

ABSTRAK

Disertasi ini mencadangkan satu taburan baru untuk pemodelan data serakan terkurang, sama dan terlebih. Taburan ini muncul sebagai kes istimewa daripada perjalanan rawak yang diubah pada satah dan taburan ini termasuk binomial, binomial negatif dan shifted binomial negatif sebagai kes-kes khasnya. Ciri-ciri, ujian hipotesis untuk serakan sama, kajian simulasi kuasa, penganggar kebolehdajian maksimum (PKM) dan kaedah jarak kuasa dua berdasarkan fungsi penjana kebarangkalian dipertimbangkan. Selain itu, taburan yang dicadangkan ini dibandingkan dengan taburan yang sedia wujud seperti taburan COM-Poisson dan taburan Generalized Poisson dalam pemodelan serakan. Taburan ini telah dibuktikan sebagai model yang fleksibel melalui aplikasi kebagusan penyuaian kepada empat set data nyata. Tambahan pula, dua taburan bivariat baru diterbitkan berdasarkan taburan yang telah dicadangkan sebagai taburan alternatif bivariat diskrit dengan menggunakan konvolusi daripada dua taburan bivariat dan kaedah klasik penurunan trivariat. Didapati, taburan bivariat baru ini mengizinkan lebih banyak fleksibiliti dalam pemodelan dan had kepada korelasi antara dua pembolehubah rawak adalah kurang. Ciri-ciri dan anggaran parameter taburan bivariat baru ini telah disediakan. Selanjutnya, inferensi statistik untuk taburan COM-Poisson dan isu-isu pengiraan turut diselidik. Ujian untuk serakan sama, kajian tentang kekuatan statistik dan anggaran parameter dengan menggunakan kaedah PKM dipelajari. Disertasi ini juga mempertimbangkan satu sukatan jarak berdasarkan fungsi penjana kebarangkalian bagi anggaran parameter. Prestasi dan keteguhan bagi statistik yang dicadangkan dalam anggaran parameter diselidik di bawah taburan binomial negatif melalui cara simulasi Monte Carlo, khususnya dibandingkan dengan kaedah PKM dan jarak Hellinger minimum.

ABSTRACT

This thesis proposed a new distribution to model under-, equi- and over-dispersion in count data. It arises as a particular case of a modified random walk on the plane, and includes the binomial, negative binomial and shifted negative binomial as special cases. Some properties, test of hypothesis for equi-dispersion, simulation study of power, parameter estimation by maximum likelihood and a squared distance method based on the probability generating function are considered. The proposed distribution is compared with existing distributions like the COM-Poisson and generalized Poisson for modelling dispersion. It is shown to be a flexible model in applications by illustrating its goodness-of-fit to four real data sets. In addition, two new bivariate distributions are derived from proposed distribution as the alternative bivariate discrete distributions by applying the convolution of two bivariate distributions and the classical trivariate reduction method. It is found that the new bivariate distributions permit more flexibility in modelling and less limitation on the correlation between the two random variables. The characteristic and the estimation of the parameters of the new bivariate distributions are provided. Furthermore, the statistical inference for the COM-Poisson distribution and computational issues are studied. Test for equi-dispersion, study of statistical power and parameter estimation by maximum likelihood are developed. This thesis also considers a probability generating function-based divergence statistic for parameter estimation. The performance and robustness of the proposed statistic in parameter estimation is studied for the negative binomial distribution by Monte Carlo simulation, especially in comparison with maximum likelihood and minimum Hellinger distance estimation.

ACKNOWLEDGEMENTS

I am heartily thankful to my supervisor, Prof. Dr. Ong Seng Huat, whose encouragement, advice, guidance and support from the initial to the final stages of this thesis, has inspired and enriched my growth as a student and a researcher. Thank you for triggering and nourishing my intellectual maturity that will benefit me for a long time to come. I am extraordinarily fortunate to have him as my supervisor.

I would like to thank my thesis examiners who have spent their precious time to read this thesis and gave crucial comments to improve it.

Thanks are also due to my family members, especially my parents who raised me with care, love and have always supported me in all my pursuits.

My special thanks go to Yap Soon Lee for his invaluable help throughout my thesis work and for keeping me organized.

Lastly, I offer my regards and blessings to those who have supported me in any way for the successful realization of this thesis.

TABLE OF CONTENTS

ABSTRAK	iii
ABSTRACT	iv
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vi
GLOSSARY OF ABBREVIATIONS	xi
LIST OF FIGURES	xiii
LIST OF ALGORITHMS	xv
LIST OF TABLES	xvi
LIST OF APPENDICES	xx
1 INTRODUCTION	
1.0 Discrete Distributions in Statistical Analysis	1
1.1 Literature Review	2
1.2 Contributions of Thesis	8
1.3 Organization of Thesis	9
2 PRELIMINARIES	
2.0 Introduction	11
2.1 Parameter Estimation	11
2.1.1 Maximum Likelihood Estimation	11
2.1.2 Minimum Hellinger Distance Estimation	13
2.1.3 Simulated Annealing	14

2.2	Hypothesis testing	16
2.2.1	Chi-Square Goodness-of-Fit Test	16
2.2.2	Rao's Score Test and Likelihood Ratio Test	17
2.3	Statistical Power Analysis	19
2.3.1	Significance Level	20
2.3.2	Reliability of Sample and Sample Size	20
2.3.3	The Effect Size	21
2.4	Construction of Bivariate Discrete Distribution	21
2.4.1	Mixtures of Convolution	21
2.4.2	Trivariate Reduction	22
2.5	Properties of the bivariate distribution	23
2.5.1	Probability Generating Function and Joint Probability Function	23
2.5.2	Marginal and Conditional Distribution	24
2.5.3	Covariance and Correlation	25
3	A DISTRIBUTION FOR MODELLING DISPERSION	
3.0	Introduction	26
3.1	Subclass of the Generalized Shifted Inverse Trinomial Distribution	28
3.2	Behaviour of the Subclass of the Generalized Shifted Inverse Trinomial Distribution, $GIT_{3,1}$	30
3.3	Extension of the $GIT_{3,1}$ to Non-Integer n	33
3.4	Log-concavity, Unimodality and Reliability Properties	33
3.5	Parameter Estimation	34

3.5.1	Maximum Likelihood Estimation	34
3.5.2	A Probability Generating Function Based Estimation	36
3.6	Test of Equi-Dispersion	36
3.6.1	Partial Derivatives	37
3.6.2	Rao's Score Test	38
3.6.3	Likelihood Ratio Test	39
3.7	Power Analysis of the Rao's Score Test and Likelihood Ratio Test	39
3.8	Applications	43
3.8.1	Over-Dispersion Data	43
3.8.2	Under-Dispersion Data	46
3.8.3	Equi-Dispersion Data	48
4	BIVARIATE DISTRIBUTION EXTENSION	
4.0	Introduction	50
4.1	Formulation of Bivariate Distribution	51
4.1.1	Convolution (Type I)	51
4.1.2	Trivariate Reduction (Type II)	52
4.2	Properties of bivariate distribution	53
4.2.1	Properties of Type I BGITD	53
4.2.2	Properties of Tye II BGITD	57
4.3	Characteristic of bivariate distribution	59
4.3.1	Characteristic of Type I BGITD	59
4.3.2	Characteristic of Tye II BGITD	61
4.4	Parameter Estimation	64
4.4.1	Maximum Likelihood Estimation	64

4.4.2	The Pgf-based Minimum Hellinger-type Distance Estimation	65
4.5	Numerical Examples	65
5	STATISTICAL INFERENCES ON THE CONWAY- MAXWELL-POISSON DISTRIBUTION	
5.0	Introduction	74
5.1	Probability Functions of COM-Poisson Distribution	75
5.2	Limitation of Computer Accuracy	77
5.3	Partial Derivatives of the Constant $Z(\lambda, \nu)$	80
5.4	Maximum Likelihood Estimation	81
5.5	Test for Equi Dispersion	82
5.5.1	Rao's Score Test	83
5.5.2	Likelihood Ratio Test	84
5.6	Statistical Power Analysis of the Rao's Score Test and Likelihood Ratio Test	84
6	A PROBABILITY GENERATING FUNCTION BASED MINIMUM HELLINGER TYRE DISTANCE ESTIMATION	
6.0	Introduction	90
6.1	Generalized Minimum Hellinger Divergence Statistic based on Probability Generating Function	92
6.2	Asymptotic Results	93
6.3	Monte Carlo Simulation Design	95
6.4	Discussion of the Monte Carlo Simulation Results	102
6.5	Examples of Data Fitting	110

7	CONCLUSION AND FURTHER WORK	114
	APPENDICES A – C	118
	REFERENCES	125

GLOSSARY OF ABBREVIATIONS

χ^2	Chi-square
$f(x_1, x_2)$	Joint probability mass function
$\varphi(t_1, t_2)$	Joint probability generating function
$\rho(X_1, X_2)$	Correlation between the random variables X_1 and X_2
$Cov(X_1, X_2)$	Covariance between the random variables X_1 and X_2
$Var(X_i)$	Variance for the random variables X_i
$E(X_i)$	Mean for the random variables X_i
${}_2F_1$	Gauss hypergeometric function
$\mu_{[r]}^{\cdot}$	r th descending factorial moment
N	Sample size
$\varphi_{X_1}(t x_2)$	The pgf of the conditional distribution of X_1 given $X_2 = x_2$
B	Binomial
BGITD	Bivariate $GIT_{3,1}$ distribution
BGPD	Bivariate generalized Poisson distribution
BNBD	Bivariate negative binomial distribution
COM-Poisson	Conway-Maxwell-Poisson
df	Degree of freedom
epgf	Empirical probability generating function
GIT	Generalization of the shifted inverse trinomial
$GIT_{3,1}$	Subclass of the generalized shifted inverse trinomial distribution
gof	goodness-of-fit

GPD	generalized Poisson distribution
HD	Hellinger distance
ID	Index of dispersion
IT	Inverse trinomial
KL	Kullback-Leibler
LR	Likelihood ratio
LRT	Likelihood ratio test
MHD	Minimum Hellinger distance
MHDE	Minimum Hellinger distance estimation
MLE	Maximum likelihood estimation
MSE	Mean square error
NTA	Neyman Type A
NB	Negative binomial
pf	Probability function
pgf	Probability generating function
PIG	Poisson inverse Gaussian
pmf	Probability mass function
POI	Poisson
RS	Rao's score
SA	Simulated annealing

LIST OF FIGURES

FIGURES	TITLE	PAGE
3.1	$p_1=0.2, p_2=0.1, p_3=0.7, n=5$	31
3.2	$p_1=0.2, p_2=0.1, p_3=0.7, n=20$	31
3.3	$p_1=0.6, p_2=0.1, p_3=0.3, n=5$	31
3.4	$p_1=0.6, p_2=0.1, p_3=0.3, n=20$	31
3.5	$p_1=0.2, p_2=0.7, p_3=0.1, n=5$	31
3.6	$p_1=0.2, p_2=0.7, p_3=0.1, n=20$	31
3.7	$p_1=0.6, p_2=0.3, p_3=0.1, n=5$	32
3.8	$p_1=0.6, p_2=0.3, p_3=0.1, n=20$	32
3.9	$p_1=0.8, p_2=0.1, p_3=0.1, n=5$	32
3.10	$p_1=0.8, p_2=0.1, p_3=0.1, n=20$	32
4.1	$p_1=0.1, p_2=0.1, q_1=0.1, q_2=0.1, n_1=16, n_2=1$	60
4.2	$p_1=0.1, p_2=0.1, q_1=0.1, q_2=0.1, n_1=42, n_2=1$	60
4.3	$p_1=0.1, p_2=0.1, q_1=0.1, q_2=0.1, n_1=42, n_2=30$	61
4.4	$p_1=0.1, p_2=0.4, q_1=0.1, q_2=0.4, r_1=0.1, r_2=0.4,$ $n_1=n_2=n_3=1$	62
4.5	$p_1=0.1, p_2=0.4, q_1=0.1, q_2=0.4, r_1=0.1, r_2=0.4,$ $n_1=5, n_2=n_3=1$	62
4.6	$p_1=0.1, p_2=0.4, q_1=0.1, q_2=0.4, r_1=0.1, r_2=0.4,$ $n_1=1, n_2=5, n_3=1$	63

4.7	$p_1 = 0.1, p_2 = 0.4, q_1 = 0.1, q_2 = 0.4, r_1 = 0.1, r_2 = 0.4,$ $n_1 = 1, n_2 = 1, n_3 = 5$	63
4.8	$p_1 = 0.1, p_2 = 0.4, q_1 = 0.1, q_2 = 0.4, r_1 = 0.1, r_2 = 0.4,$ $n_1 = 1, n_2 = 5, n_3 = 5$	63
4.9	$p_1 = 0.1, p_2 = 0.4, q_1 = 0.1, q_2 = 0.4, r_1 = 0.1, r_2 = 0.4,$ $n_1 = 5, n_2 = 5, n_3 = 1$	64
4.10	$p_1 = 0.1, p_2 = 0.4, q_1 = 0.1, q_2 = 0.4, r_1 = 0.1, r_2 = 0.4,$ $n_1 = 5, n_2 = 1, n_3 = 5$	64
6.1	MSE of parameter estimates for NB distribution ($\mu = 1.5; r = 2.0$)	96
6.2	97.5% NB ($\mu = 6.53; r = 4.0$) + 2.5% Poi ($\lambda = 30$) with sample size of 500.	97
6.3	97.5% NB ($\mu = 6.53; r = 4.0$) + 2.5% Poi ($\lambda = 65$) with sample size of 500.	97
6.4	Design of simulation study.	98
6.5	MSE of parameter estimates for NB distribution ($\mu = 1.5; r = 2.0$) with different number of Gaussian quadrature points	99

LIST OF ALGORITHMS

ALGORITHM	TITLE	PAGE
2.1	Simulated annealing	15

LIST OF TABLES

TABLE	TITLE	PAGE
3.1	Dispersion under the $GIT_{3,1}$ distribution	30
3.2	Simulated power of score test and LRT for $GIT_{3,1}$: over-dispersion	40
3.3	Simulated power of score test and LRT for $GIT_{3,1}$: under-dispersion	41
3.4	Estimated empirical level of score test and LRT for $GIT_{3,1}$	42
3.5	Frequency distribution of dicentrics for mammalian cytogenetic dosimetry: lesions in rabbit lymphoblasts induced by streptonigrin with exposure 30 ($\mu g / kg$) (NSC-45383)	44
3.6	Frequency distribution of dicentrics for mammalian cytogenetic dosimetry: lesions in rabbit lymphoblasts induced by streptonigrin with exposure 90 ($\mu g / kg$) (NSC-45383)	45
3.7	Frequency distribution of dicentrics for dose 1200	46
3.8	Frequency distribution of dicentrics for dose 2000	47
3.9	Frequency distribution of dicentrics for dose 800	48
3.10	Frequency distribution of 102 spiders under 240 boards	49
4.1	Numerical examples	54
4.2	Covariance and correlation of Type I BGITD	55

4.3	Change of the positive integers n_1 , n_2 and n_3	62
4.4	Number of accidents sustained by 122 experienced shunters over 2 successive periods of time	67
4.5	Summary statistics for Table 4.4	69
4.6	Number of patients in two boxes in a room of the critical care and emergency service in the San Agustin Hospital (Linares, Spain)	70
4.7	Summary statistics for Table 4.6	71
4.8	Number of times bacon and eggs were purchased on four consecutive shopping trips	72
4.9	Summary statistics for Table 4.8	73
5.1	Calculate $Z(\lambda, \nu)$ with $\lambda=5$ ($\nu=0.1, 0.3$ and 0.5)	78
5.2	Calculate $Z(\lambda, \nu)$ with $\lambda=5$ ($\nu=1, 2$ and 3)	78
5.3	Calculate $Z(\lambda, \nu)$ with $\lambda=30$ ($\nu=0.1, 0.3$ and 0.5)	79
5.4	Calculate $Z(\lambda, \nu)$ with $\lambda=30$ ($\nu=1, 2$ and 3)	79
5.5	Comparison between the moments obtained from the asymptotic approximation $Z(\lambda, \nu)$ and the partial derivatives of $Z(\lambda, \nu)$	81
5.6	Simulated power of score test and LRT: COM-Poisson distribution ($\lambda=5$)	85
5.7	Simulated power of score test and LRT: COM-Poisson distribution ($\lambda=10$)	86
5.8	Simulated power of score test and LRT: COM-Poisson distribution ($\lambda=20$)	87

5.9	Estimated empirical level of score test and LRT: COM-Poisson distribution	88
6.1	MSE and bias (in bracket) for proposed estimators T_3 and T_4 (with 5% outliers)	100
6.2	MSE and bias (in bracket) for proposed estimators T_3 and T_4 (without outliers)	100
6.3	Estimated parameters for α and β for different range	101
6.4	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (with 2.5% outliers), $\mu = 1.5$, $r = 2.0$, $\lambda = 30$	105
6.5	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (with 5% outliers), $\mu = 1.5$, $r = 2.0$, $\lambda = 30$	105
6.6	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (with 2.5% outliers), $\mu = 1.5$, $r = 2.0$, $\lambda = 65$	106
6.7	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (with 5% outliers), $\mu = 1.5$, $r = 2.0$, $\lambda = 65$	106
6.8	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (without outliers), $\mu = 1.5$; $r = 2.0$	107
6.9	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (with 2.5% outliers), $\mu = 6.53$, $r = 4.0$, $\lambda = 30$	107
6.10	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (with 5% outliers), $\mu = 6.53$, $r = 4.0$, $\lambda = 30$	108
6.11	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (with 2.5% outliers), $\mu = 6.53$, $r = 4.0$, $\lambda = 65$	108

6.12	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (with 5% outliers), $\mu = 6.53$, $r = 4.0$, $\lambda = 65$	109
6.13	MSE and bias (in bracket) for proposed estimators $T_1 - T_6$ (without outliers), $\mu = 6.53$, $r = 4.0$	109
6.14	Fitting NB to counts of red mites on apple leaves with MLE, MHD and proposed estimation methods	110
6.15	Fitting NB to yeast cells per square in a haemocytometer with MLE, MHD and proposed estimation methods	111
6.16	Fitting NB to soil bacteria per field with MLE, MHD and proposed estimation methods	112
6.17	Fitting NB to soil bacteria per colony with MLE, MHD and proposed estimation methods	113

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Partial derivatives and elements of information matrix for GIT _{3,1} distribution	118
B	Partial derivatives and elements of information matrix for COM-Poisson distribution	120
C	Mixture of discrete distributions	121