

DOCTORAL THESIS

New Developments in Meta-analysis with Five-number Summary

SHI, Jiandong

Date of Award:
2021

[Link to publication](#)

General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

Abstract

Meta-analysis is a statistical method for synthesizing multiple studies to achieve more comprehensive and reliable conclusions. This thesis mainly focuses on the meta-analysis with continuous outcomes, where some studies are reported with the whole or part of the five-number summary including the minimum and maximum values, the first and third quartiles, and the median. Given that most existing meta-analysis models can only handle the studies reported with the sample mean and standard deviation (SD), it is often desired to convert the five-number summary back to the sample mean and SD before synthesis, as otherwise the studies reported with the five-number summary have to be excluded from further analysis. To tackle this problem, some data transformation methods have emerged recently, and they are getting more and more popular in practice.

We note, however, that most popular methods for data transformation are built on the normality assumption, which may not always hold in practice. In particular, when a study chooses to report the five-number summary, it can be an indication that the underlying data may not be normal or symmetric. For such data, if still applying the normal-based methods for data-transformation, the final results can be misleading or even wrong in meta-analysis. Motivated by this, we propose to further enhance the meta-analysis literature on data transformation with the five-number summary. Specifically, this thesis consists of four important projects including (1) the hypothesis tests for skewness and normality, (2) the mean and variance estimation from the five-number summary of a log-normal distribution, (3) new effect size estimation methods, and (4) a new paradox in random-effects meta-analysis.

In Chapter 2, we propose three skewness tests and a normality test with the whole or part of the five-number summary. Despite of the limited data available, the information in the five-number summary, together with the sample size, is surprisingly sufficient enough to conduct the skewness tests. Moreover, when the five-number summary is fully available, we further incorporate the kurtosis information in the test for testing the normality beyond skewness. Simulation studies demonstrate that the type I error rates are well controlled and the newly proposed tests also provide

good statistical power.

In Chapter 3, we propose to estimate the mean and variance from the reported five-number summary of a log-normal distribution. For normal data, some well-performed methods are established. However, when the data are significantly skewed, there are few methods that could properly handle the problem. Motivated by this and noting that many skewed medical data are modeled with the log-normal distribution, we provide two types of estimators for the mean and variance with the five-number summary of a log-normal distribution. Their performance is demonstrated by the simulation studies and real data analysis.

In Chapter 4, we develop new methods that estimate the mean difference and standardized mean difference from the five-number summary. This is motivated by the fact that, even though the normal-based methods can be readily applied, the estimated sample mean and SD are unlikely to be the same as the true values. As a consequence, if one directly applies them as the true sample mean and SD and then applying the classical methods including the Cohen's d or Hedges' g to estimate the effect size, it may yield biased estimates so that the final meta-analytical results can be unreliable. Our new methods, as demonstrated by simulation studies, achieve a better accuracy and a higher coverage probability than the existing methods.

In Chapter 5, we introduce a new paradox in random-effects meta-analysis. Once the new paradox appears, the individual studies and the meta-analytical result are contradictory, which leads to a dilemma for clinical decision making. As found, the key reason for the paradox is the between-study heterogeneity involved in random-effects meta-analysis. In particular when the heterogeneity is large and the number of studies is small in meta-analysis, the probability of the new paradox appearing is not ignorable. Moreover, the new paradox only appears in random-effects meta-analysis but it does not exist in the common-effect and fixed-effects models. It thus raises an interesting question whether the current random-effects model is reasonable and tenable for meta-analysis, or it needs to be abandoned or further improved.

Keywords: Five-number summary, Log-normal distribution, Mean difference, Meta-analysis, Normality test, Paradox, Skewness test, Standardized mean difference

Table of Contents

Declaration	i
Abstract	ii
Acknowledgements	iv
Table of Contents	v
List of Figures	ix
List of Tables	xii
Chapter 1 Introduction	1
1.1 Evidence-based medicine	1
1.2 Meta-analysis	2
1.3 Five-number summary	4
1.4 Outline of the thesis	5
Chapter 2 Hypothesis tests for skewness and normality	9
2.1 Introduction	9
2.2 Skewness tests	12
2.2.1 Motivating examples	12
2.2.2 Skewness tests under three scenarios	14
2.2.3 Simulation studies	22
2.2.4 Real data analysis	27

2.3	The boxplot test	29
2.3.1	Classic tests for normality	32
2.3.2	Boxplot test for normality	35
2.3.3	Simulation studies	39
2.3.4	The boxplot test in the boxplot	43
2.4	Conclusion	45
2.5	Technical results	46
Chapter 3	Data transformation with the five-number summary of a log-normal distribution	57
3.1	Introduction	57
3.2	Estimating the mean and variance from a log-normal distribution	58
3.2.1	Estimation under scenario \mathcal{S}_1	60
3.2.2	Estimation under scenario \mathcal{S}_2	61
3.2.3	Estimation under scenario \mathcal{S}_3	63
3.3	Bias-corrected estimation	64
3.3.1	Bias-corrected estimation under scenario \mathcal{S}_1	64
3.3.2	Bias-corrected estimation under scenario \mathcal{S}_2	65
3.3.3	Bias-corrected estimation under scenario \mathcal{S}_3	66
3.4	Simulation study	66
3.5	Real data analysis	70
3.6	Discussion	73
3.7	Technical results	75
Chapter 4	New effect size estimation with the five-number summary	84
4.1	Introduction	84
4.2	Estimating the MD and SMD with the sample mean and SD	85
4.2.1	Estimating the MD with the sample mean and SD	86
4.2.2	Estimating the SMD with the sample mean and SD	86
4.3	Estimating the MD with the five-number summary	88

4.3.1	Estimating the MD under scenario \mathcal{S}_1	88
4.3.2	Estimating the MD under scenario \mathcal{S}_2	89
4.3.3	Estimating the MD under scenario \mathcal{S}_3	90
4.4	Estimating the SMD with the five-number summary	91
4.4.1	Estimating the SMD under scenario \mathcal{S}_1	91
4.4.2	Estimating the SMD under scenario \mathcal{S}_2	93
4.4.3	Estimating the SMD under scenario \mathcal{S}_3	95
4.5	Simulation studies	96
4.5.1	Evaluating the estimation for the MD	96
4.5.2	Evaluating the estimation for the SMD	99
4.6	Conclusion	102
4.7	Technical results	103
Chapter 5 A new paradox in random-effects meta-analysis		114
5.1	Introduction	114
5.2	The new paradox in random-effects model	117
5.2.1	The random-effects model	117
5.2.2	The new paradox with a real data example	118
5.3	Theoretical results	120
5.3.1	Algebraic results	121
5.3.2	Statistical results	127
5.4	Simulation studies	132
5.5	The new paradox does not exist in the common-effect and fixed-effects models	137
5.5.1	The common-effect model	137
5.5.2	The fixed-effects model	139
5.6	Simpson's paradox and the new paradox	140
5.7	Conclusion and discussion	143
5.8	Technical results	144

Chapter 6 Future work	146
Bibliography	148
Curriculum Vitae	156

List of Figures

1.1	A flow chart for performing meta-analysis when some studies are reported with the whole or part of the five-number summary.	6
2.1	Probability density functions of the four normal-related distributions	13
2.2	The green points represent the exact critical values under scenarios \mathcal{S}_1 , \mathcal{S}_2 and \mathcal{S}_3 respectively.	18
2.3	The type I error rates for the proposed test statistics under three scenarios for n up to 401	25
2.4	The statistical power of the proposed tests under three scenarios for n up to 401	26
2.5	The forest plot of the meta-analysis that excludes the two studies with significantly skewed data.	29
2.6	The forest plot of the meta-analysis that includes the two studies with significantly skewed data.	29
2.7	The classical boxplot with a simulated data set	30
2.8	The boxplots depict four simulated data sets	32
2.9	The green points represent the exact values of $k(n)$, and the red line represents the approximate function of $k(n)$ for n up to 401.	38
2.10	The green points represent the exact values of $c_{0.05}$, and the red line represents the approximate function of $c_{0.05}$ for n up to 401.	39
2.11	The type I error rates for the boxplot test statistic for n up to 401	40
2.12	The statistical power of the tests for n up to 401	42

2.13	Comparison of neutralising antibody titres between intervention and control groups	44
2.14	The histograms of $a+b-2m$ for $N(10, 0.37)$ and for Skew-normal(0, 1, -10)	47
3.1	The green points represent the true values of the coefficient z_1 for n from 5 to 401, and the red line represents the approximate function of z_1	62
3.2	The dashed line represents the probability density function of $LN(3, 0.3^2)$ and the solid line represents the probability function of $LN(3, 0.7^2)$. .	67
3.3	The RBs and RMSEs of three types of mean estimators under scenario \mathcal{S}_1	68
3.4	The RBs and RSLs of three types of variance estimators under scenario \mathcal{S}_1	69
3.5	The forest plot based on the normal-based (NB) estimates.	71
3.6	The forest plot based on the plug-in (PI) estimates.	72
3.7	The forest plot based on the bias-corrected (BC) estimates.	72
3.8	The RBs and RMSEs of three types of mean estimators under scenario \mathcal{S}_2	80
3.9	The RBs and RSLs of three types of variance estimators under scenario \mathcal{S}_2	81
3.10	The RBs and RMSEs of three types of mean estimators under scenario \mathcal{S}_3	82
3.11	The RBs and RSLs of three types of variance estimators under scenario \mathcal{S}_3	83
4.1	The coverage probability for the estimated CIs for n_1 up to 400 . . .	98
4.2	The relative bias and coverage probability for the SMD with the balanced data for n up to 400	100
4.3	The relative bias and coverage probability for the SMD with the unbalanced data for n_1 up to 400, and $n_2 = 0.5n_1$	101

4.4	The relative bias and coverage probability for the SMD with the unbalanced data for n_1 up to 400, and $n_2 = 2n_1$	102
4.5	The coverage probability for the estimated CIs for n_1 up to 400	106
4.6	The coverage probability for the estimated CIs for n_1 up to 400	107
4.7	The relative bias and coverage probability for the SMD with the balanced data for n up to 400	108
4.8	The relative bias and coverage probability for the SMD with the unbalanced data for n_1 up to 400, and $n_2 = 0.5n_1$	109
4.9	The relative bias and coverage probability for the SMD with the unbalanced data for n_1 up to 400, and $n_2 = 2n_1$	110
4.10	The relative bias and coverage probability for the SMD with the balanced data for n up to 400	111
4.11	The relative bias and coverage probability for the SMD with the unbalanced data for n_1 up to 400, and $n_2 = 0.5n_1$	112
4.12	The relative bias and coverage probability for the SMD with the unbalanced data for n_1 up to 400, and $n_2 = 2n_1$	113
5.1	Forest plot for a meta-analysis with five hypothetical studies.	116
5.2	Forest plot for a meta-analysis on the length of hospital stay.	120
5.3	The relationship among the three key points for a general case	122
5.4	The relationship among the three key points under common within-study variance assumption	124
5.5	The probability of the new paradox with μ varying from 3 to 6	134
5.6	The probability of the new paradox with τ^2 varying from 0 to 4	135
5.7	The probability of the new paradox with k varying from 2 to 10	136
5.8	The intuitive presentation and comparison of the Simpson's paradox and the new paradox.	142

List of Tables

2.1	The true and estimated averages (standard errors) of the sample mean and SD for the four normal-related distributions.	14
2.2	The summary statistics of the six studies included in the meta-analysis of Nnoaham and Clarke (2008)	14
2.3	The observed values of the test statistic T_1 , the critical values for the corresponding sample sizes and the decisions of the test.	27
2.4	The five-number summaries from the six groups of data for the neutralising antibody titres.	43
2.5	The summary table of the skewness tests under three scenarios.	46
2.6	The numerical values of $c_{1,0.025}$ for $1 \leq Q \leq 100$, where $n = 4Q + 1$	54
2.7	The numerical values of $c_{2,0.025}$ for $1 \leq Q \leq 100$, where $n = 4Q + 1$	55
2.8	The numerical values of $c_{3,0.05}$ for $1 \leq Q \leq 100$, where $n = 4Q + 1$	56
3.1	The normal-based mean and variance estimators under the three scenarios.	67
3.2	The summary statistics of the six studies included in the meta-analysis.	71
3.3	The recommended mean and variance estimators under the three scenarios.	74
5.1	The admission data from University of California, Berkeley in 1973	115
5.2	The success rates of two treatments for small and large stones	141
5.3	The key statistics in Figure 5.2	141