	0.968(0.77)	0.965(0.85)
	LN(0, 1)	0.961(4.93)	0.967(3.02)	0.965(2.33)	0.963(1.98)	0.965(1.80)
	EXP(1)	0.961(2.66)	0.964(1.99)	0.966(1.70)	0.959(1.56)	0.959(1.49)
	χ^2_2	0.964(2.67)	0.958(1.98)	0.965(1.70)	0.950(1.56)	0.961(1.50)
	PAR(1, 7)	0.967(3.74)	0.963(2.49)	0.962(2.02)	0.961(1.78)	0.961(1.65)
500	N(2,1)	0.972(0.41)	0.968(0.41)	0.972(0.43)	0.965(0.46)	0.958(0.50)
	LN(0, 1)	0.962(2.89)	0.959(1.80)	0.962(1.40)	0.962(1.19)	0.964(1.08)
	EXP(1)	0.961(1.60)	0.960(1.21)	0.956(1.04)	0.957(0.95)	0.960(0.91)
	χ^2_2	0.959(1.61)	0.958(1.20)	0.957(1.04)	0.959(0.95)	0.955(0.90)
	PAR(1, 7)	0.962(2.21)	0.957(1.50)	0.960(1.22)	0.957(1.08)	0.959(1.00)
1000	N(2,1)	0.972(0.28)	0.965(0.28)	0.963(0.29)	0.959(0.32)	0.959(0.35)
	LN(0, 1)	0.961(1.98)	0.960(1.24)	0.961(0.96)	0.962(0.82)	0.959(0.75)
	EXP(1)	0.956(1.11)	0.957(0.84)	0.957(0.72)	0.956(0.66)	0.956(0.63)
	χ^2_2	0.958(1.11)	0.957(0.84)	0.958(0.72)	0.955(0.66)	0.956(0.63)
	PAR(1, 7)	0.958(1.52)	0.956(1.04)	0.958(0.85)	0.958(0.75)	0.958(0.69)
5000	N(2,1)	0.960(0.12)	0.956(0.12)	0.953(0.13)	0.953(0.14)	0.954(0.15)
	LN(0, 1)	0.957(0.85)	0.954(0.54)	0.955(0.42)	0.952(0.36)	0.955(0.32)
	EXP(1)	0.957(0.49)	0.952(0.37)	0.953(0.32)	0.952(0.29)	0.953(0.28)
	χ^2_2	0.957(0.48)	0.955(0.37)	0.954(0.32)	0.950(0.29)	0.953(0.28)
	PAR(1, 7)	0.954(0.66)	0.957(0.45)	0.954(0.37)	0.955(0.33)	0.952(0.30)
10000	N(2,1)	0.955(0.08)	0.956(0.08)	0.954(0.09)	0.956(0.10)	0.952(0.11)
	LN(0, 1)	0.956(0.59)	0.951(0.38)	0.957(0.29)	0.952(0.25)	0.953(0.23)
	EXP(1)	0.953(0.34)	0.954(0.26)	0.954(0.22)	0.952(0.20)	0.955(0.19)
	χ^2_2	0.951(0.34)	0.954(0.26)	0.951(0.22)	0.950(0.20)	0.957(0.19)
	PAR(1, 7)	0.954(0.46)	0.954(0.32)	0.953(0.26)	0.952(0.23)	0.955(0.21)

Before we consider interval estimators for the AUCs, we provide simulated coverage probabilities for an interval estimator of λ_p using an estimated asymptotic variance from (10). We considered $p = \{0.05, 0.1, 0.15, 0.2, 0.25\}$ and the results are provided in Table 3. The interval provides very good coverage compared to the nominal 0.95 and the interval width decreases with increasing sample sizes. Groeneveld *et al.* (2009) recommended to use $\lambda_{p=0.05}$ since it does not ignore the tail behaviour of the distribution. Very good coverage probabilities are achieved for this p.

Simulated coverages based on 10,000 trials for interval estimators of AUC_γ, AUC_{γ^*}, AUC_λ and AUC_{λ^*} are provided in the Table 4. The interval estimators of AUC_γ, AUC_{γ^*}, AUC_λ

Table 4: Simulated coverage probabilities (and widths) for 95% confidence interval estimators for AUC_γ, AUC_λ, AUC_{γ^*} and AUC_{λ^*}.

n	Dist.	AUC_γ	AUC_λ	AUC_{γ^*}	AUC_{λ^*}
50	N(2,1)	0.997(1.84)	0.993(6.57)	1.000(0.68)	0.996(2.53)
	LN(0, 1)	0.998(2.60)	0.953(11.44)	0.999(0.89)	0.987(3.05)
	EXP(1)	0.996(2.42)	0.964(6.94)	0.999(1.50)	0.988(2.63)
	χ^2_2	0.997(1.65)	0.966(6.55)	0.999(1.80)	0.989(2.17)
	PAR(1, 7)	0.996(6.23)	0.954(7.86)	0.998(0.80)	0.988(2.67)
100	N(2,1)	0.992(0.93)	0.988(3.63)	0.995(0.58)	0.992(2.64)
	LN(0, 1)	0.994(1.43)	0.953(4.72)	0.997(0.39)	0.984(1.95)
	EXP(1)	0.991(1.42)	0.966(3.70)	0.996(0.42)	0.982(1.52)
	χ^2_2	0.992(0.96)	0.969(3.60)	0.995(0.37)	0.981(2.14)
	PAR(1, 7)	0.993(0.91)	0.962(4.00)	0.997(1.17)	0.982(2.01)
200	N(2,1)	0.987(0.73)	0.985(6.05)	0.990(0.32)	0.988(1.26)
	LN(0, 1)	0.989(0.77)	0.954(2.96)	0.992(0.35)	0.982(1.01)
	EXP(1)	0.986(0.81)	0.971(2.26)	0.991(0.32)	0.982(0.83)
	χ^2_2	0.989(0.65)	0.971(2.87)	0.992(0.28)	0.983(1.36)
	PAR(1, 7)	0.988(0.93)	0.964(3.34)	0.991(0.33)	0.982(0.85)
500	N(2,1)	0.969(0.25)	0.975(0.90)	0.977(0.09)	0.976(0.32)
	LN(0, 1)	0.974(0.24)	0.959(1.55)	0.979(0.09)	0.975(0.43)
	EXP(1)	0.974(0.24)	0.967(1.24)	0.978(0.09)	0.976(0.39)
	χ^2_2	0.969(0.25)	0.964(1.29)	0.979(0.09)	0.973(0.40)
	PAR(1, 7)	0.972(0.24)	0.959(1.30)	0.978(0.09)	0.974(0.43)
1000	N(2,1)	0.964(0.17)	0.970(0.41)	0.969(0.06)	0.973(0.15)
	LN(0, 1)	0.966(0.16)	0.962(0.95)	0.972(0.06)	0.969(0.22)
	EXP(1)	0.965(0.17)	0.960(0.69)	0.967(0.06)	0.972(0.19)
	χ^2_2	0.964(0.17)	0.959(0.70)	0.960(0.06)	0.967(0.19)
	PAR(1, 7)	0.966(0.17)	0.960(0.81)	0.967(0.06)	0.965(0.19)
5000	N(2,1)	0.957(0.07)	0.957(0.15)	0.958(0.02)	0.956(0.05)
	LN(0, 1)	0.958(0.07)	0.947(0.38)	0.958(0.02)	0.962(0.07)
	EXP(1)	0.957(0.07)	0.952(0.28)	0.955(0.02)	0.961(0.07)
	χ^2_2	0.956(0.07)	0.953(0.27)	0.959(0.02)	0.959(0.07)
	PAR(1, 7)	0.957(0.07)	0.950(0.33)	0.958(0.02)	0.957(0.07)
10000	N(2,1)	0.953(0.05)	0.953(0.10)	0.953(0.02)	0.951(0.04)
	LN(0, 1)	0.951(0.05)	0.938(0.27)	0.957(0.02)	0.955(0.05)
	EXP(1)	0.955(0.05)	0.955(0.20)	0.955(0.02)	0.953(0.05)
	χ^2_2	0.955(0.05)	0.955(0.20)	0.955(0.02)	0.952(0.05)
	PAR(1, 7)	0.951(0.05)	0.944(0.23)	0.955(0.02)	0.954(0.05)

and AUC$_{\lambda^*}$ provide good coverage compared to the nominal 0.95 for moderate to large n and the interval width decreases with increasing sample sizes. For smaller n, the coverages are conservative. Overall, the coverages for the AUC of the λ skewness measure are usually closer to nominal and approach nominal more quickly with increasing sample size. We have seen this across a broad range of distributions and the reader can verify this by using our web application detailed next.

6.1.1 A Shiny web application for the performance comparisons of the intervals

For further comparisons, we have developed a Shiny (Chang *et al.* , 2017) web application that readers can use to run the simulations with different parameter choices. This can be found at `https://lukeprendergast.shinyapps.io/meanskew/`. The user can change the distribution, parameters, sample size, probability and the number of trials according to their choices. Once the desired options are selected the 'Run Simulation' button can be pressed and the relevant estimates, coverage probability (cp) and the average width of the confidence interval (w) will be calculated according to their input choices.

6.2 Examples

We have selected two datasets as examples.

6.2.1 Computer price data

The "Computers" data set which is available in the "Ecdat" package (Croissant, 2016) in R consists of prices for 6259 personnel computer which have been obtained from a cross section from 1993 to 1995 in the United States. Figure 4 depicts the computer price distribution which is clearly positively skewed.

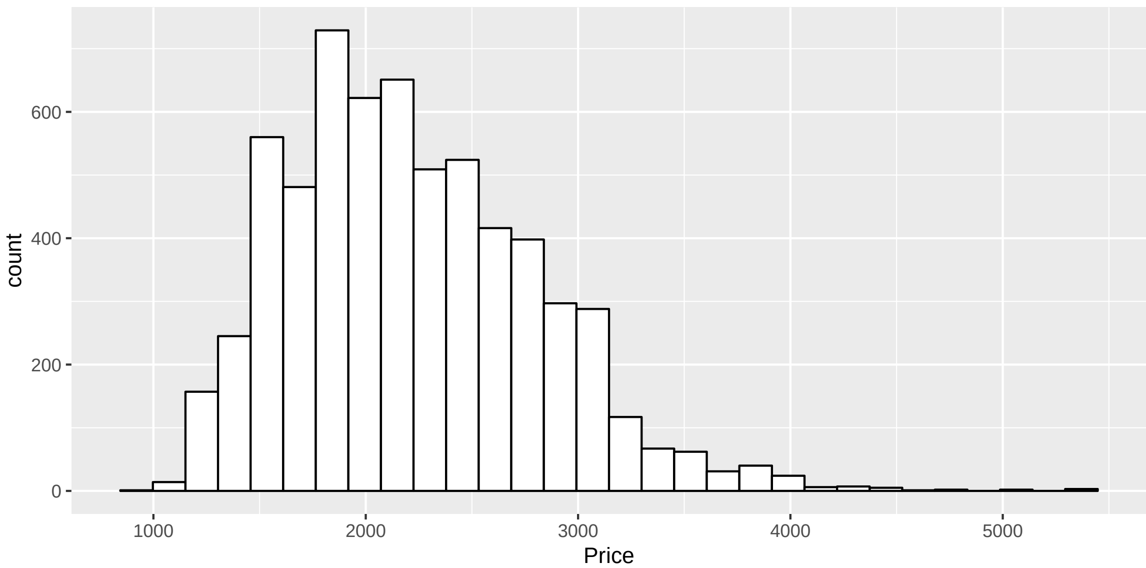


Figure 4: Histogram of the price of computers (in US dollars)

Table 5 contains the estimate and 95% confidence intervals for the six skewness measures, γ_p and λ_p for $p = \{0.05, 0.1, 0.15, 0.2, 0.25\}$ and AUC$_\gamma$, AUC$_{\gamma^*}$, AUC$_\lambda$ and AUC$_{\lambda^*}$. The confidence interval for γ_p was found as given in Staudte (2014). All intervals suggest significant skew.

Table 5: 95% confidence intervals (CIs) and estimates of the measures of skewness for the computer price data.

Measure	Estmate	CI	Measure	Estmate	CI
$\gamma_{p=0.05}$	0.1801	(0.1531, 0.2072)	$\lambda_{p=0.05}$	0.4395	(0.3589, 0.5200)
$\gamma_{p=0.1}$	0.1377	(0.1128, 0.1626)	$\lambda_{p=0.1}$	0.3194	(0.2523, 0.3865)
$\gamma_{p=0.15}$	0.1245	(0.0967, 0.1524)	$\lambda_{p=0.15}$	0.2844	(0.2117, 0.3571)
$\gamma_{p=0.2}$	0.1100	(0.0783, 0.1417)	$\lambda_{p=0.2}$	0.2472	(0.1671, 0.3273)
$\gamma_{p=0.25}$	0.1261	(0.0906, 0.1616)	$\lambda_{p=0.25}$	0.2886	(0.1957, 0.3814)
AUC_γ	0.1294	(0.0726, 0.1861)	AUC_λ	0.2885	(0.2052, 0.3718)
AUC_{γ^*}	0.0271	(0.0025, 0.0516)	AUC_{λ^*}	0.0512	(0.0257, 0.0768)

6.2.2 Doctor visits data

The doctor visits data, used as an example in Heritier *et al.* (2009), is a sub sample of 3066 individuals (987 males and 2079 females) of the AHEAD cohort born before 1924 for wave 6 (year 2002) from the Health and Retirement Study (HRS) (Heritier *et al.* , 2009). This study surveys more than 22,000 Americans over the age of 50 every 2 years. The response variable that we are interested in is the number of doctor visits in the two gender groups. The doctor visits distributions of male and female are positively skewed, and the truncated histograms can be found in Staudte (2014). There is one outlier in the female group (750 visits) and the ranges of visits, ignoring that outlier, for females is 0 to 365 and 0 to 300 for males. A complete analysis of descriptive statistics for the number of doctor visits in male and female can be found in Table 6 of Arachchige *et al.* (2019).

Table 6: Confidence intervals for the measures of skewness of the number of doctor visits of males, females and difference between males and females.

	Male		Female		Male-Female (with outlier)	
Measure	Estimate	CI	Estimate	CI	Estimate	CI
$\gamma_{0.05}$	0.5172	(0.4353, 0.5992)	0.5758	(0.5087, 0.6428)	-0.0585	(-0.1644, 0.0474)
$\gamma_{0.10}$	0.4545	(0.4089, 0.5002)	0.4545	(0.4214, 0.4877)	0.0000	(-0.0564, 0.0564)
$\gamma_{0.15}$	0.3333	(0.2654, 0.4013)	0.5238	(0.4932 ,0.5544)	-0.1905	(-0.2650, -0.1159)
$\gamma_{0.20}$	0.2308	(0.1202, 0.3413)	0.5000	(0.4542, 0.5458)	-0.2692	(-0.3889, -0.1496)
$\gamma_{0.25}$	0.2000	(0.1035, 0.2968)	0.2727	(0.2044 ,0.3412)	-0.0727	(-0.1910, 0.0455)
$\lambda_{0.05}$	2.1429	(1.4394, 2.8463)	2.7143	(1.9696, 3.4590)	-0.5714	(-1.5959, 0.4530)
$\lambda_{0.10}$	1.6667	(1.3600, 1.9733)	1.6667	(1.4440, 1.8893)	0.0000	(-0.3790, 0.3710)
$\lambda_{0.15}$	1.0000	(0.6942, 1.3058)	2.2000	(1.9302 ,2.4699)	-1.2000	(-1.6079, -0.7921)
$\lambda_{0.20}$	0.6000	(0.2264, 0.9736)	2.0000	(1.6333, 2.3667)	-1.4000	(-1.9235, -0.8765)
$\lambda_{0.25}$	0.5000	(0.1985, 0.8015)	0.7500	(0.4915 ,1.0085)	-0.2500	(-0.6471, 0.1471)
AUC_γ	0.1741	(0.1044, 0.2439)	0.2610	(0.2172, 0.3048)	-0.0869	(-0.1692, -0.0045)
AUC_{γ^*}	1.0676	(0.6717, 1.4634)	0.0296	(0.0169, 0.0424)	-0.0265	(-0.0512, -0.0018)
AUC_λ	0.0031	(-0.0180, 0.0243)	0.0296	(1.0733, 1.6754)	-0.3068	(-0.8041, 0.1905)
AUC_{λ^*}	0.0554	(0.0120, 0.0987)	0.1139	(0.0786, 0.1493)	-0.0586	(-0.1145, -0.0026)

Table 6 provides 95% confidence intervals for γ_p, λ_p, AUC_γ, AUC_{γ^*}, AUC_λ and AUC_{λ^*} for number of doctor visits for males, females and between males and females (with outlier). The intervals for each measure for males and females indicate skew. However, different conclusions about differences in skew between males and females can be made based on the different skewness measures. The intervals for λ_p and γ_p are sensitive to the choice of p. However, the intervals for the AUC measures do indicate skew with the exception of the AUC for λ.

7 Summary and future work

We have introduced more powerful alternatives to the existing measures of skewness such as γ_p and λ_p which require a choice of p. Here we introduce the integrated versions of the γ_p, λ_p, $p\gamma_p$ and $p\lambda_p$ as alternatives to measure the skewness. The simulation results show that the interval estimators perform well for all the selected distributions with moderate to large sample sizes and are typically conservative for smaller sample sizes.

While we refer to the AUCs as mean skew (i.e. mean of the skew curve over a uniform $p \in [0, 1/2]$), in truth the AUC itself is twice the mean. It is simple then to obtain point and interval estimates for the mean from the AUC estimates and vice versa. We favored AUC since it was more typically in the domain of skew values of γ_p and λ_p when p is fixed to some value between 0 and 0.25 (which would typically be done in practice). The mean skew is typically less than the skewness at fixed p since it is half the AUC and the AUC is taken over all $p \in [0, 0.5]$. An alternative would also be to consider integrating over $p \in [0, 0.25]$ and dividing by 4. This would result in a mean more like the skew values for fixed p in 0 to 0.25. We favored the AUC though since it considers the entire distribution, and not just a subset of it.

The influence function (IF Hampel, 1974) can be used to study the robustness properties and sensitivity of estimators. Groeneveld *et al.* (2009); Groeneveld (1991) computed the IFs for the quantiles based measures and in doing so established typically greater sensitivity to right skew of λ_p compared to γ_p. A study of the IFs for the AUC measures may also reveal some advantage in weighting with respect to p whereby the less weighting is applied to the extreme quantiles where estimation can be difficult. For examples of the IF, including the IF for quantiles as background, see e.g. Staudte & Sheather (1990) and Clarke (2018).

A Asymptotic variances and covariances

Asymptotic variance and covariance expressions for quantiles estimators are (e.g. see David, 1981; DasGupta, 2006),

$$\begin{aligned}
n\,\mathrm{Var}(\widehat{x}_p) &\doteq p(1-p)h^2(p)\ , \\
n\,\mathrm{Cov}(\widehat{x}_p, \widehat{x}_q) &\doteq \begin{cases} p(1-q)h(p)h(q), & 0 < p < q < 1 \\ q(1-p)h(p)h(q), & 0 < q < p < 1 \end{cases}
\end{aligned}$$

where $h(p) = 1/f(x_p)$ is known as the *quantile density function* (Tukey, 1965; Parzen, 1979). For simplicity, let $\xi_{p,q} = \mathrm{Cov}(\widehat{x}_p, \widehat{x}_q)$ and $\xi_p^2 = \mathrm{Var}(\widehat{x}_p)$. For each of the variances and covariances needed we have

$$\begin{aligned}
\mathrm{Cov}(s_p, s_q) =& \mathrm{Cov}\left(\widehat{x}_{1-p} + \widehat{x}_p - 2\widehat{x}_{0.5}, \widehat{x}_{1-q} + \widehat{x}_q - 2\widehat{x}_{0.5}\right), \\
=& \xi_{1-p,1-q} + \xi_{1-p,q} + \xi_{p,1-q} + \xi_{p,q} - 2\xi_{1-p,0.5} - 2\xi_{p,0.5} - 2\xi_{0.5,1-q} - 2\xi_{0.5,q} + 4\xi_{0.5}^2 \\
\mathrm{Cov}(s_p, r_{1,q}) =& \mathrm{Cov}\left(\widehat{x}_{1-p} + \widehat{x}_p - 2\widehat{x}_{0.5}, \widehat{x}_{1-q} - \widehat{x}_q\right) \\
=& \xi_{1-p,1-q} - \xi_{1-p,q} + \xi_{p,1-q} - \xi_{p,q} - 2\xi_{0.5,1-q} + 2\xi_{0.5,q}, \\
\mathrm{Cov}(r_{1,p}, s_q) =& \mathrm{Cov}\left(\widehat{x}_{1-p} - \widehat{x}_p, \widehat{x}_{1-q} + \widehat{x}_q - 2\widehat{x}_{0.5}\right), \\
=& \xi_{1-p,1-q} + \xi_{1-p,q} - 2\xi_{1-p,0.5} - \xi_{p,1-q} - \xi_{p,q} + 2\xi_{p,0.5}, \\
\mathrm{Cov}(r_{1,p}, r_{1,q}) =& \mathrm{Cov}\left(\widehat{x}_{1-p} - \widehat{x}_p, \widehat{x}_{1-q} - \widehat{x}_q\right) = \xi_{1-p,1-q} - \xi_{1-p,q} - \xi_{p,1-q} + \xi_{p,q},
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(s_p, r_{2,q}) =& \mathrm{Cov}\left(\widehat{x}_{1-p} + \widehat{x}_p - 2\widehat{x}_{0.5}, \widehat{x}_{0.5} - \widehat{x}_q\right)\\
=& \xi_{1-p,0.5} - \xi_{1-p,q} + \xi_{p,0.5} - \xi_{p,q} + 2\xi_{0.5,q} - 2\xi^2_{0.5},\\
\mathrm{Cov}(r_{2,p}, s_q) =& \mathrm{Cov}\left(\widehat{x}_{0.5} - \widehat{x}_p, \widehat{x}_{1-q} + \widehat{x}_q - 2\widehat{x}_{0.5}\right)\\
=& \xi_{0.5,1-q} + \xi_{0.5,q} - \xi_{p,1-q} - \xi_{p,q} + 2\xi_{p,0.5} - 2\xi^2_{0.5},\\
\mathrm{Cov}(r_{2,p}, r_{2,q}) =& \mathrm{Cov}\left(\widehat{x}_{0.5} - \widehat{x}_p, \widehat{x}_{0.5} - \widehat{x}_q\right) = \xi_{p,q} - \xi_{0.5,q} - \xi_{p,0.5} + \xi^2_{0.5}.
\end{aligned}$$

References

Arachchige, CNPG, Prendergast, LA, & Staudte, RG. 2019. Robust analogues to the coefficient of variation. *arXiv preprint arXiv:1907.01110.*

Bowley, AL. 1920. *Elements of statistics.* Vol. 2. PS King.

Chang, W, Cheng, J, Allaire, JJ, Xie, Y, & McPherson, J. 2017. *shiny: Web application framework for r.* R package version 1.0.5.

Clarke, BR. 2018. *Robustness theory and application.* John Wiley & Sons.

Croissant, Y. 2016. *Ecdat: data sets for econometrics. r package version 0.3-1.*

DasGupta, A. 2006. *Asymptotic Theory of Statistics and Probability.* New York, NY: Springer.

David, FN, & Johnson, NL. 1956. Some tests of significance with ordered variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **18**(1), 1–31.

David, HA. 1981. *Order Statistics.* New York: John Wiley & Sons.

Development Core Team, R. 2018. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Falk, M. 1986. On the estimation of the quantile density function. *Statistics & Probability Letters*, **4**(2), 69–73.

Groeneveld, RA. 1991. An influence function approach to describing the skewness of a distribution. *The American Statistician*, **45**(2), 97–102.

Groeneveld, RA, & Meeden, G. 1984. Measuring skewness and kurtosis. *The Statistician*, 391–399.

Groeneveld, RA, Meeden, G, *et al.* . 2009. An improved skewness measure. *Metron*, **67**(3), 325.

Hampel, FR. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, **69**, 383–393.

Heritier, S, Cantoni, E, Copt, S, & Victoria-Feser, Maria-Pia. 2009. *Robust methods in biostatistics.* Vol. 825. John Wiley & Sons.

Hinkley, DV. 1975. On power transformations to symmetry. *Biometrika*, **62**(1), 101–111.

Hyndman, RJ, & Fan, Y. 1996. Sample quantiles in statistical packages. *The American Statistician*, **50**, 361–365.

Oja, H. 1981. On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics*, 154–168.

Parzen, E. 1979. Nonparametric statistical data modeling. *Journal of the American Statistical Association*, **74**, 105–121.

Prendergast, LA, & Staudte, RG. 2016. Quantile versions of the lorenz curve. *Electronic Journal of Statistics*, **10**(2), 1896–1926.

Prendergast, LA, & Staudte, RG. 2016a. Exploiting the Quantile Optimality Ratio in finding Confidence Intervals for Quantiles. *Stat*, **5**, 70–81.

Prendergast, LA, & Staudte, RG. 2018. A simple and effective inequality measure. *The American Statistician*, 1–16.

Staudte, RG. 2014. Inference for quantile measures of skewness. *Test*, **23**(4), 751–768.

Staudte, RG, & Sheather, SJ. 1990. *Robust Estimation and Testing.* New York: Wiley.

Tukey, JW. 1965. Which part of the sample contains the information? *Proceedings of the National Academy of Sciences*, **53**, 127–134.

Welsh, AH. 1988. Asymptotically efficient estimation of the sparsity function at a point. *Statistics & Probability Letters*, **6**(6), 427–432.

Yule, GU. 1912. *An introduction to the theory of statistics.* London: Griffin.